NOÛS

NOÛS 00:0 (2019) 1–24
doi: 10.1111/nous.12292

Hypatia's silence

*Truth, justification, and entitlement**

MARTIN FISCHER
Ludwig-Maximilians-Universität München

LEON HORSTEN
University of Bristol

CARLO NICOLAI
Kings College London

Abstract

Hartry Field distinguished two concepts of type-free truth: scientific truth and disquotational truth. We argue that scientific type-free truth cannot do justificatory work in the foundations of mathematics. We also present an argument, based on Crispin Wright's theory of cognitive projects and entitlement, that disquotational truth can do justificatory work in the foundations of mathematics. The price to pay for this is that the concept of disquotational truth requires non-classical logical treatment.

1. Introduction

Can the concept of type-free truth play an essential role in justifying new mathematical knowledge? This question is clearly of philosophical importance, but it is also ambiguous. As argued in [Field 1994], there are (at least) *two* concepts of truth. There is no consensus in the literature about the exact content of these two concepts, nor is the terminology used to mark the distinction uniform. But there is some agreement on the existence of a salient distinction along the lines that Field suggests, and on the acceptability of the following minimal characterisation of the two concepts. The first is a concept of truth that plays some role in scientific explanations – e.g. explaining communication by specifying truth-conditions for some natural language expressions; we call this theoretical notion *scientific truth*. The second is a notion of truth that is governed by rules of semantic ascent and descent, and we call it *disquotational truth*.[1]

In this article, we leave the scientific concept of truth mostly aside and focus on a concept of truth characterized by the unrestricted principles of disquotation. We

*Thanks to Hartry Field and Vann McGee and two anonymous referees. Martin Fischer was supported by the DFG Project 'Syntactical treatments of interacting modalities'. Carlo Nicolai's work was supported by the VENI NWO Grant 275-20-057.

claim, with McGee, that disquotational truth can play a justificatory role, but we disagree with his reasons for *why* it can do this. McGee's account of the justificatory role of disquotational truth hinges on the admissibility of *stipulatively introducing* a concept of disquotational truth in certain circumstances. Against this, we will argue that the introduction of disquotational truth by a stipulative act compromises its potential for playing a justificational role. Instead, we provide an alternative account of the way in which disquotational truth can play a role in the justification of mathematical knowledge. On this account disquotational truth allows one to significantly expand a mathematical theory in a way that preserves justification, i.e. if the starting theory is justified, then so is the resulting stronger theory.

The concept of disquotational truth we prefer is *type-free* and we take the essence of disquotational truth to be that of a device for *unrestricted quotation and disquotation*. On our view, disquotational truth is a concept that is governed by unrestricted principles of disquotation; in this work we describe such principles and surrounding formalism in terms of *sequents*, namely expressions of the form $\Gamma \Rightarrow \Delta$ where Γ, Δ are finite sets of formulas of a language containing truth. In our chosen formalism the core disquotational principles amount to the sequents $A \Rightarrow \mathsf{T}\ulcorner A\urcorner$ and $\mathsf{T}\ulcorner A\urcorner \Rightarrow A$, where T is our disquotational truth predicate. McGee focuses in his discussion of disquotationalism on *typed* disquotational theories, whereas Field in his recent work concentrates on type-free disquotational truth [Field 2008]. There are good reasons for preferring a type-free concept of truth over the Tarskian typing strategy: they have been thoroughly defended elsewhere – see for instance [Kripke 1975] and [Field 2008, Ch. 3,§1]. Concepts of type-free truth are often compared with respect to their treatment of paradoxical sentences such as Liar sentences. Our notion of type-free truth will be articulated inferentially in a non-classical setting that allows us to preserve disquotational truth. With this strategy it is even coherent to remain *silent* with respect to the status of Liar-like sentences.[2]

Our discussion of the justificational role of disquotational truth is framed in the context of Wright's *cognitive projects* [Wright 2004a]: accepting the scientific notion of truth, and accepting the disquotational concept of truth as a justificatory device, are two distinct cognitive projects. These cognitive projects are in some sense in tension with each other. Science uses classical logic throughout, so scientific truth operates in a context of classical logic. By Tarski's theorem on the undefinability of truth, however, we know that full type-free disquotational truth can only function in a logic that is weaker than classical logic. So the question which concept of truth to use (for a given purpose) is intertwined with the discussion what the correct logic is.

Another ingredient of our account is the conviction that it is warranted to expand a sound mathematical theory by principles stating its soundness. This strategy of expanding theories by suitable soundness principles has its roots in Gödel's incompleteness theorems; the idea is to interpret the incompleteness results as having a positive content in that they provide a reasonable way of strengthening formal theories.[3] Some of the more famous results based on this idea are provided by Feferman and his account of reflective closure.[4] On this view, *reflection principles* for a theory S – that is, soundness statements stating that *everything that S proves, is true* – are part of what one *ought* to accept if one accepts the theory S. However,

in this paper we do not rely on the normative dimension conveyed by Feferman's 'ought'. Only the notion of *being entitled to accept* will be employed.

The innovation of our account is that we propose to view the epistemological import of reflection in the light of the distinction between the notions of entitlement and justification (see again [Wright 2004a]). The view will be that by accepting a justified theory S, one is *entitled* to accept a reflection principle for S. This results in a stronger theory that will be itself justified. This process can be repeated, so that we eventually become justified in accepting much stronger theories obtained by iteration of reflection. Another innovative aspect of our account is that we intend to combine this reflection process with a suitable concept of truth, that allows us to directly employ explicit soundness statements in the form of *global reflection principles*, rather than schematic derivatives thereof.

Putting these elements together, we arrive at an account of the justificatory force of disquotational truth that can be outlined as follows. Suppose that we are justified, to start with, in believing a given mathematical theory S, governed by classical logic. Then we are warranted in extending our conceptual repertoire with an unrestricted type-free disquotational concept of truth. This results in a theory S' in which we are entitled to believe and trust. Our trust in S' entitles us to accept reflection principles and add them explicitly to S'. This results in a stronger theory that is again justified and we can iterate the process in a reliable way. Thus we eventually come to accept much stronger theories resulting from iterated reflection. The mathematical part of this theory will again be governed by classical logic, although our truth concept is governed by non-classical principles. The resulting mathematical theory will be significantly stronger than the starting theory S. As a particular case study of this pattern we will sketch how from a justified belief in a fragment of arithmetic, disquotational truth leads to justified belief in a substantial fragment of classical analysis.

Let us now look at the details of how all of this works.

2. Two Concepts of Truth

When looked at from the outside, and in a somewhat superficial manner, contemporary research in theories of type-free truth appears to be divided into two communities. The first community of researchers concentrates on truth theories formulated in *classical logic*. The second community focuses on truth theories formulated in the context of some *non-classical logic*.[5]

Against explorations of truth in the context of non-classical logic, the following concern can be raised: *can withdrawing from classical logic ever be a sound methodological move?*[6] In particular, the following argument is proposed. The concept of truth plays a role in scientific argumentation. The concept of truth plays a fundamental role in formal semantics, for instance, which is part of linguistics, and in the foundations and philosophy of logic. The concept of truth also plays a role in the foundations of mathematics. For instance, it plays a key role in one of the neatest presentations of Predicative Analysis offered by Solomon Feferman

[Feferman 1991].[7] Classical logic is the one and only logic that governs scientific reasoning. Therefore classical logic governs the concept of truth.

At least partly in reaction to concerns of this kind, it has been argued that there are *two concepts of truth* [Field 1994], [McGee 2005b]. There is disagreement in the literature about the precise content of these two concepts, which shows that it is not easy to get the intended distinction into sharp focus. Here we give our own take on what the distinction amounts to.

The first is the concept of *scientific truth*.[8] This is the concept of truth that is used in scientific theories that are first and foremost concerned with explanations of non-semantic facts; it is governed by classical logic. Scientific truth is for instance employed in trying to understand how 'human beings communicate by language',[9] or to understand which arithmetical statements one ought to accept if one has accepted the basic axioms and rules of elementary arithmetic.[10] The scientific concept of truth is a *theoretical concept*, like the concept of force in classical mechanics, for instance. It is related to our pre-theoretical, ordinary language concept of truth. But there is no reason to think that it does or should coincide with it, just as there is no reason to expect the scientific concept of force to coincide with our pre-theoretical concept of force.

The second is the concept of *disquotational truth*.[11] This notion of truth intends to be a device of full quotation (*semantic ascent*) and disquotation (*semantic descent*). Indeed, a core part of the meaning of the truth predicate is given by principles that allow for a substitution of a sentence A by the statement of its truth $\mathsf{T}\ulcorner A\urcorner$ – i.e. the ascription of truth to the name of A – and vice versa in all extensional contexts. By formulating the disquotationalist principles as the sequents

$$\text{(T1)}\quad A \Rightarrow \mathsf{T}\ulcorner A\urcorner \qquad\qquad \text{(T2)}\quad \mathsf{T}\ulcorner A\urcorner \Rightarrow A$$

we ensure that such substitution can be also available in hypothetical contexts. In (T1) and (T2), T stands for the truth predicate, and A ranges over sentences that can include T.

The liar paradox teaches us that *if* there is a coherent concept of type-free disquotational truth, then it is governed by non-classical logic. Many philosophers have argued that truth substitution principles are fundamentally correct principles about truth. But our approach will not rely on this: we follow [Wright 2004a] and consider the coherence of the concept of disquotational truth as a *presupposition* in the cognitive project of a truth-theoretic justification of mathematical knowledge.

It is often intimated that disquotational truth *is* our ordinary language notion of truth. But there really is little evidence to support this. At any rate, we shall not take it to be so in this article.[12]

3. Entitlement to Cognitive Project

Over the past decades, the distinction between *entitlement* and *justification* has become prominent in epistemology.[13]

The notion of entitlement has been used in certain philosophical accounts of *knowledge transfer*, which appeal to entitlement to rely on basic patterns of reasoning for which one has no justification [Boghossian 2003], [Wright 2004b], [Burge 2011]. The idea is that one can be *justified* in believing a conclusion that one has inferred by means of basic logical steps from a collection of premises for which one has justification, even if one does not have justification for the claim that the basic logical steps are valid.

Consider the following scenario:

> **Antigone** knows (and thus is justified in believing in) a proposition P. From this premise, using a logical pattern such as Disjunction Introduction, she infers $P \vee Q$. Antigone does not have sufficient logical training and even does not have a sufficiently rich conceptual repertoire to justify the validity of the logical rule of Disjunction Introduction.

Boghossian claims that Antigone has an entitlement to *blind* logical reasoning that is knowledge-transferring: in the scenario under consideration, Antigone's reasoning suffices to come to *know* the proposition $P \vee Q$.

Wright defines the notion of *entitlement of cognitive project* along the following lines [Wright 2004a, 191–192]:

> . . . an entitlement of cognitive project [. . .] may be proposed to be any presupposition P of a cognitive project meeting the following additional two conditions:
>
> (i) We have no sufficient reason to believe that P is untrue
> (ii) The attempt to justify P would involve further presuppositions in turn of no more secure a prior standing . . .

Wright argues that relying on the validity of certain logical rules of inference fulfils the condition for being an entitlement of cognitive project [Wright 2004b, Section IV]. We will not go into the details of Wright's argumentation, but assume for the purposes of our discussion that his account is *basically* correct.

There are, however, a few questions that are left open by Wright's account that turn out to be important for our discussion. First, *which* rules of inference are we entitled to rely on in logical reasoning? Wright argues that Modus Ponens, for instance, is among them. But for many putative such rules (such as Disjunctive Syllogism), he is silent about this question. Second, what is the *strength* of our entitlement to logical reasoning? In particular, are we entitled to rely on logical reasoning involving sentences of any admissible extension of the language that we are currently using? Or are we entitled to use them only for the language that we are currently using, leaving it open that we may not be entitled to rely on them for certain future language extensions? For instance, might one be entitled to rely on classical logic in mathematics but not when the language of mathematics is extended by vague predicates? Another way of putting this is the following. If we agree to regard logical inference rules as *schematic*, then what is the *substitution rule* that we are entitled to use when we instantiate the rules in concrete arguments? These questions will be addressed in due course.

4. The Justificatory Role of Truth

We will now relate the distinction between scientific and disquotational truth to the question to which extent the concept of truth can play a *justificatory role*. We focus on the role that the concept of truth plays in justification in mathematics.

It has been claimed that truth is a *logical* notion.[14] If there is something to this slogan, then one may wonder whether, as in the case for the first-order logical connectives, we can be *entitled* to principles and rules governing the concept of truth without having *justification* for them. If the answer to this question is yes, then truth might be able to play a role in justificatory processes that is similar to the role that logical reasoning plays in them.

4.1 Scientific truth and justification

If the concept of *scientific* truth is understood as a foundational device for empirical sciences – e.g. in giving a good and coherent account of linguistic meaning –, the answer to our question must surely be negative. Like the logical notions, the theoretical notion of truth is part of a *package*, which is a scientific theory (and we have seen that classical logic is part of this package). The package as a whole is judged, as Quine has taught us, by the extent to which it is *successful* – in giving a good model of communication via truth conditions, for instance. Derivatively, this then also holds for the principles and rules governing the logical connectives and the truth predicate, all of which belong to this package. Under this reading, there is no room for *entitlement* (or 'warrant for nothing', in Wright's terms) to logical principles and rules or theoretical truth principles and rules: they can only be to a smaller or greater extent *justified*.[15] This, however, does not entail that scientific truth cannot play a justificatory role in non-empirical sciences such as logic itself or mathematics.

Theories of truth formulated in classical logic that may be regarded as theories of scientific truth in this broader sense come in *two kinds*. The *first kind* consists of theories of truth in which truth is closed under the rules of *Necessitation* and *Co-Necessitation*:

$$\frac{\Rightarrow A}{\Rightarrow \mathsf{T}\ulcorner A\urcorner} \qquad \frac{A \Rightarrow}{\mathsf{T}\ulcorner A\urcorner \Rightarrow}.$$

These rules are *weaker* than our initial sequents (T1), (T2). Consequently, the notion of truth that they describe is not fully transparent: in reasoning under an assumption, for instance, it is not always possible to infer $\mathsf{T}\ulcorner A\urcorner$ from A and vice versa. Nonetheless, such theories express a notion of theoretical truth that most closely approximates a notion of transparent truth. The most famous of such theories is Friedman and Sheard's theory FS.[16] The main problem with FS (and its close relatives) is that it does not preserve the intended structure of the truth bearers. It is *ω-inconsistent*, therefore it does not admit models based on the standard natural numbers. Under the assumption that natural numbers are satisfactory bearers of truth modulo isormorphism with suitable syntactic objects, this amounts to saying that FS-like theories do not apply to syntactic objects as we standardly conceive

of them. This is a sufficient reason to put this first kind of theories of classical truth aside.

The *second kind* of theory of type-free scientific truth is not closed under Necessitation and Co-Necessitation. Theories of this type can be seen as axiomatisations of certain classes of classical models that result from 'closing off' a fixed point model of the kind described in [Kripke 1975]. The most famous of these theories is Feferman's theory KF, which is obtained by closing Peano Arithmetic under a natural collection of type-free compositional truth principles in which the truth predicate never occurs in the scope of a negation symbol [Feferman 1991].

KF is based on a conception of truth that, in its essential traits, is fundamentally sound. Starting from a truth-free language $\mathcal{L}_0$, KF states that (i) atomic sentences $P(t)$ of $\mathcal{L}_0$ are true iff the value of t belongs to the extension of P, false if it does not; (ii) a disjunction is true iff at least one disjunct is true, false if both disjuncts are false; (iii) an existentially quantified sentence $\exists x \varphi$ is true iff $\varphi(t)$ is true for at least one t, false if $\varphi(t)$ is false for all t; (iv) a truth ascription $\mathsf{T}^{\ulcorner} A^{\urcorner}$ is true iff A is true, false if A is false. But being formulated in classical logic, KF cannot be completely faithful to the conception of truth that inspires it. Because of the Liar Paradox, the disquotational character of the truth predicate can only be formulated under the scope of the truth predicate: in KF, $\mathsf{T}^{\ulcorner}\mathsf{T}^{\ulcorner} A^{\urcorner\urcorner}$ is equivalent to $\mathsf{T}^{\ulcorner} A^{\urcorner}$, but it is in general *not* the case that $\mathsf{T}^{\ulcorner} A^{\urcorner}$ is equivalent to A. Therefore KF cannot be considered to be a theory of *disquotational* truth. We will see, however, that the compositionality of truth that is encompassed in clauses (i)-(iv) can be fully vindicated.

Theories of theoretical truth do not sit well with statements of their own soundness. The most natural way to express the soundness of a formal theory is via reflection principles, especially Global Reflection principles, that is principles of the form

$$\mathsf{Bew}_S(\varphi) \Rightarrow \mathsf{T}(\varphi) \qquad (\mathrm{GRFN_S})$$

stating, for a given theory S, that all theorems of S are *true*.[17] Even if Global Reflection principles cannot be assumed to be derivable in the theory itself, it appears to be a natural requirement for a trustworthy theory of truth that it is compatible with the claim of its own soundness.

Scientific notions of truth, however, are inadequate if such a requirement is adopted. Neither FS nor KF are compatible with global reflection. In one case, FS is ω-inconsistent: there is a sentence, call it γ, such that FS proves $\mathsf{T}_n{}^{\ulcorner}\gamma^{\urcorner}$ for all natural numbers n – i.e. all finite iterations of the truth predicate applied to γ – and at the same time it proves the sentence $\neg\exists x\, \mathsf{T}_x{}^{\ulcorner}\gamma^{\urcorner}$. Adding global reflection to FS enables one to transform such finite iterations of truth in the claim $\forall x\, \mathsf{T}_x{}^{\ulcorner}\gamma^{\urcorner}$, thereby producing an outright inconsistency. *But also in the case of* KF *global reflection principles lead to a form of inconsistency.* Actually, as far as KF-like systems are concerned, the situation is somewhat more complicated. On the one hand we have to distinguish different versions of KF related to different interpretation of paradoxical sentences. In one version we stay neutral and add

no additional axioms to the compositional principles: this theory allows for both truth value gaps and gluts. This is the version put forward in Feferman 1991. Alternatively, we can exclude gluts by adding a *consistency* axiom of the form

$$\mathsf{T}\neg\varphi \Rightarrow \neg\mathsf{T}\varphi; \tag{CONS}$$

or we exclude gaps by adding a *completeness* axiom

$$\neg\mathsf{T}\varphi \Rightarrow \mathsf{T}\neg\varphi \tag{COMP}$$

The latter theory, $\mathsf{KF} + \mathsf{COMP}$, has an inconsistent truth predicate. As we shall argue shortly, we think this is not compatible with the idea of truth as a justificatory device. However, Global Reflection also produces inconsistencies when combined with KF or $\mathsf{KF} + \mathsf{CONS}$. We prove a stronger result: already the global reflection principle for logic, i.e. the principle $\mathsf{Bew}_{\emptyset}(\varphi) \Rightarrow \mathsf{T}(\varphi)$ stating that all the theorems derivable from classical logic alone are true, is incompatible with such theories.

Observation 1.

(i) $\mathsf{KF} + \mathsf{GRFN}_{\varnothing}$ *is internally inconsistent, i.e.* $\mathsf{T}^{\ulcorner} A^{\urcorner}$ *and* $\mathsf{T}^{\ulcorner}\neg A^{\urcorner}$ *is derivable for some* A.

(ii) $\mathsf{KF} + \mathsf{CONS} + \mathsf{GRFN}_{\varnothing}$ *is inconsistent.*

The argument for (i) can be sketched as follows: We start with a liar sentence $\lambda \leftrightarrow \neg\mathsf{T}^{\ulcorner}\lambda^{\urcorner}$. This biconditional is logically equivalent to $(\lambda \wedge \neg\mathsf{T}^{\ulcorner}\lambda^{\urcorner}) \vee (\neg\lambda \wedge \mathsf{T}^{\ulcorner}\lambda^{\urcorner})$. The existence of such a self-referential liar sentence is derivable in a finitely axiomatized arithmetical theory, such as Q, i.e. Robinson arithmetic formulated in the language with the truth predicate. So in classical logic we can derive the conditional $\bigwedge \mathsf{Q} \rightarrow (\lambda \leftrightarrow \neg\mathsf{T}^{\ulcorner}\lambda^{\urcorner})$. Now with logical transformations we arrive at $\neg \bigwedge \mathsf{Q} \vee ((\lambda \wedge \neg\mathsf{T}^{\ulcorner}\lambda^{\urcorner}) \vee (\neg\lambda \wedge \mathsf{T}^{\ulcorner}\lambda^{\urcorner}))$. Up to this point we only used classical logic and so global reflection for classical logic will give us $\mathsf{T}(^{\ulcorner}\neg \bigwedge \mathsf{Q} \vee ((\lambda \wedge \neg\mathsf{T}^{\ulcorner}\lambda^{\urcorner}) \vee (\neg\lambda \wedge \mathsf{T}^{\ulcorner}\lambda^{\urcorner}))^{\urcorner})$. The compositional principles of KF allow us to distribute the truth predicate so that we arrive at the disjunction $\mathsf{T}(^{\ulcorner}\neg \bigwedge \mathsf{Q}^{\urcorner}) \vee (\mathsf{T}^{\ulcorner}\lambda^{\urcorner} \wedge \mathsf{T}^{\ulcorner}\neg\lambda^{\urcorner})$. The second disjunct is already an internal inconsistency and the first one can be turned into one by the fact that KF proves the finitely many axioms of Q to be true. (ii) is then an obvious consequence.

So even though adding Global Reflection for classical logic to KF itself does not lead to outright contradictions, it has inacceptable consequences. KF is not governed by a paraconsistent logic, but its combination with Global Reflection results in a flavour of dialetheism. If, by proving a statement to be true we cannot thereby exclude the possibility that it is at the same time false, then it is difficult to understand truth as a justificatory device.[18] This means that in theories of classical truth we cannot consistently hold that what they prove is true, and not false. This entails that scientific theories of truth suffer the same fate, by our assumption that only theories of classical truth can be considered theories of scientific truth. Notice that for our purposes it is sufficient to provide reasons to *doubt* the cognitive project based on classical, KF-truth. This violates one of Wright's requirements for

entitlement. The internal inconsistency of KF, and the outright inconsistency of KF + CONS, provide sufficient reasons to reject KF-truth as a justificatory device.

However, it might be objected that there are in fact *two* cognitive projects involved in the combination of KF and Global Reflection. One is the cognitive project of scientific truth as a justificatory notion. The other is Global Reflection as acceptable means to express one's trust in the starting theory. The internal or the external contradiction should suggest that something has gone wrong in one, or both, of these cognitive projects. So why doubting scientific truth, and not Global Reflection? In our view, the inconsistencies just considered do not suffice to put in doubt the cognitive project of Global Reflection as expressing one's trust in a theory. This is for at least two reasons. The first is that there is overwhelming evidence that the strategy of extending formal theories by reflection principles is both mathematically fruitful and soundness preserving [Feferman 1962, Franzén 2004]. Such principles usually do not involve truth, but they are implicitly based on a Tarskian, meta-theoretic notion of truth in a standard model. This provides strong evidence that it is indeed the notion of KF-truth that is to be put in doubt.[19] Moreover, even if one considers the object-linguistic principle of Global Reflection, there are concepts of truth that are perfectly compatible with it. One such example is the concept of disquotational truth that we will consider shortly. In §5 we will consider a further argument for the reliability of our entitlement to Global Reflection.

An additional reason for doubting the applicability of a KF-like notion of truth for a cognitive project of justification is that justifying startling conclusions in KF-like systems seems *too easy*. As an example, consider again the version of Feferman's KF that commits itself to there being no true contradictions – which Maudlin takes to be the correct theory of type-free truth [Maudlin 2004]. This theory is philosophically indeed remarkably strong. Stern has recently shown that this theory proves the elusive conclusion of the Lucas-Penrose argument, i.e., that *the human mind is not a formal system* [Stern 2018]. Yet Stern (rightly) does not in any way take this argument as a *justification* of the conclusion that the mind is not a Turing machine.

In the face of these problems, it has been suggested that we should not accept all of KF, but only those sentences which KF proves to be *true*: the collection of those sentences is called the *inner logic of* KF [Reinhardt 1986]. Much can be said in support of such a policy.[20] But withdrawing to the inner logic of KF means surrendering. The inner logic of KF is not closed under classical logic – but under a non-classical logic that we will later call FDE. Therefore the inner logic of KF cannot capture a concept of truth as it may be used in *scientific explanations*.

4.2 Entitlement to disquotational truth

Now we leave the concept of scientific truth aside and turn to the question whether we have *entitlement without justification* for the rules governing *disquotational* truth. Let the minimal theory of disquotational truth consist of our disquotational principles (T1) and (T2) from page 6, and an elementary syntax theory (which is inter-translatable with an elementary arithmetical theory), in a suitable logic.

We have seen that the correct ambient logic for disquotational truth must be non-classical. But there is no agreement about what the correct ambient logic for a theory of transparent truth is. Some advocate a paracomplete logic, others a paraconsistent logic. Here we assume a four-valued background logic known in the literature as FDE.[21] FDE is a proper sublogic of classical logic as well as of the paracomplete logic K3 and the paraconsistent logic LP. The main feature of our logic, which harmonises perfectly with the conception of truth given on page 11, is that it only involves rules of introduction and elimination for *monotone* connectives: informally, this means that truth values of complex sentences are preserved or 'increased' when we consider other complex sentences whose compounds have truth-values no smaller than the original compounds. Moreover, it is *neutral* regarding the choice between paracompleteness and paraconsistency.

The monotonicity of FDE explains why it does not take a stance on the existence of truth-values gaps or gluts, which would require at least one of the introduction or elimination rules for negation and implication, which are clearly non-monotone connectives. In particular, a feature that our logical system shares with paracomplete approaches is that the classical logical rule of *conditional introduction* on the right hand side of the sequent – corresponding to the rule of conditionalization in natural deduction presentations – has to be restricted. But no one wants to abandon conditional introduction *completely*. Typically, the rule of conditional introduction for material implication is restricted to *truth-determinate* – viz. either classically true or classically false – premises as follows [Halbach & Horsten 2006], [Field 2008]:

$$\frac{\Gamma, A \Rightarrow B, \Delta \qquad \Gamma \Rightarrow \neg A, A, \Delta}{\Gamma \Rightarrow A \rightarrow B, \Delta}$$

In FDE, conditional introduction, as well as other rules containing negative parts, are indeed restricted as indicated above. The basic structural rules and the rules for conjunction, disjunction and the quantifiers are preserved.

The restriction to a positive part of the language not only in the internal logic but also in the external logic guarantees that theories of disquotational truth are compatible with their Global Reflection Principle. In particular, if one starts with a sound theory of arithmetic S and expands the language by a truth predicate governed by the rules (T1) and (T2), obtaining in this way a theory S'. Next one adds the Global Reflection Principle $\mathsf{Bew}_{S'}(\varphi) \Rightarrow \mathsf{T}(\varphi)$ to S', obtaining a theory that is again sound, and in particular sound with respect to a model $(\mathbb{N}, X)$ where $\mathbb{N}$ is the standard interpretation of the syntactic-arithmetical part of the language of S', and the interpretation of the truth predicate X is a fixed point in the sense of [Kripke 1975]. This process can then be iterated even further in a soundness preserving fashion. Moreover, if one starts out with a theory S' that is neutral with respect to the question of truth value gaps or truth value gluts, then (iterated) Globally Reflecting on S preserves this neutrality. *Crucially*, one is not pushed towards an existence claim of gluts, in sharp contrast with our earlier criticism of KF.[22]

> Consider **Theano**, who has *not* committed herself to *full* schematic classical logic, but only to classical logic for the concepts that she already possesses. She leaves the possibility open that she may acquire concepts that are not governed by full classical logic. However, Theano does not remain completely neutral about the logical rules governing future concepts: she does commit herself to applying the rules of the minimal logic FDE to any future concepts.

Theano is entitled to rely on the rules of logic in the way that she does. Can she go on to acquire an entitlement to rely on the disquotational principles (T1) and (T2)?

One might argue that Theano can introduce a notion of disquotational truth by *stipulation*. The idea would be that, in Theano's situation, we are permitted to introduce a new predicate T, and to stipulate that (T1) and (T2) hold for it. These stipulations are to be seen as *meaning stipulations* or implicit definitions for a newly introduced concept.[23]

It is well-known that we are not always entitled to introduce a new concept by laying down inference rules for it. The stipulations for introducing Prior's *Tonk* [Prior 1960], for example, do not succeed in introducing a concept: we are not entitled to follow these stipulations. Belnap proposes that a condition for introducing a new concept C by stipulation is that the resulting theory is *proof-theoretically conservative* over the C-free fragment of the language [Belnap 1962]. But this is not sufficient to generate the required entitlement. At least we should insist on *semantic conservativeness*, which is a matter of not excluding possibilities.[24] Proof theoretical conservativeness is a weaker requirement than semantic conservativeness. So it seems, *pace* Belnap, that we should at least insist on semantic conservativeness of the proposed stipulation that introduces a new concept.

Indeed, for natural choices of non-classical logic, introducing the disquotational principles (T1) and (T2) (against the background of a syntax theory) results in a theory that is *semantically conservative* over the truth-free part of the language: every model for the original theory in the original language (which does not contain the truth predicate) can be expanded to a model of the disquotational truth theory. This is easily seen for an arbitrary model M of our arithmetical theory S, if we consider the set of all arithmetical sentences that are true in the model. We close this set under all finite iterations of (positive) compositional principles including truth introduction resulting in a suitable extension for the disquotational truth predicate.[25] This means that in the context of such a non-classical logic, the introduction of fully disquotational truth does not exclude any possibilities; rather, possibilities are fleshed out more fully with the aid of the notion of disquotational truth.

McGee argues that a compositional notion of disquotational truth can stipulatively be introduced, and that its semantic conservativeness guarantees that this notion of truth can then do justificatory work for us [McGee 2005a, Section 5, p. 94]:

> So what justifies a disquotationalist in accepting the compositional theory of truth?
>
> Again, a one-word answer: [semantic] conservativeness.

But even semantic conservativeness is not enough to guarantee that the stipulatively introduced truth concept can function in justification, as can be seen as follows. Suppose we start with Peano Arithmetic, formulated in the language of arithmetic (without a truth predicate). Now consider a theory S, consisting of the axioms of Peano Arithmetic with the truth predicate not allowed in instances of the induction scheme, and classical logic extended to sentences involving the notion of truth. Moreover, S contains one further axiom:

$$M \vee \exists\varphi \neg \mathsf{IND}(\mathsf{T}\varphi),$$

where M is some very strong arithmetical principle (asserting the consistency of ZFC plus a large cardinal axiom, say, such that even the consistency of the resulting theory is not beyond doubt), $\mathsf{IND}(\mathsf{T}\varphi)$ is the instance of the induction scheme for $\mathsf{T}\varphi$ with φ a (code of a) formula of the truth free part of the language with one free variable.[26] Then a routine model expansion argument shows that S is semantically conservative over Peano Arithmetic. So by McGee's argument, it is admissible *stipulatively* to introduce the predicate T in this manner. Moreover, it is easy to see that S plus induction for the extended language (including the truth predicate[27]) proves M. Therefore, in Wright's terminology, we *cannot* be entitled to the presupposition given by stipulatively introducing T in this manner. Contrary to the recipe given by Wright (see page 7), in fact, there are good reasons to doubt our presupposition as viciously circular: in the act of justifying M, we presupposed a conditional $A \rightarrow M$ with an antecedent A that is clearly true, given our other presupposition on the presence of truth in induction.[28]

In sum, model-theoretic conservativeness – let alone proof-theoretic conservativeness – is not sufficient to underwrite an entitlement to rely on reasoning principles governing an introduced notion. But the concept of disquotational truth is not affected by the problems just sketched, and we may embark in an epistemologically blameless way on a new cognitive project of justifying mathematical knowledge by presupposing the validity of the disquotational principles (T1) and (T2).[29] This is because we consider them as presuppositions of a cognitive project in accordance with Wright's requirements (i) and (ii). If all is well, i.e. if our presuppositions are correct and our trust in those principles is not misplaced, then there is a fully disquotational truth concept, governed by non-classical logical rules, that we are entitled to rely on in our reasoning. We can be entitled to rely on (T1) and (T2) in our arguments without having *justification* for it. This absence of justification for our truth rules does not prevent us from gaining knowledge of the conclusions we reach by relying on them and also claiming knowledge for them.

Following this line of reasoning, Theano is then entitled to expand her conceptual resources by a disquotational truth predicate governed by FDE principles. We will argue in the following section that such a disquotational truth concept is indeed suitable for playing a key role in genuine justificatory processes in mathematics.

5. Truth and Mathematics

We now finally explain how disquotational truth can play a justificatory role in the foundations of mathematics.

> **Hypatia** works in the foundations of mathematics. Her epistemic commitments are like those of Theano, except that she is in addition happy to rely on full disquotational truth in her reasoning. She is persuaded that at least a portion of arithmetic can be fully justified. In regard to "stronger" infinitistic methods she is more careful. Although she is not strictly refusing these infinitistic parts, she intends to justify them by extending her justification of arithmetic to richer areas of mathematics.

As mentioned above, Gödel's theorems did not only show that Hilbert's finitistic methods are probably too restrictive, but maybe more importantly provided a genuine way to expand sound formal systems by principles that are equally sound but not provable in the theory: one example of such principles are reflection principles. This process is also at the heart of Feferman's formulation of Predicative Mathematics on the basis of one's implicit commitments contained in the acceptance of arithmetical principles [Feferman 1991]. We intend to articulate this process by providing epistemological foundations to it; Wrights' notion of entitlement to a cognitive project introduced in the previous sections will precisely serve this purpose. We argue that expansions by reflection principles can extend the set of (mathematical) sentences we are justified in believing.[30]

The picture we want to propose is as follows. We trust basic arithmetical principles, basic logical rules and principles, and basic truth rules and principles. This is evident from the way in which we use what we establish on the basis of these rules and principles. Just as perceptual states (as representational states) are integrated in our belief system, so are our arithmetical proof states integrated in our belief system. Indeed, we indispensably use arithmetical theorems in our best explanations of physical phenomena. Similarly, elementary reasoning involving principles and rules of truth is integrated in our belief system. All this is just to say that we *trust* these principles and rules. Reflection principles *express* our trust in these rules and principles. *If* we were entitled or justified in our full and unqualified acceptance of the rules and principles we started out with, then we are moreover *entitled* to explicitly embrace our trust by coming to believe this reflection principle. In such circumstances, we are entitled to do this without providing any independent justification for the reflection principle.

Our story is related to a story of implicit commitment that has been discussed critically by [Dean 2015] and [Cieśliński 2017]. Dean for example points out that the implicit commitment thesis should not be taken as a general requirement; he showed the incompatibility of the presence of implicit commitments and certain foundational programs bound to a fixed formal system. In our story we understand it rather as a reasonable and warranted possibility on expanding a formal system, transferring our trust in the original theory to the expanded theory, in cases where the informal understanding transcends the formal system.[31] In Gödel's words: the new axioms are "just as evident and justified" as those with which we started.[32]

We can be a bit more specific and suppose that Hypatia is justified in believing in the truth of – and therefore to accept – a weak fragment of Peano Arithmetic, call it *S*. The question of what exactly the principles are that characterise Hypatia's commitments should not bother us too much. For our purposes, she might regard *S* to be Primitive Recursive Arithmetic PRA as acceptable, or she might consider primitive recursion, and even exponentiation as non-acceptable operations (like [Parsons 1998], for instance) and therefore opt for weaker systems such as *Elementary Arithmetic* EA or one of the sub-exponential arithmetical systems such as $\mathsf{I}\Delta_0 + \Omega_1$ or Buss' S^1_2 (see [Hájek and Pudlák 1998]) respectively. The details of these systems do not matter: what matters is that Hypatia can freely choose a very weak arithmetical theory as her basic standard for mathematical justification: we only assume that the weaker she goes, her commitments become more and more uncontroversial. The question now before us is: from her epistemic vantage point, can Hypatia come to be justified in believing a portion of mathematics that non-trivially surpasses her initial theory?

For the sake of definiteness, let's assume that Hypatia's justified arithmetical beliefs *S* amount to the principles of EA. As for her logical background, she is entitled to rely on FDE logic in a fully schematic form so that she knows any arithmetical sentence that can be seen to follow from the axioms of EA.[33] Now she is warranted in introducing a notion of disquotational truth. As we have seen, she cannot justify the validity of the disquotational principles (T1) and (T2); nonetheless, she is entitled to embark on a cognitive project that involves adopting them. Thus Hypatia comes to accept the theory *S* formulated in the language expansion with a truth predicate and closed under FDE logic and the disquotational rules for truth: call this theory TS_0.

When she does so, her acceptance of TS_0 includes her firm belief that all the theorems of this theory are true. She comes to accept the stronger theory obtained by reflecting on the basic disquotational theory TS_0. If all is well, she is *entitled* to embrace reflection principles or rely on reflection rules for TS_0.[34]

Hypatia is justified in believing all the mathematical theorems of this extended theory. Moreover, the reliability of the disquotational truth concept and the process of reflection allows her to believe in the truth of everything that the extension of TS_0 with reflection proves. Hypatia is then again entitled to adopt reflection principles or rules for the stronger theory and justified in accepting all the (mathematical) theorems of a further iteration of reflection over TS_0.

As already mentioned, several reflection principles and rules are discussed in the literature. The most natural candidate in a setting with the truth predicate is the Global Reflection Principle in the form $\mathsf{Bew}_T(\varphi) \Rightarrow \mathsf{T}\varphi$.[35] This reflection principle talks about the theorems and it is sufficient for a classical setting, where we have the expressive resources to rewrite a sequent $\Gamma \Rightarrow \Delta$ as an equivalent theorem $\Rightarrow \bigwedge \Gamma \rightarrow \bigvee \Delta$. In our non-classical FDE-setting we lack an object linguistic conditional to transform provable sequents into theorems. Therefore we have to distinguish properly between provable theorems of the form $\Rightarrow A$ and provable sequents $\Gamma \Rightarrow \Delta$. This turns out to be a useful and intended property as it allows us to handle also pathological sentences in our proof system.[36]

The distinction also makes it necessary to adapt the formal provability predicates. In order to express provability of sequents we employ a one-place predicate Pr_T representing the derivability-in-T of sequents, i.e., if $T \vdash \Gamma \Rightarrow \Delta$ then $\mathsf{EA} \vdash \Rightarrow \mathsf{Pr}_T(\ulcorner \Gamma \Rightarrow \Delta \urcorner)$. This leads to a reflection rule for provable sequents:

$$\frac{\Rightarrow \mathsf{Pr}_T(\ulcorner \Gamma \Rightarrow \Delta \urcorner)}{\Gamma \Rightarrow \Delta} \qquad (\mathsf{r}_T)$$

In words: if we have established that our background theory can formally recognise that the sequent $\Gamma \Rightarrow \Delta$ is provable in T, then we can conclude that this sequent holds.

A further step for our reflection process is the realization that it is not only the soundness of derivable sequents that we can reflect upon. Additionally we can also consider rules that are admissible in a proof system. In order to adopt this form of reflection we employ a two-place provability predicate $\mathsf{Pr}^2_T(\ulcorner \Gamma \Rightarrow \Delta \urcorner, \ulcorner \Theta \Rightarrow \Lambda \urcorner)$ representing the fact that it is admissible in T to infer $\Theta \Rightarrow \Lambda$ from $\Gamma \Rightarrow \Delta$. Our second reflection rule involves admissible rules:[37]

$$\frac{\Rightarrow \mathsf{Pr}^2_T(\ulcorner \Gamma \Rightarrow \Delta \urcorner, \ulcorner \Theta \Rightarrow \Lambda \urcorner) \qquad \Gamma \Rightarrow \Delta}{\Theta \Rightarrow \Lambda} \qquad (\mathsf{R}_T)$$

It states that if we can formally recognise that the rule $\frac{\Gamma \Rightarrow \Delta}{\Theta \Rightarrow \Lambda}$ is admissible in T and $\Gamma \Rightarrow \Delta$ holds, also $\Theta \Rightarrow \Lambda$ holds. It is clear that the first principle can be derived from the second, and it is in fact the latter that will be mostly employed to unfold Hypatia's commitments in what follows.

As we have discussed earlier it is unproblematic to extend a fully disquotational theory of truth such as TS_0 with a global reflection rule and it is a coherent undertaking to do so. Therefore the choice of the rules (R_T) and (r_T) in schematic (uniform) form and without explicit mention of the notion of truth can be regarded as a *technical* and not as a *conceptual* one.[38] As we shall see soon, iteration of application of these rules starting from TS_0 yields mathematically sound systems that are interesting both from a truth-theoretic and from a proof-theoretic perspective.

The situation is different in a setting based on classical logic such as the one considered in [Horsten & Leigh 2017]. There one in iterates reflection principles over theories that are not fully disquotational. In the latter case uniform reflection rules such as (r_T) and global reflection rules provably come apart. As Observation 1 discussed in §4.1 shows, the latter forces internal or external contradictions, whereas the former do not. This shows that the strategy proposed here is a significant improvement on the strategy proposed in [Horsten & Leigh 2017], especially in relation to the epistemological status of the new claims obtained by means of the reflective process.

The compositional conception of truth introduced on page 11 can now be fully recovered by Hypatia. For the compositional principles of conjunction and disjunction we need only to reflect once over the arithmetical base theory TS_0. Let's

consider for instance how (T1) and (T2) can help us to recover the compositional sequent for conjunction

$$\mathsf{T}\varphi \wedge \mathsf{T}\psi \Rightarrow \mathsf{T}(\varphi \wedge \psi) \qquad (\mathsf{T}\wedge)$$

First, (T1) and (T2) enable one to establish the compositionality of truth over conjunction for all schematic variables over sentences, namely they suffice to establish the *schema* $\mathsf{T}\ulcorner A\urcorner \wedge \mathsf{T}\ulcorner B\urcorner \Rightarrow \mathsf{T}(\ulcorner A \wedge B\urcorner)$ for all sentences A, B. The uniformity of this process – that is the possibility of formalizing this in S for all A, B – allows us then, with the help of the reflection principle R_T, to transform the schematic formulation of the principle into the *object-linguistic*, quantifiable principle ($\mathsf{T}\wedge$). In a similar fashion we are able to recover all of the compositional principles with two iterations of the rule R_T over the basic disquotational theory TS_0, a theory that we call $\mathsf{R}^2(\mathsf{TS}_0)$. In fact, there is a sense in which $\mathsf{R}^2(\mathsf{TS}_0)$ can achieve even more than what the classical theory KF has to offer.[39] $\mathsf{R}^2(\mathsf{TS}_0)$ contains commutation principles for negation in *rule-form* via the sequents

$$\mathsf{T}(\neg\varphi) \Rightarrow \neg\mathsf{T}\varphi, \qquad \neg\mathsf{T}\varphi \Rightarrow \mathsf{T}(\neg\varphi). \qquad (1)$$

In the classical Kripke-Feferman theory on the other hand, because of the Liar paradox, the truth predicate *cannot* commute with negation. The price to pay for such generality is that, in the present setting, we only obtain inferences that, by the nature of the FDE-conditional, cannot be internalized; in the case of the principles (1), this means that such sequents do not entail the corresponding conditional sentences $\Rightarrow \mathsf{T}(\neg\varphi) \rightarrow \neg\mathsf{T}\varphi$ and $\Rightarrow \neg\mathsf{T}\varphi \rightarrow \neg\mathsf{T}(\neg\varphi)$.

If both the coherence of Hypatia's entitlement to reflection and the presence of general forms of compositionality amount to compelling evidence that her cognitive project is acting on the right presuppositions, what still remains to be seen is how these new truth theoretic principles can lead to her new arithmetical knowledge. A first observation in this direction is that, although our starting theory EA features only a *restricted* form of induction, the theory $\mathsf{R}(\mathsf{TS}_0)$ obtained by closing TS_0 under R_T already gives us the *full induction schema* for the language $\mathcal{L}_\mathsf{T}$. Moreover, an additional reflection step enables us to reach the principle of transfinite induction up to and including the ordinal ω^ω. This is the principle stating that a property $P(x)$ expressed by a predicate of $\mathcal{L}_\mathsf{T}$ holds of all ordinals smaller than ω^ω if it can be naturally iterated over a standard well-ordering of the ordinals, i.e., if when $P(x)$ holds for all ordinals $\beta < \alpha$, $P(x)$ also holds of α.[40] This is already *more than* what full Peano arithmetic formulated in the language $\mathcal{L}_\mathsf{T}$ and governed by FDE can give: by a result of [Halbach & Horsten 2006], it can only prove transfinite induction for $\mathcal{L}_\mathsf{T}$ up to any ordinal of the form ω^n, with n a natural number.

So far it then seems that the reflective process Hypatia has embraced is leading her *just* beyond Peano Arithmetic, but there is no clear indication of how Hypatia's presupposition of disquotational truth might *substantially* contribute to her cognitive project of justifying mathematical claims. After all it is well-known that the step from EA to PA, even when formulated in the expanded language,

can be obtained by means of (uniform) reflection over EA alone. However, the combination of transfinite induction and full compositional truth just introduced and that one can reach in iterations of R_T over TS_0 enables Hypatia to significantly exceed the mathematical content of Peano Arithmetic in the way we now indicate.

Predicative Analysis, as characterised in [Feferman 1964], for instance, can be seen as a hierarchy of comprehension principles over Peano Arithmetic of the form

$$\exists X^{\alpha} \forall u (u \in X^{\alpha} \leftrightarrow B(u)) \qquad (\mathrm{CA}^{\alpha})$$

where α is an ordinal smaller than the Feferman-Schütte ordinal Γ_0 and where $B(x)$ is a formula that can contain quantification only over sets of level $\beta < \alpha$. (CA^{α}) essentially enables us to define sets of natural numbers via previously defined ones and without quantifying over totalities yet to be defined.[41]

Now full compositional truth and transfinite induction up to an ordinal α enable us to recover (CA^{α}) for all $\beta < \alpha$ via a hierarchy of typed truth predicates:[42] the basic idea is that the set X^{α} is interpreted as (the code of) a formula $\varphi(x)$ of $\mathcal{L}_{\mathsf{T}}$ with one free variable and containing only iterations of truth predicates of length $\beta < \alpha$; $x \in X^{\alpha}$ is then interpreted as $\mathsf{T}_{\alpha}\varphi(x)$ (see for instance [Feferman 1991]). Therefore this general pattern yields that in accepting EA, the logic FDE and full disquotational truth, Hypatia is entitled to accept all consequences of $\mathsf{R}^2(\mathsf{TS}_0)$ that, as we have just seen, include iterations of (CA^{α}) up to ω^{ω}. And the latter theory is substantially stronger than Peano Arithmetic.

Two iterations of the rule R_T, however, is not the end of Hypatia's entitlements. She can go on repeatedly to reflect on the previous stages. This results in her accepting ever larger fragments of Predicative Analysis. One of the questions is how far Hypatia is entitled to carry out this reflection process? Another question is whether this would be sufficient to accept all of Predicative Analysis.

Regarding the first question it appears reasonable to allow Hypatia iterations of the reflection process along ordinals that can be apprehended to be well-founded from the perspective of Hypatia's point of view. We follow Feferman's strategy of autonomous progressions of ordinals to make this step formally precise. This means that Hypatia will be allowed to carry out arbitrary iterations of length less than Γ_0.

To answer the second question we employ a result from [Fischer et al. 2017]. By letting ω_0 to be ω, and ω_{n+1} to be ω^{ω_n}, one has:

Proposition 2. $\mathsf{R}^{\omega_n+1}(\mathsf{TS}_0) \vdash \mathsf{TI}_{\mathcal{L}_T}(\omega_n)$.

This observation guarantees that for all ω_n there is an ordinal $\alpha < \epsilon_0$ – where ϵ_0 is the limit of all ω_n –, such that α iterations of the reflection process allow one to prove transfinite induction for the language of truth for ω_n. Since for any $\beta < \epsilon_0$ there is an ω_n, such that $\beta < \omega_n$, we have thus directly established that iterations up to ϵ_0 allow for transfinite induction up to ϵ_0.

Now in $\mathsf{R}^{\omega_n+1}(\mathsf{TS}_0)$ it is possible to well-order – with respect to arithmetical predicates such as 'x is an axiom of $\mathsf{R}^{\alpha}(\mathsf{TS}_0)$' for suitable α – recursive ordinals

even bigger than ϵ_0. By following the strategy of autonomous progressions of theories initiated by [Feferman 1964], one can allow for iterations up to Γ_0, the Feferman-Schütte ordinal. By iterating the reflection process in this way, Hypatia is able to prove transfinite induction for truth up to the limit of the autonomous progressions, i.e. $\mathsf{TI}_{\mathcal{L}_T}(< \Gamma_0)$. Thus Hypathia has moved from a very modest commitment to a portion of arithmetic to a full-blown Predicativist position. Evidently it does not matter for our argument whether or not Feferman's characterisation of Predicativism is definitive. Our point is merely that what Feferman takes to be Predicative Analysis is mathematically much stronger than Hypathia's starting point.

If all is well, then the result of this process is Hypatia *knowing* the theorems of (what Feferman takes to be) Predicative Analysis, where 'all is well' means that Hypatia was *justified* in her belief in EA in the first place, is *entitled* to rely on FDE logic, is *entitled* to rely on the inference rules that govern the concept of disquotational truth, and is *entitled* to rely on the reflection rule R_T for a suitable T. The concept of disquotational truth plays a crucial role in this process: the nominalising function of disquotational truth (semantic ascent) allows formulas to be treated as objects (sets) that can be quantified over.

The reflective process that we have described is not the way in which the bounds of mathematical knowledge are *typically* explicitly extended. Attempts in the foundations of mathematics to extend these bounds often invoke 'strong principles of infinity', or, alternatively, strong combinatorial principles. Such principles, if they can be justified, extend the limits of mathematical knowledge in much more dramatic ways than iterated reflection does.

Thus there are also other ways in which Hypatia may come to accept Predicative Analysis. For instance, she may straightaway, i.e., without going through the iterative reflection process described above, acquire a belief in Zermelo-Fraenkel set theory, perhaps by coming to understand and accept a version of the iterative conception of set. If ZF can indeed be justified from the iterative conception, then Hypatia can in this way come to know a mathematical theory that is much stronger than Predicative Analysis. This way of extending the scope of our mathematical knowledge differs structurally from extension by reflection. In order to accept a new axiom (strong principle of infinity, combinatorial principle), we need to do justificatory work, whereas no new justification is needed to adopt the global reflection rule for a theory that you are already justified to believe in. The global reflection rule for a theory is *exactly as safe* as the theory itself and there is no reason to doubt that Hypatia can know this.

The reflective process that we have described in this section is not restricted to weak theories of arithmetic but also applies (in essentially the same way) to stronger theories. In particular, it applies to our most encompassing justified mathematical theory.[43] In this way, disquotational truth plays not just a justificatory role in mathematics, but even a *foundational* role: however many principles of infinity we have come to accept, we are always implicitly committed to more than what can logically be derived from them.

6. Claiming Knowledge

The conclusion of the foregoing is that Hypatia can, from a starting point where she is justified in believing the consequences of a weak theory of arithmetic, by reflection and relying on disquotational truth, arrive at epistemically entitled belief in Predicative Analysis. By Wright's lights ([Wright 2004b, section VIII]), if Hypatia can in addition come to *know* that she is justified in believing in elementary arithmetic, then she can come to *know* that she knows what follows from the axioms of Predicative Analysis, i.e., she can "claim knowledge" of theorems of Predicative Analysis.

Before we tackle the question of justifying the logical laws we address a possible worry. According to Glanzberg, type-free truth predicates face a problem with explaining away strengthened liar reasoning: by the very lights of Hypatia's truth theory, the reflection principle $\mathsf{R}_{\mathsf{TS}_0}$ would have to be already part of her truth theory TS_0, and thereby her position is unstable. Glanzberg's argument goes as follows. The theory TS_0 is silent about the truth value of the liar sentence, and therefore does not classify it as true. The type-free truth theorist tries to dissolve the threat of strengthened liar reasoning by emphasising that only what is asserted by her truth theory is to be taken as true. So, in particular, the statement that the liar sentence is not true, should not itself be taken to be true. So for Hypatia, the reflection principle is part of the explanation of what goes wrong in the strengthened liar reasoning. But then, Glanzberg says [Glanzberg 2004, p. 294]:[44]

> this principle [$\mathsf{R}_{\mathsf{TS}_0}$] must be properly assertible. The norms of assertion require us to only assert what we take to be true. But by the very view being offered, the only ground for truth there can be is the provability of truth in [TS_0]. Hence, the explanation *requires* the provability of truth of [$\mathsf{R}_{\mathsf{TS}_0}$] in [$\mathsf{TS}_0$] for the explanation to be acceptable by its own lights.

Of course, for familiar Gödelian reasons, TS_0 cannot contain $\mathsf{R}_{\mathsf{TS}_0}$.

Against this, we maintain that Hypatia is not forced to accept that 'the only ground for truth there can be is the provability of truth in a theory', i.e., she does not and should not believe that *only* what is proved by TS_0, is acceptable. Indeed, she implicitly has the resources for acquiring more truths: she is implicitly committed to $\mathsf{R}(\mathsf{TS}_0)$, and this goes beyond the explicit content of TS_0.

Later in his paper Glanzberg acknowledges the possibility of reflection as being only implicit in the formulation, but he takes this to reveal the hierarchical nature of truth although he also maintains that it is still the same concept. So Glanzberg is right that the closure under reflection principle is crucial, but this feature is not only available to hierarchical approaches. In the end, reflection as implicit commitment is perfectly compatible with Hypatia's silence about the truth value of the liar sentence.

The question whether and how Hypatia can also come to know that her logical inference rules are valid, and that the reflection process and the disquotation truth rules are reliable, is more delicate. According to [Wright 2004b], in order for Antigone to know, for instance, that an instance of the Disjunction Introduction

rule is valid, she would have to prove the corresponding material conditional. Clearly such proofs will typically be circular, but, according to Wright, not viciously so [Wright 2004b, section VIII, p. 173].

Antigone, however, has only signed up unrestrictedly to FDE logic. This means that Conditional Introduction is not unrestrictedly available to Antigone (or Hypatia). Therefore, she is not able, according to Wright, to claim the validity of Disjunction Introduction. Similar remarks hold, *mutatis mutandis*, for the reliability of reflection and of the disquotational truth rules.

What Hypatia can do, is to prove the reliability of Conditional Introduction for the arithmetical instances of her inference rules. Moreover, she can do this, using the truth predicate, in a *uniform* manner. For instance, Hypatia can show:

$$\Rightarrow \forall \sigma, \tau \in \mathcal{L}_0 : \mathsf{T}(\sigma) \rightarrow \mathsf{T}(\sigma \vee \tau).$$

This also works for all the other rules used in the classical mathematical theory of Predicative Analysis.

We do not have to stop here: we can step by step expand the range of sentences for which the classical rules are provably valid. Basically, we can prove more and more sentences to be *grounded*. The fragment of the language for which we can do this corresponds to the amount of transfinite induction for the language containing the truth predicate is provable. We will have Conditional Introduction exactly for these initial segments of the minimal fixed point.

7. Two Cognitive Projects

We have discussed the entanglement of two types of cognitive project: one about logic, another about truth.

The first project, in its boldest form, involves the *acceptance of full classical logic* in open-ended schematic form. Some have argued that we are entitled to rely on, and *must* rely on, particular unrestricted inference rules governing particular logical concepts because we could not have the concepts without relying on the rules [Boghossian 2003]. In particular, our entitlement to rely on Conditional Introduction has been defended in those terms. However, there are strong reasons for rejecting this line of reasoning. For any logical concept and any logical rule governing it, it is possible to understand the concept without accepting the logical inference rule [Williamson 2003]. It seems then that acceptance of logical inference rules is never *inevitable*; one can never be *forced* to do so. But this does not mean that one ought not to fully engage in this cognitive project: we may well be entitled to accept, in an epistemologically blameless way, full classical logic in open-ended schematic form.[45]

The second project consists in *fully embracing a notion of type-free disquotational truth*. We have argued that from a place where one has not yet signed up to Conditional Introduction in open-ended schematic form, one can come to accept such a notion.

Nonetheless, the two cognitive projects clash with each other. One cannot *fully rely on* classical material implication and on type-free disquotational truth at the same time—even though there is no problem whatsoever in *understanding* both concepts at the same time. Having signed up to one of these two projects, one can of course always reconsider, retrace one's steps, and embark on the other project instead. But it is impossible to *exercise* or *practice* both concepts at the same time. In this light it might be interesting to investigate, in more detail than has been done so far, the *relations* between cognitive projects in general (and the entitlements that go with them).

It does not follow from anything that we have said that one of the two projects is somehow flawed, or epistemologically blameworthy. It is just that engaging in a cognitive project imposes limitations: *choosing is losing*. Both the practicer of material implication and the user of disquotational truth may well have their own 'warrants for nothing'. But if you open-endedly rely on classical material implication, then you cannot also use full disquotational truth. If you rely on a notion of full disquotational truth, then you cannot fully rely on the inference patterns of classical material implication.

Incidentally, there is a connection here with the literature on abstraction principles. *Irenicity* is a rational acceptability condition on theories of abstraction that has attracted fairly wide levels of support.[46] This condition says, roughly, that any two abstraction principles that are judged to be admissible by themselves, should also be judged to be jointly admissible. What we are suggesting here is that a corresponding rationality condition should not be taken to hold for cognitive projects.

At any rate, both cognitive projects that we have discussed have their benefits and drawbacks. On the one hand, the absence of full Conditional Introduction undeniably makes mathematical argumentation cumbersome and restricts its power, whereas mathematically reasoning in classical logic is perfectly natural.[47] On the other hand, the concept of scientific truth can do no justificatory work in the foundations of mathematics, whereas we have argued that Hypatia's silences allow her to acquire a fully disquotational truth concept which can do justificatory work for her in the foundations of mathematics.

Notes

[1] Throughout the article we accept Field's distinction in this rough-and-ready way without arguing for it. Observe that one can accept Field's distinction while being sceptical of the credentials of one of the concepts. Horwich, for instance, agrees that correspondence notions of truth aim at being useful in science, but he disputes that these notions can live up to their promise [Horwich 1998]. Field himself has over the years also become sceptical about the usefulness of what we call the scientific concept of truth. McGee accepts Field's distinction, but argues that only disquotational truth can play a fundamental role in justifying new mathematical principles [McGee 2005a], [McGee 2005b]. Horwich, on the other hand, holds that disquotational truth can play no essential justificatory role [Horwich 1998].

[2] Such a strategy has been suggested in [Horsten 2011].

[3] For example, Gödel's remarks in his Gibb's lectures can be seen interpretable in this way: "Hence he has a mathematical insight not derivable from his axioms" in [Gödel 1990] p.309.

[4] See [Feferman 1962], [Feferman 1991] and also [Franzén 2004].

[5] For a survey of the theories studied by the first community, see [Halbach 2014]. For a survey of the non-classical theories, see [Field 2008].

[6] This worry goes back at least to [McGee 1991, Objection 3, p. 102–106].

[7] Similarly, it plays a fundamental role in Aczel's reconstruction of Frege's logicism [Aczel 1980].

[8] McGee calls this notion *correspondence truth*; Field calls it *inflationary truth*.

[9] Cf. [McGee 2010, p. 423].

[10] See [Feferman 1991, p. 2].

[11] Field speaks of *deflationary truth*; at times McGee also uses this term. In the literature this notion is often labeled as *transparent truth* or *naive truth*.

[12] The ordinary language concept of truth may perhaps be seen as a *third* truth concept. This concept may be *inconsistent* [Burgess & Burgess 2011].

[13] We assume that the reader is somewhat familiar with this distinction. Two seminal articles are [Burge 1993] and [Wright 2004a].

[14] Or rather, at best it is a *logico-mathematical* notion. For a discussion, see for instance [Horsten 2011, Chapter 10].

[15] If Quine's confirmational holism is rejected, then there may be room for entitlement for relying on the rules of logic. See section 7.

[16] See for [Friedman & Sheard 1987] and [Halbach 2014].

[17] In our formalism, Bew_S expresses provability in the theory S in a canonical way, φ is an *object-linguistic* (not schematic!) variable ranging over sentences of the language $\mathcal{L}_S$ of S.

[18] As pointed out by a referee, KF is internally consistent for *arithmetical* sentences. One might therefore think that this is sufficient for justificatory work in the foundations of mathematics. It is correct that KF is internally consistent for arithmetical sentences, but not obviously so. An argument is needed to establish this fact, especially because some of the arithmetical theorems of KF necessarily make use of non-arithmetical statements in their derivation. So the KF principles cannot be used in a cognitive project as they rely on a prior justification. Moreover, internal consistency is primarily concerned with first-order arithmetical statements, whereas, as we shall see later on, for a justification of second-order theories such as predicative analysis we also make use of the truth vocabulary for interpreting sets of natural numbers.

[19] Although we only focus on the case of theories in which our trust is also reasonable, it might be interesting to investigate a more general conception of entitlement that would also cover cases of misplaced trust. We postpone this more general investigation of the notions of entitlement and trust itself for further work.

[20] It is defended in [Soames 1999].

[21] See for example [Priest 2008] for a presentation. In [Fischer et al. 2017] the logic is labelled as *Basic De Morgan logic*, BDM following Field's terminology in [Field 2008]. The differences are minor. For a full description of *Basic De Morgan logic*: see [Fischer et al. 2017].

[22] For a precise treatment of these issues see [Fischer et al. 2017, Section 2.3].

[23] The suggestion for stipulating the meaning of a truth notion by giving a partial implicit definition of it is not without precedent. Soames, for instance, claims that the meaning of our truth predicate is given by axioms of a system in the KF family, and that these axioms can be taken to be a *partial implicit definition* of the meaning of our notion of truth [Soames 1999].

[24] A theory S' in a richer language is semantically conservative over a theory S in the background language iff every model of S can be expanded to a model of S'.

[25] For details see [Fischer et al. 2017].

[26] Informally, $\mathsf{IND}(\mathsf{T}\varphi)$ reads: if φ is true of 0, and for every n, if φ is true of n then it's true of $n+1$, then φ is true of every n.

[27] McGee holds that our commitment to mathematical induction is open-ended, and, anyway, reflection principles to which we are implicitly committed take one from induction without the truth predicate to mathematical induction for the whole language including the truth predicate.

[28] A similar but more sophisticated example involves the axioms of the theory of truth KF discussed above: they are model theoretically conservative in the absence of full induction, but remarkably stronger in its presence.

[29] [Belnap 1962] would emphasise that because the rules introducing disquotational truth do not pin down the reference of T uniquely, we have not introduced *the* concept of disquotational truth, but only *a* concept of disquotational truth, which is fair enough.

[30] A related account is already articulated in [Horsten & Leigh 2017, Section 6]. However, the strategy employed here goes beyond the account of [Horsten & Leigh 2017] in that the implicit commitment is connected with a trustworthy theory of truth that allows for a cognitive project of justification.

[31] For example in [Tait 1981, p. 545], Kreisel's analysis of finitism is criticised on the grounds that a finitist cannot recognise the validity of PRA because she cannot rely on the notion of function. But our situation is different. We do not claim that Hypatia is a finitist and so she can come to accept the validity of the theory TS_0 in full generality.

[32] See for [Gödel 1990], p.151.

[33] We can take our arithmetical theory to be formulated in FDE; as it is shown in [Fischer et al. 2017], in fact, as long as we focus on arithmetical theorems, classical arithmetic and arithmetic in FDE coincide.

[34] Giving a detailed epistemological analysis of the process of coming explicitly to accept a reflection principle is beyond the scope of this article. First steps in this direction are taken in [Horsten 2019].

[35] An alternative is the scheme of uniform reflection $\mathsf{Bew}_T(\ulcorner Ax \urcorner) \rightarrow A(x)$, for all formulas $A(v)$.

[36] It can be seen from within our non-classical framework that if the liar sentence, for instance, is asserted, a contradiction ensues, and likewise if the negation of the liar sentence is asserted, a contradiction follows. In this sense the liar sentence can be labelled pathological in our non-classical setting. Thanks to one of the referees for pointing out the need to clarify this sense of pathologicality.

[37] A further clarification is appropriate. In our reflection principles we intend to use uniform versions, i.e., versions for formulas with free variables. For details see [Fischer et al. 2017].

[38] Actually in the context of fully disquotational truth one can even show that suitable rules of global and uniform reflection coincide: see [Fischer et al. 2017, Prop. 1].

[39] See [Fischer et al. 2017, Lemma 4].

[40] For details, see [Fischer et al. 2017, Proposition 3].

[41] Alternatively, but equivalently, one can describe Predicative Analysis via the second-order system ATR_0 that is one of the 'big-five' systems in Reverse Mathematics. Our description can be seen as the result of iterating Π^0_1-comprehension up to Γ_0. This system and ATR_0 are proof-theoretically equivalent and therefore equally good to characterize Predicative Analysis. We choose the stratified formulation because it compares better with iterations of truth ascriptions.

[42] For a general approach on how to obtain typed truth predicates in theories of disquotational truth in FDE we refer to [Nicolai 2018].

[43] *Pace* [Dean 2015, p. 56]. For details of how this phenomenon generalises in straightforward ways to stronger theories such as second-order number theory or set theory, see [Fujimoto 2012].

[44] Glanzberg is focusing on a type-free truth theory that is somewhat different from TS_0, but this does not affect the argument.

[45] This question is too large for us to tackle in this article.

[46] See [Ebert & Rossberg 2016].

[47] See [Halbach 2014, chapter 20], [Halbach & Nicolai 2018], [Nicolai 2018].

References

P. Aczel, Frege Structures and the Notions of Truth and Proposition, in J. Barwise and H. J. Keisler and K. Kunen (eds.), *The Kleene Symposium*, North-Holland, 1980.

J.C. Beall & B. Armour-Garb (eds.) *Deflationary truth*. Open Court, 2005.

Belnap, N. Tonk, plonk and plink. *Analysis* 22(1962), p. 130–134.

Boghossian, P. Blind reasoning. Proceedings of the Aristotelian Society Supplemental Volume LXXVII, p. 225–248, 2003.

Burge, T. Content preservation. *Philosophical Review* 102(1993), p. 457–488.

Burge, T. Self and self-understanding. *The Journal of Philosophy* 108(2011), p. 287–383.

Burgess, A. & Burgess, J. *Truth*. Princeton University Press, 2011.

Cezary Cieśliński, *The Epistemic Lightness of Truth: Deflationism and its Logic*, Cambridge University Press, 2017.
Dean, W. Arithmetical reflection and the provability of soundness. *Philosophia Mathematica* 23(2015), p. 31–64.
Ebert, P. & Rossberg, M. (eds) *Abstractionism*. Oxford University Press, 2016.
Feferman, S. Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic* 27(1962), p. 259–316.
Feferman, S. Systems of predicative analysis. *Journal of Symbolic Logic* 29(1964), p. 1–30.
Feferman, S. Reflecting on incompleteness. *Journal of Symbolic Logic* 56(1991), p. 1–49.
Feferman, S. Axioms for determinateness and truth. *Review of Symbolic Logic* 1(2008), p. 204–217.
Field, H. Deflationist views of meaning and content. *Mind* 103(1994), p. 249–285.
Field, H. *Saving truth from paradox*. Oxford University Press, 2008.
Fischer, M., Nicolai, C., & Horsten, L. Iterated reflection over full disquotational truth. *Journal of Logic and Computation* 27(2017), p. 2631–2651.
Franzén, T. *Inexhaustibility*, ASL Lectures Notes in Logic, CRC Press, 2004.
Friedman, H. & Sheard, M. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic* 33(1987), p. 1–21.
Fujimoto, K. Classes and truths in set theory. *Annals of Pure and Applied Logic*, 163(2012), p. 1484–1523.
Gödel, K. *Collected Works, Vol. II. Publications 1938-1974*, Oxford University Press, 1990.
Glanzberg, M. Truth, reflection, and hierarchies. *Synthese* 142(2004), p. 289–315.
Hájek, Petr and Pudlák, Pavel. *Metamathematics of First–Order Arithmetic*. Springer Verlag, 1998.
Halbach, V. *Axiomatic theories of truth*. 2nd edition, Cambridge University Press, 2014.
Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic* 71(2006), p. 677–712.
Halbach, V. & Nicolai, C. On the costs of nonclassical logic. *Journal of Philosophical Logic* 47(2018), p. 227–257.
Horsten, L. *The Tarskian Turn. Deflationism and axiomatic truth*. MIT Press, 2011.
Horsten, L. *On reflection*. Submitted, 2019.
Horsten, L. & Leigh, G. Truth is simple. *Mind* 216(2017), p. 195–232.
Horwich, P. *Truth*. Second edition. Clarendon Press 1998.
Kripke, S. Outline of a theory of truth. *Journal of Philosophy* 72(1975).
Martin, R. (ed). *Recent essays on truth and the liar paradox*. Oxford University Press, 1984.
Maudlin, T. *Truth and paradox. Solving the riddles*. Oxford University Press, 2004.
McGee, V. Truth, Vagueness, and Paradox. *Hackett*, 1991.
McGee, V. Two conceptions of truth? — Comment. *Philosophical Studies* 124(2005a), p. 71–104.
McGee, V. Afterword: Trying (with limited success) to demarcate the disquotation-correspondence intuition. In [Beall & Armour-Garb 2005, p. 142–152], 2005b.
McGee, V., Field's logic of truth. *Philosophical Studies* 147: 421–432, 2010.
Nicolai, C. Provably true sentences across axiomatizations of Kripke's theory of truth. *Studia Logica* 106(2018), p. 101–130.
Parsons, C. (1998). Finitism and intuitive knowledge. In M. Schirn (Ed.), *The philosophy of mathematics today* (pp. 249–270). Oxford: Oxford University Press.
Priest, G. *An Introduction to Non-classical Logic*. 2nd edition, Cambridge University Press (2008), p. 38–39.
Prior, A. The runaway inference ticket. *Analysis* 21(1960), p. 38–39.
Reinhardt, W. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic* 15(1986), p. 219–251.
Soames, S. *Understanding truth*. Oxford University Press, 1999.
Stern, J. Proving that the Mind is not a Machine? *Thought* 7(2018), p. 81–90.
Tait, W. Finitism. *Journal of Philosophy* 78(1981), p. 524–546.
Williamson, T. Understanding and inference. Proceedings of the Aristotelian Society Supplemental Volume LXXVIII, p. 249–293, 2003.
Wright, C. Warrant for nothing (and foundations for free)? Proceedings of the Aristotelian Society Supplemental Volume LXXVIII, p. 167–212, 2004a.
Wright, C. Intuition, entitlement and the epistemology of logical laws. *Dialectica* 58(2004b), p. 155–175.