Erschienen in: Modes of Truth : The Unified Approach to Truth, Modality, and Paradox / Nicolai, Carlo; Stern, Johannes (Hrsg.). - New York : Routledge, 2021. - S. 69-87. - ISBN 978-0-367-14109-7
http://dx.doi.org/10.4324/9780429030208-4

4 Truth, Reflection, and Commitment

Leon Horsten and Matteo Zicchetti

4.1 Introduction

Proof-theoretic reflection principles have been discussed in proof theory ever since Gödel's discovery of the incompleteness theorems. But these reflection principles have not received much attention in the philosophical community. The aim of the present chapter is to survey some of the principal meta-mathematical results on the iteration of proof-theoretic reflection principles, and also to investigate these results from a logico-philosophical perspective; we will concentrate on the epistemological significance of these technical results and on the epistemic notions involved in the proofs. In particular, we will focus on the notions of *commitment to* and *acceptance of* a theory. Special attention is given to the connection between proof-theoretic reflection and axiomatic truth theories.

After distinguishing between different types of proof-theoretic reflection principles, we review some proof-theoretic results concerning extensions of formal theories by (iterated) reflection principles. As basis theories we concentrate on standard arithmetical and elementary axiomatic truth theories. We then go on to explore the epistemological significance of these results. In this investigation we aim at showing that epistemic notion of acceptance of (or commitment to) a theory plays a crucial role in the philosophical argumentation for reflection principles and their iteration.

The structure of this chapter is as follows. In Sections 4.2 and 4.3, iterated reflection over arithmetical theories is discussed. In Section 4.4, we discuss reflection over axiomatic truth theories—here we concentrate on theories of disquotational and of compositional truth. The philosophical background for our discussion in Sections 4.3 and 4.4 is given by Feferman's theory of implicit commitment. However, as we will show, the epistemic notions involved in the investigation of reflection principles presented in Sections 4.2, 4.3, and 4.4 are never made explicit; they are employed only informally in the philosophical argumentation for reflection principles. In Section 4.5 we turn to Cieśliński's formal analysis of the process of reflection on implicit acceptance of a formal theory.

As we will show, in this approach the epistemic notion of acceptance of a theory is made fully explicit via the use of a modal predicate. We will analyse Cieśliński's approach and indicate some problems and questions. We close this chapter with some general philosophical remarks on the nature and role of reflective processes in mathematics.

We try to keep our notation as standard as possible. Concerning proof-theoretical background, we presuppose some familiarity with a few basic formal systems of arithmetic, such as Peano Arithmetic (PA, and its language $\mathcal{L}_{\mathsf{PA}}$) and Elementary Arithmetic (EA). Moreover, although we will present some basic facts about Kleene's notation system $\mathcal{O}$, we will presuppose some familiarity with ordinal notations, the Veblen hierarchy, and related notions. Concerning truth theory, we assume a passing acquaintance with a handful of the main truth theories, such as the compositional theory CT, the Kripke-Feferman system KF, and the Partial Kripke-Feferman system PKF. Nevertheless, for the benefit of readers who are not familiar with these systems, we include footnote references to places where they are defined and discussed.

4.2 Reflection Principles and Progressions of Theories

We concentrate on theories that are formulated in the language of first-order arithmetic or an extension thereof, and at least as strong as Elementary Arithmetic (EA). We are interested in the iteration of proof-theoretic reflection principles over these theories, where a proof-theoretic reflection principle for a given theory S is a formalised soundness statement for S: it expresses that everything provable in S is also true.

By Tarski's theorem of the undefinability of truth, the language of arithmetic does not contain its own truth predicate. So in the language of arithmetic this guiding idea can only be approximated to varying degrees. We can distinguish the following types of reflection principles (for a given theory S):

(i) Con_S (consistency)
(ii) $Prov_S \ulcorner \varphi \urcorner \rightarrow \varphi$ (local reflection)
(iii) $Prov_S \ulcorner \varphi(\dot{x}) \urcorner \rightarrow \varphi(x)$ (uniform reflection)

Here $Prov_S$ is a standard provability predicate for the given theory S. The formula Con_S expresses the consistency of S in terms of $Prov_S$: it can be taken to be the formula

$$Prov_S \ulcorner 0 = 1 \urcorner \rightarrow 0 = 1.$$

Local reflection for a theory S is denoted as Rfn_S, and uniform reflection is denoted as RFN_S. Restricted versions for these principles are also considered: one can consider Rfn_S (RFN_S) for sentences (formulae) of a

specific syntactic complexity. Π^0_1-*Rfn*$_S$, for instance, is local reflection for the Π^0_1 fragment of S and is equivalent to *Con*$_S$.

We can iterate the procedure of adding a reflection principle to a given theory S. For a given theory S and a given reflection principle $\mathcal{R}$ we let $\mathcal{R}[S]$ mean "the reflection principle $\mathcal{R}$ over S". Then we can define the iteration of reflection in the following way by letting:

$\mathcal{R}^0[S]$ be S;
for α a successor ordinal, $\mathcal{R}^{\alpha+1}[S]$ be $\mathcal{R}[\mathcal{R}^\alpha[S]]$;
for λ a limit ordinal, $\mathcal{R}^\lambda[S]$ be the union of all $\mathcal{R}^\alpha[S]$ for $\alpha < \lambda$.

The first proof-theoretic results that we will discuss concern *progressions of theories* generated via iteration (into the transfinite) of reflection principles.

However, before presenting the notion of a progression of theories and the results, we introduce a few notions concerning Kleene's $\mathcal{O}$. We call $|a|$ the ordinal denoted by an ordinal notation a in Kleene's notation system $\mathcal{O}$, which is partially ordered by the relation $<_\mathcal{O}$. We have $a <_\mathcal{O} b$, for two ordinal notations a and b, if and only if $|a| < |b|$.

A *path* P is a subset of $\mathcal{O}$ such that (i) for any a, $b \in P$ either $a \leq_\mathcal{O} b$ or $b \leq_\mathcal{O} a$, (ii) if $b \in P$ and $c \leq_\mathcal{O} b$ then $c \in P$. For any $a \in \mathcal{O}$, a set $P = \{ b \mid b <_\mathcal{O} a\}$ is called a *path within* $\mathcal{O}$. The *length* of a path P is the ordinal of the restriction of $<_\mathcal{O}$ to P. For any path P within $\mathcal{O}$, the order type of P, denoted as $|P|$, is less than ω_1^{CK}. A path P is a path *through* $\mathcal{O}$ if $|P| = \omega_1^{CK}$, where $\omega_1^{CK} = \sup\{|a| : a \in \mathcal{O}\}$. The relation $<_\mathcal{O}$ is not recursively enumerable; indeed, it is Π^1_1-complete. However, for any a, the restriction of $<_\mathcal{O}$ to $\{ b \mid b <_\mathcal{O} a\}$ is recursively enumerable.

Now we introduce the notion of a progression of theories. A progression of a theory S is a primitive recursive mapping taking any ordinal notation a in some path in Kleene's ordinal notation system $\mathcal{O}$ to a Σ^0_1-formula φ_a that recursively enumerates the axioms of a theory S_a, such that

1. $S_0 = S$;
2. $S_{suc(a)} = S_a + \mathcal{R}^a[S]$;
3. $S_{lim(a)} = \bigcup_{b < a} S_b$.

In words: the starting theory S_0 is just S, the successor stage of the progression is, for any notation a, just the previous theory S_a plus a reflection principle $\mathcal{R}^a$ for S_a, and at limit stages we take unions.

Any transfinite progression yields a *progressive reflection sequence*, which is a sequence of theories of the form

$$S_0, S_1, \ldots S_\omega, S_{\omega+1}, \ldots S_\alpha, \ldots,$$

where $S_{\alpha+1}$ is an extension by reflection of S_α, and S_λ, for limit ordinals λ, has as axioms the union of the axioms of earlier theories.

In the following section we will survey two main results: Turing's completeness theorem for consistency progressions and Feferman's completeness theorem for uniform reflection progressions. Moreover, we will briefly touch upon Feferman's results about autonomous progressions of formal theories.

4.3 Mathematical Reflection

Turing used consistency progressions in an attempt to reduce incompleteness in arithmetic. He proved the following theorem:

THEOREM 4.3.1 (Turing, 1939). For any true Π^0_1 sentence φ there is an $a \in \mathcal{O}$ such that $|a| = \omega + 1$ and $S_a \vdash \varphi$. Moreover, there is a primitive recursive function that associates such an a with each true Π^0_1 sentence φ.

Turing (1939) suggests that the transition from a theory S_a to $S_{suc(a)}$ invokes some sort of reflection:

> We were able, however, from a given system to obtain a more complete one by the adjunction as axioms of formulae, seen intuitively to be correct, but which the Gödel theorem shows are unprovable in the original system; from this we obtained a yet more complete system by a repetition of the process, and so on.
>
> (p. 198)

However, the epistemological import of Turing's completeness theorem is limited. Theorem 4.3.1 only tells us that for any true Π^0_1 sentence φ there is a consistency progression with length $\omega + 1$, such that $S_{\omega + 1}$ proves φ. As Franzén (2004b, Section 6) already pointed out, Turing's result does not provide us with a method of *recognising*, for any true Π^0_1 sentence φ, that it is true. Turing's proof indeed associates with every true Π^0_1 sentence φ a consistency reflection sequence of length $\omega + 1$ that ends in a theory $S_{\omega + 1}$ that proves φ. However, the axioms of S_ω have a non-canonical definition; the trick of Turing's proof consists in defining S_ω in such a way that its consistency entails that φ is true. Even though Turing's clever definition of ω and "canonical" definitions of ω extensionally coincide, no S_n proves that this is so.[1]

Feferman realised that in order to strengthen Turing's completeness result, uniform reflection progressions rather than consistency or local reflection progressions are needed. He proved:

THEOREM 4.3.2 (Feferman, 1962). There is a uniform reflection progression based on PA such that for any true arithmetical sentence φ there is an $a \in \mathcal{O}$ such that $|a| \leq \omega^{\omega^{\omega+1}}$ with $S_a \vdash \varphi$.[2]

This is known as Feferman's completeness theorem. His proof generates a *path* P within $\mathcal{O}$ of length $\omega^{\omega^{\omega+1}}$ such that the union of

all theories associated with the notations in this path is arithmetically complete.

As with Turing's completeness theorem, and for the same reasons, the epistemological import of Feferman's completeness proof is limited. Following Franzén, we can see that it would be wrong to say that Turing's and Feferman's results show that we will eventually obtain every arithmetical truth by iterating reflection principles.[3]

4.3.1 Autonomous Progressions

The proof of Turing's completeness theorem (and the proof of Feferman's completeness theorem) shows that there is a sense in which progressions as defined in the previous section fail to capture how systems of a higher ordinal level are warranted "from below". For this reason, Kreisel (1958) argued that progressions should satisfy an additional *autonomy* requirement: for every S_a that is in a progression, it should be provable in some S_b with $b <_{\mathcal{O}} a$ that a is in $\mathcal{O}$. A progression that satisfies this additional criterion is called an *autonomous progression*. Before surveying the results about the autonomous progressions, we will introduce briefly the notions of Veblen functions and Veblen hierarchy.

Veblen functions are a hierarchy of normal functions (continuous strictly increasing functions from ordinals to ordinals). If φ_0 is any normal function, then for any ordinal $\alpha > 0$, φ_α is the function enumerating the common fixed points of φ_β for $\beta < \alpha$. These functions are all normal. In the special case when $\varphi_0(\alpha) = \omega^\alpha$ this family of functions is known as the Veblen hierarchy. The function φ_1 is the same as the ε function: $\varphi_1(\alpha) = \varepsilon_\alpha$. The first ε ordinal number ε_0 is $\sup\{1, \omega, \omega^2, \ldots, \omega^\omega, \ldots, \omega^{\omega^{\omega^{\cdots}}}\}$ and is the least fixed point of φ_0, so that $\omega^\alpha = \alpha$. And then $\varphi_2(0)$ is the least ordinal α, such that $\varepsilon_\alpha = \alpha$.

The Feferman–Schütte ordinal Γ_0 can be defined as the smallest ordinal that cannot be obtained by starting with 0 and using the operations of ordinal addition and the Veblen functions $\varphi_\alpha(\beta)$. That is, it is the smallest α such that $\varphi_\alpha(0) = \alpha$. Feferman (1964) and Schütte (1964, 1965) investigated autonomous progressions of predicative theories of analysis. In particular, Feferman (1964) investigated autonomous progressions via uniform reflection based on the systems H and R,[4] determining the *limit* of predicative reasoning. In a nutshell, he showed that the ordinal Γ_0 is the least ordinal that "cannot be reached" predicatively. Or in other words, it is the least ordinal greater than all autonomous a in the progression.[5]

But one can also consider autonomous reflection progressions over first-order arithmetic. The following is a typical result, which is apparently "folkore":[6]

THEOREM 4.3.3. The autonomous uniform reflection progression based on Peano Arithmetic is the first-order fragment of the system of Ramified Analysis up to (but not including) level ω, and the length of this progression is $\varphi_2(0)$.

These theorems are epistemologically more significant than the completeness theorems of Turing and Feferman. In contrast to the non-autonomous progressions, the autonomy condition assures that *we recognise* by means of a proof in a previous stage of the progression that for a limit a, a is an ordinal notation. In this sense, propositions such as Theorem 4.3.3 show *what we can come to know* in reflection progressions. Of course a strong idealisation is involved here: *we* are only able to go through a finite number of stages of an autonomous progression before we die.[7]

In this chapter we are interested in the informal notions involved in the transition from a theory S_a to $S_{suc(a)}$, that is, in the addition of the reflection principles. Like Turing, Feferman (1962) claims that the transition from a theory S_a to $S_{suc(a)}$ is obtained via a process of reflection. He states that our acceptance of a reflection principle for our base theory (and iterating this procedure) rests on our pre-theoretic *attitude*:

> In contrast to an arbitrary procedure for moving from A_K to A_{K+1}, a reflection principle provides that the axioms of A_{K+1} shall express a certain *trust* [our emphasis] in the system of axioms A_K.
>
> (p. 261)

We observe that Feferman's appeal to trust differs from Turing's appeal to mathematical intuition; if we look at the previous quote by Feferman, we see that a reflection principle for a theory S does not only express the soundness of S, but has also an *epistemic* component. In later work, Feferman (1991) continued to emphasise in that reflection principles have an epistemic component:

> Gödel's theorems show the inadequacy of single formal systems [for the purpose of formal analysis of mathematical thought]. However at the same time they point to the possibility of systematically generating larger and larger systems whose *acceptability is implicit in acceptance of the starting theory*.
>
> (p. 2, our emphasis)

Feferman here sketches an epistemological route from knowledge of the axioms of a weaker system to knowledge of the axioms of a stronger system. One starts by believing the axioms of a system S. If one's reasons for doing so are good and S is true, then these beliefs amount to knowledge of the axioms of S. When one is in such a situation, one is implicitly committed to reflection principles for S, such as Con_S. By explicitly endorsing such implicit commitments, one can come to accept, and perhaps even to know, the axioms of a stronger system S'.

4.4 Reflecting on Truth

We will now leave reflection over purely arithmetical theories behind, and concentrate on the iteration of reflection principles over theories

of truth (and falsity) that are formulated in an expansion of the language of PA or EA with a fresh truth (and falsity) predicate.

4.4.1 Axiomatic Truth Theories

Pioneers of the investigation of proof-theoretic reflection principles pointed out that the concept of *truth* is involved in the concept of reflection:

> By a "reflection principle" for a formal system S we mean, roughly, the formal assertion stating the soundness of S:
>
> > *If a statement φ (in the formalism S) is provable in S then φ is valid.*
>
> (Kreisel and Lévy, 1968, p. 98)

This was regarded as a problem:

> Literally speaking, the *intended* reflection principle cannot be formulated in S itself by means of a single statement. This would require a *truth definition* T_S, with a variable a over (Gödel numbers of, or, simply, over) formulas of S, and a definition of the proof relation $Prov_S(p, a)$ (read: p is (the Gödel number of) a proof of a in S). The reflection principle for S would be
>
> $$\forall p \forall a [Prov_S(p, a) \rightarrow T_S(a)].$$
>
> Such a truth definition T_S, does not exist.
>
> (Kreisel and Lévy, 1968, p. 98)

This difficulty can be (and was) circumvented by *approximating* the intended reflection principle by means of the purely arithmetical principles Rfn_S and RFN_S. But this is not the only possible way forward. Instead, a primitive truth predicate T can be added to the language of arithmetic, thus generating the language $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$, and new axioms governing the behaviour of the truth predicate can be added to the background arithmetical theory. This is what some proof theorists started to do in the late 1970s. Moreover, the resulting formal systems were related to a philosophical discussion about the function or role of the concept of truth.

One important role for the concept of truth is to express and reason with generalisations over statements. For this purpose, the use of the truth predicate as a device of quotation and of disquotation is essential. This means that Tarski-biconditionals, i.e., formulae of the form $T \ulcorner \varphi \urcorner \leftrightarrow \varphi$, play a pivotal role in truth theory.

A distinction is made between *typed* and *untyped* (or type-free) Tarski-biconditionals. In the typed case, the truth predicate is not itself allowed

to occur in φ. If we start with PA as a base theory and add to PA the collection of all *typed* Tarski-biconditionals $T \ulcorner\varphi\urcorner \leftrightarrow \varphi$ for $\varphi \in \mathcal{L}_{\mathsf{PA}}$, the resulting theory is called TB.[8] If one wants to add to PA a collection of *untyped* Tarski-biconditionals, then, in order to avoid the liar paradox, one can either weaken the background logic, or restrict the collection of Tarski-biconditionals and preserve full classical reasoning. One consistent way of weakening the logic that keeps the full Tarski-biconditionals is to work in *Basic De Morgan* logic (*BDM*).[9] The untyped truth theory formulated in *BDM*, where the Tarski-biconditionals are completely unrestricted, is called TS_0 and is discussed in Fischer et al. (2017).

If one wants to preserve classical logic, then there are different options for restricting the Tarski-biconditionals to avoid inconsistency. Here we discuss two such possible restrictions. One possibility is to restrict the Tarski-biconditional scheme to the sentences φ in which the truth predicate only occurs *positively* (i.e., in the scope of an even number of negation symbols). If we add this collection to PA, the resulting truth theory is called PTB.[10] A natural way of extending this theory is to expand the language of the truth theory ($\mathcal{L}_T$) with a primitive *falsity* predicate, thus generating the language $\mathcal{L}_{T,F}$. We then consider the sublanguage $\mathcal{L}^+_{T,F}$, which is obtained by allowing the negation symbol from $\mathcal{L}_{T,F}$ only to prefix atomic arithmetical formulas. Moreover, we consider the truth biconditionals $T \ulcorner\varphi\urcorner \leftrightarrow \varphi$ with φ restricted to $\mathcal{L}^+_{T,F}$, and the falsity biconditionals $F \ulcorner\varphi\urcorner \leftrightarrow \bar{\varphi}$, where $\bar{\varphi}$ is the *dual* of φ. We can define duals recursively: the dual of an atomic arithmetical formula is its negation; the dual of an atomic formula of the form Tt is Ft, and vice versa, the dual of $A \wedge B$ is the disjunction of the dual of A and the dual of B, and so on.[11] PA plus these two collections of biconditionals is called TFB.

4.4.2 Compositionality and Implicit Commitment

The philosophical question now arises whether the *content of the concept of truth* is given by some such collection of Tarski-biconditionals. An affirmative answer to this question is defended, for instance, in Horwich (1990), Halbach (2001), and Horsten and Leigh (2016). This position is called *disquotationalism*, as it asserts that the content of the concept of truth is captured by a relatively simple and natural collection of Tarski-biconditionals, i.e., by a disquotational theory of truth. If disquotationalism is correct, then the concept of truth really is at bottom merely a device for quotation and disquotation, as Quine maintained.

A standard objection against this, which traces back to Davidson, is that truth is *compositional*. According to this view, truth theories

should be able to prove intuitive semantic principles, for instance that any conjunction is true if and only if its conjuncts are both true, and so forth. But these compositional truth clauses cannot be derived from a set of Tarski-biconditionals. In this way it seems that disquotationalist views fall short of capturing the content of the concept of truth.

The standard typed compositional truth theory is called CT.[12] The most popular compositional type-free truth theory in classical logic is KF; the most popular type-free compositional truth theory in non-classical logic is PKF.[13] The Davidsonian objection against disquotational truth theories applies to all the theories mentioned above: the message is that compositional typed (type-free) truth outstrips disquotational typed (type-free) truth by proving *core* principles governing the concept of truth that disquotational theories cannot prove. Without further resources, it seems that there is no way out for the disquotationalist.

At this point, reflection principles enter the philosophical debate. The idea is that the compositional principles might be *implicit* in some collection of Tarski-biconditionals and that *reflection* can bridge the gap between disquotational and compositional truth.

This is indeed the case. In the typed context, Halbach observed that iterating uniform reflection over TB twice recovers typed compositional truth (Halbach, 2001, Section 4):

THEOREM 4.4.1. $RFN^2[\mathsf{TB}] \vdash \mathsf{CT}$.

This phenomenon extends to the classical type-free context (Horsten and Leigh, 2016, Theorem 7):

THEOREM 4.4.2. $RFN^2[\mathsf{TFB}] \vdash \mathsf{KF}$.

Theorem 4.4.2 has to be taken, however, with a grain of salt. Even though the version of KF that is used by Horsten and Leigh (2016) is closely related to the usual formulations of KF (for instance, the version given in Halbach, 2014, Definition 15.2), it is not outright equivalent to them. In *Pos*(KF) (*positive* KF), the version of KF derivable via two iterations of reflection from TFB, the compositional axioms are restricted to the positive fragment of the language, whereas in the case of the usual KF the compositional axioms are completely unrestricted. Therefore, although these two versions of KF are equivalent for the arithmetical part of the language, their truth predicate behaves somewhat differently. In Zicchetti (2020) it has been shown that TFB and the version of KF adopted in Horsten and Leigh (2016), i.e., the version of KF that we obtain in Theorem 4.2.2 *via reflection from the theory* TFB, can be consistently closed under unrestricted rules of *Necessitation* and *Conecessitation* for the truth and falsity predicates to the theory *Pos*(KF)*, whereas the version of KF given in Halbach (2014) is inconsistent with the addition of the two rules.

The recovery of compositionality through reflection also extends to the type-free non-classical context (Fischer et al., 2017, Corollary 1, Section 3.2):

THEOREM 4.4.3. $\mathcal{R}^2[\mathsf{TS}_0] \vdash \mathsf{PKF}$,

where the uniform reflection principle $\mathcal{R}$ is formulated as a rule instead of an axiom. The reflection principle used in the proof of Theorem 4.3.3 is the following:

$$\frac{\Rightarrow \mathit{Prov}^*_{\mathsf{TS}_0}\ulcorner \Gamma(\dot{x}) \Rightarrow \Delta(\dot{x}), \Phi(\dot{x}) \Rightarrow \Psi(\dot{x})\urcorner \quad \Gamma(x) \Rightarrow \Delta(x)}{\Phi(x) \Rightarrow \Psi(x)} \quad (\mathcal{R})$$

where the $\mathit{Prov}^*_{\mathsf{TS}_0}$ expresses that the rule from $\Gamma(x) \Rightarrow \Delta(x)$ to $\Phi(x) \Rightarrow \Psi(x)$ is an admissible rule of TS_0.

Again, following the general idea that the acceptance of a theory generates the possibility to accept stronger theories of which the acceptability is implicit in the acceptance of the weaker theory, we can see that, if we commit ourselves to disquotational typed (type-free) truth theories, then we *implicitly* commit ourselves to compositional typed (type-free) truth theories.[14]

However, iterating reflection does not only recover compositional principles from disquotational ones. As it is shown in Leigh (2016, Theorem 1.4, Theorem 1.5, Section 1), iterating the process of reflection also increases the amount of provable transfinite induction.

We fix a natural notation system for ordinals up to and not including Γ_0 that can be presented as an *elementary ordinal notation system* in the sense of Rathjen (1997), and call it $\mathbf{O}$. Then both $\mathbf{O}$ and the ordering relation $\prec$ on ordinals defined by elements of $\mathbf{O}$ are definable in first-order arithmetic.

DEFINITION 4.4.4 [Transfinite induction]. Let A be a formula.

1. Transfinite induction for A up to any $\alpha < \Gamma_0$, denoted as $TI(A, \alpha)$, is the formula

   $$\mathit{Prog}(\lambda x A) \rightarrow A(t),$$

   where t is a notation in $\mathbf{O}$ for α, and $\mathit{Prog}(\lambda x A)$ states that A is progressive along $\prec$, i.e.,

   $$\forall x \in \mathbf{O}[\forall y \prec x A(y/x) \rightarrow A(x)].$$

2. For a language $\mathcal{L}$ and ordinal $\alpha < \Gamma_0$, the schema of transfinite induction up to α, $TI_{\mathcal{L}}(< \alpha)$, is the collection of formulae

   $$\{TI(A, \beta) \mid A \in \mathcal{L} \wedge \beta < \alpha\}.$$

DEFINITION 4.4.5. For a theory S and an (elementary) ordinal κ, let S^{κ} denote the extension of S by $TI_{\mathcal{L}}(< \kappa)$.

DEFINITION 4.4.6. For a theory S and (elementary) ordinal κ, let $RFN^{\kappa}[S]$ denote the theory EA + κ times iterated uniform reflection over S.

Now suppose that we start from a disquotational theory that is based on the weak arithmetical theory EA instead of on full PA. In particular, let TB_0, TFB_0 be just like TB, TFB, respectively, except that they have EA instead of PA as their arithmetical background component. Then we have (Leigh, 2016, Theorem 1.4):

THEOREM 4.4.7. For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:

1. $\mathsf{CT}^{\varepsilon_\kappa} = RFN^{1+\kappa}[\mathsf{TB}_0]$;
2. $\mathsf{KF}^{\varepsilon_\kappa} = RFN^{1+\kappa}[\mathsf{TFB}_0]$.

Moreover, if we look at the consequences of these theories for the restricted language $\mathcal{L}_{\mathsf{PA}}$, then we have the following result (Leigh, 2016, Theorem 6.24):

THEOREM 4.4.8. For all $\kappa \in \mathbf{O}$ with $\kappa > 0$:

1. If A is an $\mathcal{L}_{\mathsf{PA}}$-formula provable in $RFN^{1+\kappa}[\mathsf{TB}_0]$, $RFN^{\kappa}[\mathsf{CT}]$, or $\mathsf{CT}^{\varepsilon_\kappa}$, then A is a theorem of $\mathsf{EA} + TI(< \varepsilon_{\varepsilon_\kappa})$.
2. If A is an $\mathcal{L}_{\mathsf{PA}}$-formula provable in $RFN^{1+\kappa}[\mathsf{TFB}_0]$, $RFN^{\kappa}[\mathsf{KF}]$, or $\mathsf{KF}^{\varepsilon_\kappa}$, then A is a theorem of $\mathsf{EA} + TI(< \varphi_{\varepsilon_\kappa}(0))$.

The situation in the non-classical settings is similar. In Fischer et al. (2017, Proposition 3.3.3) it is shown that two acts of uniform reflection over the theory called *Basic*, which is EA formulated in the language with the truth predicate $\mathcal{L}_T$ with an induction rule for Δ_0^0-formulae and in *BDM* logic,[15] proves the principle of transfinite induction for the language $\mathcal{L}_T$ for all ordinals up to and including ω^{ω}:

THEOREM 4.4.9. $\mathcal{R}^2[\mathit{Basic}] \vdash TI_{\mathcal{L}_T}(\omega^{\omega})$.

Iterating reflection into the transfinite proves even more transfinite induction, as it is shown in Fischer et al. (2017, Corollary 3, Subsection 3.3):

THEOREM 4.4.10. $\mathcal{R}^{\omega}[\mathit{Basic}] \vdash TI_{\mathcal{L}_T}(< \omega^{(\omega^2)})$.

In other words, transfinitely many iterations of uniform reflection over a non-classical truth theory still proves much less transfinite induction than just two iterations of uniform reflection over classical logic. This is because *Basic* is formulated in the non-classical logic *BDM*. Some interpret this as a defect of (truth) theories in non-classical logic: they cannot reproduce (possibly not even with reflection) the same mathematical reasoning that classical theories offer (Halbach and Nicolai, 2018).

4.4.3 Global Reflection

The reflection principles involved in the theorems that have been discussed so far merely *approximate* the correct way of formalising soundness. This correct way of formalising soundness was already articulated by Kreisel and Lévy (1968):[16] it is the *Global Reflection Principle* (*GRP*), which can be defined as follows:

DEFINITION 4.4.11. The global reflection principle for a theory S, denoted as GRP_S, is the formula

$$\forall x[Sents(x) \wedge Provs(x) \rightarrow T(x)].$$

From a "typed" perspective on truth, one mark against global reflection is the fact that already one iteration of global reflection over a typed truth theory violates typing. But from a "type-free" perspective, GRP_S may be a plausible way of making the commitment that is implicit in accepting type-free truth theory S explicit.

If we look at theories formulated in non-classical logic such as TS_0, then we get (Fischer et al., 2017, Proposition 1):

THEOREM 4.4.12. The uniform reflection principle and the global reflection principle are provably equivalent over TS_0.

Since TS_0 is arithmetically sound when uniform reflection is added, global reflection over TS_0 is likewise sound. Moreover, this procedure can then consistently be repeated. In other words, TS_0 is *coherent* with its implicit commitment.

The situation in classical logic is different. The closure of classical truth theories under *GRP* for the whole language often forces some kind of inconsistency. This can either be outright inconsistency, or what is called *internal inconsistency*, i.e., the existence of a sentence φ, such that it is provable that $T\varphi \wedge \neg\varphi$. In Halbach (2014) it is shown that FS is inconsistent with $GRP_{\mathsf{FS}}[\mathsf{FS}]$; in Fischer et al. (forthcoming, p. 8) it is observed that the standard axiomatisation of KF is internally inconsistent with $GRP_{\mathsf{KF}}[\mathsf{KF}]$.[17] Indeed, KF is internally inconsistent even with GRP_{FOL}, where *FOL* is first-order logic formulated in $\mathcal{L}$. This phenomenon has been interpreted by some to indicate that standard theories of type-free truth in classical logic are implicitly incoherent.

In our discussion so far, we have taken the implicit acceptance of or commitment to a theory S to be made explicit via the addition (and iteration) of reflection principles. However, in the previous approaches the epistemic notion of acceptance had been only made indirectly explicit via the notions of provability and truth. In what follows, we will discuss a different procedure to make the implicit acceptance of a theory explicit.

4.5 Reflecting on Acceptance

Instead of taking for granted the idea that proof-theoretic reflection principles express trust or acceptance, one might decide to investigate the notion of acceptance of a given theory T directly, with the aim of spelling it out without the help of reflection principles or the concept of truth. In this case, the concept of *accepting a theory* T should be made precise.

An attempt at doing this was made by Galinon (2014), who focusses on the weakest reflection principle: consistency. In his explication of the reflection process, Galinon uses two key principles. The first of these is the *Principle of* (*first-person*) *Responsibility*:

> If a rational agent accepts a collection T of propositions, then she must accept "T is acceptable".
>
> (Galinon, 2014, p. 328)

Second, he endorses the following principle:

> A rational agent must accept that if a collection propositions is acceptable, then that collection is coherent.
>
> (Galinon, 2014, p. 325)

Using these principles, Galinon (2014) develops the following argument for the acceptance of consistency statements. Suppose a rational agent unconditionally accepts a mathematical theory T. Then, using the Principle of Responsibility, she must accept "T is acceptable". And from this, using the second principle, the agent is rationally obliged to infer that T is consistent (p. 329).

In this chapter we cannot do justice to the philosophical complexity of the issues that are relevant here, so we restrict ourselves to a brief discussion of one of Galinon's key principles.[18] The Principle of Responsibility seems a demanding requirement. One might wonder if reflecting on one's acceptance of T might not, in some cases, lead one to abandon rather than to accept one's acceptance of T. Of course this does not exclude that there are cases where we reflect on our acceptance of a theory T and *legitimately* conclude that T is acceptable. If that is so, then maybe Galinon and Feferman go too far when they claim that one is *rationally obliged* to accept reflection principles for theories that one accepts. Perhaps the claim should rather be that there are cases where an agent is *rationally permitted* to accept, on the basis of reflecting on a theory T that she already accepts, reflection principles for T.[19]

Cieśliński (2018, 2017) provides an alternative analysis of reflection on one's mathematical beliefs. He first spells out which informal

notion of acceptance of S is relevant, and then proposes the following informal understanding of acceptance of S:

> For any sentence φ, if I believed that φ has a proof in S and I had no independent reason to disbelieve φ, then I would be ready to accept φ.
> (Cieśliński, 2018, p. 1087, notation has been adapted to ours)

Cieśliński (2018) provides an axiomatic theory of believability that employs the informal notion of acceptance presented in the quote above. He makes this notion of acceptance of S explicit by extending S to a new theory S^+, which captures the informal notion expressed above. He does this by presenting a theory of *believability*, which extends the theory S that we accept with a fresh predicate $B(x)$ for believability and with axioms that govern its behaviour.

The thought is that when a person reflects on the implicit commitments involved in her acceptance of a theory K, she comes to accept a theory of believability $Bel(K)^-$ over K.[20] Cieśliński explains how this process is structured, and he spells out $Bel[K]^-$ as an axiomatic theory (Cieśliński, 2018, p. 254).

Suppose we start with a theory K, formulated in a language $\mathcal{L}_K$. Let $\mathcal{L}_{K,B} = \mathcal{L}_K \cup \{B\}$. And let KB be the theory which is just like K except that its schemata range over all formulas of $\mathcal{L}_{K,B}$. The theory of believability $Bel[K]^-$ is an extension of KB with the following axioms and rules (Cieśliński, 2018, Definition 13.4.1):[21]

(Ax_1) $\forall \psi \in \mathcal{L}_{K,B}[Prov_{KB}(\psi) \rightarrow B(\psi)]$,
(Ax_2) $\forall \varphi, \psi \in \mathcal{L}_{K,B}[(B(\varphi) \wedge B(\varphi \rightarrow \psi)) \rightarrow B(\psi)]$,

$$(NEC)\ \frac{\vdash \varphi}{\vdash B(\varphi)} \qquad (GEN)\ \frac{\vdash \forall n :\ B(\varphi(n))}{\vdash B(\forall x \varphi(x))}$$

Let us now apply Cieśliński's general theory to a concrete example. Consider the "weak" typed disquotational truth theory TB^-, which is like the disquotational theory TB except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept TB^-. Then if we make the acceptance of TB^- explicit via $Bel[\mathsf{TB}^-]^-$, we recover compositional principles for typed truth (Cieśliński, 2018, p. 264):

THEOREM 4.5.1. $Bel\ [\mathsf{TB}^-]^- \vdash B(\mathsf{CT})$,

where $B(\mathsf{CT})$ consists of all sentences $B(\varphi)$ such that φ is an axiom of CT. In particular we obtain the believability of mathematical induction for $\mathcal{L}_T$ from a situation where we only accepted induction for $\mathcal{L}_{\mathsf{PA}}$.

Analogous results hold in type-free settings. Consider the typed disquotational truth theory TFB^-, which is like TFB except that the truth predicate is not allowed to occur in the induction schema. Suppose that we accept TFB^-. Then if we make the acceptance of TFB^- explicit

via $Bel[\mathsf{TFB}^-]^-$, we recover compositional principles for type-free truth (Cieślińskip, 2018, p. 266):

THEOREM 4.5.2. $Bel\ [\mathsf{TFB}^-]^- \vdash B(\mathsf{KF})$.

So, taking stock: if we are committed to typed (type-free) disquotational truth and if this commitment is made explicit via a theory of believability, then this theory proves that the compositional principles for typed (type-free) truth are indeed believable.

The believability theory over the disquotational truth theory does not contain a factivity principle or rule ("*B*-Out") for the believability predicate *B*. Indeed, the inference from the believability of a statement to the statement itself is a *defeasible* rule. For this reason, we do not have $Bel[\mathsf{TB}^-]^- \vdash \mathsf{CT}$. Nonetheless, according to Cieśliński's informal definition of acceptance of a theory, this then means that, in the absence of independent reasons for disbelieving compositional principles of typed (type-free) truth, we should be ready to accept them. In this sense Cieśliński's results provide and argument for the thesis that our commitment to compositional truth principles is not greater than the commitment to disquotational truth principles.

It would take us too far to give a detailed evaluation of Cieśliński's position, so again we confine ourselves to a few cautiously critical remarks. Cieśliński argues that processes of reflection on one's acceptance of a theory *K* can be described as proofs in a believability theory $Bel[K]^-$ for *K*. But it is not clear that all principles of $Bel[K]^-$ are in all circumstances correct. In particular, for the same reasons as why Galinon's Principle of Responsibility might not in all cases be correct, it is not clear that axiom Ax_1 of $Bel[K]^-$ is always true. Might there not be circumstances where the agent starts out by accepting *K*, but by reflecting on *K* comes to abandon parts of *K*—perhaps because in the reflective process she comes to realise that *K* is actually quite strong—rather than to judge that *K* is believable? It seems to us that a deeper phenomenological analysis of reflection processes than has been given thus far is needed to decide this question.[22]

4.6 Reflective Processes

The reflection principles that we have discussed in the previous sections take the form of conditional statements. These conditional statements express the result of *reflective processes*, which have an argumentative structure. They aim systematically to draw out consequences from hypothetical situations. The resulting formal reflection principles intend to express a necessary connection between the "input" of a reflection process and the "output" of that process.

Because of this, reflection principles have played a role in debates in the foundations of mathematics about the justification of mathematical theories. However, the extent to which proof theoretic reflection

principles can play a justificatory role in this context, is contested. On the one hand, Horsten and Leigh (2016) argue that if accepting a theory S is justified, then accepting a proof-theoretic reflection principle for S is also epistemically warranted.[23] On the other hand, Dean (2014) urges caution. He argues that even in a context where accepting a theory S is justified, justification for proof-theoretic reflection principles for S must be obtained before we are warranted to accept them. Getting to the bottom of this requires deeper philosophical reflection on the nature of proof-theoretic reflection than has been carried out so far. Indeed, we believe that reflection processes that underpin formal reflection principles deserve more attention from philosophers of mathematics than has hitherto been accorded to them.

In this chapter we have concentrated on reflection principles that are connected with reflective processes that start from hypothetical facts about provability in a formal system. Some such reflective processes terminate in propositions that attribute truth to statements (Section 4.4); others terminate in propositions about rational believability (Section 4.5). However, there exists a class of reflection principles that are related to reflective processes that do not terminate in, but rather start from, hypothetical propositions that attribute truth to statements. Such principles are called *set theoretic reflection principles*.[24]

It can be argued that proof-theoretic reflection principles are related to set theoretic reflection principles.[25] Consider, for instance, local reflection for a theory S. For theories S that prove the completeness theorem, Rfn_S is equivalent to the scheme

$$\varphi \rightarrow \exists M : M \models S + \varphi,$$

which is a set theoretic reflection principle.[26] Of course this principle is so weak that it is hardly mentioned in discussions of set theoretic reflection. Indeed, the weakest set theoretic reflection principle that is widely discussed is Montague-Levy reflection. The Montague-Levy reflection principle is provable in ZFC. Nonetheless, the fact that it has proof-theoretic strength is shown by the fact that over the remaining axioms of ZFC, it is equivalent to the axiom of infinity plus the axiom of replacement.

It is commonly assumed that "set theoretic reflection principles can be very strong, but proof-theoretic reflection principles are always weak". But in an absolute sense, this is not quite correct, as can be seen as follows.[27] The axiom MC, which expresses that there exists a measurable cardinal, can be expressed as an embedding principle (the existence of a non-trivial embedding from Gödel's L to L). And such embedding principles are often (but not always) informally described as set theoretic reflection principles. But even though ZFC + MC proves the consistency of ZFC, it is easy to see that ZFC + MC $\nvdash$ ZFC + Rfn_{ZFC}. So there is a sense in which even local reflection is strong.

The discussion of set theoretic reflection principles falls outside the scope of this chapter. The same holds for the discussion of the nature of our epistemic warrant for set theoretic reflection principles. We restrict ourselves here to observing that it should not automatically be assumed that our epistemic warrant for even moderately strong set theoretic reflection principles is of the same nature as our warrant for proof theoretic reflection principles. We have seen that our warrant for a proof theoretic reflection principle for a theory *S* is often taken somehow to be implicit in our warrant for *S*. But it is hard to see how something like this might be true for set theoretic reflection principles, since even the modest ones (such as Montague-Levy reflection) make no explicit reference to a background theory.

Acknowledgments

Thanks to Kentaro Fujimoto, Karl-Georg Niebergall, and the editors of this book for valuable comments on an early version of this chapter. The second author also thanks the South, West & Wales Doctoral Training Partnership for (grant reference AH/L503939/1) the financial and ideal support for his research project *Theories of Truth and Foundations for Mathematics: Epistemic Warrants and Reflection Principles*.

Notes

1. For more on the philosophical significance of the use of non-canonical definitions, see Franzén (2004a,b).
2. Feferman's completeness theorem can be strengthened. Using the notion of *smooth progression* developed in Beklemishev (1995) it can be shown that the length of this path can be shortened to ω^{ω^2+1}. For an idea of the proof of this improvement, see Franzén (2004b).
3. It is also known that completeness depends on the choice of the path in $\mathcal{O}$. Feferman and Spector (1962) showed for instance that there are paths *through* $\mathcal{O}$, such that corresponding uniform reflection progression does not even prove every true Π^0_1 sentence.
4. *H* is the extension of first-order *Peano Arithmetic*, PA, with Kreisel's *hyperarithmetic comprehension rule* (*HCR*): see Fererman (1964, p. 17) for Feferman's original formulation of the system *H* and of *HCR*. *R* is a system of *Ramified analysis*: see Feferman (1964, pp. 21–22).
5. See Feferman (1964, p. 23, Theorem 6.10) for Feferman's original formulation of the theorem.
6. The claim has been made in Feferman (1964). Thanks to Kentaro Fujimoto for pointing this out to us.
7. For a discussion of the role of idealisation in the epistemological discussion of transfinite progressions of formal theories, see Antonutti Marfori and Horsten (2019).
8. In TB the induction scheme is extended to allow also formulae that contain the truth predicate.
9. Of course there are also other non-classical logics that one can opt for, such as Strong, Weak Kleene Logic, etc. For background on these non-classical logics, see for instance Priest (2008).

10. See Halbach (2014, Section 19.3).
11. See Leigh (2016, Section 5).
12. See Halbach (2014, chapter 8).
13. See Halbach and Horsten (2006) and Halbach (2014, chapters 15, 16).
14. Although, as we pointed out, in the classical case a restricted version of compositionality is obtained, starting with positive biconditionals.
15. See Fischer et al. (2017, Section 2.2) for more details.
16. See Section 4.4.1 above.
17. No claim of originality for this result is made in this chapter. Indeed, this elementary observation is folklore.
18. Galinon argues for the Principle of Responsibility on the basis of norms of rationality (Galinon, 2014, Section 7), and he argues for the second principle on the basis of a "Gödelian Dutch book argument" (Galinon, 2014, Section 5).
19. This stance is taken in Fischer et al. (forthcoming).
20. Cieśliński also considers a believability theory $Bel(K)$ over K that is stronger than $Bel(K)^{-}$. We do not discuss this stronger theory $Bel(K)$ here.
21. In the interest of readability we are sloppy with the Gödel coding in what follows.
22. An attempt to provide such an analysis is given in (Horsten, forthcoming).
23. In this connection, see also Fischer et al. (forthcoming).
24. In the literature on predicativity, reflection principles are considered that take facts about *definability* as input: see Lorenzen (1958). Discussion of these principles falls outside the scope of this chapter.
25. Kreisel and Levy are undecided whether proof theoretic and set theoretic reflection are related: see Kreisel and Lévy (1968, p. 101).
26. Thanks to Kentaro Fujimoto for putting it this way.
27. Thanks to Karl-Georg Niebergall for pointing this out to us.

References

Antonutti Marfori, M. and Horsten, L. (2019). Human-effective computability. *Philosophia Mathematica*, 27(1): 61–87.

Beklemishev, L. (1995). Iterated local reflection versus iterated consistency. *Annals of Pure and Applied Logic*, 75(1): 25–48.

Cieśliński, C. (2017). *The Epistemic Lightness of Truth. Deflationism and its Logics*. Cambridge University Press.

Cieśliński, C. (2018). Minimalism and the generalisation problem: On Horwich's second solution. *Synthese*, 195: 1077–1101.

Dean, W. (2014). Arithmetical reflection and the provability of soundness. *Philosophia Mathematica*, 23(1): 31–64.

Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *The Journal of Symbolic Logic*, 27(3): 259–316.

Feferman, S. (1964). Systems of predicative analysis. *The Journal of Symbolic Logic*, 29(1): 1–30.

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56(1): 1–49.

Feferman, S. and Spector, C. (1962). Incompleteness along paths in progressions of theories. *Journal of Symbolic Logic*, 27(4): 383–390.

Fischer, M., Nicolai, C., and Horsten, L. (2017). Iterated reflection overfull disquotational truth. *Journal of Logic and Computation*, 27(8): 2631–2651.

Fischer, M., Nicolai, C., and Horsten, L. (forthcoming). Hypathia's silence. Truth, justification, and entitlement. *Noûs*.

Franzén, T. (2004a). *Inexhaustibility: A Non-exhaustive Treatment*. Association of Symbolic Logic.

Franzén, T. (2004b). Transfinite progressions: A second look at completeness. *The Bulletin of Symbolic Logic*, 10(3): 367–389.

Galinon, H. (2014). Acceptation, cohérence et responsabilité. In *Liber Amicorum Pascal Engel*. J. Dutant, D. Fassio, and A. Meylan, editors, Université de Genève.

Halbach, V. (2001). Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66(4): 1959–1973.

Halbach, V. (2014). *Axiomatic Theories of Truth*. Cambridge University Press.

Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, 71(2): 677–712.

Halbach, V. and Nicolai, C. (2018). On the costs of nonclassical logic. *Journal of Philosophical Logic*, 47: 227–257.

Horsten, L. (forthcoming). On reflection. *Philosophical Quarterly*.

Horsten, L. and Leigh, G. E. (2016). Truth is simple. *Mind*, 126(501): 195–232.

Horwich, P. (1990). *Truth*. Clarendon Press.

Kreisel, G. (1958). Ordinal logics and the characterization of informal concepts of proof. In *Proceedings of the International Congress of Mathematicians* (1958), pages 289–299. J. A. Todd, editor, Cambridge University Press, 1960.

Kreisel, G. and Lévy, A. (1968). Reflection principles and their use for establishing the complexity of axiomatic systems. *Mathematical Logic Quarterly*, 14: 97–142.

Leigh, G. E. (2016). Reflecting on truth. *IFCoLog Journal of Logics and their Applications*, 3: 557–593.

Lorenzen, P. (1958). Logical reflection and formalism. *The Journal of Symbolic Logic*, 23(3): 241–249.

Priest, G. (2008). *An Introduction to Non-Classical Logic From If to Is*. Cambridge University Press.

Rathjen, M. (1997). The realm of ordinal analysis. In Cooper, S. B. and Truss, J. K., editors, *Sets and Proofs*, pages 219–279. Cambridge University Press.

Schütte, K. (1964). Eine Grenze für die Beweisbarkeit der transfiniten Induktion in der verzweigten Typenlogik. *Archiv für Mathematische Logik und Grundlagenforschung*, 7: 45–60.

Schütte, K. (1965). Predicative well-orderings. In Crossley, J. and Dummett, M., editors, *Formal Systems and Recursive Functions*, volume 40 of Studies in Logic and the Foundations of Mathematics, pages 280–303. Elsevier.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, s2–45(1): 161–228.

Zicchetti, M. (2020). *Truth, Trustworthiness and Reflection*. Submitted for publication.