Feferman and the Truth

Andrea Cantini, Kentaro Fujimoto and Volker Halbach

Abstract We outline some of Feferman's main contributions to the theory of truth and the motivations behind them. In particular, we sketch the role truth can play in the foundations of mathematics and in the formulation of reflection principles, systems of ramified truth, several variants of the Kripke–Feferman theory, a deflationist theory in an extension of classical logic, and the system for determinate truth.

Keywords Axiomatic theories of truth · Ramified truth · Kripke–Feferman theory · Reflective closure · Determinate truth

2010 Mathematics Subject Classification 03F03 · 03F25 · 03F35 · 3F40 · 03A05.

1 Truth and the Foundations of Mathematics

Theories of comprehension and satisfaction are closely intertwined. To say that a is an element of the class $\{x : \varphi(x)\}$ for a formula $\varphi(x)$ seems tantamount to saying that the formula $\varphi(x)$ is satisfied by a or that the formula $\varphi(x)$ is true of a. Using this observation, one can reduce class theories to theories of satisfaction or those

Dedicated to the memory of Solomon Feferman.

A. Cantini (✉)
Dipartimento di Lettere e Filosofia, Universitá di Firenze, via Bolognese 52, 50139 Firenze, Italy
e-mail: andrea.cantini@unifi.it

K. Fujimoto
Department of Philosophy and School of Mathematics, University of Bristol, Cotham House, Bristol BS6 6JL, England
e-mail: fujimotokentaro@gmail.com

V. Halbach
University of Oxford, New College, Oxford OX1 3BN, England
e-mail: volker.halbach@philosophy.ox.ac.uk

G. Jäger and W. Sieg (eds.), *Feferman on Foundations*, Outstanding Contributions to Logic 13, DOI 10.1007/978-3-319-63334-3_11

of truth: Quantification over classes is translated to quantification over formulae or propositional functions. The basic strategy can be traced back at least to Russell [51] (see Schindler [53]).

In the good old days of logical positivism, however, the concept truth was considered with suspicion. It had been stained by too many dodgy philosophical theories. So the reduction of mathematical theories to theories of satisfaction or truth didn't look too attractive. Why should one want to replace a respectable mathematical theory with a 'philosophical' theory of truth or satisfaction? As foundational concepts membership and classes seemed much more suitable than truth and satisfaction.

More desirable was a reduction in the other direction, that is, the reduction of truth and satisfaction to a theory of classes. Consequently many philosophers were much more interested in Tarski's [58] definition of truth in a type theory over a theory of syntax. This reduction vindicated truth as a respectable notion for philosophers like Popper (see Leitgeb [43]).[1] Tarski added further reasons not to treat truth or satisfaction as a primitive notions to the neopositivist qualms: He thought that either the theory of truth will lack 'deductive power', if truth is axiomatized by the typed T-sentences, or the axioms for truth will have an 'accidental character' (see Halbach [33, Sect. 7]). Hence mathematical theories with comprehension axioms retained their conceptual priority over axiomatic theories of truth in foundational discussion. Of course some work on truth was done, but truth wasn't much used in the foundations of mathematics.

It took a long time for truth to recover from the neopositivist qualms and Tarski's verdict that any axiomatization of truth is either weak or arbitrary. Even philosophers beyond any suspicion of neopositivist convictions have endorsed the conceptual priority of some class theory over a theory of truth or satisfaction. For instance, the type-free truth theories advanced by Kripke [42], Gupta [30] or Herzberger [35] follow the pattern of Tarski's theory in a crucial aspect. On their accounts set theory is employed as the basic framework; then a semantics for a language with the truth predicate is defined within set theory. The concept of truth these authors are interested in is more worrisome than Tarski's, because type-free truth is prone to paradox. Consequently it is not a surprise that these authors rely on set theory as ultimate framework and show that semantics for languages with a type-free truth predicate can be developed within set theory. Again a theory of classes serves as the bedrock foundational framework on which a theory of truth can rest.

Truth, however, conceived as a primitive, undefined notion, does have a potential for use in the foundations of mathematics and the formal sciences. One use where the need for a truth predicate is obvious are the proof-theoretic reflection principles. As Gödel [29] had shown, a system S cannot prove even very weak consequences of the soundness of the theory. In particular, S cannot prove the consistency of S, if the latter is expressed in a natural way. However, by endorsing or accepting a system S, one is also committed to the soundness of S and thus to all consequences of its soundness such as the consistency of S. One can try to add statements expressing the soundness

[1] For those who have doubts about the success of Tarski's reduction, we add that worries had been raised early on and more recently by Field [22] and Putnam [48].

of S to the system S in order to obtain a system that features commitments implicit in the acceptance of S as explicit theorems. As a very early reaction to Gödel's incompleteness theorems, Turing [59] had tried to add such principles to S and to iterate this procedure even along transfinite ordinals.

Kreisel and Lévy [41, p. 98] describe a reflection principle for a system S as 'the formal statement stating the soundness of S'. As a soundness principle the consistency statement is very weak. More powerful are the local reflection principle $\mathrm{Bew}_{\mathsf{S}}(\ulcorner\varphi\urcorner) \rightarrow \varphi$ and the uniform reflection principle $\forall x\,(\mathrm{Bew}_{\mathsf{S}}(\ulcorner\varphi(\dot{x})\urcorner) \rightarrow \varphi(x))$. But even these schemata fall short of expressing full soundness.

Full soundness is the statement that all (closed) theorems of S are true. Kreisel and Lévy [41, p. 98] call a version of this the *global reflection principle*. They write:

> 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} […]

They go on to point out that such a truth definition doesn't exist because of Tarski's theorem on the undefinability of truth. Hence, under the usual assumptions, the 'intended reflection' cannot be expressed in the language of S. It is also hardly an option to pass from S to another stronger system containing S with the resources for defining a predicate T_{S}, because it is only soundness that is to be added to the system and not, for instance, more comprehension axioms or other axioms sufficient for defining T_{S}.

In order to avoid the addition of new mathematical resources for the definition of T_{S}, one can add a primitive symbol for truth in S. This approach goes directly against the neopositivist qualms against truth. Thus it doesn't come as a surprise that initially a primitive formal truth predicate wasn't used in the discussion about reflection principles, reflective closure, recursive progressions and related topics. Turing's early work was continued without the use of a truth predicate in the object language. Of course, Feferman is the single most influential author in the wake of Turing's approach. In particular, Feferman [9] contained amazing results on the iterated addition of proof-theoretical reflection principles in the sense of Kreisel and Lévy [41] to Peano arithmetic. These progressions of theories were all formulated in the language of arithmetic without any additions.

In the 1970s more logicians were less restrained by worries about truth and started working on formal truth theories with truth as a predicate in the object language. They also became bolder and turned to type-free notions of truth. They explored other ways of solving or blocking the paradoxes than Tarski's restrictive method involving the distinction between an object and a metalanguage. Especially after the publication of Kripke [42], more philosophers and logicians turned their attention to truth. Feferman recognized their potential for adding strong soundness claims and making explicit assumptions implicit in the acceptance of a theory.

2 Ramified Truth and Reflection

In 1979 Feferman gave a talk entitled *Gödel's incompleteness theorems and the reflective closure of theories* at the meeting of the Association of Symbolic Logic in San Diego in 1979 ([45] and [17, p. 3]). This paper, which was to become the paper [17], had been circulated for quite some time and other logicians including Reinhardt [50] and Cantini [5] had published on Feferman's ideas. Feferman's talk and the subsequent papers by him and others mark the return of truth as an undefined, primitive notion of foundational significance. Let $\mathcal{L}_{\mathsf{PA}}$ be the language of first-order arithmetic.

The starting point is the 'compositional' axioms for truth. Given a *system* satisfying certain natural conditions, axioms for a truth predicate T can be added that correspond to the clauses in Tarski's definition of satisfaction. When one works in a theory that can encode every object as a closed term within it one can dispense with satisfaction and use a unary truth predicate – at least if the theory satisfies certain further assumptions. Feferman [17] gave fairly general account that applies to a variety of theories such as set theory for instance (see Fujimoto [28]). This generality is significant and not just a trivial extension of the special case of Peano arithmetic as starting theory.

However, here we use Peano arithmetic as our base theory. This will allow us to keep the presentation sufficiently simple. Let $\mathcal{L}_T$ be $\mathcal{L}_{\mathsf{PA}} \cup T$. For PA the compositional axioms can be chosen as follows:

T1 $\forall s\, \forall t \left(T(s \doteq t) \leftrightarrow s^\circ = t^\circ\right)$ and similarly for other predicates other than $=$, except for the special predicate T
T2 $\forall x \left(\mathrm{Sent}_{\mathsf{PA}}(x) \rightarrow (T(\dot{\neg} x) \leftrightarrow \neg T(x))\right)$
T3 $\forall x\, \forall y \left(\mathrm{Sent}_{\mathsf{PA}}(x \mathbin{\dot{\wedge}} y) \rightarrow (T(x \mathbin{\dot{\wedge}} y) \leftrightarrow T(x) \wedge T(y))\right)$
T4 $\forall v\, \forall x \left(\mathrm{Sent}_{\mathsf{PA}}(\dot{\forall} v x) \rightarrow (T(\dot{\forall} v x) \leftrightarrow \forall t\, T(x[t/v]))\right)$

Here and in what follows we use quantifiers $\forall s$ and $\forall t$ to range over the codes (or Gödel numbers) of closed $\mathcal{L}_{\mathsf{PA}}$-terms; namely, the expression $\forall t$ is short for $\forall x\, (\mathrm{ClTerm}(x) \rightarrow \ldots)$, where $\mathrm{ClTerm}(x)$ represents the set of the codes of closed $\mathcal{L}_{\mathsf{PA}}$-terms. The symbol $^\circ$ is a representation in PA of a recursive function that takes a code of a closed $\mathcal{L}_{\mathsf{PA}}$-term and returns its value in the standard model; e.g., $\ulcorner 0 + 0 \urcorner^\circ = 0$. Hence the axiom T1 expresses that a closed equation $s = t$ is true iff the values of the closed terms s and t coincide. There won't be a function symbol $^\circ$ in the language, but the function can be expressed using suitable formulae. The formula $\mathrm{Sent}_{\mathsf{PA}}(x)$ represents that x is a code of a sentence of the language of PA. The symbol $\dot{\neg}$ is a representation in PA of a recursive function that takes a code of a sentence and returns the code of its negation; e.g., $\dot{\neg} \ulcorner 0 = 0 \urcorner = \ulcorner 0 \neq 0 \urcorner$. Hence the axiom T2 expresses that the negation of a sentence of PA is true iff the sentence is not true. The symbol $\dot{\wedge}$ is a representation in PA of a recursive function that takes two codes of sentences and returns the code of their conjunction; e.g., $\ulcorner 0 = 0 \urcorner \mathbin{\dot{\wedge}} \ulcorner 0 \neq 0 \urcorner = \ulcorner 0 = 0 \wedge 0 \neq 0 \urcorner$. Hence the axiom T3 expresses that the conjunction of two sentences of PA is true iff both of the conjuncts are true. The symbol $\dot{\forall}$ is a representation in PA of a recursive function that takes a code of

a variable and a code of a formula and returns the universal quantification of the formula with respect to the variable; e.g., $\forall\,(\ulcorner v\urcorner, \ulcorner\varphi\urcorner) = \ulcorner\forall v\varphi\urcorner$ for a variable v and a formula φ. The operation $x[y/z]$ yields the code of the result of substituting a term encoded by y for a variable encoded by z in a formula encoded by x; e.g., $\ulcorner\varphi(v)\urcorner[\ulcorner t\urcorner/\ulcorner v\urcorner] = \ulcorner\varphi(t)\urcorner$ for a formula φ, a term t, and a variable v. Hence axiom T4 expresses that a universally quantified sentence of PA is true iff all its substitution instances (with closed terms) are true. A more detailed explanation of the notation can be found in Halbach [33]. But we hope that our notation should be largely self-explanatory. There are also some minor deviations from Feferman's definition of this theory in [17, p. 14].

The axioms T1–T4 are adjoined to those of Peano arithmetic. Crucially, the induction schema is expanded to the new language with the truth predicate. The resulting theory T(PA) or also called CT (for 'compositional truth') proves the soundness of Peano arithmetic, that is, it proves the global reflection principle

$$\forall x \left(\mathrm{Sent}_{\mathsf{PA}}(x) \wedge \mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow T(x) \right) \tag{2.1}$$

Thus the expressive resources needed for stating the soundness of PA already imply the soundness of PA.[2] The global reflection principle doesn't have to be added as an additional axiom. As we have discussed, the uniform reflection principle is derived from (2.1) with the help of the axioms T1–T4; furthermore, CT is strong enough to derive the iterated reflection principle along any constructive ordinal α provably well-founded in CT in the sense that all the $\mathcal{L}_T$-instances of transfinite induction along α are derivable in CT.

If the commitment to the soundness of PA is implicit in the acceptance of PA, then also the resources needed for expressing the soundness claim are implicit in the acceptance of PA. Thus the acceptance of PA commits one to CT. Thus CT makes explicit some implicit commitments of the acceptance of PA. However, the implicit commitment in the acceptance of PA is not exhausted by CT and (2.1), and the iteration procedure further continues. For, once CT is explicitly accepted, one is then committed to the soundness of CT and thus to a truth predicate for CT, which is needed to express the soundness of CT. To this end one can add a further truth predicate T_1 that applies to all sentences formulated in the language of arithmetic expanded by T. The new predicate T_1 is then axiomatized in the same way as T except that T is treated as one of the non-special predicate symbols in T1. Moreover in T2–T4 quantification over sentences of the arithmetical language is replaced with quantification over sentences of the arithmetical language with T. This procedure can be iterated and an axiomatization of Tarski's hierarchy of languages is obtained. The exact specification of the procedure requires some more detail; but a general

[2] CT contains the expanded induction schema, and this expansion is indeed crucial in deriving (2.1), since CT without the expanded induction schema is conservative over PA and thus does not yield (2.1). The question whether the expanded induction schema is an essential part of 'the expressive resources needed for stating the soundness of PA' is a subtle issue and gave rise to lively debates in the context of deflationism; see a debate between Shapiro [56] and Field [23] for instance.

recipe for the transition from one level to the next can be specified. The procedure fits Feferman's [9, p. 274] description of a reflection principle:

> By a *reflection principle* we understand a description of a procedure for adding to any set of axioms A certain new axioms whose validity follow from the validity of the axioms A and which formally express, within the language of A, evident consequences of the assumption that all the theorems of A are valid.

This characterization of reflection principles precedes Feferman's work on truth by more than one and a half decades.

The addition of truth predicates and accompanying axioms can be iterated into the transfinite. A truth predicate T_λ at a limit level λ is axiomatized in the same way as the truth predicates at successor level with the exception that an axiom is added that says that a sentence is true iff it is true at one of the previous levels. For details see Halbach [33, Sect. 9.1] and Fujimoto [28].

The need for an ordinal notation system is obvious: The truth predicates need to be indexed and all levels of the hierarchy need to be axiomatized. Technically it is no problem to define theories up to any recursive ordinal. To this end one can pick a path through Kleene's $\mathcal{O}$. However, this would betray the original motivation for considering these iterated theories of truth: their purpose is to make explicit assumptions implicit in the acceptance of PA. Using our general mathematical machinery we can prove that there is a well-founded ordering of natural numbers whose order-type is any ordinal below the first non recursive ordinal ω_1^{CK}. But this theorem is not implicit in the acceptance of PA in any way. In particular, PA doesn't prove that these orderings are well-founded. Kreisel [40] suggested to employ autonomous iterations in situations of this kind. That is, one iterates a procedure of this kind in unfolding the implicit commitments in accepting a theory only if the theory in question can prove transfinite induction up to that ordinal level.

Since PA proves transfinite induction for any ordinal up to ε_0, the truth theories are iterated up to that point. This new theory, however, proves transfinite induction for longer wellorderings. Hence the truth theories are iterated even further until a point is reached where the hierarchy of truth theories is iterated to a point Γ_0 that proves transfinite induction for all ordinals smaller than Γ_0. This ordinal Γ_0 is the so-called Feferman–Schütte ordinal that had figured prominently in Feferman's earlier work [10] on predicativity.

The iterated truth theories very much resemble the systems of predicative analysis, which had been studied thoroughly by Feferman [10] and Schütte [54] in the 1960s. Thus, in a sense, the results on iterated truth theories are formally not extremely exciting. Perhaps this is the reason why they figure less prominently in the published paper [17] than in the draft version [16].

From a foundational point of view, however, we think that the iterated truth theories are significant. They are a very convincing way of carrying out the programme of determining the reflective closure of PA, that is, of characterizing the theory that makes explicit what is implicit in the acceptance of PA.

The formulation of the systems of iterated truth is technically awkward. The specification of the language already requires an ordinal notation system. Then the

motivation of the terminal ordinal ε_0 or Γ_0 relies on some deeper results. Moreover, it is highly specific to PA.

Feferman has made various attempts at characterizing the reflective closure of theories in a more elegant way. The reasons for seeking a more succinct characterization are not only of an aesthetic nature. A method of defining the reflective closure of a theory that is less reliant on ordinal notation systems and an explicit appeal to proof-theoretic techniques and notions, should also be more generally applicable; moreover, it would also be philosophically less prone to the objection that it depends on arbitrary stipulation; a more elegant system would depend on a 'natural' ordinal notation system and arithmetization.

Feferman has tried various methods for characterizing the reflective closure of a theory. The various approaches should not be seen so much as competing but rather as different characterization of the same concept. The situation is similar to that in recursion theory: The different characterizations of computability do not exclude each another, rather their equivalence assures us that we have found a stable concept.

3 Kripke–Feferman

The first method of characterizing the reflective closure of PA is fairly close to the iterated truth theories. But it shuns already the need for an ordinal notation system. Somewhat metaphorically speaking, the ordinals emerge from the theory itself and are not imposed on it from the outside.

The theory has been dubbed the *Kripke–Feferman theory* or KF for short. Presumably Reinhardt was the first to publish on the theory in [49, 50] and by the time Feferman published his paper [17], the label KF had already been established. Unfortunately, since no authoritative version had been published by Feferman, different authors formulated KF in slightly different ways. Here we try to stick to the formulation chosen by Feferman; but we use a slightly different notation.

Feferman [17] formulated KF in the language of arithmetic augmented with two new unary predicates T and F for truth and falsity. Let $\mathcal{L}_{\mathsf{KF}} := \mathcal{L}_{\mathsf{PA}} \cup \{T, F\}$ denotes the language of KF. As the axiom K4 below indicates, the falsity predicate F is actually not necessary and could be understood as defined notion, because the falsity of a sentence coincides with the truth of its negation. Hence for the sake of simplicity, we will identify $\mathcal{L}_{\mathsf{KF}}$ with $\mathcal{L}_T$. The axioms of KF comprise those of PA and the following truth-theoretic axioms:

K1 $\forall s\, \forall t \left(\left(T(s \doteq t) \leftrightarrow s^\circ = t^\circ \right) \wedge \left(F(s \doteq t) \leftrightarrow s^\circ \neq t^\circ \right) \right)$, and similarly for other predicates other than $=$, except for the special predicate T;

K2 $\forall s \left((T(\dot{T}\, s) \leftrightarrow T(s^\circ)) \wedge (F(\dot{T}\, s) \leftrightarrow F(s^\circ)) \right)$;

K3 $\forall s \left((T(\dot{F}\, s) \leftrightarrow F(s^\circ)) \wedge (F(\dot{F}\, s) \leftrightarrow T(s^\circ)) \right)$;

K4 $\forall x \left(\mathrm{Sent}_{\mathsf{KF}}(x) \rightarrow (T(\dot{\neg} x) \leftrightarrow F(x)) \wedge (F(\dot{\neg} x) \leftrightarrow T(x)) \right)$;

K5 $\forall x\, \forall y \left(\mathrm{Sent}_{\mathsf{KF}}(x \mathbin{\dot{\wedge}} y) \rightarrow (T(x \mathbin{\dot{\wedge}} y) \leftrightarrow T(x) \wedge T(y)) \wedge (F(x \mathbin{\dot{\wedge}} y) \leftrightarrow F(x) \vee F(y)) \right)$;

K6 $\forall v \forall x \left(\mathrm{Sent}_{\mathsf{KF}}(\dot{\forall} v x) \rightarrow (T(\dot{\forall} v x) \leftrightarrow \forall t\, T(x[t/v])) \wedge (F(\dot{\forall} v x) \leftrightarrow \exists t\, F(x[t/v]))\right)$.

The axiom K1, exactly like T1, says that a closed equation $s = t$ is true (false) iff the value of the terms s and t agree (disagree, resp.). The formula $\mathrm{Sent}_{\mathsf{KF}}(x)$ expresses that x is a sentence of the language of (i.e. $\mathcal{L}_T$); hence the axiom K5 and K6 express essentially the same compositional axioms as T3 and T4 but extended to sentences of the larger language $\mathcal{L}_T$. The axioms K2 and K3 describes the iterative self-applicative characteristic of the truth and falsity predicates in KF; K2 says that it is true (false) that a sentence is true iff the sentence is true (false, resp.) and K3 says its dual. Finally, K4 defines the falsity of a sentence to be the truth of its negation. There are different ways to motivate the axioms of KF and Halbach [33] develops a fuller picture. KF can be seen as a generalization of CT, which is a subtheory of KF, or, more naturally, as a generalization of a theory PT of a positive inductive definition of truth and falsity [33, Sect. 8.7].

Although KF derives transfinitely iterated uniform reflection principles for its base theory (i.e., PA in the current setting), KF cannot derive its own soundness due to Gödel's incompleteness theorem: namely, there is an $\mathcal{L}_T$-sentence φ (e.g., $0 = 1$) such that

$$\mathsf{KF} \nvdash \left(\mathrm{Bew}_{\mathsf{KF}}(\ulcorner \varphi \urcorner) \rightarrow \varphi\right).$$

Hence KF doesn't derive the global reflection principle for itself:

$$\mathsf{KF} \nvdash \forall x \left(\mathrm{Sent}_{\mathsf{KF}}(x) \wedge \mathrm{Bew}_{\mathsf{KF}}(x) \rightarrow T(x)\right).$$

This fact may suggest one to iterate KF-truth as in the case of RT_α; c.f., [27]. However, Feferman [17] gave an argument against such iteration, and thereby explain why KF (or $Ref^*(\mathsf{PA}(P))$ defined below in Sect. 3.2.3) is to be called reflective *closure*, that is, why KF is to be seen as exhausting 'what notions and principles one ought to accept if one accepts the basic notions and principles of the theory' [18, p.205]. Iterating KF would mean to adopt principles that go beyond what is implicit in the acceptance of the base theory, that is, PA in the case considered here.

We sketch Feferman's argument against using iterations of KF in order to define the reflective closure of PA. If one were to add axioms for a further truth predicate T', that is, a truth predicate for the language of KF including the truth predicate T, one would specify axioms for T' analogous to those for T. To this end, one would now quantify over sentences in the full language with T' in axioms K4–K6; moreover, one would treat T as just like another predicate of the base language and therefore add the following axiom in analogy to K1 for the predicate T [17, p.40]:

$$\forall s \left((T'(\underset{\cdot}{T}\, s) \leftrightarrow T(s^\circ)) \wedge (F'(\underset{\cdot}{T}\, s) \leftrightarrow \neg T(s^\circ))\right) \tag{3.1}$$

From the logical truth $\forall s\, (T(s^\circ) \vee \neg T(s^\circ)$ and this axiom, we can derive the following 'totality' claim:

$$\forall s \left(T'(\dot{T}\, s) \vee F'(\dot{T}\, s) \right) \tag{3.2}$$

Therefore, by endorsing the new axiom (3.1), we would treat T as a 'total' or 'determined' predicate, just like the predicates of the base language. Feferman [17, p.40] thereby concludes that 'when iterating reflective closure we thus vitiate the informal idea behind the use of partial predicates of truth and falsity.' Hence, for Feferman, KF is closed under the process of making explicit what are implicitly accepted in accepting the basic notions and principles.

KF is a quite rich theory. Indeed, it is intimately related to an approach to predicativity, to the theme of unfolding (see Feferman and Strahm [20], Feferman and Strahm [21]), and it can be regarded as characteristic of a fruitful interaction between the logical *development of non-extensional concepts* (classification, operation) and *semantical investigations*. Further, it has a sort of mathematical appeal: it hinges upon well-known lattice theoretical facts (Knaster–Tarski theorem), and it naturally implies the existence of fixed points of arithmetically definable monotone operators. One can show that KF and the standard fixed point theory $\widehat{\mathsf{ID}}_1$ (see Feferman [14]) are mutually interpretable, thus reducing the classification of proof theoretic strength of KF to that of $\widehat{\mathsf{ID}}_1$. $\widehat{\mathsf{ID}}_1$ leads towards the foundations of intuitionistic type theory and predicatively reducible subsystems of analysis. Below we summarize a few important facts and results on KF.

3.1 Inner Logic and Outer Logic

The theory KF is sometimes criticized for yielding a discrepancy of its outer and inner logic. Let us call $\{\varphi \mid \varphi \text{ is an } \mathcal{L}_T\text{–sentence and } \mathsf{KF} \vdash \varphi\}$ the *outer theory* of KF and $\{\varphi \mid \varphi \text{ is an } \mathcal{L}_T \text{ -sentence and } \mathsf{KF} \vdash T(\ulcorner \varphi \urcorner)\}$ the *inner theory* of KF, and also call the logic governing the former the outer logic of KF and the logic governing the latter the inner logic of KF. [3] The inner logic of KF is strong Kleene logic whereas its outer logic is classical logic; also the inner theory of KF fails to coincide with the outer theory of KF. Hence, KF does not meet one of the desiderata that Leitgeb [44] suggests for a theory of truth.

In order to develop a theory that is an axiomatization of Kripke's theory with strong Kleene logic that avoids this discrepancy, Halbach and Horsten [34] presented a system based on strong Kleene logic. The resulting theory PKF has exactly the same inner theories and outer theories. Halbach and Horsten [34] showed that PKF proves significantly less arithmetical statements than the classical system KF. They suggested that thus KF should not be seen as a formal device for generating theorems in strong Kleene logic: the use of the classical outer logic is indispensable for proving

[3] The notion of inner logic thus defined is ambiguous, because it is not clear enough how to extract logic from a given set of sentences, and one sometimes simply identify outer/inner theories and outer/inner logics. At any rate, the intended inner logic of KF is strong Kleene logic, and Halbach and Horsten's [34] result can be construed to have 'shown' that the inner logic of KF is indeed strong Kleene logic.

certain arithmetical theorems. Hence it is unlikely that a purely 'instrumentalist' interpretation of KF in the spirit of Reinhardt [50] is a viable option.

The difference between inner and outer logic is adumbrated in the Sect. 6.1.2 of Feferman [17, p.40] on the informal interpretation of partial-self-applicable truth predicate. There, considering the question as to whether Kripke's [42] construction corresponds to a (more or less) clear informal notion, Feferman finds it 'reasonably convincing' that in the formulation of KF, $T(\ulcorner\varphi\urcorner)$ expresses that φ is a *grounded truth*, that is, that the denotation of T is given by Kripke's least fixed-point construction, understood in its full generality.[4] As a consequence, the analogy with partial computable predicates becomes helpful: Truth should be relative to *given rules of computation* or *inductive rules*, and ought to be distinguished from our everyday informal notion of truth with which *classical logic* is justified.

On the formal side, this leads to define for an $\mathcal{L}_T$-sentence φ to have a *determined* truth-value, when $\mathsf{KF} \vdash T(\ulcorner\varphi\urcorner) \leftrightarrow \neg F(\ulcorner\varphi\urcorner)$ holds, and to choose D to be the set of all sentences with a determined truth-value, that is the set of all sentences that are true or false, but not both (see the next subsection for a formal elaboration). This set D is interpreted to apply to those sentences whose truth-value is determinable by the rules of *internal* truth and falsity, while the complement of D is interpreted to apply to those whose truth-value is *not determinable* by those rules, but still *definite* in the sense that classical logic is justified with it. Hence, for instance, the statement $\lambda \vee \neg\lambda$ is not an internal truth but true in our informal and, say, naïve platonistic sense.

Those sentences with determined truth-value in this sense enjoy some nice properties. First, the laws of classical logic are provably true for all determined $\mathcal{L}_T$-sentences[5]: namely, if $\varphi_1, \ldots, \varphi_k$ have determined truth-value, then each logical axiom of classical logic involving these sentences is provably true in KF. Second, more importantly, the T-schema holds when restricted to those sentences even with the untyped truth predicate of KF: namely, it holds that, for all $\mathcal{L}_T$-sentence φ, $\mathsf{KF} \vdash D(\ulcorner\varphi\urcorner) \rightarrow \big(T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi\big)$.

3.2 Kripke–Feferman: The Proof Theoretic Side

The main result which is proven in Feferman's paper on reflective closure is the following:

Theorem 1

$$\mathsf{KF} \equiv (\Pi^0_1\text{-}CA)_{<\varepsilon_0} \tag{3.3}$$

[4] He adds that the 'facts', on which T and F are grounded, may be representable as true (false) sentences of *any system* (arithmetic, set theory, etc.) we come to accept as basic.

[5] As well as for total predicates, see Lemma 1.

In general $(\Pi^0_1\text{-CA})_{<\alpha}$ is the theory of iterated jump up to any $\beta < \alpha$.[6]
Let us outline the proof of the theorem, possibly following different routes.

3.2.1 Lower Bound

We exploit the foundational side of truth, as a way of interpreting non-extensional predicate application, in order to show the main result of Feferman [17].

In order to enhance readability, we adopt the following shortenings:

(i) $a(x)$ stands for the the term representing the operation of substituting the first free variable of the expression encoded by a with the xth-numeral.
(ii) $D(a) := \mathrm{Sent}_{\mathsf{KF}}(a) \wedge (T(a) \vee T(\dot{\neg} a)) \wedge \neg(T(a) \wedge T(\dot{\neg} a))$; $D(a)$ means 'a is determined';
(iii) non-extensional membership: $x \in a := T(a(x)) \wedge \neg T(\dot{\neg} a(x)))$; non-extensional co-membership: $x \bar{\in} a := T(\dot{\neg} a(x)) \wedge \neg T(a(x))$
(iv) the class(ification) predicate $Cl(a) := \forall x D(a(x))$.

Then it is easily seen that the language $\mathcal{L}_2$ of second order arithmetic can be regarded as a sublanguage of $\mathcal{L}_T$.

In order to relate KF with standard subsystems of second order arithmetic, we need a few additional definitions:

(i) A formula $\varphi(x, \vec{z})$ of $\mathcal{L}_T$ is *elementary in* $\vec{z} := z_1, \ldots z_n$ iff it is inductively generated from atoms of the form $t = s, t \in z_i, (1 \leq i \leq n)$ by means of $\wedge$, $\neg$, $\forall u$, with $u \notin \{z_1, \ldots z_n\}$.
(ii) We say that Cl is closed under *elementary comprehension* iff for every $\varphi(x, \vec{z})$ elementary in $\vec{z}$, then there exists a primitive recursive term $s_\varphi(\vec{z})$ such that, provably in KF,

$$Cl(\vec{z}) \rightarrow Cl(s_\varphi(\vec{z})) \wedge \forall x(\varphi(x, \vec{z}) \leftrightarrow x \in s_\varphi(\vec{z})). \tag{3.4}$$

(iii) *A family g of classes indexed by a class a* is simply an index g of a partial recursive function $\lambda n.\{g\}(n)$,[7] such that $\forall y(y \in a \rightarrow Cl(\{g\}(y))$.
(iv) Finally, we say that Cl is *closed under join* iff, whenever f is a family of classes indexed by a class a, there exists a term $j(a, f)$ whose elements are exactly those ordered pairs (u, v) such that that $u \in a$ and $v \in \{f\}(u)$.

Lemma 1 *The collection of classes is closed – provably in* KF*- under elementary comprehension and join.*[8]

---

[6] For a precise definition, see Feferman [17]. One can replace $(\Pi^0_1\text{-CA})_{<\alpha}$ in the statement of the theorem and below with ramified analysis up to any level $< \alpha$ in a fixed formalization, provided α has the form ω^β, $\beta \geq \omega$.

[7] We assume a standard formalization of standard recursion theory via the Kleene bracket relation.

[8] The statement can be used to interpret into KF a basic system of Feferman's Explicit Mathematics, see Feferman [13].

Proposition 1

$$(\Pi^0_1\text{-}CA)_{<\varepsilon_0} \leq \mathsf{KF} \tag{3.5}$$

Proof First of all, KF proves, for each $\alpha < \varepsilon_0$, the transfinite induction scheme $TI(\varphi, < \alpha)$, for arbitrary formulas φ in the full language. This is well-known (Gentzen–Schütte–Feferman) and it follows insofar as KF has the full number theoretic induction schema (see Lemma 4.3.2 in Feferman [17]). Then apply $TI(\varphi, < \alpha)$ in conjunction with the previous lemma. □

The fact that the jump hierarchy is available in KF up to any $\alpha < \varepsilon_0$ is crucial in order to carry out a wellordering proof for each initial segment of the standard wellordering of $\phi\varepsilon_0 0$ (see Feferman [17], appendix and Feferman [11]).

Lastly, it is not difficult to prove that KF can directly interpret the fixed point theory $\widehat{\mathsf{ID}}_1$ (see Cantini [5, Sect. 3.11] and Halbach [33, Sect. 19.5]).

3.2.2 Upper Bound

We sketch two strategies for proving the upper bound direction of the theorem.

The first route: Kripke–Feferman as a fixed point theory. This route is essentially Feferman's route in Feferman [17] and consists of a reduction of KF to classical subsystems of known strength.

The theory $\widehat{\mathsf{ID}}_1$ contains– besides the usual axioms for Peano arithmetic and induction schema for the whole language– fixed point axioms FP asserting the existence of fixed points I_φ for arbitrary elementary positive operators $\varphi(x, P)$[9]

$$\forall x(\varphi(x, I_\varphi) \leftrightarrow I_\varphi(x)) \tag{3.6}$$

but $\widehat{\mathsf{ID}}_1$ has no minimality schema.

By Aczel [2] it is known that $\widehat{\mathsf{ID}}_1$ has the same arithmetical theorems as PA with the schema of transfinite induction for each initial segment of the canonical primitive recursive wellordering of type $\phi\varepsilon_0 0$. The idea is to show that there are (not necessarily minimal) Σ^1_1-solutions to the fixed point Eq. (3.6) by standard diagonalization, provably in the subsystem Σ^1_1-AC_0.[10] The argument works in second order arithmetic with arithmetical comprehension except that Σ^1_1-AC_0 is also required in order to show that the result of replacing the parameter P by means of a Σ^1_1-predicate in a positive elementary operator can be still made equivalent to a Σ^1_1-formula (see Feferman [17], 4.2). The argument also holds when induction is restricted to arith-

[9] So $\varphi(x, P)$ is a formula in the language of PA expanded with the new predicate symbol P and positive in P.

[10] Concerning this subsystem and the corresponding one with full induction, see Simpson [57]. There are in the literature several strategies for classifying its proof-theoretic strength, which apply either non-standard models (as in H. Friedman's original proof) or some kind of proof theoretic machinery (in papers by several authors, among them Feferman himself).

metical formulae. For any system S, let $\mathsf{S}\!\upharpoonright$ be the system *with the induction axioms restricted to the language of arithmetic*.

Theorem 2

(i) $\mathsf{KF} \leq \Sigma^1_1\text{-}\mathsf{AC}_0$
(ii) $\mathsf{KF}\!\upharpoonright \leq \Sigma^1_1\text{-}\mathsf{AC}_0\!\upharpoonright$

This is sufficient for calibrating KF and its subsystems.

The second route. The previous proof is simple but it has a disadvantage: it cannot be adapted to deal with *truth consistency* (or minimality), i.e. if we want to consider KF plus CONS:

$$\forall x(\mathrm{Sent}_{\mathsf{KF}}(x) \rightarrow \neg T(x) \vee \neg F(x)) \qquad (\mathsf{CONS})$$

Of course, Axiom CONS rules out truth value gluts. That is, the axiom excludes models where the extension and the antiextension of the truth predicate overlap and sentences are simultaneously true and false. This restriction to consistent fixed points is in line with Kripke's [42] original account.

In order to analyze KF plus CONS, we can easily apply methods from predicative proof theory. Thus we devise a suitable sequent calculus version of KF, $\mathsf{KF}\!\upharpoonright$, ..., choosing to rephrase $\mathsf{KF}\!\upharpoonright$ as *a fixed point theory with consistency*. This means that we can easily find a formula $\mathcal{T}(x, P)$, formalizing the closure properties of the truth predicate T (see Halbach [33], p. 281), i.e. such that $\forall x(\mathcal{T}(x, T) \leftrightarrow T(x))$.

Since we like to formalize KF and $\mathsf{KF}\!\upharpoonright$ as Tait calculi, the basic *positive atoms* have the form: $t = s$, $T(t)$, and the *negative atoms* are obtained by negating the positive ones. An *atom* is simply a positive or a negative atom and we stipulate that $\neg\neg\varphi := \varphi$ (φ atom). Formulas are inductively generated from atoms by closing under disjunction, conjunction, unbounded quantification. If φ is an arbitrary formula, $\neg\varphi$ is the formula which results from the negation normal form of $\neg\varphi$ by erasing each even sequence of occurrences of negation in front of atoms.

Let us expand $\mathcal{L}_T$ with a new predicate symbol P. If $Q := P, T$, a formula φ of $\mathcal{L}_T$ is Q-*positive* (Q-*negative*) if every occurrence of Q in φ occurs within positive (negative) atoms of the form $Q(t)$ ($\neg Q(t)$). A formula φ is Q-*separated* if φ is Q-positive or Q-negative. A formula φ is Q-*free* if Q does not occur in φ. A Q-free formula can be regarded as both Q-positive and Q-negative.

Definition 1 The system $\mathsf{TKF}^{\mathsf{c}}$ consists of:

- logical axioms of the form
$$\Gamma, \neg\varphi, \varphi$$
$$\Gamma, \neg t = s, \varphi[x := t], \varphi[x := s]$$
where φ is an atom (according to the previous definitions);
- axioms of the form Γ, Δ where Δ is an e-atom or a finite set of e-atoms; Δ formalizes the standard axioms for zero, successor, or the defining equations for the function symbols of $\mathcal{L}_T$;

- standard logical rules (see Schwichtenberg and Wainer [55]) for introducing $\wedge$, $\vee$, $\forall$, $\exists$ and the cut rule:

$$\frac{\Gamma, \varphi \quad \Gamma, \neg\varphi}{\Gamma}$$

- T-consistency:

$$\Gamma, \neg \mathrm{Sent}_{\mathsf{KF}}(t), \neg T(t), \neg T(\dot{\neg} t)$$

- induction rule for *classifications*: if t is an arbitrary term,

$$\frac{\Gamma, Cl(a) \quad \Gamma, 0 \in a \quad \Gamma, \forall x(x \in a \to (x+1) \in a)}{\Gamma, t \in a}$$

- T-closure:

$$\frac{\Gamma, \mathcal{T}(t, T)}{\Gamma, T(t)}$$

- T-soundness:

$$\frac{\Gamma, \neg\mathcal{T}(t, T)}{\Gamma, \neg T(t)}$$

TKF is the calculus which is obtained from $\mathsf{TKF}^{\mathsf{c}}$ by replacing the induction rule for *classes* by full number-theoretic induction rule, that is, if $\varphi(x)$ is an arbitrary formula of $\mathcal{L}_T$ an arbitrary term,

$$\frac{\Gamma, \varphi(0) \quad \Gamma, \forall x(\varphi(x) \to \varphi(x+1))}{\Gamma, \varphi(t)}$$

If $S := \mathsf{TKF}^{\mathsf{c}}, \mathsf{TKF}$, we can inductively define a derivability relation with explicit tree height and suitable cut complexity, so that the following holds:

Lemma 2 (Partial Cut-elimination) *Every $\mathsf{TKF}^{\mathsf{c}}$-derivation $\mathcal{D}$ can be effectively transformed into a $\mathsf{TKF}^{\mathsf{c}}$-derivation of the same end-sequent, where cut–formulas are T-separated.*

The argument is standard; it essentially depends on the fact that the active formulas in the axioms and in the conclusions of the mathematical inferences (in particular, number theoretic induction) are T-separated. It is also obvious that the system $\mathsf{TKF}^{\mathsf{c}}$ (TKF) proves the sequents corresponding[11] to the theorems of $\mathsf{KF}\restriction(\mathsf{KF})$ plus CONS.

Approximating truth by its finite levels. By Lemma 2 it is then possible to approximate truth by its finite levels and hence to eliminate truth from $\mathcal{L}_T$.

[11] Under the obvious translation of the language of KF, $\mathsf{KF}\restriction$ into the Tait framework.

Definition 2 Let $\bot = 0 = 1$; then

$$T^0(t) = 0 = 1 \tag{3.7}$$
$$T^{m+1}(t) = \mathcal{T}(t, T^m)) \tag{3.8}$$

Clearly *each formula in the sequence belongs to the language* $\mathcal{L}_{\mathsf{PA}}$.

If φ is any formula in negation normal form, let $\varphi[m, n]$ be obtained from φ by replacing each atom of the form $T(t)$ $(\neg T(t))$ by $T^n(t)$ $(\neg T^m(t))$.

Finally, a derivation $\mathcal{D}$ of $\mathsf{TKF^c}$ is *quasi-normal* provided all cut–formulas in $\mathcal{D}$ are T-separated.

Theorem 3 *Let* $\mathcal{D}$ *be a quasi-normal* $\mathsf{TKF^c}$*-derivation of* Γ *with height* k*. Then, provably in* PA*, for every* $m > 0$,[12] *if* $H(m) := m + 2^k$*:*

$$\Gamma[m, H(m)] \tag{3.9}$$

The proof is by induction on the height k of the given derivation. The soundness of the asymmetric interpretation – with respect to PA-provability, relies on the fact that cut formulas are always T-separated and on standard persistence properties.[13] On the other hand, restriction to classes has the effect that number theoretic induction for $\mathcal{L}_{\mathsf{PA}}$-formulas is enough, as a consequence of the fact that for each given m, $T^m(t)$ is an arithmetical formula[14] As to the verification of consistency, it reduces to verify by outer induction on m

$$\forall x(\mathrm{Sent}_{\mathsf{KF}}(x) \to \neg T^m(x) \vee \neg T^m(\dot{\neg} x)) \tag{3.10}$$

Since the transform $\varphi \mapsto \varphi[m, n]$ is the identity function on formulas of $\mathcal{L}_{\mathsf{PA}}$, we eventually conclude that

Corollary 1 *If* $\mathsf{TKF^c} \vdash \varphi$ *and* $\varphi \in \mathcal{L}_{\mathsf{PA}}$*, then* $\mathsf{PA} \vdash \varphi$.

Theorem 4

(i) $\mathsf{PA} \equiv \mathsf{KF}\restriction + \mathsf{CONS}$
(ii) $(\Pi^0_1\text{-}CA)_{<\varepsilon_0} \equiv \mathsf{KF} + \mathsf{CONS}$

As to $\mathsf{KF} + \mathsf{CONS} \leq (\Pi^0_1\text{-CA})_{<\varepsilon_0}$: by lifting the previous method to the case of systems with full number theoretic induction. This can be carried out by standard embedding into systems with ω-rule (see Schwichtenberg and Wainer [55]).

[12] In general, if $\Gamma := \{\varphi_1, \ldots, \varphi_q\}$, $\Gamma[m, n] := \{\varphi_1[m, n], \ldots, \varphi_q[m, n]\}$.

[13] This means that, if $0 < m_2 \leq m_1 \leq k_1 \leq k_2$, Γ is a set of formulas such that $\Gamma[m_1, k_1]$, Δ is derivable in PA, then $\Gamma[m_2, k_2]$, Δ is also PA-derivable, leaving height and cut complexity unchanged.

[14] Note, however, that its logical complexity increases with m.

3.2.3 From Ordinary Reflective Closure to Schematic Reflective Closure

Equation 3.4 means that the ordinary reflective closure of arithmetic is in fact equivalent to predicative analysis of any level up to the first ϵ-number ε_0 i.e. roughly to the fragment of second order arithmetic, which is based upon transfinite iteration up to ε_0 of the standard arithmetical comprehension

$$(\exists X)(\forall u)(u \in X \leftrightarrow \varphi(u)),$$

φ being a formula containing only number theoretic quantifiers and possibly free second order variables.

The notion of ordinary reflective closure can then be extended *by a suitable substitution rule* in order to answer the problem of which *schemata* can be regarded as implicit in accepting a given list of *schematic axioms and rules*. The substitution rule has the form

$$\frac{\varphi(P)}{\varphi(\hat{x}\psi(x))}$$

where φ is a formula of $\mathcal{L}_{\mathsf{PA}}$ with an additional predicate symbol P, ψ is a formula of the language $\mathcal{L}_T$ expanded with P, and $\varphi(\hat{x}\psi(x))$ is the formula obtained by replacing the atoms of the form $P(t)$ in $\varphi(P)$ by $\psi(t)$.[15] Informally, the rule allows us to make inferences from schemata accepted in the original arithmetical language to schemata of the language with reflective means (i.e. self-referential truth). It turns out that the resulting notion of *schematic reflective closure* $Ref^*(\mathsf{PA}(P))$ yields *an alternative unramified characterization of predicative analysis* (see Feferman [17]):

$$Ref^*(\mathsf{PA}(P)) \equiv (\Pi^0_1\text{-CA})_{<\Gamma_0};$$

this proof-theoretic[16] equivalence still holds even when we add consistency (CONS) to $Ref^*(\mathsf{PA}(P))$. The theorem witnesses that investigations deriving from formal semantics and paradoxes have reached a high level of integration with different areas of foundational investigations; for instance, while the lower bound result is actually a refinement of typical predicative well-ordering techniques, the upper bound theorem can be achieved by techniques and results from reductive proof theory (see Feferman [12]).

Remarks

(i) The main result about reflective closure (either ordinary or schematic) still holds once consistency (CONS) is replaced by the *completeness* axiom, i.e.

[15] With the obvious proviso ensuring that no clash of variables occurs.

[16] Γ_0 is the first strongly critical ordinal, which is known to be the limit of predicative provability in the sense of Feferman and Schütte.

$$(\forall x)(\mathrm{Sent}_{\mathsf{KF}}(x) \rightarrow (T(x) \vee F(x)).$$

This axiom, which is true in the greatest fixed point of the Kripkean operator for self-referential truth, rules out truth value gaps in the same way as the consistency axiom (CONS) rules out truth value gluts. Kripke [42] considered only consistent fixed point (that is, fixed points without overlapping extension and antiextension of truth). But later Visser [60] and others generalized the approach to fixed points where extension and antiextension are allowed to overlap and thus a sentence can be simultaneously true and false.

(ii) It turns out that *suitable variants of 'reflective closure' can be profitably applied as tools for proof-theoretic investigations*: for instance, as an intermediate step for computing the proof theoretic strength of transfinitely iterated fixed point theories (see Jäger et al. [38]).

3.2.4 Digression: KF with Minimality

Many philosophers think that the minimal fixed point model of Kripke's [42] theory is the most natural. It gives a picture of grounded truth: The truth and falsity of any sentence ultimately depends on the truth and falsity of non-semantic sentences, that is, sentences that do not contain the truth or falsity predicates. In other fixed points that are not minimal ungrounded sentences such as a truth teller sentence can be true and false as well.

KF is the theory of *all fixed point models* of the monotone operator specifying the clauses for reflective truth. Consequently, KF doesn't decide the truth teller sentence. If one aims at a theory of the minimal fixed point and thus of grounded truth, one can try to add axioms that exclude fixed point models that are not minimal. Of course, this is conceptually relevant for the whole enterprise of truth: think of the important distinctions that arise from considering the plurality of truth predicates (grounded sentence, paradoxical sentence, intrinsic, etc. see Kripke [42]).

Definition 3

(i) KF + GID: see Cantini [5]. GID is the schema

$$\forall x(\varphi(x, \hat{x}\psi(x)) \rightarrow \psi(x)) \rightarrow \forall x(x \in I_\varphi \rightarrow \psi(x)) \tag{3.11}$$

where ψ is an arbitrary formula, I_φ is obtained by diagonalization (Lemma 3.9, [5]), $\varphi(x, P)$ is an elementary positive (in P) operator in the language of $\mathcal{L}_T$ with an additional predicate variable, P positive.[17]

(ii) KF_μ (Truth with minimality, see Burgess [4]): it is the fragment of KF$\restriction$ with

(i) only the *composition principles*, e.g. $\forall x(\mathcal{T}(x, T) \rightarrow T(x))$;

[17] Of course, represented in $\mathcal{L}_T$, so that $P(t)$ is translated into $t \in p$, p fresh variable; $\varphi(x, \psi)$ is obtained by the substitution $t \in y \mapsto \psi(t)$. The schema (3.11) claims that I_φ represents the least fixed point of the monotone operator defined by $\varphi(x, P)$ in a given arithmetical model.

(ii) the schema: if ψ is an arbitrary formula,

$$\forall x(\mathcal{T}(x, \hat{x}\psi(x)) \to \psi(x)) \to \forall x(T(x) \to \psi(x)) \tag{3.12}$$

Then KF_μ proves the decomposition axioms and the consistency axiom. Also, it explicitly refutes statements that fail in the least fixed point model. Clearly $\mathsf{KF}_\mu\restriction$ has an inner model in $\mathsf{KF}\restriction + GID$ (see proposition 3.12, Cantini [5]).

As to the upper bound, we can apply the proof-theoretic methods of Cantini [7] to $\mathsf{KF}_\mu\restriction$.

As to the lower bound, let $\mathsf{ID}_1^{\mathsf{acc}}$ be the theory of accessibility inductive definitions over PA, which is known to be proof-theoretically equivalent to the theory of elementary inductive definitions. Then we can lift to the present context a suitable version of the so-called *bar-induction* schema, which goes back to Kreisel:

Theorem 5 *If $\prec$ is a binary relation which is* determined,[18] *then the schema of transfinite induction on the largest well-founded part $W(\prec)$ of $\prec$ holds for arbitrary formulas φ, provably in KF_μ.*

The proof takes inspiration from an analogous result of Friedman and Sheard [25]; it is also exploited by Burgess [4]. Hence we have by the previous theorem:

Corollary 2 $\mathsf{ID}_1^{\mathsf{acc}}$ *is interpretable in* KF_μ.

3.3 Related Works Inspired by KF

KF is nowadays considered to have brought the birth to the subject called axiomatic theories of truth as an area of logical research in its own right. A variety of formal theories of truth not in a semantic form but in an axiomatic form have been presented since KF. Some try to axiomatize known semantic theory or construction of truth as KF does for Kripke's theory of truth; others take more purely axiomatic approach by considering what combinations of truth-theoretic principles are possible (consistent) and plausible. Friedman and Sheard [25] listed nine natural and naïvely correct postulates of truth, which are inconsistent altogether, and determined the maximal consistent combinations of them. The other two theories of Feferman's own presented below would be counted in the latter 'purely axiomatic' type of axiomatic theories of truth as well. Cantini [6] presented a system that can be regarded as an axiomatization of Kripke's least fixed-point truth with supervaluation schema. Horsten et al. [36] made an attempt to axiomatize Gupta–Belnap–Herzberger revision theoretic truth. There are far more examples that we could list here, and the subject is still lively developing.

[18] I.e. total in the sense of T, F, see Sect. 3.2.1.

4 Type-Free T-Schema with Reinterpreted Biconditional

The second theory of truth by Feferman [15] was originally intended to provide a uniform type-free treatment for set-comprehension and truth-predication, both of which notoriously yield a contradiction with their naïve formulations. Feferman [18] described the purpose of the theory as 'pragmatic'; particularly, on the side of set-comprehension, the theory was presented for the purpose of suitably and consistently dealing with natural type-free statements that mathematicians would like to make, for instance, in category theory. However, he later gave it a reformation into a theory of deflationist truth in Feferman [19].

The core idea behind the theory is to formally interpret the informal biconditional 'if and only if' in the Comprehension Axiom and T-schema, not as the material biconditional $\leftrightarrow$, but as another new primitive logical connective $\equiv$ so that $\equiv$ preserves certain essential connotations or characteristics of the informal biconditional but still avoids a contradiction. Let us restrict ourselves to his theory of truth. The T-schema is expressed in English as

$$\ulcorner \varphi \urcorner \text{ is true if and only if } \varphi \tag{4.1}$$

for arbitrary $\varphi \in \mathcal{L}_T$. If we interpret 'if and only if' by the standard material biconditional $\leftrightarrow$, then the resulting translation of (4.1) is inconsistent. Hence, Feferman introduces a new binary connective $\equiv$ into the syntax of a theory, and add the following as a new axiom schema that meant to interpret (4.1):

$$T(\ulcorner \varphi \urcorner) \equiv \varphi,$$

for arbitrary $\varphi \in \mathcal{L}_T$.

Since $\equiv$ is a newly introduced connective separate and independent of any other standard connectives such as $\rightarrow$ and $\neg$, a suitable axiomatic characterization should be provided for it, and also such a characterization should be so made that $\equiv$ is seen as representing the informal biconditional 'if and only if' at least in the context of our discourse involving truth.

Feferman [15] initially gave a semantic characterization of $\equiv$, where he describes how to define the extension of a truth predicate and the semantic evaluation rule for the new connective $\equiv$ over a given model of a base theory. The construction goes as follows. We start with an arbitrary model $\mathfrak{M}$ with a domain M of a base theory B over a language $\mathcal{L}$. For the simplicity of argument, we take $\mathcal{L}_{\mathsf{PA}}$ and PA to be $\mathcal{L}$ and B respectively in what follows. Recall that we defined $\mathcal{L}_T = \mathcal{L}_{\mathsf{PA}} \cup \{T\}$. Let $\mathcal{L}_T(\equiv)$ be a language obtained by augmenting $\mathcal{L}_T$ with the connective $\equiv$. We first give the Kripkean least fixed-point construction of truth over $\mathfrak{M}$ with a suitable partial logic such as strong Kleene or 3-valued Łukasiewicz logic, and let $X \subset M$ be the extension of the predicate T in thus constructed fixed-point semantics over $\mathcal{L}_T$ with the partial logic we have chosen. Now, for each $\mathcal{L}_T(\equiv)$-formula φ, let us denote by φ^* the $\mathcal{L}_T$-formula obtained by replacing all the occurrences of $\equiv$ in φ with the

material biconditional $\leftrightarrow$. Then, we expand the base model $\mathfrak{M}$ to an $\mathcal{L}_T(\equiv)$-structure $\mathfrak{M}(\equiv)$ by putting

- $\mathfrak{M}(\equiv) \models T(a)$, iff $a = \ulcorner\sigma\urcorner$ for some $\mathcal{L}_T(\equiv)$-sentence σ and σ^* is contained in thus constructed Kripkean truth, and
- $\mathfrak{M}(\equiv) \models \varphi \equiv \psi$, iff φ^* and ψ^* have the same semantic value in the fixed-point semantics;

note that $\equiv$ is now changed to $\leftrightarrow$ in φ^* and ψ^* and it is evaluated according to the partial logic we chose in the fixed-point semantics; the evaluation of $\mathcal{L}_{\mathsf{PA}}$-atomics in $\mathfrak{M}(\equiv)$ are the same as $\mathfrak{M}$, and the other logical connectives and quantifiers are classically evaluated in $\mathfrak{M}(\equiv)$. Hence, an informal reading of $\varphi \equiv \psi$ is 'the semantic values of φ and ψ coincides in the Kripke's least fixed-point semantics', and thus since the Kripkean construction gives the same semantic value to $T(\ulcorner\varphi\urcorner)$ and φ for all $\mathcal{L}_T$-sentences φ it holds that

$$\mathfrak{M}(\equiv) \models T(\ulcorner\varphi\urcorner) \equiv \varphi, \text{ for all } \mathcal{L}_T(\equiv)\text{–sentences } \varphi. \tag{4.2}$$

Under this interpretation $\equiv$ enjoys several desired properties. Let $\mathbf{t} := 0 = 0$ and $\mathbf{f} := 0 \neq 0$: namely, $\mathbf{t}$ and $\mathbf{f}$ respectively stand for definitely true and false sentences. Then we define $\mathcal{D}(\varphi)$ to abbreviate $(\varphi \equiv \mathbf{t}) \vee (\varphi \equiv \mathbf{f})$, which informally expresses 'φ is determinate.' Then, we can show that $\mathfrak{M}(\equiv)$ is a model of the following.

S1 $T\ulcorner\varphi\urcorner \equiv \varphi$ for all $\mathcal{L}_T(\equiv)$-sentences φ
S2 $\equiv$ is an equivalence relation
S3 $\mathbf{t} \not\equiv \mathbf{f}$
S4 $\equiv$ preserves all the connectives and quantifiers including $\equiv$ itself; e.g., $((\varphi \equiv \psi) \wedge (\sigma \equiv \chi)) \rightarrow ((\varphi \equiv \sigma) \equiv (\psi \equiv \chi)) \wedge ((\varphi \vee \sigma) \equiv (\psi \vee \chi))$.
S5 $\mathcal{D}(\varphi)$ for all $\mathcal{L}_{\mathsf{PA}}$-sentences.
S6 $\mathcal{D}$ is closed under $\neg$, $\vee$, and $\forall$.
S7 $(\varphi \equiv \mathbf{t} \rightarrow \varphi) \wedge (\varphi \equiv \mathbf{f} \rightarrow \neg\varphi)$.

Proof S1 immediately follows from (4.2). S2 and S4 are obvious from the evaluation for $\equiv$. S3 and S5 hold since the fixed-point semantics does not change the evaluations of $\mathcal{L}_{\mathsf{PA}}$-sentences. S6 is due to the evaluation schema of the logic we chose. Finally we can show S7 by induction on φ. □

Later Feferman [19] adopted these seven properties of $\equiv$ as the axioms of a deflationist theory of truth which he calls S. There he interprets deflationism in the form that the proposition $T\ulcorner\varphi\urcorner$ is evaluated in the same way as φ and they are thereby equated even if they do lack a determined truth value[19]; hence the aforementioned informal reading of $\equiv$ well fits in this interpretation. Furthermore, since any model of PA can be expanded to a model of S, S is conservative over B; it is even 'semantically

[19]This intuition yields a model in a suitable infinitary combinatory logic, as detailed in Aczel and Feferman [3].

conservative' over B in the sense of Craig and Vaught [8]. Thus it satisfies the 'conservativeness requirement' for deflationist theories of truth.[20]

Remark 1 Several extra axioms can be still conservatively added to S. For instance, we may consider the following further axioms for S.

S8 $(\varphi \vee \psi) \equiv \mathbf{t} \leftrightarrow \varphi \equiv \mathbf{t} \vee \psi \equiv \mathbf{t}$.
S9 $\exists x \varphi(x) \equiv \mathbf{t} \leftrightarrow \exists x (\varphi(x) \equiv \mathbf{t})$.
S10 $(\varphi \equiv \psi) \equiv \mathbf{t} \leftrightarrow \varphi \equiv \psi$.
S11 $(\varphi \equiv \psi) \equiv \mathbf{t} \leftrightarrow \mathcal{D}(\varphi) \wedge \mathcal{D}(\psi) \wedge (\varphi \equiv \psi)$.
S12 $\mathcal{D}$ is strongly compositional with respect to the connectives and quantifiers except $\equiv$; e.g., $\mathcal{D}(\varphi \vee \psi) \leftrightarrow \mathcal{D}(\varphi) \wedge \mathcal{D}(\psi)$.

Which of them can be conservatively added to S depends on which partial logic we choose in the construction of the Kripkean fixed-point semantics over $\mathfrak{M}$. For instance, if we choose 3-valued Łukasiewicz logic, then we can show that the resulting semantics $\mathfrak{M}(\equiv)$ is a model of S8–S10 and thus they can be conservatively added to S; if we choose strong Kleene logic instead, S8–S9 and S11 are satisfied in $\mathfrak{M}(\equiv)$ and thus can be conservatively added to S; by choosing weak Kleene logic, we can also see that S11–S12 can be conservatively added to S.

5 Determinate Truth

In the system S of the previous section one has the unrestricted T-schema. However, the biconditional in the T-schema is no longer the classical biconditional $\leftrightarrow$ but rather the nonclassical $\equiv$. This is the price for the unrestricted T-schema. Instead of having the unrestricted T-schema and a nonclassical biconditional, one can restrict the T-schema to well-behaved sentences and then have the full classical T-schema and all other truth-theoretic axioms one might want to have. Feferman [18] pursued this strategy with his system DT.

Of course the restriction needs to be well motivated. According to Russell [52], every predicate P has a *domain of significance* D and it makes sense to apply P only to objects in D and the principles which are supposed to characterize (or axiomatize) the concept expressed by P are meant to be only about the objects in D. Hence, in the case of truth, if such a domain is appropriately given and if no contradictory sentences, such as the liar sentence, are contained in D, the naïve truth-theoretic principles, such as Tarski's T-schema, can be safely postulated. For Feferman such

[20] Deflationism is a claim that truth is a 'metaphysically thin and insubstantial' notion and a merely logico-linguistic device for generalization and implicit endorsement. It is often argued that deflationist theory of truth should be conservative over a base theory, since otherwise the addition of a truth predicate would yield something that was not obtained without the help of it in the base theory. See also footnote 2 above. Shapiro [56] and McGee [47] discuss model-theoretic or semantic conservativeness.

a domain D for the case of truth consists of the sentences that are *meaningful* and *determinate*,[21] that is, have a definite truth value, true of false.

Let us assume that our base theory is PA as before. A theory of truth based on the above view have two predicate T and D representing truth and its domain of significance; namely, the language $\mathcal{L}_{\mathsf{DT}}$ of DT is $\mathcal{L}_{\mathsf{PA}} \cup \{T, D\}$. It is plausible for him to require that the condition on the domain D of significance for any predicate P should be prior to the conditions on P; in the current context, an axiomatic characterization of D should be given without appealing to T. Feferman takes *strong compositionality* as the axiomatic characterization of the domain D of significance for T. Namely, D is closed under the propositional operations and quantifiers, and, conversely, a sentence is meaningful only if all of (the substitution instances of) its parts are meaningful. Thus conceived *strong compositionality* can be axiomatized as follows:

D1 $\forall x \left(\mathrm{AtSent}_{\mathsf{PA}}(x) \to D(x)\right)$
D2 $\forall s \left(D(\dot{T}\, s) \leftrightarrow D(s^\circ)\right)$
D3 $\forall x \left(\mathrm{Sent}_{\mathsf{DT}}(x) \to \left(D(\dot{\neg} x) \leftrightarrow D(x)\right)\right.$
D4 $\forall x \forall y \left(\mathrm{Sent}_{\mathsf{DT}}(x \mathbin{\dot{\vee}} y) \to \left(D(x \mathbin{\dot{\vee}} y) \leftrightarrow (Dx \wedge Dy)\right)\right)$
D5 $\forall v \forall x \left(\mathrm{Sent}_{\mathsf{DT}}(\dot{\forall} v.x) \to \left(D(\dot{\forall} v.x) \leftrightarrow \forall t\, D(x[t/v])\right)\right)$

The formula $\mathrm{AtSent}_{\mathsf{PA}}$ stands for the set of all atomic $\mathcal{L}_{\mathsf{PA}}$-sentences, and $\mathrm{Sent}_{\mathsf{DT}}$ for the set of sentences of $\mathcal{L}_{\mathsf{DT}}$. The axiom D1 does not come from the strong compositionality requirement in question but from the common and plausible view that the sentences of the base language, which contains no semantic predicates, should be non-paradoxical, determinate, and counted in the domain of the significance of T.

The T-schema and other truth-theoretic principles are accordingly restricted to such sentences. For instance, the T-schema is formulated as follows:

$$D(\ulcorner \varphi \urcorner) \to (T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi), \quad \text{for all } \mathcal{L}_{\mathsf{DT}}\text{-sentences } \varphi. \tag{5.1}$$

Feferman points out that thus formulated T-schema yields the following intimate connection of D and T as a special phenomenon particular to the case of truth:

$$D(\ulcorner \varphi \urcorner) \leftrightarrow \left(T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \neg \varphi \urcorner)\right) \tag{5.2}$$

Its proof goes as follows. Suppose $D(\ulcorner \varphi \urcorner)$ for one direction. By D3 we have $D(\ulcorner \neg \varphi \urcorner)$. Hence, (5.1) yields $T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$ and $T(\ulcorner \neg \varphi \urcorner) \leftrightarrow \neg \varphi$. Finally, the law of excluded middle $\varphi \vee \neg \varphi$ entails $T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \neg \varphi \urcorner)$. Suppose $(T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \neg \varphi \urcorner))$ for the converse. Since T only applies to its domain of significance D, we have $D(\ulcorner \varphi \urcorner) \vee D(\ulcorner \neg \varphi \urcorner)$. Hence $D(\ulcorner \varphi \urcorner)$ follows from D3. Feferman then generalizes (5.2) to a universal statement and postulates

D0 $\forall x \left(D(x) \leftrightarrow (T(x) \vee T(\dot{\neg} x))\right)$

[21] Feferman [17] adopts the term 'determined' for a similar notion; see p. 20.

Thereby, specially in the case of truth, the domain predicate D becomes definable in terms of T and redundant. Indeed, Feferman [18] does not take a predicate D as a primitive predicate for this reason. Instead he formulates DT over $\mathcal{L}_T$ and introduces Dx as a mere abbreviation of $T(x) \vee T(\dot{\neg} x)$. Therefore let us assume $\mathcal{L}_{\mathsf{DT}} = \mathcal{L}_T$ in what follows.

Given these characterizations of D, Feferman formulates the axioms for T as the the compositional Tarskian clauses restricted to D:

D6 $\forall s \forall t \left(T(s \doteq t) \leftrightarrow s^\circ = t^\circ\right)$, and similarly for other predicates of $\mathcal{L}_{\mathsf{PA}}$
D7 $\forall s \left(D(s^\circ) \rightarrow \left(T(\dot{T}\, s) \leftrightarrow T(s^\circ)\right)\right)$
D8 $\forall x \left(\mathrm{Sent}_{\mathsf{DT}}(x) \wedge D(x) \rightarrow \left(T(\dot{\neg} x) \leftrightarrow \neg T(x)\right)\right.$
D9 $\forall x \forall y \left(\mathrm{Sent}_{\mathsf{DT}}(x \dot{\vee} y) \wedge D(x \dot{\vee} y) \rightarrow \left(T(x \dot{\vee} y) \leftrightarrow (T(x) \vee T(y))\right)\right)$
D10 $\forall v \forall x \left(\mathrm{Sent}_{\mathsf{DT}}(\dot{\forall} v.x) \wedge D(\dot{\forall} v.x) \rightarrow \left(T(\dot{\forall} v.x) \leftrightarrow \forall t T(x[t/v])\right)\right)$

Let us denote by DT^{w} the axiomatic system comprising PA plus D0–D10 augmented by the expanded induction for $\mathrm{Sent}_{\mathsf{DT}}$. We can easily show by induction on formulae that (5.1) is a theorem of DT. This DT^{w} is a straightforward axiomatization of Feferman's view described so far and is already proof-theoretically pretty strong; DT^{w} is proof-theoretically equivalent to KF and indeed identical as theories with $\mathsf{WKF} + \mathsf{CONS}$, the Kripke–Feferman system for weak Kleene logic plus the axiom of consistency; see [26] for more details. However, he sees some defects in DT^{w} as a theory of truth and slightly amends it.

We can standardly show that DT^{w} proves the global reflection principle as KF does: i.e.,

$$\mathsf{DT}^{\mathsf{w}} \vdash \forall x \left(\mathrm{Bew}_{\mathsf{PA}}(x) \rightarrow T(x)\right).$$

However, in contrast to KF, DT^{w} does not prove *the truth of the global reflection principle*. Feferman requires for an adequate theory of truth to derive it and thereby suggests to amend DT^{w}.

Let us first see how its proof goes in KF. In addition to the global reflection principle, both KF and DT^{w} derive $\forall s \left(T(\dot{T}\, s) \leftrightarrow T(s^\circ)\right)$ as well as

$$\forall x \left(\left(T(\ulcorner \mathrm{Bew}_{\mathsf{PA}}(\dot{x}) \urcorner) \leftrightarrow \mathrm{Bew}_{\mathsf{PA}}(x)\right) \wedge \left(T(\ulcorner \neg \mathrm{Bew}_{\mathsf{PA}}(\dot{x}) \urcorner) \leftrightarrow \neg \mathrm{Bew}_{\mathsf{PA}}(x)\right)\right).$$

Hence, it follows from these that both KF and DT^{w} derive

$$\forall x \left(T(\ulcorner \mathrm{Bew}_{\mathsf{PA}}(\dot{x}) \urcorner) \rightarrow T(\dot{T}\, \dot{x})\right).$$

By using the equivalence of $\varphi \rightarrow \psi$ and $\neg\varphi \vee \psi$, we obtain

$$\forall x \left(T(\ulcorner \neg \mathrm{Bew}_{\mathsf{PA}}(\dot{x}) \urcorner) \vee T(\dot{T}\, \dot{x})\right). \tag{5.3}$$

However, one has to verify in DT^{w} that $\forall x D(Tx)$, in order to proceed from (5.3) to $\forall x(T\ulcorner\neg \mathrm{Bew}_{\mathsf{PA}}(\dot{x}) \vee Tx\urcorner)$, but $\forall x D(Tx)$ is incompatible with the view behind DT^{w} and indeed underivable in DT^{w}; in contrast, KF can immediately derive it via the axioms K4 and K5 and thereby the truth of the global reflection principle. Given this observation, Feferman proposes to treat $\rightarrow$ as separate basic propositional connectives from $\vee$ and $\neg$, not identifying $\varphi \rightarrow \psi$ with $\neg\varphi \vee \psi$, and then postulates an independent axiom for the meaningfulness and determinateness of the sentences of the form $\varphi \rightarrow \psi$. Namely, he postulate the following two axioms for the conditional:

$$\text{D11}\quad \forall x \forall y \left(\mathrm{Sent}_{\mathsf{DT}}(x \dot{\rightarrow} y) \rightarrow \left(D(x \dot{\rightarrow} y) \leftrightarrow \left(D(x) \wedge (T(x) \rightarrow D(y)) \right) \right) \right)$$

$$\text{D12}\quad \forall x \forall y \left(\mathrm{Sent}_{\mathsf{DT}}(x \dot{\rightarrow} y) \wedge D(x \dot{\rightarrow} y) \rightarrow \left(T(x \dot{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)) \right) \right)$$

It is true, as Feferman himself points out, strong compositionality in the aforementioned sense is not met with respect to the conditional '$\rightarrow$', since D11 indicates that if φ is false then $D(\ulcorner \varphi \rightarrow \psi \urcorner)$ may hold even when $D(\ulcorner \psi \urcorner)$ fails, but he argues that it is rather natural since we do not care whether $D(\ulcorner \varphi \rightarrow \psi \urcorner)$ holds when φ is definitely false. Thereby the system DT of determinate truth is defined to be PA plus D0–D12 augmented by the expanded induction schema for $\mathcal{L}_{\mathsf{DT}}$.

A closer inspection reveals that DT is very intimately related to KF. KF is an "axiomatization" of Kripke's fixed-point semantics with strong Kleene logic. Naturally we may consider a similar "axiomatization" of the Kripkean fixed-point semantics with other logics. We have already mentioned such a theory for weak Kleene logic, which is related to DT^{w}. As a matter of fact, DT is identical with such a KF-style axiomatization of the Kripkean fixed-point semantics with what we call Aczel–Feferman logic, plus the axiom of consistency CONS; namely, the inner logic of the KF-style system is Aczel–Feferman logic; see Fujimoto [27] for details.[22] Here we mean by Aczel–Feferman logic a variant of weak Kleene logic with an independent evaluation for the conditional '$\rightarrow$' than that for $\vee$ and $\neg$; in Aczel–Feferman logic, '$\rightarrow$' is not defined away in terms of '$\neg$' and '$\vee$' and the evaluation of '$\rightarrow$' is determined by the following separate truth table:

$\rightarrow$	T	U	F
T	T	U	F
U	U	U	U
F	T	T	T

Hence, from a purely technical point of view, Feferman's theory of determinate truth essentially boils down to his own axiomatization of Kripkean fixed-point semantics, and therefore DT is proof-theoretically equivalent to KF. Also, if we extend DT in the same way that we extend the reflective closure of PA, i.e., KF, to its *schematic reflective closure*, the resulting theory has the strength of the predicative limit Γ_0.

[22] What Aczel [1] does in his construction of Frege structure essentially amounts to the construction of Kripkean fixed-point semantics with Aczel–Feferman logic.

Remark 2 Feferman [18] raised two conjectures about the strength of the theories of determinate truth over PA, and the conjectures have been first settled by constructing a direct interpretation of them in KF and $Ref^*(\mathsf{PA}(P))$ in Fujimoto [26], but indeed the construction of Sect. 3.2.2 can also be adapted to the system DT (over PA), in order to produce an alternative proof of the first conjecture.

6 The Impact of Feferman's Work on Truth

We have mentioned already some further research inspired by Feferman's work. However, we don't aim at a survey of the impact of his work, because the impact is far too wide-ranging to be summarized in a few pages. Moreover, an assessment of the reception of Feferman's work is made harder by the fact that his work has been studied by different communities: In mathematical logic proof theorists have worked most extensively on Feferman's system and their variants; but set theorists have also availed themselves to Feferman's ideas (see Koellner [39]).

Also in philosophy Feferman's work has been fruitful not only for one topic. For once, Feferman advanced the use of truth as a primitive concept in the foundations of mathematics. Of course, this was his main motivation in the seminal *Reflecting on Incompleteness*. Generally Feferman didn't put so much emphasis on his theories as attempts to resolve the liar paradox; rather he was after theories of truth (and comprehension) that prove useful for foundational purposes. Of course, the paradoxes have to be addressed in some way, but the analysis of the paradoxes themselves doesn't figure very prominently in Feferman's work. Philosophers – among them Reinhardt [50], McGee [46] and Field [24] – have discussed Feferman's systems from the perspective of the theory of paradoxes. In this respect they also proved very fruitful.

One aspect of Feferman's work that hasn't been fully exploited is its relevance for the discussion about deflationism. Of course in some way, deflationists have employed ideas from Feferman's papers. In particular, Field's [24] recent research programme is very much motivated by deflationist or disquotationalist considerations and Field acknowledges Feferman's influence.

In particular, much of the work of Field and the work inspired by it can be seen as a continuation of Feferman's [15]: The challenge is to obtain the full T-schema in an extension of classical logic with a reasonably behaved conditional (or biconditional).

Deflationists have often seen truth as a device for expressing generalizations and infinite conjunctions (see Horwich [37] and Halbach [31]). Feferman's proof-theoretic analysis – more specifically the proof of the lower bound in Feferman [17, Sects. 4.3 and 5.3] – proceeds in terms of infinite conjunctions. While the discussion about the purpose of truth as a device of expressing infinite conjunctions usually remains somewhat metaphorical in the literature on deflationist, Feferman's analysis offers a precise conceptual analysis on the usability of truth for expressing infinite conjunctions. This is just one aspect of Feferman's work that offers scope for future philosophical research.

References

1. Aczel, P.: Frege structures and the notions of proposition, truth and set. In: Barwise, J., Keisler, H., Kunen, K. (eds.) The Kleene Symposium, pp. 31–59. North Holland, Amsterdam (1980)
2. Aczel, P.: The strength of Martin-Löf's intuitionistic type theory with one universe. In: Mietissen, S., Väänänen, J. (eds.) Proceedings of the Symposiums in Mathematical Logic in Oulu 1974 and Helsinki 1975, pp. 1–32. Department of Philosophy, University of Helsinki, Report no. 2 (1977)
3. Aczel, P., Feferman, S.: Consistency of the unrestricted abstraction principle using an intensional equivalence operator. In: Seldin, J.P., Hindley, J.R. (eds.) To H.B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism, pp. 67–98. Academics Press, New York (1980)
4. Burgess, J.P.: Friedman and the axiomatization of Kripke's theory of truth. In: Tennant, N. (ed.) Foundational Adventures: Essays in Honor of Harvey M. Friedman, pp. 125–148. College Publications, London (2014)
5. Cantini, A.: Notes on formal theories of truth. Zeitschrift für mathematische Logik und Grundlagen der Mathematik **35**, 97–130 (1989)
6. Cantini, A.: A theory of formal truth arithmetically equivalent to ID_1. J. Symb.Logic **55**, 244–259 (1990)
7. Cantini, A.: Logical Frameworks for Truth and Abstraction: An Axiomatic Study. Number 135 in Studies in Logic and the Foundations of Mathematics. Elsevier, Amsterdam, Lausanne, New York, Oxford, Shannon and Tokyo (1996)
8. Craig, W., Vaught, R.: Finite axiomatizability using additional predicates. J. Symb. Logic **23**, 289–308 (1958)
9. Feferman, S.: Transfinite recursive progressions of axiomatic theories. J.Symb. Logic **27**, 259–316 (1962)
10. Feferman, S.: Systems of predicative analysis I. J. Symb.Logic **29**, 1–30 (1964)
11. Feferman, S.: Systems of predicative analysis II. J. Symb. Logic **33**(2), 193–220 (1968)
12. Feferman, S.: A more perspicuous formal system for predicativity. In: Lorenz, K. (ed.) Konstruktionen versus Positionen, vol. I, pp. 68–93. de Gruyter, Berlin (1978)
13. Feferman, S.: Constructive theories of functions and classes. In: Boffa, M., van Dalen, D., McAloon, K. (eds.) Logic Colloquium '78: Proceedings of the Colloquium held in Mons, August 1978, pp. 159–224. North-Holland, Amsterdam, New York and Oxford (1979)
14. Feferman, S.: Iterated inductive fixed-point theories: application to Hancock's conjecture. In: Metakides, G. (ed.) Patras Logic Symposion, pp. 171–195, Amsterdam, North Holland (1982)
15. Feferman, S.: Towards useful type-free theories I. J. Symb. Logic **49**, 75–111 (1984)
16. Feferman, S.: Reflecting on incompleteness (outline draft). handwritten notes (1987)
17. Feferman, S.: Reflecting on incompleteness. J. Symb. Logic **56**, 1–49 (1991)
18. Feferman, S.: Axioms for determinateness and truth. Rev. Symb. Logic **1**, 204–217 (2008)
19. Feferman, S.: Axiomatizing truth: Why and how. In: Berger, U., Diener, H., Schuster, P. (eds.) Logic, Construction, Computation, 185–200. Ontos, Frankfurt (2012)
20. Feferman, S., Strahm, T.: The unfolding of non-finitist arithmetic. Rev. Symb. Logic **3**, 665–689 (2000)
21. Feferman, S., Strahm, T.: Unfolding finitist arithmetic. Ann. Pure Appl. Logic **104**, 75–96 (2010)
22. Field, H.: Tarski's theory of truth. J. Philos. **69**, 347–375 (1972)
23. Field, H.: Deflating the conservativeness argument. J. Philos. **96**(10), 533–40 (1999)
24. Field, H.: Saving Truth From Paradox. Oxford University Press, Oxford (2008)
25. Friedman, H., Sheard, M.: An axiomatic approach to self-referential truth. Ann. Pure Appl. Logic **33**, 1–21 (1987)
26. Fujimoto, K.: Relative truth definability of axiomatic truth theories. Bull. Symb. Logic **16**, 305–344 (2010)
27. Fujimoto, K.: Autonomous progression and transfinite iteration of self-applicable truth. J. Symb. Logic **76**, 914–945 (2011)

28. Fujimoto, K.: Classes and truths in set theory. Ann. Pure Appl. Logic **163**, 1484–1523 (2012)
29. Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik **38**, 173–198 (1931)
30. Gupta, A.: Truth and paradox. J. Philos. Logic **11**, 1–60 (1982)
31. Halbach, V.: Disquotationalism and infinite conjunctions. Mind **108**, 1–22 (1999)
32. Halbach, V.: How not to state the T-sentences. Analysis **66**, 276–280 (2006). Correction of printing error in 67, 268
33. Halbach, V.: Axiomatic Theories of Truth, revised edition. Cambridge University Press, Cambridge (2014). first edition 2011
34. Halbach, V., Horsten, L.: Axiomatizing Kripke's theory of truth. J. Symb. Logic **71**, 677–712 (2006)
35. Herzberger, H.G.: Notes on naive semantics. J. Philos. Logic **11**, 61–102 (1982)
36. Horsten, L., Leigh, G., Leitgeb, H., Welch, P.: Revision revisited. Rev. Symb. Logic **5**, 642–664 (2012)
37. Horwich, P.: Truth, 2nd edn. Oxford University Press, Oxford (1998). first edition 1990
38. Jäger, G., Kahle, R., Setzer, A., Strahm, T.: The proof-theoretic analysis of transfinitely iterated fixed point theories. J. Symb. Logic **64**(1), 53–67 (1999)
39. Koellner, P.: On reflection principles. Ann. Pure Appl. Logic **157**, 206–219 (2009)
40. Kreisel, G.: Ordinal logics and the characterization of informal notions of proof. In: Todd, J.A. (ed.) Proceedings of the International Congress of Mathematicians, Edinburgh 1958, pp. 289–299, Cambridge University Press, Cambridge (1960)
41. Kreisel, G., Lévy, A.: Reflection principles and their use for establishing the complexity of axiomatic systems. Zeitschrift für mathematische Logik und Grundlagen der Mathematik **14**, 97–142 (1968)
42. Kripke, S.: Outline of a theory of truth. J. Philos. **72**, 690–712 (1975). reprinted in Martin (1984)
43. Leitgeb, H.: Poppers Wahrheitstheorie(n). In: Morscher, E (ed.) Was wir Karl R. Popper und seiner Philosophie verdanken: Zu seinem 100. Geburtstag, ProPhil · Projekte zur Philosophie, pp. 185–217. Academia Verlag, Sankt Augustin (2002)
44. Leitgeb, H.: What theories of truth should be like (but cannot be). In: Blackwell Philosophy Compass 2/2, pp. 276–290. Blackwell (2007)
45. Manaster, A.B., Payne, T.H., Harrah, D.: Meeting of the Association of Symbolic Logic, San Diego 1979. J. Symb. Logic **46**, 199 (1981)
46. McGee, V.: Truth, Vagueness, and Paradox: An Essay on the Logic of Truth. Hackett Publishing, Indianapolis and Cambridge (1991)
47. McGee, V.: In praise of the free lunch: Why disquotationalists should embrace compositional semantics. In: Hendricks, V.F., Pedersen, S.A. (eds.) Self-Reference, vol. 178, pp. 95–120. CSLI Publications, Stanford (2006)
48. Putnam, H.: A comparison of something with something else. New Lit. Hist. **17**(1), 61–79 (1985). Reprinted in (Putnam 1994, 330-50)
49. Reinhardt, W.: Remarks on significance and meaningful applicability. In: de Alcantara, L.P. (ed.) Mathematical Logic and Formal Systems: A Collection of Papers in Honor of Professor Newton C.A. Da Costa. Lecture Notes in Pure and Applied Mathematics, vol. 94, pp. 227–242. Marcel Dekker inc. (1985)
50. Reinhardt, W.: Some remarks on extending and interpreting theories with a partial predicate for truth. J. Philos. Logic **15**, 219–251 (1986)
51. Russell, B.: On some difficulties in the theory of transfinite numbers and order types. Proc. Lond. Math. Soc. **4**, 29–53 (1906)
52. Russell, B.: Mathematical logic as based on the theory of types. Am. J. Math. **30**, 222–262 (1908)
53. Schindler, T.: Type-free truth. Ph.D. Thesis, Ludwig-Maximilians-Universität München (2014)
54. Schütte, K.: Proof Theory. Springer, Berlin (1977)
55. Schwichtenberg, H., Wainer, S.: Proofs and Computations. Perspectives in Logic. Association of Symbolic Logic and Cambridge University Press, Cambridge (2012)

56. Shapiro, S.: Proof and truth: through thick and thin. J. Philos. **95**, 493–521 (1998)
57. Simpson, S.: Subsystems of Second Order Arithmetic. Springer, Berlin (1998)
58. Tarski, A.: Der Wahrheitsbegriff in den formalisierten Sprachen. Studia Philosophica Commentarii Societatis Philosophicae Polonorum **1**, 261–405 (1935). translation of Tarski (1936) by L. Blaustein, translated as 'The Concept of Truth in Formalized Languages' in (Tarski 1956, 152–278)
59. Turing, A.: Systems of logic based on ordinals. Proc. Lond. Math. Soc. **45**, 161–228 (1939)
60. Visser, A.: Four-valued semantics and the liar. J. Philos. Logic **13**, 181–212 (1984)

Author Biographies

Andrea Cantini is Professor of Logic at the University of Florence. His research interests include proof theory, axiomatic theories of truth, theories of operations and classes, history and philosophy of contemporary logic. Among his publications, the monograph *Logical Frameworks for Abstraction and Truth* (Elsevier/North–Holland 1996).

Kentaro Fujimoto is Lecturer in Mathematical and Philosophical Logic at the University of Bristol. His research interests include proof theory, axiomatic theories of truth, philosophy of logic and mathematics. He earned his master's degree from Stanford University under the supervision of Professor Solomon Feferman.

Volker Halbach is Professor of Philosophy at the University of Oxford and Fellow of New College. His fields of research are logic, philosophy of mathematics, philosophy of language, and epistemology. Among his publications is the monograph *Axiomatic Theories of Truth* (Cambridge University Press, 2011).