Citation for published version (APA):
Nicolai, C., Fischer, M., & Piazza, M. (in press). A reflection principle for Kripke-Feferman truth. *JOURNAL OF PHILOSOPHY*.

A Reflection Principle for Kripke-Feferman truth

Carlo Nicolai, Martin Fischer, and Mario Piazza

Forthcoming in *The Journal of Philosophy*[*]

Abstract

The Kripke-Feferman theory of truth is arguably the most discussed formal approach to primitive truth. KF is a classical axiomatization of fixed-point semantics and features reasonable mathematical strength. However, KF has been harshly criticized: KF can prove claims that are, according to the Kripke-Feferman theory itself, unsound. Examples of this phenomenon involve both logical and non-logical axioms of KF. In a thought-provoking paper, Reinhardt acknowledges this challenge to KF and offers strategies to overcome it. Reinhardt argues that the non-significant sentences of KF can be justified by invoking a very interesting principle that shares several features with set-theoretic reflection. Yet Reinhardt does not discuss the details of this proposal. In the paper, we provide precise renderings of Reinhardt's reflection principle; we show that some of these renderings can be consistently incorporated into KF, while others are provable within KF itself. We elucidate how the principles can be used to realize Reinhardt's project of justifying non-significant theorems of KF.

1 The Project

The Kripke-Feferman theory of truth is arguably the most discussed formal approach to primitive truth. It has found application in various theoretical contexts, such as Feferman's characterization of predicativity [Fef91], philosophy of language and semantics [Gla15], analysis of the Liar paradox [Mau04, Rei86], with some forays even in the philosophy of mind [Koe18, Ste18]. In all of these areas, the Kripke-Feferman theory is treated both as a useful logical tool and as a way of characterizing some intuitive features of the concept of truth itself.

Kripke-Feferman truth does not commonly refer to a single theory, but to a collection of (slightly) different theories. For example, some versions of the theory enforce the truth

[*] Martin Fischer would like to thank the audience of the 'Luxemburger Zirkel' workshop on logic and philosophy of mathematics. Carlo Nicolai acknowledges funding from a Marie Sklodowska-Curie action funded by the EU under the Horizon Europe Research and Innovation Programme, the PLEXUS project (Grant agreement No 101086295) and from the UK Arts and Humanities Research Council (project AH/V015516/1).

predicate to be consistent, while others require no truth-value gaps. In this work we are mostly concerned with a version of Kripke-Feferman that does not demand truth to be consistent or complete. When we use the label KF, we are referring to this version. There are good reasons for employing this version of Kripke-Feferman truth. One of them will be discussed shortly: the consistency or completeness axioms decide some paradoxical sentences, such as the Liar. In doing so, however, they intensify the already existing tension between the notions of provability and truth of KF. The other reason concerns a lack of formal robustness of the notions of consistency and completeness added to KF; as discussed in [Nic22], consistency and completeness are theoretically equivalent over KF.

KF captures the class of fixed-point ω-models as defined in [Kri75], and serves as an axiomatization of fixed-point semantics. Notably, it is a *classical* axiomatization of a natural, type-free notion of truth. As demonstrated in Feferman's work, this feature endows KF with reasonable mathematical strength. This has been shown to be closely linked to the classicality of the theory, since the mathematical power of non-classical axiomatizations of fixed-point semantics is more limited [HN18].

For the philosophical understanding of truth, two further features of KF stand out. The first is that KF delivers a principled restricted version of the T-schema to sentences that are, in the terminology employed by [Rei86, p. 231], *significant*:

$$\mathtt{S}(\ulcorner A \urcorner) \rightarrow (\mathtt{T}\ulcorner A \urcorner \leftrightarrow A), \tag{1}$$

Here $\mathtt{S}(x)$ is a defined predicate for 'x is true or $\neg x$ is, while x and $\neg x$ are not both true'; it expresses that x has a determinate truth value. (1) can be seen as a realization of Kripke's original idea that the T-schema needs to be restricted to "meaningful" sentences. In addition, S validates entirely plausible principles. The second feature relates to the axiomatic approach itself; if axioms are given instead of a model-theoretic construction, one does not need to ascend to a more expressive metalanguage, nor is one required to restrict quantifiers to a specific set.[1]

Kripke-Feferman truth has also been harshly criticized. A well-known problem concerns the provability of theorems that are in a sense incompatible with its truth predicate. For example, in a version of Kripke-Feferman featuring a consistent truth predicate, one can prove sentences that are deemed untrue by the theory. The Liar sentence λ is an example: by proving λ, one also proves $\neg\mathtt{T}\ulcorner\lambda\urcorner$. McGee [McG91, p. 106] summarizes the problem as follows:

> If we accept the Kripke-Feferman theory, this simple connection between truth and proof will be broken. In the Kripke-Feferman theory, we can prove things that are, according to the Kripke-Feferman theory, untrue.

In KF, the situation is less dramatic. KF only derives some counter-intuitive claims, such

[1] For a more extensive discussion of this point, see [Hor11, Ch. 2].

as the disjunction

$$(2) \qquad (\lambda \wedge \neg \mathtt{T}\ulcorner\lambda\urcorner) \vee (\neg\lambda \wedge \mathtt{T}\ulcorner\lambda\urcorner)$$

Each horn of the disjunction witnesses, again, an asymmetry between proof and truth.

Prima facie, (2) does not appear to be worrisome. After all, any theory T that proves the diagonal lemma, and for which a provability predicate $\mathrm{Prov}_T(x)$ satisfies the usual meaning postulates,[2] will entail

$$(3) \qquad (\gamma_T \wedge \neg\mathtt{Prov}_T(\ulcorner\gamma_T\urcorner)) \vee (\neg\gamma_T \wedge \mathtt{Prov}_T(\ulcorner\gamma_T\urcorner)),$$

where γ_T is a Gödel sentence for T. (3) can hardly be regarded as breaking the link between internal and external provability for a mathematical theory.

A stronger version of the objection is as follows. McGee's thesis that provability should be a guide to truth can be expressed in the language of KF by means of a reflection principle

(4) everything provable in KF is true.

The combination of (2) with (4) and the KF axioms leads to inconsistency (see Fact 1). This results in an outright contradiction if one assumes the consistency axiom, whereas in KF it leads to an *internal* inconsistency, meaning the existence of sentences that are both true and false. So, the objection goes, KF cannot be closed under standard soundness assertions. That being said, there are soundness assertions that *are* compatible with KF. One is obtained by reformulating (4) as

(5) every significant theorem of KF is true.

We are not advocating (5) in this paper, although it appears to be a promising alternative to standard proof-theoretic reflection in the context of KF, as well as an interesting starting point for a plausible reaction to (this form of) the unsoundness challenge.

The phenomenon affects not only the Liar and other well-known paradoxical sentences. Any universally quantified claim that is provable in KF, but has non-significant instances, cannot itself be significant in KF. This sense of "universally quantified" can be quite broad, encompassing both meta-theoretical quantification over instances of schemata and object-theoretical quantification. This category includes both logical and non-logical axioms of KF. An unsoundness affecting the theory's axioms is particularly worrysome. The Liar and other paradoxical sentences may be seen as quirks or singularities to be properly contained. The KF-axioms, however, lie at the heart of Kripke-Feferman truth. *Anyone who advocates KF as a theory of truth should be prepared to justify the axioms of KF in light of their paradoxical instances. This is one of the main tasks of this paper.*

[2] One does not formally require any principle for $\mathtt{Prov}_T$ to be derivable in the theory, but it makes sense for our discussion that T knows something about $\mathtt{Prov}_T$.

In the insightful paper [Rei86], Reinhardt recognizes many of these challenges to KF and proposes strategies to justify some of its fundamental consequences. He initially proposes an instrumentalist reading of KF-theorems: the internal or provably true statements of KF are regarded as the significant ones. This proposal has been extensively studied, but it also has well-known drawbacks: as shown in [HH06], the provably true sentences of KF rely essentially on the non-significant axioms of KF for their proofs.[3] To provide a justification for the non-logical axioms of KF, Reinhardt puts forward an additional proposal, which is less well-known and will be the focus of our discussion. Reinhardt argues that the non-significant sentences of KF can be justified by invoking a principle that is similar to set-theoretic reflection:

> If A, then there is an internal interpretation of the truth predicate of A which makes A *true*. [...] The principle is a reflection principle: because something is true (in the absolute sense, a partial predicate), there are models (with ordinary, total predicates) which reflect this. ([Rei86], p. 237.)

Reinhardt does not discuss the details of this proposal; he also does not indicate whether this reflection principle can be expressed in the language of KF.

However, Reinhardt's principle captures some fundamental features of KF that can be used to justify its axioms. The *intended* range of the truth predicate of KF is the collection of grounded sentences, i.e. those sentences whose truth or falsity only depends on arithmetical facts (more on this in Section 6).[4] In addition, the statements provably true in KF are guaranteed to be grounded. Although the KF axioms as they stand are not grounded, a suitable regimentation of Reinhardt's principle will enable us to extract the grounded content from the KF axioms, thereby providing formal and conceptual tools for their justification. In the paper, we provide different formulations of Reinhardt's reflection principle; we show that some formulations can be consistently added to KF, and others are provable in KF. We then explain how the principles can be used to justify the non-logical axioms of KF and appropriate restrictions of the logical ones. We consider our results and subsequent discussion as a vindication of KF against the unsoundness charges described above.

2 Kripke-Feferman and its "Unsoundness"

We start with the language of arithmetic with its standard signature $\{0, \mathtt{S}, +, \times\}$. We then extend this language by adding a finite number of predicates and function symbols for primitive recursive functions and properties. The primitive predicates are intended to be shorthand for the corresponding characteristic primitive recursive functions. This expanded language is referred to as $\mathcal{L}_{\mathbb{N}}$. In particular, we will need primitive predicates

[3] There are strategies of defending Reinhardt's use of the provably true sentences of KF. See for instance [Nic17, CS23].

[4] This is recognized by Feferman himself, cf. [Fef12].

$\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(x)$ to denote the notion of a sentence in the language $\mathcal{L}_{\mathtt{T}} := \mathcal{L}_{\mathbb{N}} \cup \{\mathtt{T}\}$, and $\mathtt{ct}(x)$ to denote the notion of a closed term in $\mathcal{L}_{\mathbb{N}}$. We will write $\mathtt{val}(x)$ for a suitable formula expressing in $\mathcal{L}_{\mathtt{T}}$ a primitive recursive evaluation function of these finitely many additional primitive functions. We will also assume function symbols for the numeral and substitution functions whose properties can be verified in Peano Arithmetic ($\mathtt{PA}$), our theory of syntax: $\mathtt{PA} \vdash \mathtt{val}(\mathtt{num}(x)) = x$ and $\mathtt{PA} \vdash \ulcorner A(t) \urcorner = \ulcorner A(v) \urcorner(\ulcorner t \urcorner / \ulcorner v \urcorner)$. We broadly follow the conventions in [Hal14], including the dot notation for formal syntactic functions and operations. The expression $\ulcorner \varphi(\dot{x}) \urcorner$ abbreviates the result of substituting, in $\ulcorner \varphi(v) \urcorner$, the variable $\ulcorner v \urcorner$ by the code for the x-th numeral. We abbreviate quantified expressions such as $\forall x(\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(x) \rightarrow ...$ and $\forall x(\mathtt{ct}(x) \rightarrow ...$ by $\forall \varphi(...$ and $\forall t(...$, respectively.

We work with a Hilbert-style calculus for classical predicate logic with equality where Modus Ponens is the only rule of inference [End01]. Our assumptions mean that the axioms of $\mathtt{PA}$ include the primitive recursive equations for the additional primitive recursive functions, as well as an induction schema.

($\mathtt{IND}(\mathcal{L}_{\mathtt{T}})$) $\quad A(0) \wedge \forall y(A(y) \rightarrow A(\mathtt{S}(y))) \rightarrow \forall y\, A(y) \quad \text{for } A(v) \in \mathcal{L}_{\mathtt{T}}.$

The Kripke-Feferman system $\mathtt{KF}$ we will be concerned with is formulated in the language $\mathcal{L}_{\mathtt{T}}$. It extends classical logic with equality with the basic axioms of $\mathtt{PA}$ (including recursive equations for finitely many primitive recursive functions not in $\{0, \mathtt{S}, \times, +\}$) and $\mathtt{IND}(\mathcal{L}_{\mathtt{T}})$. The truth-theoretic axioms of $\mathtt{KF}$ amount to the clauses of the positive inductive definition of the Strong-Kleene truth-conditions:

($\mathtt{KF1}$) $\quad \forall \varphi \forall s, t(\varphi = (s \underset{\cdot}{\neq} t) \rightarrow (\mathtt{T}(\varphi) \leftrightarrow \mathtt{val}(s) \neq \mathtt{val}(t)))$

($\mathtt{KF2}$) $\quad \forall \varphi \forall s, t(\varphi = (s \underset{\cdot}{=} t) \rightarrow (\mathtt{T}(\varphi) \leftrightarrow \mathtt{val}(s) = \mathtt{val}(t)))$

($\mathtt{KF3}$) $\quad \forall \varphi(\mathtt{T} \ulcorner \mathtt{T} \dot{\varphi} \urcorner \leftrightarrow \mathtt{T} \varphi)$

($\mathtt{KF4}$) $\quad \forall \varphi(\mathtt{T} \ulcorner \neg \mathtt{T} \dot{\varphi} \urcorner \leftrightarrow \mathtt{T} \underset{\cdot}{\neg} \varphi)$

($\mathtt{KF5}$) $\quad \forall \varphi \forall \psi(\mathtt{T}(\varphi \underset{\cdot}{\wedge} \psi) \leftrightarrow (\mathtt{T} \varphi \wedge \mathtt{T} \psi))$

($\mathtt{KF6}$) $\quad \forall \varphi \forall \psi(\mathtt{T} \underset{\cdot}{\neg}(\varphi \underset{\cdot}{\wedge} \psi) \leftrightarrow (\mathtt{T} \underset{\cdot}{\neg} \varphi \vee \mathtt{T} \underset{\cdot}{\neg} \psi))$

($\mathtt{KF7}$) $\quad \forall v \forall \varphi(\mathtt{T} \underset{\cdot}{\forall} v \varphi \leftrightarrow \forall t\, (\mathrm{Sent}_{\mathcal{L}_{\mathtt{T}}}(\varphi(t/v)) \rightarrow \mathtt{T} \varphi(t/v)))$

($\mathtt{KF8}$) $\quad \forall v \forall \varphi(\mathtt{T} \underset{\cdot}{\neg} \underset{\cdot}{\forall} v \varphi \leftrightarrow \exists t\, (\mathrm{Sent}_{\mathcal{L}_{\mathtt{T}}}(\varphi(t/v)) \wedge \mathtt{T} \underset{\cdot}{\neg} \varphi(t/v)))$

($\mathtt{KF9}$) $\quad \forall x(\mathtt{T}(x) \rightarrow \mathrm{Sent}_{\mathcal{L}_{\mathtt{T}}}(x))$

($\mathtt{KF10}$) $\quad \forall \varphi(\mathtt{T}(\underset{\cdot}{\neg} \underset{\cdot}{\neg} \varphi) \leftrightarrow \mathtt{T} \varphi)$

The axiomatization of $\mathtt{KF}$ provided differs from the usual formulation given for instance in [Hal14], mainly due to the formulation of $\mathtt{KF3}$ and $\mathtt{KF4}$, which are now restricted to sentences only. However, our version is proof-theoretically equivalent to Halbach's version. Therefore, for our purposes we can safely reason with the restriction to $\mathtt{KF3}$ and $\mathtt{KF4}$. It can be shown that our version of $\mathtt{KF}$ is able to define all Tarskian truth predicates up to any ordinal α smaller than ε_0 – compare with [Hal14, Lemma 15.24]. Therefore, our version of $\mathtt{KF}$ proves at least the same arithmetical sentences as the version of $\mathtt{KF}$ with the general version of $\mathtt{KF3}$ and $\mathtt{KF4}$. Moreover, it is easy to see that it cannot prove more, as our version of $\mathtt{KF}$ is a subtheory of the one with unrestricted $\mathtt{KF3}$ and $\mathtt{KF4}$.

The internal theory of the version of $\mathtt{KF}$ we presented, that is the class of $\mathcal{L}_{\mathtt{T}}$-sentences A such that $\mathtt{KF} \vdash \mathtt{T}\ulcorner A \urcorner$, is governed by a four-valued logic called First-Degree Entailment $\mathtt{FDE}$ [AB63].[5] In other words, the truth predicate of $\mathtt{KF}$ allows for interpretations of $\mathtt{T}$ that admit truth-value gaps or gluts. To force only gaps or only gluts, one needs to add, respectively,

(CONS) $\qquad \forall \varphi(\mathtt{T}\neg\varphi \to \neg\mathtt{T}\varphi),$

(COMP) $\qquad \forall \varphi(\neg\mathtt{T}\varphi \to \mathtt{T}\neg\varphi).$

$\mathtt{KF}$ axiomatizes fixed-point semantics in the sense of [Kri75]. In particular, the version of $\mathtt{KF}$ we are considering is sound with respect to the four-valued models obtained by applying the Tarski-Knaster theorem to the monotone operator associated with the Kleene evaluation schema – see for instance, [Hal14, p. 206] – and to the complete lattice given by $\mathcal{P}(\omega)$ of possible extensions of $\mathtt{T}$. The sense in which $\mathtt{KF}$ axiomatizes a suitable class of fixed points is customarily explained with reference to a form of "completeness", or $\mathbb{N}$-categoricity [FHKS15]:[6]

(6) $\qquad (\mathbb{N}, S) \vDash \mathtt{KF}$ iff S is a fixed point of the Kleene operator.

The minimal and maximal elements of the lattice of fixed points – that we dub $\mathcal{I}$ and $\mathcal{G}$, respectively – amount to natural models of $\mathtt{KF} + \mathtt{CONS}$ and $\mathtt{KF} + \mathtt{COMP}$, respectively. It will be useful later on to look at the ordinal "stages" of the construction of $\mathcal{I}$, the collection of all grounded sentences of $\mathcal{L}_{\mathtt{T}}$ in the sense of Kripke; $\mathcal{I}$ can in fact be seen as the result of the closure of the empty set under the Kleene operator. With Φ^α standing for α-many iterations of such an operator, we have

$$\mathcal{I} = \bigcup_{\alpha \in \mathrm{Ord}} \Phi^\alpha,$$

where

$$\Phi^\alpha = \Phi(\Phi^{<\alpha}) \text{ and } \Phi^{<\alpha} = \bigcup_{\beta<\alpha} \Phi^\beta.$$

For A a sentence of $\mathcal{L}_{\mathtt{T}}$, its inductive norm $|A|_\Phi$ is defined as the minimal stage for which $A \in \Phi^\alpha$ – and it is set to be $\omega_1^{\mathtt{CK}}$, the first non-recursive ordinal, if $A \notin \mathcal{I}$.[7]

[5] A full proof of this claim can be found in [Nic17], where it is shown that the internal theory of $\mathtt{KF}$ is identical to the $\mathtt{FDE}$-version of the theory $\mathtt{PKF}$ from [HH06] extended with a suitable rule of transfinite induction.

[6] This fact was already observed in [Fef91].

[7] It is worth remarking that no grounded sentence in the sense of [Kri75] can have ordinal norm $\omega_1^{\mathtt{CK}}$,

In Section 1 we claimed that $\mathtt{KF}$ is affected by forms of unsoundness. The next observation provides a first assessment of this unsoundness phenomenon, as it tells us that $\mathtt{KF}$ is incompatible with standard, formal soundness assertions.

Fact 1.

(i) *Some instances of the logical axiom schemata cannot be proved true in* $\mathtt{KF}$*;*[8]

(ii) $\mathtt{KF}$ *cannot prove the truth of any of* $\mathtt{KF}$3-$\mathtt{KF}$8*,* $\mathtt{KF}$10*;*[9]

(iii) $\mathtt{KF}$ *cannot establish that Modus ponens (and with it that the material conditional) is truth preserving. That is, there are some sentences A, B of $\mathcal{L}_{\mathtt{T}}$ such that, within* $\mathtt{KF}$*:*

$$\mathtt{T}\ulcorner A\urcorner, \mathtt{T}(\ulcorner A \to B\urcorner) \nvdash \mathtt{T}\ulcorner B\urcorner$$

(iv) $\mathtt{KF}$ *cannot prove that all instances of* $\mathrm{IND}_{\mathcal{L}_{\mathtt{T}}}$ *are true.*

As an immediate corollary, we obtain that the result of extending $\mathtt{KF}$ with each of the following principles

(GRP(KF)) $\quad \forall\varphi(\mathtt{Prov}_{\mathtt{KF}}(\varphi) \to \mathtt{T}\varphi)$

(ARP(KF)) $\quad \forall\varphi(\mathtt{Ax}_{\mathtt{KF}}(\varphi) \to \mathtt{T}\varphi)$

(NEC) $\quad$ if $\mathtt{KF} \vdash \varphi$, then $\vdash \mathtt{T}\varphi$.

results in internal inconsistency.[10]

Beside these forms of "unsoundness" of $\mathtt{KF}$, there are also (restricted) soundness assertions that are fully compatible with it. As mentioned, the internal theory of $\mathtt{KF}$ is

$$\mathtt{IKF} = \{A \in \mathcal{L}_{\mathtt{T}} \mid \mathtt{KF} \vdash \mathtt{T}\ulcorner A\urcorner\}.$$

$\mathtt{IKF}$ *is* sound in the sense that every A that $\mathtt{KF}$ proves to be true is in the extension of the minimal fixed point with $|A|_{\Phi} < \varepsilon_0$. In fact, Reinhardt already mentions the possibility of consistently adding a reflection principle for $\mathtt{IKF}$ to $\mathtt{KF}$ [Rei86, p. 234]. He notices that even the truth of the following proof-theoretic reflection for $\mathtt{IKF}$,

$$\forall x(\mathtt{Prov}_{\mathtt{IKF}}(x) \to \mathtt{T}(x)), \tag{7}$$

because of general facts concerning inductive definability.

[8] More formally, if $\Phi(\mathcal{A}_1, \dots, \mathcal{A}_n)$ is a logical axiom schema of the Hilbert-style calculus assumed– where $\Phi(\cdot)$ is the schema template, and the $\mathcal{A}_i$'s are a meta-variables for $\mathcal{L}_{\mathtt{T}}$ formulae –, then there are formulae $A_1, \dots, A_n$ such that $\Phi(A_1/\mathcal{A}_1 \dots, A_n/\mathcal{A}_n)$ cannot be proved true by $\mathtt{KF}$. This generalizes to other natural Hilbert-style axiomatizations of classical logic.

[9] $\mathtt{KF}$1, $\mathtt{KF}$2, by contrast, are provably true.

[10] The proof of all claims is analogous. Since (GRP(KF)) entails (ARP(KF)) and (NEC), it suffices to establish the claim for the latter principles. For (NEC), one reasons as in item (i) of the previous Fact. For (ARP(KF)), we can reason as in (ii) in Fact 1.

can be consistently added to $\mathtt{KF}$. It is fairly straightforward to check that the addition of (7) to $\mathtt{KF}$ does not exclude any ω-model. In fact, the restriction to significant sentences discussed in the previous section reduces to (7). Let

$$\mathtt{S}(x) :\leftrightarrow (\mathtt{T}x \vee \mathtt{T}(\neg x)) \wedge \neg(\mathtt{T}x \wedge \mathtt{T}(\neg x)).$$

Then:

$$\forall\varphi(\mathtt{Prov}_{\mathtt{KF}}(\mathtt{S}\dot{\varphi} \wedge \varphi) \to \mathtt{T}\varphi) \tag{8}$$

is provably equivalent over $\mathtt{KF}$ to (7): the equivalence follows directly from the formalization in $\mathtt{KF}$ of (1); that is, $\mathtt{KF}$ proves that

$$\forall\varphi\, \mathtt{Prov}_{\mathtt{KF}}(\mathtt{S}\dot{\varphi} \to (\mathtt{T}\dot{\varphi} \to \varphi)). \tag{9}$$

3 KF and Reinhardt's proposal

Reinhardt's project in [Rei86] is to address the logical and semantic paradoxes. He is inspired by scattered remarks made by Gödel, who suggested reconsidering the assumption that concepts like truth or predication can be meaningfully applied to all arguments [Göd95]. This idea is already implicit in the theory of types, but Gödel believed that typing was ultimately inadequate as it would omit some "intuitive" logical claims.[11]

To implement Gödel's idea, Reinhardt employs Kripke's theory of truth as outlined in [Kri75]. Reinhardt highlights that, within Kripke's fixed-point models, truth functions as both a partial *predicate* and total *class*. In the classical metatheory in which fixed-point semantics is formulated, the sentences

$$\mathtt{T}\ulcorner A \urcorner \vee \mathtt{T}\ulcorner \neg A \urcorner \tag{10}$$

$$\mathtt{T}\ulcorner A \urcorner \vee \neg\mathtt{T}\ulcorner A \urcorner \tag{11}$$

are not equivalent. Take for instance $\mathcal{I}$: (10) may not be satisfied by the model $(\mathbb{N}, \mathcal{I})$ – so, $\mathtt{T}$ can be seen as partial –, whereas (11) trivially is – so, $\mathtt{T}$ can also be seen as total. It is this phenomenon that gives rise to the possibility of axiomatizing Kripke's construction *externally*, via $\mathtt{KF}$, and *internally*, via theories in nonclassical logic such as Halbach and Horsten's $\mathtt{PKF}$.[12]

[11] For instance, as reported by Hao Wang, according to Gödel

> [the type-theoretic hierarchy] cannot satisfy the condition of including the concept of concept which applies to itself or the universe of all classes that belong to themselves. To take such a hierarchy as the theory of concepts is an example of trying to eliminate the intensional paradoxes in an arbitrary manner. [Wan96, p. 278]

[12] Actually, nonclassical axiomatizations fixed point semantics based on $\mathtt{FDE}$ and variants thereof may

Reinhardt's idea is to employ KF to provide a *theory* whose theorems are all significant. This theory is IKF, introduced earlier. The emphasis on 'theory' is key: by looking at axiomatic systems, Reinhardt attempts to sidestep the need for a set-theoretic, classical meta-theory. By selecting only sentences that are *provably true* in KF, IKF addresses the mismatch between proof and truth affecting KF. Halbach and Horsten [HH06] have shown that the use of some non-significant axiom of KF is required to obtain IKF. Another feature of IKF is that it does not come with a natural recursive axiomatization in classical logic. One way to extract a recursive set from the recursively enumerable definition of IKF is via Craig's reaxiomatization theorem. However, it is far from clear that such a re-axiomatization renders the classical principles of KF *eliminable*, just like Craig's theorem does not make superfluous the theoretical terms of a scientific theory [Put65]. It should also be mentioned that Nicolai in [Nic17] has shown that PKF with the addition of transfinite induction for any ordinal smaller than ε_0 is a rather natural *nonclassical* axiomatization of IKF.

However, we are interested in another proposal contained in Reinhardt's paper. After introducing IKF, the significant core of KF, he attempts to justify also the *non-significant* consequences of KF. Among the non-significant consequences of KF, there are of course the logical and non-logical axioms of KF. Here is the relevant passage from Reinhardt's paper:

> But we all know that formalists are tricky folks, always tending to sneak in some interpretation by the back door, while pretending to make merely formal manipulations. Don't we assign some truth to the axioms of KF, for example? [...] In particular, what justifies the use of such sentences, along with classical syntax and logic in the formal theory? (Especially since the application of classical logic and syntax to sentences involving partial predicates already leads to such non-significant sentences, for example $\forall x(\mathtt{T}x \vee \neg\mathtt{T}x)$.) There may in this case be some subtle proof theoretic justification; this is an interesting question. I wish to suggest a simple general principle, however. The principle is a reflection principle: because something is true (in the absolute sense, a partial predicate) there are models (with ordinary total predicates) which reflect this. *I do not attempt to state this precisely and significantly here.*[13] Here I only wish to state the following formal principle which may be added to our formal theory. Writing A^{T_0,F_0} for the result of replacing T by T_0, etc. in A, the principle is
>
> $$\text{(II)} \qquad A \to \exists T_0, F_0(\text{total})A^{T_0,F_0}$$

not be fully faithful to the *reasoning* available in fixed points. In PKF, one has inferences of form $\lambda \Rightarrow \lambda$ that involve non-significant sentences. Some sub-structural options fare better; [NR23] develop a substructural version of PKF, called RKF, in which every step in a proof is significant.

[13] Our emphasis.

> [...] Since all total predicates have total truth predicates, this means that
>
> $$A \to \exists T_0, F_0 T[A^{T_0,F_0}]$$
>
> Thus the formal theory allows for its own significant interpretation (at least piecewise). ([Rei86], p. 237)

In what follows, we continue Reinhardt's project, and state *precisely and significantly* the principle (II).

Our guiding intuition is as follows. A substantial class of non-significant theorems of KF – following Reinhardt's terminology – display some significant content. By refining Reinhardt's principle, we can, in favorable cases, *extract* the significant content from these non-significant theorems. These favorable cases are the ones involving a uniform distribution of truth-theoretic content such as the non-logical axioms of KF.

More specifically, when introducing KF, we stated that the intended range of its truth predicate is the collection of *grounded sentences.* In a general claim such as axiom KF10, quantification over sentences ranges over grounded instances as well as ungrounded ones. Those grounded instances can be extracted by means of a regimentation of Reinhardt's principle to yield a restricted quantified principle that features only *total* truth predicates. More generally, the procedure implicit in our version of Reinhardt's principle (II) will uncover significant content by restricting quantifiers in suitable KF-theorems to *their grounded instances.* The procedure just sketched crucially employs total truth predicates: these will take the form of type-free adaptations of *Tarskian* truth predicates that keep track of the dependency structure of semantic content from basic arithmetical facts.

4 From Set Theory to Arithmetic

Reinhardt's principle is inherently set-theoretic, and it displays a striking similarity with a reflection principle due to Bernays.[14] In this section, we will elaborate on the connection between (II) and class-theoretic reflection to arrive at our formulation of (II) in $\mathcal{L}_{\mathtt{T}}$, which will then be developed in the next section.

Reinhardt explicitly refers to the principle (II) as a set- or class-theoretic reflection principle. Reflection principles in set theory embody the idea that the universe of sets cannot be characterized by a unique property expressible in the language of set theory (including its higher-order extensions), in the sense that any A true in $\mathbb{V}$ is already true in some $\mathbb{V}_\alpha$. By restricting our attention to second-order parameters only, this idea can

[14] See [Ber76]: the class-theoretic reflection principle in question is,

$$A^X \to \exists x(\text{transitive } x \text{ and } A^{\mathcal{P}(x)}(X \cap x)).$$

Here $A^{\mathcal{P}(x)}$ relativizes class quantifiers in A to subsets of x.

be expressed by the schema:

$$\mathbb{V} \vDash A(X) \Rightarrow \exists\beta\, \mathbb{V}_\beta \vDash A^{\mathbb{V}_\beta}(X \cap \mathbb{V}_\beta), \tag{12}$$

where $A^{\mathbb{V}_\beta}$ relativizes quantifiers in A to $\mathbb{V}_\beta$ and $\mathbb{V}_{\beta+1}$ (if second-order quantification is allowed), respectively. All instances of first-order reflection (with second-order parameters) are provable in ZF. If A is taken to be second-order, one can derive the existence of large cardinals.[15]

Reinhardt does not make the full extent of the analogy between (II) and reflection principles in set theory explicit. In footnote 12 of his paper, he mentions a principle 'closely related' to (II) and expressed in a set-theoretic language:

$$(V_\kappa, \in, T_\kappa) \models A \rightarrow \exists\alpha\; A^{V_\alpha, T_\kappa \cap V_\alpha}, \tag{13}$$

for A a sentence of a set-theoretic language featuring a truth predicate. In (13), κ is an inaccessible cardinal and T_κ is the result of carrying out a fixed-point construction over V_κ. The claim is a consequence of the strong "reflective" properties of inaccessibles: satisfaction in $(V_\kappa, \in, T_\kappa)$ can always be approximated by an unbounded sequence of models based on V_α's, for $\alpha < \kappa$.[16] In the footnote Reinhardt continues with a cryptic remark about the totality of the sets $T_\kappa \cap V_\alpha$ versus the inability of the model to recognize the totality of the corresponding predicate '$T(x) \wedge x \in V_\alpha$'. It is unclear what notions of totality are at play for Reinhardt. In a straightforward interpretation, the partial nature of Kripke's minimal fixed point justifies immediately the non-totality of the predicate $T(x) \wedge x \in V_\alpha$, whereas any set, including $T_\kappa \cap V_\alpha$, is total because set-membership is. However, this interpretation is problematic because it could have been already applied to T_κ, so (13) would become superfluous in the analysis.

In addition, it is difficult to see how Reinhardt's principle (13) in its current form could help to justify the axioms of KF, or a suitable version of it formulated in a set-theoretic language. For instance, let A be KF3. Since A is true in $(V_\kappa, \in, T_\kappa)$, by (13) its relativization to $T_\kappa \cap V_\alpha$ should also be true. However, in one interpretation of (13) which echoes the worry above, the truth predicate will be interpreted via T_κ and quantifiers will be restricted to V_α. So, although the relativization of KF3 will also be satisfied, V_α will contain codes of many ungrounded sentences, so $T_\kappa \cap V_\alpha$ will fail to be total in the strong sense required, contrary to the original motivation. On another reading, $T_\kappa \cap V_\alpha$ will restrict T_κ itself to the level α. However, on this interpretation, some care is required to make sure that the internal structure of codes of sentences is handled via a suitable translation. For instance, let φ be the code of a sentences with parameters from V_α, then

$$(\mathtt{T}\varphi \rightarrow \mathtt{T}\ulcorner \mathtt{T}\dot{\varphi} \urcorner)^{T_\kappa \cap V_\alpha}$$

[15] The proof of first-order reflection can be found in any standard set theory textbook, whereas for the claim about second-order reflection, see [Kan09, Ch. 1,§6].

[16] Compare Kanamori [Kan09], Lemma 6.1 p.57.

won't be satisfied by $(V_\kappa, \in, T_\kappa)$ unless one suitably restricts also the range of sentences to which the truth predicate can apply: in $\mathtt{T}\ulcorner\mathtt{T}\dot{\varphi}\urcorner$ the external truth predicate is now $\mathtt{T}_\alpha$, but the internal one would be unrelativized.

At any rate, Reinhardt's analogy specifically pertains to the language of set theory, and it becomes desirable to establish the non-significant consequences of KF independently of the base theory. We see Reinhardt's (II) as offering a general recipe that should also be applicable within the arithmetical setting of Kripke's original work. This certainly shares some fundamental intuition with set-theoretic reflection, but it also needs to be faithful to the specific challenges one faces in the standard truth-theoretic setting.[17]

It is well known that a truth predicate can reproduce set- or class-membership in certain contexts by understanding $u \in X$ as $\mathtt{T}\varphi_X(\dot{u})$, where φ_X is a formula of $\mathcal{L}_\mathtt{T}$ with one free variable – intuitively, φ_X is true of u. So, a certain amount of class-membership, directly employed in (12) and in (13), can be mimicked in $\mathcal{L}_\mathtt{T}$; this also shows that (definable) second-order parameters, implicitly used in the reflection strategy, can be handled adequately.

However, reference to inaccessible cardinals, with their strong closure properties, is clearly out of reach for the arithmetical language. It is instead more illuminating to focus on $\omega_1^{\mathtt{CK}}$, the first non-recursive ordinal, as a suitable replacement. In fact, $\omega_1^{\mathtt{CK}}$ resembles the least inaccessible in several ways. It is the least recursively regular ordinal and has strong closure principles, and it represents the "limit" of the inductive construction of one of the intended models of KF, the minimal fixed point.

It is then plausible to adapt the template (12) to the language $\mathcal{L}_\mathtt{T}$. By defining $u \in X$ via $\mathtt{T}\varphi_X(\dot{u})$ as prescribed above, (12) becomes

$$\mathcal{I} \vDash A(X) \Rightarrow \exists \alpha < \omega_1^{\mathtt{CK}}\, \Phi^\alpha \vDash A(X) \tag{14}$$

In words: if $A(X)$ is true in the minimal fixed point, then there is some stage in the construction of the minimal fixed point that satisfies $A(X)$. As with (13), (15) cannot hold in general without further adjustments. Again, sentences such as $\forall\varphi(\mathtt{T}\varphi \rightarrow \mathtt{T}(\mathtt{T}(\varphi)))$, which are satisfied in $\mathcal{I}$, cannot be satisfied by any of the Φ^α.

A way to address this issue is, of course, to not only reflect satisfaction in the minimal fixed point to satisfaction in a stage of its construction, but to relativize the quantifiers over sentences as well. By writing A^α for the result of restricting sentential quantifiers in A to sentences that contain at most α embeddings of T,

$$\mathcal{I} \vDash A(X) \Rightarrow \exists \alpha < \omega_1^{\mathtt{CK}}\, \Phi^\alpha \vDash A^\alpha(X) \tag{15}$$

Of course, a uniform method to generate the appropriate restriction $A^\alpha(X)$ in (15) involves some additional adjustments, which will be explained shortly.

At any rate, it is evident that (15) is essentially formulated in the metalanguage. This

[17] That being said, the project of clarifying the scope and feasibility of Reinhardt's remark within a set theoretic framework is worthwhile and remains unexplored.

is not satisfactory, since having a truth predicate in someone's object-language should enable them to express semantic facts in it (or at least all semantic facts that can be expressed in the object-language, should be expressed in it). The language of set theory can express principles in the vicinity of an object-linguistic formulation of (15), which do not require the existence of inaccessibles. Some of them may seem to provide a direct route towards an object-linguistic formulation of (15). More specifically, in the context of admissible set theory, a significant role is played by the so-called principle of Σ-reflection[18]

$$A \to \exists a A^{(a)} \tag{16}$$

The principle states that if A is a formula of a suitable restricted complexity (Σ refers to the presence of one unrestricted existential quantifier in A, not necessarily the outermost one as in Σ_1-formulae), then there is a set a such that $A^{(a)}$, i.e. the result of restricting all quantifiers in A by a, holds.[19] Σ-formulae display a strong connection to the expressive resources of $\mathcal{L}_{\mathtt{T}}$: The sets definable by partial predicates in the language of type-free truth over the minimal fixed point are the same as the sets definable by a Σ_1-formula of set theory over the admissible set $\mathtt{HYP}_{\mathbb{N}}$, i.e. the collection of all hyperarithmetical sets. Moreover, we have that sets definable by total predicates in the language of type-free truth correspond to the elements of $\mathtt{HYP}_{\mathbb{N}}$.[20]

Although this points towards an object linguistic reflection principle, not relying explicitly on the ordinals, it quickly becomes clear that (16) only provides a restricted form of reflection for $\mathcal{L}_{\mathtt{T}}$-sentences, which does not suffice for an adequate regimentation of Reinhardt's (II). To see this, we have just mentioned that the assertion made by an $\mathcal{L}_{\mathtt{T}}$-sentence of form $\mathtt{T}\varphi(\overline{n})$ – namely that n belongs to the set defined by φ – can be replicated by a Σ-sentence of the language of set theory. Thus, by (16), this set-theoretic claim can be reflected down, so that an object-linguistic version of (15) in $\mathcal{L}_{\mathtt{T}}$ could be achieved along the following lines:

$$\mathtt{T}\varphi \to \exists \psi(\mathsf{tot}(\psi) \wedge \mathtt{T}\varphi^{\mathtt{T}\psi}). \tag{17}$$

In (17), $\mathsf{tot}(\psi) := \forall x(\mathtt{T}\psi(\dot{x}) \vee \mathtt{T}\neg\psi(\dot{x}))$ and $\mathtt{T}\varphi^{\mathtt{T}\psi}$ expresses that $\mathtt{T}\varphi$ is relativized by $\mathtt{T}\psi$.[21] In short, one is allowed to reflect down truth ascriptions, but not all formulae.

So, a full regimentation of Reinhardt's principle (II) in the object-language is a nontrivial matter. In the next section we will show that such a regimentation is indeed possible. Our principle (RR_γ) will incorporate the desiderata just discussed and extend

[18] Compare Barwise [Bar75], p.16. [Bar75] is also a standard reference for admissible set theory.

[19] The principle (16) is provable in Kripke-Platek set theory ($\mathtt{KP}$) for Σ-formulae of the language of set theory.

[20] These claims were first stated in [Kri75], made more explicit in [Bur86] and then generalized to nonstandard models in [Can89].

[21] In the relativization the formula $\mathtt{T}\psi(\dot{x})$ plays the same role as the hyperarihmetic set a in (16) and is intended to restrict the quantifiers in $\mathtt{T}\varphi$ accordingly.

the truth-theoretic version of Σ-reflection. In particular, it will feature the key reflection step from partial to total truth predicates prescribed by (II), while restricting the range of sentences truth applies to as in (15). The principle will then be used in the final section to provide additional justification for the axioms of $\mathtt{KF}$.

5 A Formal Proposal

This section contains the formal results of the paper. To regiment Reinhardt's (II), we'll employ Tarskian truth predicates definable in $\mathtt{KF}$. We then provide effective translations taking a formula A of $\mathcal{L}_{\mathtt{T}}$ to one in which occurrences of the partial predicate $\mathtt{T}$ are translated into occurrences of Tarskian ones; the translations also restrict quantifiers over sentences of $\mathcal{L}_{\mathtt{T}}$ in a uniform way. We will show that, under our understanding of (II), the principle becomes consistent and even provable for natural classes of sentences of $\mathcal{L}_{\mathtt{T}}$. This provides a precise formal rendering to the informal procedure of extracting the significant content from $\mathtt{KF}$-theorems outlined above.

We start with defining the Tarskian sub-languages of $\mathcal{L}_{\mathtt{T}}$. We will start with a general definition for arbitrary recursive ordinals. In order to guarantee uniformity we choose a path through Kleene's $\mathcal{O}$ such that, for every γ on this path, $\prec_\gamma$—the restriction of the natural ordering of recursive ordinals to γ—is recursive. Later, we will focus solely on translations for fixed ordinal notations for ε_0 or Γ_0, thereby removing the noneffective aspects of the translation. Informally, the Tarskian languages are built from the arithmetical language by 'keeping-track' how the *well-founded* iterations of the truth predicate $\mathtt{T}$ are contained in the relevant formulae. Those languages don't include Liar or Truth-teller sentences, for instance, which cannot be obtained by a sentence of $\mathcal{L}_{\mathbb{N}}$ by iterating $\mathtt{T}$ and combining it with the logical connectives. Intuitively, a language $\mathcal{L}_\alpha$ – a sub-language of $\mathcal{L}_{\mathtt{T}}$ – contains sentences involving iterations of $\mathtt{T}$ up to and including the countable ordinal α. $\mathtt{Sent}_\alpha(x)$ will then denote the collection of sentences of the language $\mathcal{L}_\alpha$; $\mathtt{Sent}_{<\alpha}(x)$ denotes the collection of sentences to which the "α^{th} truth predicate" will be applied, containing only sentences with *less-than* α iterations of $\mathtt{T}$; finally, the definition of the Tarskian truth predicate makes use of these sublanguages of $\mathcal{L}_{\mathtt{T}}$: $\mathtt{T}_\alpha$ applies to a sentence φ if and only if φ is a sentence in $\mathtt{Sent}_{<\alpha}$ and $\mathtt{T}\varphi$. Precise definitions can be found in Appendix A3.

[Hal97] translated the language $\mathcal{L}_{\mathtt{T}}$ into a Tarskian hierarchical language in such a way that 'grounded' sentences were assigned Tarskian truth predicates matching their ordinal norms in the Kripkean minimal fixed point. We slightly modify Halbach's translation to provide a precise formulation of Reinhardt's suggestion.

For any γ on our chosen path through $\mathcal{O}$ we will provide a translation $h_\gamma(k, A)$. The translation takes a (codified) ordinal k in γ and a formula from $\mathcal{L}_{\mathtt{T}}$ as input and provides a formula of the Tarskian language $\mathcal{L}_\gamma$ as an output. By our previous definition the Tarskian languages are themselves sublanguages of $\mathcal{L}_{\mathtt{T}}$. The translation is the identity function on arithmetical formulae and commutes with the logical symbols. As for the $\mathtt{T}$-

iterations, the translation attempts to assign a correct Tarskian level to it; if the formula contains less-than k iterations of truth, the translation adequately reflects the sentence's grounded structure. If this is not possible, then the formula gets assigned an arithmetical falsity.[22] Since our chosen γ is such that $\prec_\gamma$ is recursive, the procedure for the iterations is well-defined. A specific deviation from Halbach's translation is that we take $\mathtt{Sent}_{\mathcal{L}_\mathtt{T}}$ as a primitive, translating it by $\mathtt{Sent}_{<k}$ for the input k. The details of the translation are provided in Appendix A3.

There is a slight ambiguity in our use of the translation functions h_γ, which should not cause any problems. On the one hand we intend to have a direct translation of formulae of the type-free truth language into the Tarskian typed language. On the other hand, we rely on an application of Kleene's recursion theorem for the existence of the function $\underset{\dot{}}{h}\,_\gamma$, which presupposes a translation function on the Gödel codes and not the expressions themselves. Since there is a close correspondence between the use of h_γ for the two layers, the ambiguity is harmless and a disambiguation of all the occurrences in the paper within the respective contexts is possible.

With the translations h_γ in hand, we are in a position to formulate a template for a reflection principle in Reinhardt's sense:

> Reinhardt Reflection Template. For any sentence A of $\mathcal{L}_\mathtt{T}$ we can find an ordinal $\gamma < \omega_1^{\mathtt{CK}}$ such that
>
> $$A \to (\forall\alpha)(\exists\beta)(\alpha \prec_\gamma \beta \wedge h_\gamma(\beta, A)) \tag{RR$_\gamma$}$$
>
> holds.[23]

Our interpretation of Reinhardt's proposal is not only consistent, but it also has nice models, as it is satisfied by the minimal fixed point $\mathcal{I}$, one of the most natural models of $\mathtt{KF}$.

Theorem 1. *For any sentence A of $\mathcal{L}_\mathtt{T}$, we can find a $\gamma < \omega_1^{\mathtt{CK}}$ such that (RR$_\gamma$) is true in $(\mathbb{N}, \mathcal{I})$, and therefore consistent with* $\mathtt{KF}$.

The theorem, whose proof can be found in the Appendix, also entails the consistency of an apparently stronger claim involving truth. This turns out to be a precise formulation in $\mathcal{L}_\mathtt{T}$ of a variant of the principle (II) also considered by Reinhardt [Rei86, p. 236].

Corollary 1. *For any sentence A of $\mathcal{L}_\mathtt{T}$, we can find a $\gamma < \omega_1^{\mathtt{CK}}$ such that the sentence*

$$A \to (\forall\alpha)(\exists\beta)(\alpha \prec_\gamma \beta \wedge \mathtt{T}\ulcorner h_\gamma(\dot{\beta}, A)\urcorner) \tag{RR'$_\gamma$}$$

can be consistently added to $\mathtt{KF}$.

[22] Given this feature, the translation $h_\gamma(k, A)$ of A is not always faithful to the original content of A.

[23] For ease of readability, we don't use different variables for ordinals and their codes. Strictly speaking, in RR$_\gamma$ we are quantifying over codes of ordinals $< \gamma$, specifically members of a recursive set of natural numbers codifying such ordinals. For this reason, quantifiers over $\alpha, \beta, \ldots$ are implicitly bounded by γ.

The specific formulation of ($\textsc{rr}_\gamma$) contains some non-effective elements. Although it is formulated schematically for formulae A, it includes a second component γ that depends on A, making it non-effective. To see this, assume that there is a recursive set Σ of instances of ($\textsc{rr}_\gamma$). Then there is an arithmetical formula χ that defines the collection of pairs (A, γ) occurring in instances of Σ. However, then the set of all γ occurring in the instances is also arithmetically definable, and is a subset of Kleene's $\mathcal{O}$. According to the boundedness theorem (cf. Theorem 7.2.8, p. 82 in [Poh96]), any arithmetically definable subset of $\mathcal{O}$ is bounded in $\omega_1^{\mathtt{CK}}$. However, we also know, by the construction of the minimal fixed-point $\mathcal{I}$, that the set of inductive norms of grounded sentences is unbounded in $\omega_1^{\mathtt{CK}}$.

The generality in the formulation of ($\textsc{rr}_\gamma$) has the drawback of being too complex to be specified explicitly. If we focus on the specific primitive recursive ordinal notation system given by the Cantor normal form the situation changes.

By a slight adaption of Gentzen's original argument we know that KF proves the principle of transfinite induction for $\mathcal{L}_{\mathtt{T}}$-properties up to any ordinal smaller than ε_0:

$$(\mathtt{TI}_\alpha(A)) \qquad \mathtt{Prog}(A) \Rightarrow \forall \xi \prec_{\varepsilon_0} \alpha\, A(\xi), \text{ for all } \alpha < \varepsilon_0, \text{ and } A(v) \in \mathcal{L}_{\mathtt{T}},$$

where

$$\mathtt{Prog}(A) := \forall \eta (\forall \zeta \prec_{\varepsilon_0} \eta\, A(\zeta) \to A(\eta)).$$

We let $\mathtt{TI}_\alpha = \{\mathtt{TI}_\alpha(A) \mid A \in \mathcal{L}_{\mathtt{T}}\}$.

The following lemma makes precise the claim that the translation h_{ε_0} involves sentences belonging to the relevant Tarskian languages for the ordinals that are provably well-ordered in KF. Its proof involves an easy transfinite induction.

Lemma 1. *For all $\alpha < \varepsilon_0$, $\mathtt{KF} \vdash \forall x(\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(x) \to \mathtt{Sent}_\alpha(\ulcorner h_{\varepsilon_0}(\alpha, x)\urcorner))$.*

Although consistent, ($\textsc{rr}_\gamma$) has not been shown to be provable. It is then natural to ask whether we could find a version of Reinhardt's reflection schema that is provable. The answer turns out to be positive.

Proposition 1. KF *is closed under the following rule of inference:*

$$\textit{if } \mathtt{KF} \vdash \mathtt{T}\ulcorner A \urcorner, \textit{ then } \mathtt{KF} \vdash (\forall \alpha)(\exists \beta)(\alpha \prec_{\varepsilon_0} \beta \wedge \mathtt{T}(\ulcorner h_{\varepsilon_0}(\beta, A)\urcorner)).$$

Proposition 1 follows from the proof-theoretic analysis of KF from [Can89], which tells us that

$$\mathtt{KF} \vdash \mathtt{T}\ulcorner A \urcorner \Rightarrow \mathtt{KF} \vdash \mathtt{T}_\alpha \ulcorner A \urcorner, \quad \text{for some } \alpha < \varepsilon_0$$

Since $\mathtt{T}_\alpha \ulcorner A \urcorner$ and $\alpha \prec_{\varepsilon_0} \beta$, implies $\mathtt{T}_\beta \ulcorner A \urcorner$ we can infer the admissibility of inference rule in KF by Lemma 1. This proposition immediately entails that, for sentences A that are provably significant in Reinhardt's sense, that is such that $\mathtt{KF} \vdash \mathtt{S}(\ulcorner A \urcorner)$, *the corresponding instances of ($\textsc{rr}_\gamma$) are provable in* KF.

We now show that there are other classes of $\mathcal{L}_{\mathtt{T}}$-formulae for which ($\textsc{rr}_\gamma$) (for suitable γ) becomes provable in KF. Fact 1 shows that the basic principles of KF – mainly its logical and non-logical axioms – feature ungrounded instances: thus, such principles cannot be significant in Reinhardt's sense. However such claims can, in a precise sense, be made significant by ($\textsc{rr}_\gamma$); what's more, such principles will be made significant by the *schema* ($\textsc{rr}_{\varepsilon_0}$), which fixes the specific countable ordinal ε_0 in Reinhardt's Reflection Template thereby making the translation involved primitive recursive.

The first noticeable class of formulae to which Reinhardt's reflection can be applied are the *non-logical axioms of* KF. These even include candidate non-logical axioms such as the axioms of consistency (CONS) and completeness (COMP), and the principle

$$\forall\varphi\forall\psi(\mathtt{T}(\varphi\to\psi)\land\mathtt{T}\varphi\to\mathtt{T}\psi) \tag{18}$$

expressing that Modus Ponens – the sole rule of inference in our formulation of KF – is truth-preserving. Since, as mentioned in Section 1, the charge of unsoundness of KF is particularly effective when involving these nonlogical axioms of KF, this amounts to a considerable step in the direction of Reinhardt's proposed vindication of KF.[24]

A second noticeable class of formulae is what we call *truth-theoretic generalizations*. To motivate them, we can consider the logical axioms of the classical logical calculus we have assumed. They involve for instance an axiom schema $B\to B$, for B an $\mathcal{L}_{\mathtt{T}}$-formula. It is well-known that the truth predicate can be used to generalize over first-order schemata and turn meta-theoretic (universal) quantification into first-order (universal) quantification; in the case at hand, this would amount to the first-order *sentence* $\forall\varphi\,(\mathtt{T}\varphi\to\mathtt{T}\varphi)$. The class of truth-theoretic generalizations extends this pattern to $\mathcal{L}_{\mathtt{T}}$-formulae obtained by taking $\mathcal{L}_{\mathbb{N}}$-formulae $A(P_1,\dots,P_n)$, where $P_1,\dots,P_n$ are free second-order variables, and replacing $P_1,\dots,P_n$ with truth-ascriptions over arbitrary $\mathcal{L}_{\mathtt{T}}$-sentences (potentially with parameters):

$$\forall\varphi_1\dots\forall\varphi_n\,A(\mathtt{T}\varphi_1/P_1,\dots,\mathtt{T}\varphi_n/P_n).$$

Truth-theoretic generalizations include truth-theoretic formulations of the logical axioms of classical predicate logic. For a list of noticeable instances of truth-theoretic generalizations, we refer to Appendix A2.

The formulae just presented share the feature of having *grounded content* – see discussion in the concluding section; as such, Reinhardt's reflection principles can separate this grounded content from the ungrounded one. The next theorem summarizes the status of ($\textsc{rr}_{\varepsilon_0}$) in our framework; its proof can be found in Appendix A1. As mentioned, the first item in the theorem follows immediately from Proposition 1.

Theorem 2. *($\textsc{rr}_{\varepsilon_0}$) is provable in* KF *for:*

(i) *$\mathcal{L}_{\mathtt{T}}$-formulae $A(\vec{x})$ such that* $\mathtt{KF}\vdash\mathtt{S}\ulcorner A(\dot{\vec{x}})\urcorner$;

(ii) *non-logical axioms of* KF, *and candidate axioms such as* CONS, COMP, *and* (18);

[24] Cf. also [Fie08, §7.3].

(iii) *truth-theoretic generalizations.*

As a corollary, we obtain that the formulae just described in (ii)-(iii) can be in a sense be taken to be true (via their translation), and therefore significant.

Corollary 2. *($\mathrm{RR}'_{\varepsilon_0}$) is provable in* KF *for A belonging to the classes (i)-(iii) from Theorem 2.*

It is worth noting that the proof of Theorem 2 depends essentially on the extent of transfinite induction for $\mathcal{L}_{\mathtt{T}}$ provable in KF. Therefore, on the background of KF one can generalize the result to any amount of transfinite induction available in a suitable primitive recursive ordinal notation system $(O, <)$.

Proposition 2. *Let $(O, <)$ be a primitive recursive ordinal notation system. For any $\gamma \in O$ and A belonging to the classes (i)-(iii) from Theorem 2:*

$$\mathtt{KF} + \mathtt{TI}_\gamma \vdash A \to (\forall\alpha)(\exists\beta)(\alpha \prec_\gamma \beta \wedge \mathtt{T}(\ulcorner h_\gamma(\beta, A)\urcorner)).$$

6 Vindicating KF

6.1 Reinhardt's Conjecture

Reinhardt's quote reported in Section 3 ends with what we may call *Reinhardt's Conjecture*, namely the idea that the 'formal theory [KF] allows for its own significant interpretation (at least piecewise)' [Rei86, p.237]. We now assess the impact of our work on the conjecture, and then connect our main findings to more traditional themes in the philosophical justification of Kripke-Feferman truth.

We decided to test Reinhardt's task in the original language $\mathcal{L}_{\mathtt{T}}$ in which KF is formulated. Already the task proved to be non-trivial, and required sophisticated logical tools to formulate suitable forms of reflection compatible with Reinhardt's idea. However, our first main finding is that Reinhardt's conjecture *can be formulated* by means of suitable reflection principles even in the language of arithmetic. Specifically, Reinhardt suggests that the significant interpretation of a sentence of $\mathcal{L}_{\mathtt{T}}$ should be connected with a procedure of replacing partial (nonclassical) truth predicates with total (classical) ones. We have provided the details of how such a procedure can work in an arithmetical context.

Moreover, what we called Reinhardt Reflection Template (p. 15) enables one to formulate different reflection principles that are relevant to Reinhardt's Conjecture. The Conjecture is in fact vague enough to sanction different precisifications. What does it mean for KF *to allow* for its significant interpretation? In one sense, this can be made precise by requiring a suitable incarnation of Reinhardt's Reflection Template to be *compatible with a nice model of* KF, and hence being consistent with KF. Theorem 1 clearly establishes that *this version of Reinhardt's conjecture is indeed true*: the minimal, closed-off fixed point model of KF is compatible with the statement (RR_γ) expressing that if a sentence A holds, then its truth predicates can be replaced by suitably total ones, thereby making A significant.

There may be more stringent ways for KF 'to allow' its own significant interpretation. A plausible one is to require suitable instantiations of Reinhardt's Reflection Template to be *provable* in KF. In this sense, our analysis deems Reinhardt's Conjecture *false*: there are sentences A such that for no countable ordinals γ the relevant instance of (RR_γ) is provable in KF. One such A can be for instance the truth-teller sentence τ; others consist in classical tautologies involving ungrounded content such as $\tau \vee \neg\tau$.

However, if by 'formal theory' one only focuses on the *axioms* of KF, then the picture changes. Theorem 2 establishes the *truth* of Reinhardt's conjecture at least for the non-logical axioms of KF, and truth-theoretic versions of its non-logical axioms.

All in all, we believe that our results support a broadly positive outlook on Reinhardt's conjecture in an arithmetical context. The results also provide insights on traditional philosophical applications of the Kripke-Feferman theories, to which we now turn.

6.2 Groundedness.

Our results suggest that, although KF cannot eliminate altogether the ungrounded content from its theorems, the theory can, in a precise sense and in a large class of relevant cases, *tolerate* them. Reinhardt's reflection provides a uniform way to extract the grounded content from some noticeable consequences of KF, such as its non-logical axioms. The principles provide insight on the nature of ungrounded theorems of KF; theorems with exclusively ungrounded content, or *singularities* in Gödel's terminology, are the ones that cannot be reduced to theorems with exclusively grounded content via Reinhardt reflection.

In the introduction, we reported Feferman's idea that the intended range of the truth predicate of KF is the collection of grounded sentences. Here's Feferman's quotation in full:

> First of all, the distinction between outer and inner logics is only a problem if one conflates two notions of truth, namely the notion of *grounded truth* given by Kripke's least fixed-point construction, and our everyday notion of truth not tied to any particular semantical construction or theory. Thus, in KF, $\mathtt{T}(A)$ expresses that the sentence A is a grounded truth while A itself, if provable, is counted as true in the informal sense. So on that reading there is no conflict between accepting both $\neg\mathtt{T}(\lambda \vee \neg\lambda)$ and $\lambda \vee \neg\lambda$ for a formal liar sentence λ. [Fef12, p. 189]

KF is already a highly regimented reasoning environment and based on very specific assumptions. So, even if it may well be the case that provability in KF counts as truth 'in the informal sense', there is something more to say to link theoremhood in KF and the notions of groundedness and significance.

It is here that our formulation of Reinhardt's reflection principles – specifically ($\textsc{rr}_\gamma$) and ($\textsc{rr'}_\gamma$) – enters the picture. Take the closed-off minimal fixed-point model $(\mathbb{N}, \mathcal{I})$. Corollary 1 provides a link between sentences that hold at $(\mathbb{N}, \mathcal{I})$ and the *extension* $\mathcal{I}$ of $\mathtt{T}$: in particular, it provides a uniform procedure to extract the *grounded* content from sentences satisfied at $(\mathbb{N}, \mathcal{I})$. There are some sentences, such as the Liar or truth-teller sentences, for which there is no grounded content to extract. In such cases, Reinhardt's conditional is trivially satisfied because such sentences are not satisfied in the minimal closed-off fixed point and they appear as antecedents in such a conditional. However, and most crucially, the reflection principle is able to extract the grounded content from many sentences of *mixed* status, for instance non-logical axioms of KF, in which one quantifies both over grounded and ungrounded instances. What's more, our formulation of the principle enables us to capture grounded instances of such quantified claims of arbitrary ordinal norm relative to the minimal fixed point $\mathcal{I}$.[25]

The semantic link between satisfaction in a grounded model and membership in the collection of grounded truths provided by Reinhardt's reflection has a proof-theoretic counterpart. Corollary 2 shows that, for suitably defined recursively enumerable classes of $\mathcal{L}_\mathtt{T}$-formuale, corresponding informally to collection of sentences that display *some* grounded content, one can link the external and the internal theories of KF. In the specific case of the logical and non-logical axioms of KF, this means that KF is able to disregard their ungrounded content, mitigating considerably its alleged unsoundness.

For example, an axiom such as

$$(\mathtt{KF10}) \qquad \forall\varphi(\mathtt{T}(\neg\neg\varphi) \leftrightarrow \mathtt{T}\varphi)$$

has grounded and ungrounded content implicit in the range of the universal quantifier. The unsoundness challenge to KF would point to the ungrounded instance $\mathtt{T}\ulcorner\neg\neg\lambda\urcorner \leftrightarrow \mathtt{T}\ulcorner\lambda\urcorner$ of KF10. However, via Reinhardt's reflection, KF is able to filter out "singularities" such as $\mathtt{T}\ulcorner\neg\neg\lambda\urcorner \leftrightarrow \mathtt{T}\ulcorner\lambda\urcorner$ and to (provably) establish a link between KF10 and its intended range of grounded instances.

6.3 Reinhardt's instrumentalism.

Our result can also help vindicating KF against its alleged unsoundness by means of a fuller realization of Reinhardt's programme. Reinhardt identified IKF as the significant core of KF. In Reinhardt's picture, indeed, KF serves as an *indispensable tool* for reasoning about IKF. However, the unsoundness challenge poses a threat to Reinhardt's programme as well since, as shown by [HH06], this necessary detour via KF also necessarily involves sentences that are not in IKF and are not grounded. To address this issue, Reinhardt introduced his principle (II) precisely to provide a 'significant interpretation'

[25] Notice that this is possible because we are allowing our translations to work relative to fixed paths through $\mathcal{O}$ with suitable properties. As we mentioned, this is also the source of the ineffective nature of translations.

for these sentences that is available within KF itself.

Our formal development in Section 5 realizes Reinhardt's proposal by providing two levels of justification for KF. On a first level, one can formulate in $\mathcal{L}_{\mathtt{T}}$ a *full* version of Reinhardt's reflection principle – in our notation ($\textsc{rr}'_{\gamma}$) as stated in Corollary 1 – that is consistent with KF. The sense in which ($\textsc{rr}'_{\gamma}$) expresses a significant interpretation of KF-theorems, on this first level of justification, is that the grounded model $(\mathbb{N}, \mathcal{I})$ validates the principle.

Of course, on this first level, the link between KF-theorems and their significant content is expressed in $\mathcal{L}_{\mathtt{T}}$ but established at the meta-theoretic level. However, by restricting oneself to the instances of ($\textsc{rr}'_{\varepsilon_0}$) in the classes of $\mathcal{L}_{\mathtt{T}}$-formulae (i)-(iii) from Theorem 2, this link becomes fully accessible to KF. This is the content of Corollary 2, in which it is shown that Reinhardt's reflection principle ($\textsc{rr}'_{\varepsilon_0}$) bridges, via the proposed translation, KF-provably significant formulae, non-logical axioms, and truth-theoretic generalizations. On this second level of justification, there is a principled path from an important class of KF-theorems to IKF. It is worth noting that, according to Reinhardt – cf. again in Reinhardt's passage in Section 3 –, the intended target of Reinhardt's justification are the axioms of KF. In this sense, our proposal fully vindicates Reinhardt's idea. An important consequence of ($\textsc{rr}'_{\varepsilon_0}$) is that the conjunction of axioms KF1-KF10 can always be translated into a significant statement. So, in the proof of any KF-theorem, ($\textsc{rr}'_{\varepsilon_0}$) provides a stable justification for the use of KF-truth axioms in proofs.

The procedure extends, in a precise sense, to noticeable classes of KF-proofs. Suppose $\mathtt{KF} \vdash \mathtt{T}\ulcorner A \urcorner$, for some $\mathcal{L}_{\mathtt{T}}$-sentence A. Then, not only the conjunction of the finitely many non-logical axioms of KF, but also the instances of $\mathcal{L}_{\mathtt{T}}$-induction and logical axioms employed in this proof can *always* be interpreted in accordance with ($\textsc{rr}'_{\varepsilon_0}$), so that their use can be fully internalized within IKF. In this sense, for any concrete proof of a significant statement, KF can offer an interpretation of its axioms used in the derivation: uniformly for non-logical truth axioms, and in a piece-wise manner for logical and induction axioms.[26]

6.4 Reflective closure

A third proposed vindication of KF based on our results is closer to Feferman's original motivation for introducing KF. In a recent article, Cantini, Fujimoto, and Halbach reconstruct the conceptual step underlying the transition from transfinite iterations of Tarskian truth predicates:

> From a foundational point of view, [the iterated Tarskian theories] are a very

[26] Another piece-wise justification of KF-proofs may be given via Cantini's proof-theoretic analysis of KF in [Can89]. Cantini's asymmetric interpretation establishes that for each specific KF-proof of $\mathtt{T}\ulcorner A \urcorner$ one can find a specific $\alpha \prec \varepsilon_0$ such that $\mathtt{T}_{\alpha}$ can act as the "translation" of T in $\mathtt{T}\ulcorner A \urcorner$. However, it is worth noting that the piece-wise justification of proofs via the principle ($\textsc{rr}'_{\varepsilon_0}$) is a special case of Theorem 2, whereas the asymmetric interpretation can only be applied to the specific case of proofs of theorems of form $\mathtt{T}\ulcorner A \urcorner$.

> 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. [CFH17, pp. 292-93]

To assess the quote, let's focus on the simple case of the iterated Tarskian theories up to any $\alpha < \varepsilon_0$, that we call $\mathtt{CT}_\alpha$. A plausible way to spell out the reason why $\bigcup_{\alpha<\varepsilon_0} \mathtt{CT}_\alpha$ may characterize the reflective closure of PA is by appeal to the PA-provable well-ordering of such α's. The soundness of PA is naturally expressed by the Global Reflection Principle GRP(PA). The Tarskian theory $\mathtt{CT}_1$ proves GRP(PA). Given that the ordinals α less than ε_0 are provably well-ordered in PA, it seems justified to iterate the process along these α's: consider the property '$\mathtt{CT}_\alpha$ proves $\mathtt{GRP}(\mathtt{CT}_{<\alpha})$'. The property is clearly progressive. So, since we are entitled to transfinite induction for $\mathcal{L}_\mathtt{T}$ – and the Tarskian truth predicates $\mathtt{T}_\alpha$ can be expressed in $\mathcal{L}_\mathtt{T}$ – one is justified to preserve the property up to any $\alpha < \varepsilon_0$. $\bigcup_{\alpha<\varepsilon_0} \mathtt{CT}_\alpha$ is the limit of this process.[27]

The next step is to realize that KF expresses the reflective closure of PA without explicitly appealing to ordinal notations and their properties provable in specific base theories. However, [Fef91] already noticed that such properties are implicit in KF; the truth predicates of the theories $\mathtt{CT}_\alpha$, for $\alpha < \varepsilon_0$, can all be defined in KF via the predicates 'x is true and is a sentence containing $< \alpha$ iterations of T'. However, Feferman's translation does not provide a link between the axioms of KF and the reflective closure of PA expressed in terms of iterations. If, as presented in the passage above, KF is only a 'nicer presentation' of the *more direct* formulation of the reflective closure for PA given by $\bigcup_{\alpha<\varepsilon_0} \mathtt{CT}_\alpha$, it seems that its foundational relevance is only derivative. For instance, what is the relationship of the KF axioms with the explicit presentation of the reflective closure of PA?

The principle of Reinhardt reflection ($\text{RR'}_{\varepsilon_0}$) introduced above helps in providing the required link. The principle delivers a way to connect the general formulation of the KF-axioms and their content involving the explicit presentation of the reflective closure

[27] If one accepts iterations along ordinals that are provably well-ordered in the $\mathtt{CT}_\alpha$'s themselves, one is allowed to iterate along any $\alpha < \Gamma_0$. We stick to ε_0 for ease of presentation. Our discussion, in virtue of Proposition 2, can be transferred to Γ_0 and to the schematic formulation of KF.

of PA. Quantification over all $\mathcal{L}_{\mathrm{T}}$-sentences, required for the neat and parsimonious formulation of KF, is replaced in a uniform way by quantification over sentences involved in the formulation of the $\mathtt{CT}_\alpha$'s.

6.5 Objections

Before we close the discussion we address two potential objections.

Overgeneration. One may object that our strategy overgenerates. Consider the examples of (CONS) and (COMP). Although KF is not consistent with both axioms, nevertheless (RR') can be applied to both. Another way of putting the objection is that the set collection of axioms whose grounded content can be extracted via (RR') is inconsistent over KF. Is this a problem for the proposal?

We don't think so, because the principles of Reinhardt reflection are conditional, and are intended to *explain*, within a specific theory that already has some independent motivation, the link existing between a sentence and its grounded content. For instance, in the case of KF+CONS, the relevant information in need of explanation is the relationship between CONS and its grounded content. The fact that the translation of COMP happens to be provable does not undermine this.

KF and groundedness. Another potential objection may concern KF and its relationship with grounded sentences. It is well-known that KF is, in a sense, a theory of *all* fixed points: Feferman showed that the ω-models of KF are all the fixed points of Φ (cf. (6) above). Therefore, so the objection goes, it would be inadequate to extract the grounded content from the KF-axioms as this would not account for the intended range of KF-truth. Perhaps the project of employing Reinhardt reflection principles is more suited for theories that feature a minimality condition such as Burgess' $\mathtt{KF}\mu$ introduced in [Bur14].

The objection can be resisted in at least two ways. First, as we mentioned, Feferman himself – who proved (6) – refers to the grounded truths as the intended range of the truth predicate of KF. In fact, Feferman's instrumentalist reading of KF as the reflective closure of PA strongly connects KF to the iteration of Tarskian truth over PA, which are all grounded sentences. Moreover, it is clear that the truth predicate of KF, in virtue of (6), does have some grounded content as it is sound with respect to the minimal fixed point of Φ, and the truths provable of KF are all grounded. Finally, our framework can be applied to $\mathtt{KF}\mu$ as well: Proposition 1 can be adapted without modification to yield the consistency of (RR_γ) with $\mathtt{KF}\mu$, and Proposition 2 entails that ($\mathrm{RR'}_\gamma$) is provable in $\mathtt{KF}\mu$ for suitable ordinals provably well-ordered in $\mathtt{KF}\mu$.

References

[AB63] Alan Ross Anderson and Nuel D Belnap. First degree entailments. *Mathematische Annalen*, 149(4):302–319, 1963.

[Bar75] Jon Barwise. *Admissible Sets and Structures.* Springer Verlag, Berlin, 1975.

[Ber76] Paul Bernays. On the problem of schemata of infinity in axiomatic set theory. In *Studies in Logic and the Foundations of Mathematics*, volume 84, pages 121–172. Elsevier, 1976.

[Bur86] John P. Burgess. The truth is never simple. *The Journal of Symbolic Logic*, 51:663–681, 1986.

[Bur14] John P. Burgess. Friedman and the axiomatization of Kripke's theory of truth. In Neil Tennant, editor, *Foundational adventures: essays in honor of Harvey M. Friedman*, pages 125–148. College publications, 2014.

[Can89] Andrea Cantini. Notes on formal theories of truth. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 35:97–130, 1989.

[CFH17] Andrea Cantini, Kentaro Fujimoto, and Volker Halbach. Feferman and the truth. *Feferman on Foundations: Logic, Mathematics, Philosophy*, pages 287–314, 2017.

[CS23] Luca Castaldo and Johannes Stern. KF, PKF and Reinhardt's program. *The Review of Symbolic Logic*, 16(1):33–58, 2023.

[End01] Herbert B. Enderton. *A Mathematical Introduction to Logic.* Harcourt/Academic Press, 2 edition, 2001.

[Fef91] Solomon Feferman. Reflecting on incompleteness. *The Journal of Symbolic Logic*, 56:1–47, 1991.

[Fef12] Solomon Feferman. Axiomatizing truth: why and how? In *Logic, construction, computation*, volume 3 of *Ontos Math. Log.*, pages 185–200. Ontos Verlag, Heusenstamm, 2012.

[FHKS15] Martin Fischer, Volker Halbach, Jönne Kriener, and Johannes Stern. Axiomatizing semantic theories of truth? *The Review of Symbolic Logic*, 8(2):257–278, 2015.

[Fie08] H. Field. *Saving Truth from Paradox.* Oxford University Press, 2008.

[Gla15] Michael Glanzberg. Complexity and hierarchy in truth predicates. In *Unifying the philosophy of truth*, pages 211–243. Springer, 2015.

[Göd95] Kurt Gödel. Some basic theorems on the foundations of mathematics and their implications. *Collected Works*, pages 304–323, 1995.

[Hal97] Volker Halbach. Tarskian and Kripkean truth. *Journal of Philosophical Logic*, 26:69–80, 1997.

[Hal14] Volker Halbach. *Axiomatic Theories of Truth.* Cambridge University Press, Cambridge, UK, revised edition, 2014.

[HH06] Volker Halbach and Leon Horsten. Axiomatizing Kripke's theory of truth. *The Journal of Symbolic Logic*, 71:677–712, 2006.

[HN18] Volker Halbach and Carlo Nicolai. On the costs of nonclassical logic. *Journal of Philosophical Logic*, 47(2):227–257, 2018.

[Hor11] Leon Horsten. *The Tarskian Turn. Deflationism and Axiomatic Truth.* MIT Press, Cambridge, MA, 2011.

[Kan09] Akihiro Kanamori. *The Higher Infinite.* Springer, 2009.

[Koe18] Peter Koellner. On the question of whether the mind can be mechanized, II: Penrose's new argument. *The Journal of Philosophy*, 115(9):453–484, 2018.

[Kri75] Saul Kripke. Outline of a theory of truth. *The Journal of Philosophy*, 72:690–716, 1975.

[Mau04] Tim Maudlin. *Truth and Paradox. Solving the Riddles.* Oxford, 2004.

[McG91] Vann McGee. *Truth, Vagueness and Paradox.* Hackett Publishing Company, Indianapolis, 1991.

[Nic17] Carlo Nicolai. Provably true sentences across axiomatizations of Kripke's theory of truth. *Studia Logica*, 106(1):101–130, 2017.

[Nic22] Carlo Nicolai. Gaps, gluts, and theoretical equivalence. *Synthese*, 200(5):366, 2022.

[NR23] Carlo Nicolai and Lorenzo Rossi. Systems for Non-reflexive Consequence. *Studia Logica*, 2023. Forthcoming.

[Poh96] Wolfram Pohlers. Computability theory of hyperarithmetical sets. lecture notes, online, 1996.

[Put65] Hilary Putnam. Craig's theorem. *The Journal of Philosophy*, 62(10):251–260, 1965.

[Rei86] William N. Reinhardt. Some remarks on extending and interpreting theories with a partial predicate for truth. *The Journal of Philosophical Logic*, 15:219–251, 1986.

[Ste18] Johannes Stern. Proving that the mind is not a machine? *Thought: A Journal of Philosophy*, 7(2):81–90, 2018.

[Wan96] Hao Wang. *A Logical Journey: From Gödel to Philosophy.* Bradford, 1996.

Appendix A

A1: Proofs of Theorems

Proof of Fact 1. Let us define the Liar sentence λ as $\neg\mathtt{T}l$, where the base theory PA proves that $l = \ulcorner\neg\mathtt{T}l\urcorner$.

For (i) we can consider for example the instance $\lambda \to \lambda$ of a logical axiom. Suppose, seeking a contradiction, that $\mathtt{T}\ulcorner\neg(\lambda \wedge \neg\lambda)\urcorner$. The KF-axioms entail that $\mathtt{T}\ulcorner\lambda\urcorner$. By KF6 and KF10, we obtain $\mathtt{T}\ulcorner\neg\lambda\urcorner \vee \mathtt{T}\ulcorner\lambda\urcorner$. Each disjunct, however, leads to $\mathtt{T}\ulcorner\lambda\urcorner \wedge \mathtt{T}\ulcorner\neg\lambda\urcorner$, by KF3, KF4, and again KF10. Since $(\mathbb{N}, \mathcal{I})$ is a model of KF, this is a contradiction. By an analogous use of the Liar sentence we can obtain internal inconsistencies, and therefore the unprovability, of the truth of the other logical axioms.

For (ii), we can consider for instance KF3. Assuming its truth, we would have

$$\mathtt{T}\ulcorner\mathtt{T}l \to \mathtt{T}\ulcorner\mathtt{T}l\urcorner\urcorner,$$

that is

$$\mathtt{T}\ulcorner\neg\mathtt{T}l \vee \mathtt{T}\ulcorner\mathtt{T}l\urcorner\urcorner.$$

We can then reason as in the previous case.

For (iii) one can consider a fixed point model $(\mathbb{N}, S) \models \mathtt{KF}$ where $\lambda, \neg\lambda \in S$ – we know that such models exist because of (6). It is then immediate that both $\lambda \in S$ and $(\neg\lambda \vee 0 = 1) \in S$, but of course $0 = 1 \notin S$: this suffices to establish that T is not closed under the material conditional, as prescribed by (iii).

Symmetrically for (iv) one considers a model $(\mathbb{N}, R)$ where $\lambda, \neg\lambda \notin R$ – e.g. $(\mathbb{N}, \mathcal{I})$. If all instances of induction were true, also the instance concerning the formula $\lambda \wedge x = x$ would be. However, this instance cannot be satisfied in $(\mathbb{N}, R)$. □

Proof of Theorem 1. Let $\mathcal{I}$ be as above. We show that for all formulae $A \in \mathcal{L}_{\mathtt{T}}$ there is a $\gamma < \omega_1^{\mathtt{CK}}$ such that the instance of $(\textsc{rr}_\gamma)$ is validated in $\mathcal{I}$ by induction on the positive complexity of A.

- If A is an identity $s = t$ or $s \neq t$, then the claim is obvious.

- If A is $\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(t)$, then one employs the fact that, for any suitable path through $\mathcal{O}$ γ,
$$\mathbb{N} \models \forall\alpha\forall x(\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(x) \to \mathtt{Sent}_\alpha(\,\d{h}\,_\gamma(\dot{\alpha}, x))).$$
If A is $\neg\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(t)$, then the claim follow from the fact that, for all suitable α, $\mathtt{Sent}_\alpha(t) \to \mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(t)$.

- If A is $\mathtt{T}(t)$, then either $t^{\mathbb{N}} \notin \mathcal{I}$ or $t^{\mathbb{N}} \in \mathcal{I}$. If the former, then the claim is trivially obtained. If the latter, let δ be the ordinal norm of $t^{\mathbb{N}}$. Then we can choose our γ to be some limit ordinal, such that $\delta < \gamma < \omega_1^{\mathtt{CK}}$. The levels of $\mathcal{I}$ are increasing, so $t^{\mathbb{N}} \in \mathcal{I}^\eta$ for $\delta \prec_\gamma \eta$. Therefore, we can let β be δ, if $\alpha \prec_\gamma \delta$ and $\alpha + 1$ if $\delta \prec_\gamma \alpha$.

- If A is $\neg\mathtt{T}(t)$. If $t^{\mathbb{N}} \in \mathcal{I}$ the claim is trivially obtained. If $t^{\mathbb{N}} \notin \mathcal{I}$, we rely on the claim

(19) $$t^{\mathbb{N}} \notin \mathcal{I} \Rightarrow \neg\exists\delta < \omega_1^{\mathtt{CK}} \, t^{\mathbb{N}} \in \mathcal{I}^{\delta}.$$

- If A is $B \wedge C$, then by induction hypothesis there are γ, γ' and so we can take the maximum.
- If A is $\forall x B$, then the induction hypothesis yields, for all $n \in \mathbb{N}$, some γ_n. So, for $\gamma = \sup\{\gamma_n \,|\, n \in \mathbb{N}\}$, the claim holds. That $\gamma < \omega_1^{\mathtt{CK}}$ follows by the fact that $\omega_1^{\mathtt{CK}}$ is recursively regular.
- If A is $\neg(B \wedge C)$, then by induction hypothesis we have γ, γ' for which $\neg B \rightarrow \forall\alpha\exists\beta(\alpha \prec_{\gamma} \beta \wedge h_{\gamma}(\beta, \neg B))$ and $\neg C \rightarrow \forall\alpha\exists\beta(\alpha \prec_{\gamma'} \beta \wedge h_{\gamma'}(\beta, \neg C))$. Then we can choose one of γ, γ' and employ the definition of h.
- If A is $\neg\forall x B(x)$, then $\neg B(\overline{n})$ is true in $\mathcal{I}$ for some $n \in \mathbb{N}$. By induction hypothesis we have a γ_n such that $\neg B(\overline{n}) \rightarrow \forall\alpha\exists\beta(\alpha \prec_{\gamma_n} \beta \wedge h_{\gamma_n}(\beta, \neg B(\overline{n}))$. The claim then follows by definition of h.

□

Proof of Theorem 2. The proof strategy is analogous for all A's: one shows indeed that the right-hand side of the claim is derivable in KF for all $\alpha \prec \varepsilon_0$. As an example, we verify the claim for CONS. Reasoning in KF, we want to establish

(20) $$(\forall\alpha)(\exists\beta)(\alpha \prec_{\varepsilon_0} \beta \wedge h_{\varepsilon_0}(\beta, \forall\varphi(\mathtt{T}\neg\varphi \rightarrow \neg\mathtt{T}\varphi))).$$

Fixing an arbitrary $\beta \prec \varepsilon_0$, by Lemma 1 it suffices to establish, for all $\delta \prec \beta$,

(21) $$\mathtt{Sent}_{<\beta}(x) \rightarrow (\neg\mathtt{T}_{\beta}\ulcorner h_{\varepsilon_0}(\delta, \dot{x})\urcorner \leftrightarrow \mathtt{T}_{\beta}\ulcorner h_{\varepsilon_0}(\delta, \neg\dot{x})\urcorner).$$

That is, for all $\beta < \varepsilon_0$, the predicates $\mathtt{T}_{\beta}$ are consistent and complete. Claim (21) is obtained by a transfinite induction up to β and a sub-induction on the complexity of sentence x. It is important to notice that, given our assumptions on the language $\mathcal{L}_{\mathtt{T}}$, the base case in the proof of (21) needs to include also the case in which the formula is of form $\mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(x)$.

□

A2: Truth-Theoretic Generalizations

The class of truth-theoretic generalizations is obtained by taking $\mathcal{L}_{\mathbb{N}}$-formulae $A(P_1, \ldots, P_n)$, where $P_1, \ldots, P_n$ are free second-order variables, and replacing $P_1, \ldots, P_n$ with truth-ascriptions over arbitrary $\mathcal{L}_{\mathtt{T}}$-sentences (potentially with parameters):

$$\forall\varphi_1 \dots \forall\varphi_n\, A(\mathtt{T}\varphi_1/P_1, \dots, \mathtt{T}\varphi_n/P_n).$$

Among some noticeable members of this class, we list the universal closure of the following truth-theoretic version of the axioms of classical logic:

(22) $\mathtt{T}\varphi \to \mathtt{T}\varphi$

(23) $\mathtt{T}\varphi \to (\mathtt{T}\psi \to \mathtt{T}\varphi)$

(24) $(\mathtt{T}\varphi \to (\mathtt{T}\psi \to \mathtt{T}\chi)) \to ((\mathtt{T}\varphi \to \mathtt{T}\psi) \to (\mathtt{T}\varphi \to \mathtt{T}\chi))$

(25) $\mathtt{T}\varphi \to (\mathtt{T}\varphi \vee \mathtt{T}\psi)$

(26) $\mathtt{T}\psi \to (\mathtt{T}\varphi \vee \mathtt{T}\psi)$

(27) $(\mathtt{T}\varphi \to \mathtt{T}\psi) \to ((\mathtt{T}\chi \to \mathtt{T}\psi) \to (\mathtt{T}\varphi \vee \mathtt{T}\chi \to \mathtt{T}\psi))$

(28) $\mathtt{T}\varphi \wedge \mathtt{T}\psi \to \mathtt{T}\varphi$

(29) $\mathtt{T}\varphi \wedge \mathtt{T}\psi \to \mathtt{T}\psi$

(30) $\mathtt{T}\varphi \to (\mathtt{T}\psi \to (\mathtt{T}\varphi \wedge \mathtt{T}\psi))$

(31) $\neg\mathtt{T}\varphi \to (\mathtt{T}\varphi \to \mathtt{T}\psi)$

(32) $\neg\neg\mathtt{T}\varphi \to \mathtt{T}\varphi$

(33) $\forall x\mathtt{T}\varphi(\dot{x}) \to \mathtt{T}\varphi(\dot{t})$ with t free for substitution

(34) $\forall x(\mathtt{T}\varphi(\dot{x}) \to \mathtt{T}\psi(\dot{x})) \to (\forall x\mathtt{T}\varphi(\dot{x}) \to \forall x\mathtt{T}\psi(\dot{x}))$

(35) $\mathtt{T}\varphi(\dot{x}) \to \forall x\mathtt{T}\varphi(\dot{x})$ with t free for substitution

(36) $x = y \to (\mathtt{T}\varphi(\dot{x}) \to \mathtt{T}\varphi(\dot{y}))$

A3: Tarskian Languages and Translations

Definition 1. For $\gamma < \omega_1^{\mathtt{CK}}$, let:

$$\begin{aligned}
\mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\overline{0}, x) &:\leftrightarrow \mathtt{Sent}_{\mathcal{L}_{\mathbb{N}}}(x),\\
\mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta \hat{+} 1, x) &:\leftrightarrow \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta, x) \vee\\
&\qquad (\exists y \leq x)(x = \ulcorner \mathtt{T}\dot{y} \urcorner \wedge \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta, y)) \vee\\
&\qquad (\exists y < x)(x = (\underset{\cdot}{\neg} y) \wedge \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta \hat{+} 1, y)) \vee\\
&\qquad (\exists y, z < x)(x = (y \underset{\cdot}{\wedge} z) \wedge \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta \hat{+} 1, y) \wedge \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta \hat{+} 1, z)) \vee\\
&\qquad (\exists v, y < x)(x = (\underset{\cdot}{\forall} v y) \wedge \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta \hat{+} 1, y)),\\
\mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\lambda, x) &:\leftrightarrow \exists\, \zeta \prec_{\gamma} \lambda\, \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta, x) \quad \text{for } \lambda \text{ limit.}
\end{aligned}$$

The following abbreviations are decribed informally in the text but precisely defined here:

$$\begin{aligned}
\mathtt{Sent}_{\alpha}(x) &:\leftrightarrow \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\alpha, x)\\
\mathtt{Sent}_{<\alpha}(x) &:\leftrightarrow \exists \zeta \prec \alpha\, \mathtt{Sent}^{\gamma}_{\mathcal{L}_{\mathtt{T}}}(\zeta, x),\\
\mathtt{T}_{\alpha}(x) &:\leftrightarrow \mathtt{Sent}_{<\alpha}(x) \wedge \mathtt{T}(x).
\end{aligned}$$

Let γ be on our chosen path through $\mathcal{O}$, with $(\mathrm{ON}_{\gamma}, \prec_{\gamma})$ the ordinal notation system. We define a translation function $h_{\gamma} : \mathrm{ON}_{\gamma} \times \mathcal{L}_{\mathtt{T}} \to \mathcal{L}_{\mathtt{T}}$.

$$\begin{aligned}
h_{\gamma}(k, A) &:\leftrightarrow A, \quad \text{if } A \in \mathcal{L}_{\mathbb{N}}\\
h_{\gamma}(k, \neg A) &:\leftrightarrow \neg h_{\gamma}(k, A)\\
h_{\gamma}(k, A \wedge B) &:\leftrightarrow h_{\gamma}(k, A) \wedge h_{\gamma}(k, B)\\
h_{\gamma}(k, \forall v A) &:\leftrightarrow \forall v\, h_{\gamma}(k, A)\\
h_{\gamma}(k, \mathtt{Sent}_{\mathcal{L}_{\mathtt{T}}}(t)) &:\leftrightarrow \mathtt{Sent}_{<k}(\underset{\cdot}{h_{\gamma}}(k, t))\\
h_{\gamma}(0, \mathtt{T}t) &:\leftrightarrow \bot\\
h_{\gamma}(k, \mathtt{T}t) &:\leftrightarrow \mathtt{T}_{k}\, \underset{\cdot}{h}_{\gamma}(\dot{c}, t), \text{ for } k = \mathtt{Suc}_{\gamma}(c)\\
h_{\gamma}(k, \mathtt{T}t) &:\leftrightarrow \exists c \prec_{\gamma} k\, \mathtt{T}_{k}\, \underset{\cdot}{h}_{\gamma}(\dot{c}, t), \text{ for } \mathtt{Lim}_{\gamma}(k)
\end{aligned}$$