A theory of implicit commitment

Mateusz Łełyk\textsuperscript{1} · Carlo Nicolai\textsuperscript{2}

Received: 23 August 2021 / Accepted: 28 January 2022 © The Author(s) 2022

Abstract

The notion of implicit commitment has played a prominent role in recent works in logic and philosophy of mathematics. Although implicit commitment is often associated with highly technical studies, it remains an elusive notion. In particular, it is often claimed that the acceptance of a mathematical theory implicitly commits one to the acceptance of a Uniform Reflection Principle for it. However, philosophers agree that a satisfactory analysis of the transition from a theory to its reflection principle is still lacking. We provide an axiomatization of the minimal commitments implicit in the acceptance of a mathematical theory. The theory entails that the Uniform Reflection Principle is part of one’s implicit commitments, and sheds light on why this is so. We argue that the theory has significant epistemological consequences in that it explains how justified belief in the axioms of a theory can be preserved to the corresponding reflection principle. The theory also improves on recent analyses of implicit commitment based on truth or epistemic notions.

Keywords Implicit commitment · Reflection principles · Mathematical justification · Incompleteness · Epistemic stability · Entitlement

1 Introduction

Let a theory be given by a collection $X$ of axioms in a language $\mathcal{L}$ and some consequence relation $\vdash$. Suppose that some idealized, logically omniscient agent $A$ is justified in believing $X$ and in the trustworthiness of the principles governing $\vdash$. We can say that the agent is \textit{explicitly committed} to some sentence $\varphi$ of $\mathcal{L}$ if and only if $X \vdash \varphi$. If there is an effective proof system for $\vdash$, this is equivalent to saying that $A$ is
explicitly committed to \( \varphi \) if and only if there is a proof of \( \varphi \) from \( X \).\(^1\) Is this all there is to say about \( A \)’s commitments? In particular, are all of \( A \)’s commitments explicit?

The phenomenon of incompleteness in mathematical theories provides clear-cut case studies to address these questions. From the results of Gödel (1931) one can infer that if \( A \) is justified in believing the axioms of a sufficiently powerful formal mathematical system \( S \), then they will not be explicitly committed to the various articulations of the soundness of \( S \) in the language \( L_S \), including ‘S is consistent’, or ‘all theorems of \( S \) are true’ (unless of course \( S \) happens to be inconsistent). Several authors have proposed strategies to extend \( S \) by taking those very soundness assertions as additional axioms, on the grounds that—although not provable—they are somewhat justified from the perspective of \( S \) (Cieśliński, 2017; Feferman, 1962; Franzén, 2004; Turing, 1939).\(^2\)

Such projects are based on what has been referred to as the *implicit commitment thesis* by Dean (2015), according to which ‘anyone who accepts the axioms of a mathematical theory \( S \) is thereby also committed to accepting various additional statements which are expressible in the language of \( S \) but which are formally independent of its axioms.’

The implicit commitment thesis involves the notion of acceptance. In this paper we follow Fischer et al. (2019) and frame the discussion of implicit commitment in terms of belief and justification. However, once suitable bridge principles are in place, our analysis applies to the notion of acceptance as well (cf. Sect. 2.1 below).

Our task will then be twofold. Starting with justified belief in \( S \), we will investigate what are the commitments implicit in this justified belief. We shall argue for the following, familiar thesis on novel formal and philosophical grounds:

**IMPLICIT COMMITMENT THESIS (ICT):** Anyone who is justified in believing a sufficiently powerful, consistent mathematical formal system \( S \) is also implicitly committed to various additional statements which are expressible in the language of \( S \) but which are formally independent of its axioms.

Our second task will be to investigate whether there is, and if yes what are its fundamental principles, an epistemic attitude that relates whoever is justified in believing \( S \) to such implicit commitments. We will argue that justified belief in \( S \) will be transferred to its (minimal) implicit commitments.

---

1. Our terminology, and more generally the terminology employed in the debate relevant for the present paper, differs from the one employed in some recent literature on ontological commitment (Peacock, 2011; Krämer, 2014), where explicit commitment is not taken to be closed under logical consequence, and implicit commitment results from this closure.

2. In particular, Turing (1939, p. 198) claims that the addition of consistency assertions to a system enables one:

   …to obtain a more complete one [system] by the adjunction as axioms of formulae, seen intuitively to be correct, but which the Gödel’s theorem [sic] shows are unprovable in the original system.

   Similarly, Feferman (1962) analyses the result of adding a reflection principle to a given system in the following way:

   In contrast to an arbitrary procedure for moving from \( A_k \) to \( A_{k+1} \), a reflection principle provides that the axioms of \( A_{k+1} \) shall express *a certain trust* [our emphasis] in the system of axioms \( A_k \).
The importance of ICT for logic and the philosophy of mathematics is undeniable. It directly or indirectly motivates Turing’s work on ordinal logics, Feferman’s foundations of predicative mathematics, the extensions of Feferman’s techniques to theories of truth and the ordinal analysis of mathematical systems (Beklemishev & Pakhomov, 2019; Cieśliński, 2017; Fischer et al., 2017; Franzén, 2004; Halbach, 2001; Horsten & Leigh, 2017). However, there are at least three problems with current analyses of implicit commitment for mathematical theories.

First, we lack an epistemological analysis of the process of reflection underlying ICT—although Horsten (2021) takes initial steps in this direction by analysing extensions by consistency statements. Second, although we have formal theories capturing the outcomes of endorsing ICT, such as various extensions of formal theories by reflection principles, we lack a formal analysis of the basic principles of implicit commitment itself. Third, ICT has recently come under attack. It has been argued that it cannot be true in general because some restrictive foundational standpoints are incompatible with it.

In this paper we address the problems above by introducing an axiomatization of implicit commitment for reasonable mathematical theories based on two simple principles: one states that implicit commitments are preserved under recognizable proof-transformations, the other that implicit commitments for a theory include anything that the theory internally and uniformly recognizes as axioms. We propose an epistemological analysis of the structure of implicit commitment based on such principles, according to which justified belief in a mathematical theory is transferred to its implicit commitments. Our analysis also sheds light on the recent approaches to the process of reflection in terms of epistemic entitlement (Horsten & Leigh, 2017), believability (Cieśliński, 2017) and truth (Nicolai & Piazza, 2019). The theory will also give new insights on Dean’s non-uniformity objections to ICT: the notion of epistemic stability, on which Dean’s critique is based, is called into question.

## 2 Principles for implicit commitment

We aim to articulate necessary conditions for the collection of statements one is implicitly committed to when justified in believing a mathematical theory. In what follows, when referring to a theory, we will refer to a specific formula of the language of arithmetic representing a consistent axiom set. As a consequence, axioms will be identified with their Gödel codes. It will be important that the formula representing the axiom set is sufficiently simple, in a sense to be made precise in later sections.

### 2.1 Acceptance and justified belief

A distinction between acceptance and belief has been drawn in the context of the epistemology of science and constructive empiricism (van Fraassen, 1980). Accepting a scientific theory $S$ involves belief in the empirical adequacy of the basic principles of $S$, but also a pragmatic component, for instance the possibility of using the theory as a guide for action. In the context of formal systems of relevance for the foundations of
mathematics, the empirical content of theories is negligible. By accepting the axioms of a mathematical theory $S$ of this kind, one typically believes the axioms of $S$. In this particular context, we can then follow Horsten (2021) and consider acceptance to entail belief, but not vice versa. As explained by Horsten, acceptance may not be full, just like belief. Non-full acceptance of $S$ entails a non-maximal degree of belief in $S$. And full acceptance of $S$ entails full belief in the axioms of $S$.

In this paper we will focus on the doxastic component of acceptance. Moreover, we consider only full, justified belief in a mathematical theory $S$. Foundational mathematical theories are typically believed with justification. This is certainly the case of the treatment of Primitive Recursive Arithmetic in Tait (1981), or Feferman’s justification of PA based on his conceptual structuralism (Feferman, 2010). Given our understanding of ‘mathematical theory’, justified belief in a theory is a relation between an agent and a collection of axioms presented in a concrete way. Therefore, when an agent is justified in believing a specific theory, they do not need to be aware that two presentations single out the same theory, even if they actually do. This appears to be faithful to the idea that justified belief rests on features that are dependent on the agent’s resources, and individuation conditions for mathematical objects—such as theories—do not immediately coincide with extensional identity in a strong metatheory.

2.2 The principles of invariance and axiomatic reflection

Our theory of implicit commitment is motivated by two basic principles. We call them invariance and axiomatic reflection.

The principle of invariance is based on the agent’s justified belief in certain proof procedures: if such an agent is able to establish that the consequences of two sets of sentences obtained from their chosen proof system are the same, then the two sets should be ‘equally good’ for them. Crucially, this obviously entails that the agent should entertain the same attitude towards the commitments of both theories.

For instance, consider the standard presentation of Peano Arithmetic (PA), and its alternative presentation as the union of restricted systems based on the induction schemata $\bigcup_n I\Sigma_n$. Suppose our agent is justified in believing a given proof-system for first-order classical logic as well as $\bigcup_n I\Sigma_n$, and has at their disposal a simple procedure to transform each proof from the axiom set PA to a proof in $\bigcup_n I\Sigma_n$. Then, it’s plausible to claim that the agent is implicitly committed to—and, as we shall argue, also justified in believing—the consequences of PA as well. This is the informal reasoning behind a principle of invariance that underlies our theory:

Principle of Invariance: justified belief in a theory $\tau$ (and associated proof-system) commits one to theories that are reducible to $\tau$ in a sufficiently simple way.

3 These are all examples of intrinsic justifications of foundational theories in the sense of Gödel (1947). Extrinsic justifications are also possible, although arguably less relevant to the debate on ICT.

4 More precisely, PA is given by the recursive equations for basic arithmetic operations plus the full schema of first-order induction, and $\bigcup_n I\Sigma_n$ is given by the union of the systems $I\Sigma_n$, i.e. the extension of the same recursive equations with induction scheme restricted to sentences in a $\Sigma_n$-form.
In what follows we will make precise the meaning of ‘sufficiently simple’: acceptable proof transformations will be ones in which proofs are transformed by means of elementary functions—although, as we shall argue, this can be even reduced to feasible functions.

The second principle underlying our theory concerns the reflective capabilities of the agent justifiedly believing a theory. Throughout the paper, we will assume that a textbook formalization of syntax can be developed within a weak theory of arithmetic. This can be realized as follows: one starts with primitive syntactic symbols as urelements in a weak set theory, and then construct complex syntactic objects (finite sequences) by means of standard axioms for finite set theory. Definition of syntactic functions by recursion can also be developed in a standard way. For definiteness, a set theory capable of coping with all elementary functions, and therefore that suffices for this task, can be found in Pettigrew (2009). In practice, we then work directly in a natural interpretation of this system in our preferred arithmetical system (Elementary Arithmetic EA suffices, and will play an important role in what follows). Importantly, the resources required to carry out this interpretation do not go beyond weak arithmetic, as the interpretation we are assuming can be in principle formalized in such weak arithmetical systems.5

Now suppose that our agent is justified in believing PA. As we have just described, PA can talk about its syntactic structure via arithmetization. In particular, since we treat PA as a specific elementary formula, if it is indeed the case that the code $\varphi(n)$ is one of the members of the axiom set represented by PA, then both $\text{PA} \vdash \varphi(n)$ and $\text{PA} \vdash \varphi(n)$ hold. Indeed, for a specific sentence $\varphi(n)$, the internal and external representations of PA can equally well recognize whether (or not) it is an axiom. Our second guiding principle generalizes this scenario: if one is justified in believing a given elementary presentation $\tau$ of a theory such as PA, and has decisive evidence that for every $n$, $\varphi(n)$ is an axiom of the theory, (i.e. a weak theory of syntax proves that every object $x$ is such that $\tau(\varphi(x))$), then the implicit commitments of the theory $\tau$ include $\forall x \varphi(x)$.

The example generalizes to our principle of axiomatic reflection:

**Axiomatic Reflection:** justified belief in $\tau$ (and associated proof-system) commits one to universal claims whose instances are uniformly and uncontroversially recognized as axioms of $\tau$.

The formal theory presented below will also make precise the way in which a sentence is ‘uniformly and uncontroversially’ recognized as an axiom: we will assume that such facts need to be established in a weak theory of formal syntax and therefore, by the procedure just described, in a weak arithmetical system.

We will also assume a further, basic principle for implicit commitment. It will not be part of the theory, because we deem it uncontroversial. It is a closure principle stating that implicit commitments are closed under logical consequence: if one’s implicit commitments logically entail some proposition, then this proposition is an explicit

---

5 This isn’t an unprecedented assumption. Cieśliński (2017), for instance, takes an analogous stance to argue that standard arithmetization assumptions are acceptable for the truth-theoretic deflationist.
commitment of one’s implicit commitments, and therefore part of one’s implicit commitments.

In the next section we present a formal theory of the necessary conditions for implicit commitment directly inspired by the two principles above. Given a theory $\tau$ and an associated proof-system, the implicit commitments stemming from justified belief in $\tau$ will contain the consequences of $\tau$ extended with formal counterparts of the principles of invariance and axiomatic reflection.

### 3 The formal theory

#### 3.1 Formal preliminaries and their significance

As customary in formal and philosophical studies of proof-theoretic reflection, the theories we consider are based on the arithmetical signature with $+, \cdot, 1, 0, \exp, \leq$ as primitive symbols (the intended interpretation of $\exp(x)$ is $2^x$). This language will be denoted with $L_\mathbb{N}$. The arithmetical hierarchy for formulae of $L_\mathbb{N}$ is then defined in the standard way (Beklemishev, 2005, p. 201).

Throughout the paper, we assume that our agents can reason in the weak arithmetical theory EA. Given the minimal mathematical resources it requires, we also assume that our agents have justified belief in the consequences of its axioms. The axioms of EA are

(i) Statements to the effect that $\leq$ is a linear discrete order with 0 as the least element and $x + 1$ the immediate successor of $x$,

(ii) Recursive equations for $+, \cdot, \exp$,

(iii) Induction scheme for $\Delta_0$-formulae of $L_\mathbb{N}$.

For each $n$, $I\Sigma_n$ denotes the extension of EA with induction axioms for $\Sigma_n$ formulae. Peano Arithmetic (PA) is given by EA + Ind($L_\mathbb{N}$).

EA plays a significant role in the foundations of complexity theory, since its provably total functions coincide with the so-called Kalmár Elementary functions, an important subset of primitive recursive functions. Foundational debates hardly question that primitive recursive functions are simple enough. With the exception of ultrafinitists (Nelson, 1986), the use of primitive recursive functions is justified by the light of virtually all foundational standpoints. By working with EA, we adopt an even safer environment for our study.

A crucial property of EA is that it enables us to develop in a natural way a theory of the syntax of formal systems. Following Hájek & Pudlák (1998, §1.1), a standard set-theoretic development of syntactic notions and operations can be carried out in EA—specifically via the arithmetical representation of the Ackermannian membership relation. In practice, we will work in the interpretation in EA of a direct axiomatization of finite set theory with syntactic urelemente. For definiteness, we can work in an interpretation of the system EA of Pettigrew (2009) plus finitely many urelemente within EA. We would like to emphasize that the Ackermann interpretation is so natural that it gives rise to a strong form of theoretical equivalence between EA.

---

6 As we will see shortly, EA even exceeds what is required by our formal results.
and EA*, definitional equivalence. For a given \( \mathcal{L}_{\mathbb{N}} \)-formula \( \varphi \), \( \Gamma \varphi \) denotes the canonical numeral naming its Gödel number. For an arbitrary formula \( \varphi(z) \), \( y = \Gamma \varphi(x) \) denotes the canonical formalization of the relation ‘\( y \) results from \( \varphi(z) \) by substituting the canonical numeral naming \( x \) for every free occurrence of variable \( z \)’. Given such a \( \varphi(z) \), EA proves that there is a unique \( y = \Gamma \varphi(x) \); we can then adopt the functional notation and treat \( \Gamma \varphi(x) \) as if it was a term of one free variable. We will abuse of this notation and employ the dot notation also ‘internally’ in coded environments, that is for the elementary term formalizing such a substitution operation in EA (cf. for instance (2), Proposition 1).

Throughout the whole paper we identify theories with their elementary presentations. In particular, a theory is a \( \Delta_0 \)-formula \( \tau(x) \) such that

\[
\text{EA} \vdash \forall x \left( \tau(x) \rightarrow \text{Sent}_{\mathcal{L}_{\mathbb{N}}}(x) \right),
\]

where \( \text{Sent}_{\mathcal{L}_{\mathbb{N}}}(x) \) is the definable predicate naturally expressing that \( x \) is a code of an arithmetical sentence. The choice of representing theories as elementary formulae is often made, although rarely made explicit.\(^7\) However, from the epistemological point of view, the difference between \( \Sigma_1 \) (r.e.) and elementary presentations is crucial: unlike the former, the latter come with an explicit verification procedure that allows the agent to check whether or not a given sentence is an axiom of their theory. Such a representation is fully transparent to the mathematical agent. In the terminology of later sections, when we will tackle the epistemology of reflection, whenever an agent justifiedly believes a certain elementarily presented theory, they are in a position to know that the theory they reason about (via an elementary presentation) is the theory they trust.

If \( \tau \) is a theory, then Proof\(_\tau(y, x)\) denotes the canonical elementary formula representing the relation ‘\( y \) is a proof of \( x \) from the axioms of \( \tau \)’. From the proof predicate we define the notions of provability and consistency (restricted and unrestricted) in the standard way:

\[
\begin{align*}
\text{Prov}_{\tau}(x) & :\leftrightarrow \exists y \text{ Proof}_{\tau}(y, x), \\
\text{Con}_{\tau}(x) & :\leftrightarrow \forall y < x \neg \text{Proof}_{\tau}(y, \Gamma 0 = 1^\gamma), \\
\text{Con}_{\tau} & :\leftrightarrow \forall y \neg \text{Proof}_{\tau}(y, \Gamma 0 = 1^\gamma).
\end{align*}
\]

We say that \( \tau \) is \( \Sigma_1 \)-complete, provably in EA, if for every \( \Sigma_1 \) formula \( \varphi(x) \)

\[
\text{EA} \vdash \forall x \left( \varphi(x) \rightarrow \text{Prov}_{\tau}(\Gamma \varphi(x)) \right).
\]

It can be shown that any extension of Robinson’s arithmetic \( Q \) is \( \Sigma_1 \)-complete, provably in EA.

The Principle of Invariance introduced in the previous section made reference to “simple” proof transformations. The next definition fills in the details.

\(^7\) There are exceptions: Beklemishev, in his influential survey Beklemishev (2005), defines explicitly theories as elementary formulae.
Definition 1 Suppose that $\tau$ and $\tau'$ are two theories. We say that $\tau$ is \textit{elementarily reducible} to $\tau'$, denoted $\tau \leq_{er} \tau'$, iff there exists an EA-provably total elementary function $f$ such that

$$
\text{EA} \vdash \text{Proof}_\tau(y, x) \to \text{Proof}_{\tau'}(f(y), x).
$$

The relation of elementary reducibility is a refinement of the better-known relation of proof-theoretic reducibility extensively studied by Feferman (1993).

Although the theory EA and elementary transformations play a central role in this paper, this role could be played by a much weaker first-order arithmetic, such as the theory $S^1_2$ from, Buss (1986) which does not even prove the totality of the exponential function. $S^1_2$ is to feasible (P-Time) computations as EA is to computations whose length is bounded by a Kalmár-elementary function. Specifically, if one proves in $S^1_2$ that every theorem of a P-Time theory $\tau$ is a theorem of another P-Time theory $\tau'$, then one can also infer the existence of a P-Time computable function transforming proofs in $\tau$ to proofs with the same conclusions in $\tau'$. Additionally, $S^1_2$ is relatively interpretable in Robinson’s arithmetic $Q$, which makes the theory admissible for a wider spectrum of stances in the philosophy of mathematics—such as Nelson’s ultra-finitism (Nelson, 1986). However, the non-totality of the exponential function in $S^1_2$ renders the encoding of syntax more involved and less accessible. This is why we prefer to present our framework based on EA, although nothing fundamental rests on this choice: \textit{all instances of ‘elementary’ could be replaced by ‘feasible’ throughout the paper.}

Another important tool for our purposes is proof-theoretic reflection. Proof-theoretic reflection principles are schemata that express forms of soundness of a formal theory. In particular, we will focus on Uniform Reflection. Suppose that $\tau$ is a theory. The Uniform Reflection Principle for $\tau$, denoted $RFN(\tau)$, is the following collection:

$$
\{ \forall z (\text{Prov}_\tau(\neg \varphi(z)) \to \varphi(z)) \mid \varphi(v) \in \mathcal{L}_N \}.
$$

Over EA, $RFN(\tau)$ entails $\text{Con}_\tau$, and hence its addition amounts to a proper extension of $\tau$.

Typically, Uniform Reflection for $\tau$ is presented in the form of an axiom schema. However, to be formally precise, a theory $\sigma$ is given by an axiom schema (in short, \textit{schematic}) if there is a first order formula $\varphi(P)$ with a free second order variable $P$ such that $\sigma(x)$ says:

$$
\text{There is a } \psi(y) \text{ such that } x \text{ is the result of replacing } P(y) \text{ with } \psi(y) \text{ in } \varphi.
$$

Therefore, $RFN(\tau)$ does not conform to this definition, because $\varphi(z)$ is both used and mentioned (via the Gödel code of $\varphi(z)$) in its instances.

\footnote{See Ferreira and Ferreira (2012) which surveys various other results on interpretability in $Q$ as well.}
3.2 Axioms for implicit commitment

We introduce two simple axioms corresponding to the principles of Invariance and Axiomatic Reflection introduced above. We axiomatize an operator $\mathcal{I}$ on theories, which takes a concrete axiom set and associated proof-system and returns a set of sentences, intended to be the necessary part of the implicit commitments of the theory (or more precisely, the necessary part of the implicit commitments of someone who justifiably believes $\tau$). As such, our operator $\mathcal{I}$ may not exhaust the implicit commitments of the theory.

In the definition, we fix a standard proof system for predicate logic with equality. We denote derivability in such a proof system with $\vdash$. It’s important to notice that we do not rely on a specific choice of the proof apparatus: any sound and complete proof-system for predicate logic with equality would work. For this reason we shall only apply the operator $\mathcal{I}$ to sets of sentences, and omit reference to the proof system.

**Definition 2** (Principles for implicit commitment) Let $\tau, \tau'$ be $\Delta_0$-formulae:

if $\tau' \leq_{er} \tau$, then $\mathcal{I}(\tau') \subseteq \mathcal{I}(\tau)$; \hspace{2cm} (**INVARINACE**)  

if $\text{EA} \vdash \forall x \tau(\Gamma \varphi(x) \&)$, then $\forall x \varphi(x) \in \mathcal{I}(\tau)$. \hspace{2cm} (**REFLECTION**)  

Even though we aim to characterize necessary conditions for implicit commitment, it will be useful in what follows to slightly abuse of notation and write $\mathcal{I}(\tau)$ for the minimal operator on theories satisfying invariance and reflection. Notice that reflection immediately entails that any theory $\tau$ is included in its implicit commitments, since it is allowed for quantifiers in it to be vacuous.

**Invariance** states that if $\tau'$-proofs can be elementarily transformed into $\tau$-proofs in a way that $\tau$ recognizes as correct, then the implicit commitments of $\tau$ will include all implicit commitments of $\tau'$. For instance, one might consider PA, i.e. the $\Delta_0$-presentation of Peano Arithmetic as $\text{EA} + \text{Ind}(\mathcal{L}_\mathbb{N})$, and the $\mathcal{L}_\mathbb{N}$-formula

$$\text{PA}_I(x) := \text{PA}(x) \lor x = \Gamma 0 = 0 \&.$$

$\text{PA}(x)$ and $\text{PA}_I(x)$ satisfy the premise of invariance (in both directions), even though they isolate different sets of $\mathcal{L}_\mathbb{N}$-sentences. The sentence $0 = 0$ is in fact an immediate consequence of $\text{PA}(x)$. Therefore, invariance entails that $\mathcal{I}(\text{PA}_I) = \mathcal{I}(\text{PA})$.

By contrast, if we consider the formula

$$\text{PA}_{II}(x) := \text{PA}(x) \lor (\exists y \leq x)(x = \Gamma \text{Con}_{\text{ZFC}}(\hat{y}) \&),$$

things change (here ZFC is the standard, schematic axiomatization of Zermelo–Fraenkel set theory with choice). If $\text{EA}$ trivially proves that $\text{PA}(x)$ is contained in $\text{PA}_{II}(x)$, and therefore its commitments are included in those of $\text{PA}_{II}(x)$, the same does not hold for the converse.

**Observation 1** $\text{PA}_{II}$ is not elementary reducible to $\text{PA}$.  

\[\text{Springer}\]
The proof of Observation 1 is given in the Appendix. As a consequence, one cannot use invariance to equate the implicit commitments of PA(x) and PA_{II}(x). Of course, if one replaced EA with a theory that proves Con(ZFC) in the notion of reducibility, I(PA) and I(PA_{II}) would turn our to be equivalent. However, this would be a highly controversial choice for a theory of syntax. Similarly, invariance of implicit commitments depends on the agent’s initial choice of the proof system for their accepted set of sentences. For example, the accepted proof system may contain some additional rules of reasoning: by assuming a suitably modified notion of elementary reducibility, more presentations of PA might be identified in this case. These formal properties capture the idea that justified belief is a relation to a class of theories whose individuation conditions are not purely extensional, in the sense that membership in the class is dependent on the agent’s mathematical resources (i.e. their metatheory). 9

REJECTION says that if EA can establish the syntactic claim: ‘For every object x, \( \varphi(x) \) is a member of \( \tau \)’ (where \( x \) is a name for x), then it follows from \( \tau \)’s commitments that every object satisfies \( \varphi(x) \). REJECTION shares some structure with Kleene’s rule—i.e. the rule version of Uniform Reflection—but it is indeed weaker than Uniform Reflection. In the first place, unlike standard reflection rules, it concerns the notion of ‘being an axiom’ and not of ‘being provable from some set of axioms’. In other words, if Uniform Reflection can be seen as an \( \omega \)-rule with recursively enumerable premisses, REJECTION is an \( \omega \)-rule with \textit{elementarily} recognizable premisses. Moreover, as we shall see shortly, there is a precise technical sense in which REJECTION for a theory \( U \) is not stronger than \( U \) itself.

It is important to highlight the scope of the quantifiers in REJECTION. In particular, the premise of REJECTION shouldn’t be read as

for every \( n \), EA proves that \( \varphi(n) \) is a member of \( \tau \).

With such a premise, REJECTION would amount to

\[
\text{if } \forall n \text{ EA } \vdash \tau(\overline{\varphi(n)}) \text{ then } I(\tau) \vdash \forall x \varphi(x). \tag{1}
\]

However, (1) is not a satisfactory principle because it would render \( I(\tau) \) highly dependent on one’s metatheory. For example, consider

\[
\text{PA}_{III}(x) := \text{PA}(x) \lor \exists y \left( x = \overline{\text{Con}_{ZFC}(y)} \land \text{Con}_{ZFC}(y) \right).
\]

Then we have:

**Observation 2** PA_{III} is elementarily reducible to PA.

The proof of Observation 2 is given in the Appendix. We note that if one assumes Con_{ZFC}, and that implicit commitments are closed under (1) and INVARANCE, then we would conclude that the implicit commitments of PA include Con_{ZFC}—more generally, any such notion of implicit commitment will reflect all the \( \Pi^0_2 \)-consequences of one’s metatheory. Intuitively, this is because the specific formulation of (1) allows EA to access ‘external’, metatheoretic facts (such as Con_{ZFC}) to satisfy its premise.

\[\text{We thank an anonymous referee for raising this issue.}\]
The premise of REFLECTION, by contrast, can only be satisfied by appealing to the uncontroversial resources of EA. This prevents our characterization of implicit commitment to entail strong (and unintuitive) consequences such as ConZFC. As we will see shortly, \( I(\text{PAIII}) \) will not include ConZFC.

Finally, the specific formulation of REFLECTION given above rests on the availability of names for all objects in the domain of discourse of quantifiers. This assumption is not necessary and can be relaxed. In particular, instead of identifying theories with elementary predicates, we could conceive of them as binary, elementary relations whose arguments are an elementary predicate and an object. We could then write \( \tau^*(\vec{\varphi}(v), x) \) for ‘the result of applying the elementary predicate \( \varphi \) to the object \( x \) is a member of the “theory” \( \tau^* \).’ REFLECTION would then be reformulated as:

if EA \( \vdash \forall x \; \tau^*(\vec{\varphi}(v), x) \), then \( \forall x \varphi(x) \in I^*(\tau^*) \).

All the formal properties of the theory of implicit commitment that we will present below will transfer to this more general setting with only little modification. Specifically, we require that \( \tau^* \) satisfies the following natural condition

\[ \forall x (\exists y (\text{name}(x) = y \land \tau^*(\vec{\varphi}(v)[y/v], \emptyset)) \rightarrow \tau^*(\vec{\varphi}(v), x)) \],

where \( \vec{\varphi}(v)[y/v] \) stands for the result of formally replacing \( v \) with the name \( y \) of \( x \) in the formula \( \varphi \). The above bridge principle does not require that we have names for all objects, but only clarifies the connection between an object and its name when the object can indeed be named.

### 3.3 Main properties

We now turn to the main properties of our theory of implicit commitment. We show that INVARIANCE and REFLECTION, once taken together, are strong enough to deliver the uniform reflection principle for a theory. We also show that, if taken individually, each principle does not force additional logical strength as it admits interpretations that are conservative over the underlying theory.

**Proposition 1** If \( \tau \) extends EA, then RFN(\( \tau \)) \( \subseteq I(\tau) \).

**Proof** The proof follows Feferman’s reasoning that, for a theory \( \tau \), closure of \( \tau \) under the Kleene’s rule is equivalent to the uniform reflection over \( \tau \) (Beklemishev, 2005).

Given an arbitrary \( \varphi(v) \in \mathcal{L}_\text{N} \), we shall first show that EA \( \vdash \forall x \text{Prov}_\tau(\vec{\varphi}(\vec{x})) \), where \( \theta(x) \) is the so-called small reflection principle for \( \varphi \) and \( \tau \), defined as

\[ \theta(x) := \forall y_1, y_2 (y_1 = (x)_1 \land y_2 = (x)_2 \land \text{Proof}_\tau(y_1, \vec{\varphi}(y_2)) \rightarrow \varphi(y_2)) \].

10 The projections \( y_1 \) and \( y_2 \) of the finite sequence \( x \) are needed because, whereas we have introduced REFLECTION for \( \tau \) as featuring one variable only, the claim we are interested in is proved by applying REFLECTION to a formula featuring two variables. This is why we resort to the EA-definable (total) pairing function to code up pairs of variables in one.
Working in EA, we fix an arbitrary \( x \) and let \( y_1 = (x)_1 \) and \( y_2 = (x)_2 \). If \( \text{Proof}_\tau(y_1, \neg \varphi(y_2)) \), then \( \text{Prov}_\tau(\neg \varphi(y_2)) \) and the claim follows by logical reasoning inside the provability predicate for \( \tau \). Similarly, if \( \neg \text{Proof}_\tau(y_1, \neg \varphi(y_2)) \), then provable \( \Sigma_1 \)-completeness entails that

\[
\text{Prov}_\tau(\neg \exists y_1, \varphi(y_2)).
\]

(2)

Therefore, in either case, the claim follows.

Now, for \( \theta \) as above, we define

\[
\tau'(x) := \text{EA}(x) \lor \exists y \leq x \ x = \neg \theta(y).
\]

Then the previous argument shows that \( \tau' \leq_{er} \tau \). So by INVARiance

\[
\mathcal{I}(\tau') \subseteq \mathcal{I}(\tau).
\]

(3)

However, by the definition of \( \tau' \), we obtain that \( \text{EA} \vdash \forall x \tau'(\neg \theta(x)) \). Then, by REFLECTION, we have \( \forall x \theta(x) \in \mathcal{I}(\tau') \). Hence, by the definition of \( \theta \) and the fact that \( \text{EA} \subseteq \tau' \subseteq \mathcal{I}(\tau') \), we can conclude that

\[
\mathcal{I}(\tau') \vdash \forall x \forall y \left( \text{Proof}_\tau(x, \varphi(y)) \rightarrow \varphi(y) \right),
\]

which entails the uniform reflection axiom for \( \varphi \). Therefore, by (3) and the closure of implicit commitments under derivability, we obtain that \( \text{RFN}(\tau) \subseteq \mathcal{I}(\tau) \). \( \square \)

Proposition 1 gives us necessary conditions for implicit commitment. Given a theory \( \tau \), \( \mathcal{I}(\tau) \) cannot be weaker than \( \tau + \text{RFN}(\tau) \). This claim can be made even sharper: the implicit commitments of a theory \( \tau \) afforded by INVARiance and REFLECTION alone cannot surpass what is provable in \( \tau + \text{RFN}(\tau) \). Recalling that \( \mathcal{I}(\tau) \) is assumed to be the minimal operator closed under INVARiance and REFLECTION, we have:

**Proposition 2** \( \mathcal{I}(\tau) \subseteq \tau + \text{RFN}(\tau) \).

The proof of Proposition 2 is given in the Appendix.

A core feature of our account is that it breaks down the notion of implicit commitment into two simple clauses. As we will now show, each clause is logically weak if taken in isolation, and yet it produces substantial consequences when coupled with the other. Let’s consider INVARiance first.

An operator on theories \( \mathcal{I}_1 \) satisfying INVARiance would not rule out trivial interpretations. For instance, one can let \( \mathcal{I}_1(\tau) \) to be \( \tau \) itself. Since EA is arithmetically sound, the assumption \( \tau' \leq_{er} \tau \) would immediately entail that \( \mathcal{I}_1(\tau') \subseteq \mathcal{I}_1(\tau) \).

Also REFLECTION, by itself, does not force any logical strength. Unlike a reflection principle, it involves instances of single axioms and not theorems, and it can be shown that REFLECTION is properly weaker than a reflection principle. Consider an arithmetically sound theory such as PA (the standard presentation of Peano Arithmetic). One
can define a functor \( I \) satisfying \textsc{reflection} which is nonetheless conservative over PA. It suffices to let

\[
I(\tau) = \{ \forall x \varphi \mid \text{EA } \vdash \forall x \tau(\varphi(x)) \}
\]

Then \( I(PA) \) is deductively equivalent to PA. More generally, we have:

**Proposition 3** If \( \tau \) is schematic, then \( I(\tau) \) is deductively equivalent with \( \tau \).

The definition of \( \tau \) being schematic is given at the end of Sect. 3.1. The proof of Proposition 3 can be found in the Appendix. It is worth noticing that other choices of \( \tau \) may lead to much stronger \( I(\tau) \). For instance, \( I(PA_{II}) \) will prove the consistency of ZFC.

### 4 Justified belief

Now we turn to the epistemological analysis of the agent’s implicit commitments. We argue that justified belief in \( \tau \) is preserved to each instance of the Uniform Reflection Principle for \( \tau \). This will be done by showing that elementary reducibility and \textsc{reflection} preserve justified belief. Proposition 1 will then entail that justified belief in \( \tau \) is preserved to all instances of Uniform Reflection for \( \tau \). Our framework, therefore, validates a very strong reading of the Implicit Commitment Thesis.

Philosophers and logicians have recently started to pay attention to the epistemology of proof-theoretic reflection principles (Ciesiński, 2017; Fischer, 2021; Fischer et al., 2019; Horsten, 2021; Horsten & Leigh, 2017). They focused on the difference between entitlement and justification in the context of reflection. We will employ the term ‘warrant’ to refer to both kinds of epistemic support: entitlement and justification are warrants that differ in some key aspects (Burge, 1996; Wright & Davies, 2004). Justification typically requires self-evidence, or a deductive or inductive rule acting on warranted premisses.\(^{11}\) Entitlement doesn’t (Graham, 2020). For instance, perceptual beliefs such as the one expressed by ‘that one is a sphere’ are typical examples of entitlements, whereas propositions that are obtained by combining justified premisses via logical reasoning are typical examples of justifications. In the context of mathematical foundational theories, justification via self-evidence is often referred to as intrinsic justification. Mathematical axioms can also come to be accepted by their fruitfulness; in this case one often speaks of extrinsic justification (Maddy, 2011, p. 47). We can speak more generally of foundational support as a source of justification for mathematical axioms.

Justification for a mathematical theory \( \tau \) does not immediately transfer to Uniform Reflection for \( \tau \). First, there is no deductive \textit{logical} principle that warrants it, as Uniform Reflection is unprovable in \( \tau \). Also, inductive rules have little role to play in such abstract contexts as mathematical theories. However, one may argue that, given a reliable process of formalization in the background, justified belief in \( \tau \) may

\(^{11}\) Although philosophers discussed cases of failure of justification transfer in deduction (Wright, 2003), in our mathematical case-studies we can safely assume that deduction preserves justification.
be ‘expressed’ by Uniform Reflection, and this would guarantee its self-evidence. This conclusion should be resisted. As argued in Dean (2015), one can justifiedly believe fragments of first-order arithmetic such as Primitive Recursive Arithmetic or some systems of Bounded Arithmetic (Parsons, 2007; Tait, 1981), and yet consider Uniform Reflection as unwarranted given the equivalence of Uniform Reflection and full number-theoretic induction.

Another possibility is to resort to ideological expansions of the theory, for instance by means of a truth predicate. Justified belief in \( \tau \) would entail a justified belief in a theory of truth of \( \tau \), ideally one that proves RFN(\( \tau \)) (Franzén, 2004; Ketland, 2005; Shapiro, 1998). However, this only shifts the required preservation of justified belief from the reflection principle for \( \tau \) to a theory of truth for \( \tau \). And it is far from clear that justified belief in \( \tau \) warrants a justified belief in a non-conservative theory of truth for \( \tau \). Moreover, in the context of theories of truth, the addition of Uniform Reflection is more problematic than in purely arithmetical context (Fischer et al., 2017).

Here is our first main claim: elementary reducibility preserves justified belief, in the sense that justified belief in \( \tau \) transfers to any \( \tau' \) that is elementary reducible to it. The mathematical theories \( \tau \) under consideration contain a fair amount of formalized metamathematics, in particular we stipulate that EA \( \subseteq \tau \). Hence, the proof-transformations that are required by the notion of elementary reducibility, if available, are clearly formalizable in \( \tau \). Formalization in EA preserves justification, given the uncontroversial nature of our weak syntactic metatheory. Under the assumption that justified belief in \( \tau \) entails justified belief in the logical consequences of \( \tau \), the assertion that each consequence of \( \tau' \) is a consequence of \( \tau \) amounts to a justified belief.

We now turn to the claim that \textsc{reflection} preserves justified belief, that is the claim that whenever we establish that a formula \( \varphi(x) \) uniformly belongs to \( \tau \) on the basis of a justification \( J \) for \( \tau \), then we can conclude that \( J \) counts as a justification for \( \forall x \varphi \). We argue that justification is preserved via \textsc{reflection} meta-inferentially: \textsc{reflection} for a theory \( \tau \) supervenes on the inferential structure of \( \tau \) supervenes in such a way that a given justification for the inferential apparatus of \( \tau \) suffices to warrant \textsc{reflection} for \( \tau \). Three features of \textsc{reflection} support our claim: (i) \textsc{reflection} is simple in computational terms; (ii) \textsc{reflection} mirrors \( \tau \)-provability, and nothing more; (iii) \textsc{reflection} can be conservatively interpreted in \( \tau \).

The premiss of \textsc{reflection} involves an elementarily decidable property—being a member of the set \( \tau \). This contrasts with Uniform Reflection, which involves a significantly more complex property, in particular a recursively enumerable complete property. Unlike recursively enumerable properties, elementary properties are completely transparent to any choice of \( \tau \): there is always an elementary procedure for

---

12 Although, as argued in Sect. 3.1, EA can be weakened to a system of feasible arithmetic such as S12, and ‘elementary’ can be uniformly replaced with ‘feasible’.
13 Given our assumption that EA \( \subseteq \tau \), any justification \( J \) of \( \tau \) counts as a justification of EA \( \subseteq \tau \).
14 It’s worth emphasizing that properties (i)–(iii) need not be available to the mathematical agent we are discussing. In this sense, these conditions may not be luminous. This property highlights some externalist features of the notion of mathematical justification we are discussing. We thank an anonymous referee for asking for clarifications on this point.
deciding whether or not some sentence is a member of $\tau$ or not, and therefore this procedure is always available in $\tau$ itself.

Second, the addition of REFLECTION to $\tau$ is fully grounded in $\tau$-provability and does not allow for any new proofs obtained from the combination of $\tau$ rules with REFLECTION. Since REFLECTION cannot be iterated, it only adds one layer of proofs to the pre-existing structure of $\tau$-proofs. This is in stark contrast with standard procedures for extending theories such as $\tau$ with new axioms or new rules of inference.

Finally, unlike Uniform Reflection for $\tau$, REFLECTION for $\tau$ can be conservatively interpreted in $\tau$ (cf. Proposition 3). This shows that the deductive apparatus of $\tau$ is sufficient to represent the logical structure of REFLECTION, and that there is no sharp logical gap between $\tau$ and REFLECTION alone. In particular, this means that REFLECTION can be interpreted in $\tau$ in such a way that the primitive concepts of $\tau$ are preserved. As argued in (Fischer et al., 2019, §4.3), conservativeness (semantic or proof-theoretic) is no sufficient condition for justification. However, it is its combination with the other features that makes REFLECTION special.

These three properties of REFLECTION support the following picture of justification transfer. Suppose that someone is justified in believing a theory $\tau$. That is, the agent justifiedly believes every sentence in the elementary set $\tau$: more specifically, the agent has justification for $\tau$ as expressed in $\mathcal{L}_\mathbb{N}$ by a concrete $\Delta_0$-formula. Suppose now that the agent does not have justification for the claim $\forall x \varphi$, and that $\text{EA} \subseteq \tau$ proves that every numerical instance of $\varphi(x)$ is a member of $\tau$. Given the agent’s justification for $\tau$, every instance of $\varphi$ is a member of $\tau$ (modulo formalization), and therefore it is justified. Therefore, $\langle \varphi(a) \rangle$ is a member of $\tau$ for an arbitrary $a$ (i.e. an object on which no specific assumption whatsoever has been made), and given the agent’s justification in $\tau \supseteq \text{EA}$, the agent justifiedly believes that this is indeed the case for such an arbitrary $a$. However, it is a basic assumption of our framework that formalization of elementarily decidable sets via elementary formulae is trustworthy (see introduction to Sect. 2). Therefore, for an arbitrary object $a$, if one is justified in believing that $\langle \varphi(a) \rangle$ belongs to an elementarily presentable set of axioms, she is also justified in believing $\varphi(a)$. So, given the arbitrary nature of $a$, $\forall x \varphi$ is justified on the basis of $\tau$.

Horsten (2021) discusses the phenomenology of one’s acceptance of a consistency statement given belief in a base theory such as Peano arithmetic. Horsten claims that belief in a theory warrants (via entitlement) a consistency statement. To conclude this section, we provide a similar reconstruction based on Proposition 1 and on our discussion of justification transfer.

Suppose that the agent justifiedly believes a mathematical theory $\tau$. For each $\varphi(x)$, they can define a total procedure $f_{\varphi}$ such that for arbitrary numbers $n, k$, $f_{\varphi}(n, k)$ is a $\tau$-proof of the sentence

$$\text{Proof}_\tau(n, \langle \varphi(k) \rangle) \rightarrow \varphi(k).$$

(\(n\) denotes the canonical numeral naming \(n\)). The above sentence (*) expresses: “If \(n\) is a $\tau$-proof of the sentence $\varphi(k)$, then $\varphi(k)$”. The description of $f_{\varphi}$ is given purely syntactically in $\text{EA}$; therefore, we are fully justified in believing it. Given the trustworthiness of $\text{EA}$ with respect to elementary sets, for any numbers $n, k$ the agent is justified (on the basis of $\tau$) to believe that if $n$ codes the proof of $\varphi(k)$, then $\varphi(k)$.
That’s how they come to a justified belief that for every \( k \), if there is a \( \tau \)-proof of \( \varphi(k) \), then \( \varphi(k) \). And this is precisely uniform reflection for the arbitrary formula \( \varphi(x) \).

One may object that the justification transfer between \( \tau \) and \text{REFLECTION} for \( \tau \) may fail for the following reason. The move from a claim involving syntactic content (the premiss of \text{REFLECTION}) to number-theoretic claims involves higher-order resources—i.e., an understanding of the structural similarity between syntax and numbers—that may fail to be justified on the basis of weak arithmetic, EA in particular. A view on arithmetization of this form is put forth for instance in Isaacson (1987).\(^{15}\) Given our set up, this objection may be resisted. We are assuming that a fairly natural axiomatization of the resources required for syntax theory is interpretable in EA, and when reasoning syntactically in EA we are reasoning via this interpretation. The interpretation is by no means artificial, since when combined with its inverse it gives rise to the definitional equivalence of syntax and arithmetic. Moreover, the syntax theory is finitely axiomatizable,\(^{16}\) and therefore the correctness of the interpretation can be seen in EA by the provability of a single sentence (and, in fact, the definitional equivalence just mentioned can be verified in EA). We do not require higher-order claims about similarities between the structure of syntax and numbers to understand the syntactical content of \text{REFLECTION}: we take instead a direct approach where syntax is directly (via a natural interpretation) carried out within EA.

### 5 Epistemic stability

Dean (2015) has recently proposed the notion of epistemic stability to unify different foundational standpoints by emphasizing analogies between their epistemic commitments. A formal mathematical theory is said to be \textit{epistemically stable} if ‘there exists a coherent rationale for accepting [it] which does not entail or otherwise oblige a theorist to accept statements which cannot be derived from [its] axioms’ (Dean, 2015). Epistemic stability intends to explain how the finitist, the predicativist, the first-orderist,\(^{17}\) although advocating systems that greatly differ in strength and scope, can nonetheless share a common attitude towards their preferred formal systems.

Our theory of implicit commitment sheds new light on the notion of epistemic stability. On a straightforward understanding of Dean’s definition of epistemic stability—supported also by the reconstruction of the debate in Horsten (2021)—, the agent is aware of a coherent rationale for \textit{believing in} the mathematical theory under consideration. What has been said in the previous section, however, can be adapted to show that such a coherent rationale is preserved to the consequences of \text{INVARINACE} and \text{REFLECTION} for the theory. In addition, we take to be a basic presupposition of this coherent rationale that whoever possesses it possesses also the capability of rec-

\(^{15}\) A recent application of this view can be found in Mount and Waxman (2021) in the context of the theories of truth studied in Leigh and Nicolai (2013).

\(^{16}\) It suffices to notice that, for instance, the theory \( \text{EA}^\ast \) in Pettigrew (2009) is definitionally equivalent with the finitely axiomatizable theory \( \text{EA} \), and therefore finitely axiomatizable.

\(^{17}\) We are employing here the terminology from Dean (2015): a first-orderist is someone who believes that Peano Arithmetic is sound and complete with respect to finite mathematics. This position is close to what is known as \textit{Isaacson’s Thesis}. 
ognizing whether or not a syntactic object is an axiom of the theory. Proposition 1 then tells us that all instances of Uniform Reflection for the theory can be warranted by this rationale. This seems to be strong evidence that by accepting a theory, one is bound to accept uniform reflection for the theory. As a consequence, it seems that there cannot be any epistemically stable theory after all.

For instance, let’s consider the example of a first-orderist as depicted by Isaacson (1987). According to Isaacson, PA is sound and complete with respect to finite mathematics. This is essentially because the axioms of PA are ‘first-orderizations’ of the second-order arithmetical axioms, whose truth can be directly perceivable on the basis of our grasp of the structure of natural numbers. This rationale is taken by Dean to witness the epistemic stability of PA, and it ‘stands in conflict with the version of ICT which holds that acceptance of a theory [PA] always entails commitment to principles such as RFN(PA)’ (Dean, 2015, p. 59).

Now, let’s not question whether Isaacson’s account of PA actually amounts to a coherent rationale that fits Dean’s definition of epistemic stability. For our purposes, it suffices to point out that it would not be possible to articulate a coherent rationale to accept PA without possessing a general notion of what an axiom of PA is. And this for us is sufficient to transfer acceptance of PA on the basis of this rationale to the consequences of INVARIANCE and REFLECTION for PA. And this, as shown above, include several PA-unprovable statements in LN.

We now consider some possible reactions prima facie available to the advocate of epistemic stability. One may insist, following the objection to the formulation of reflection discussed above, that there is a coherent rationale for accepting, say, PA, without accepting the resources needed to carry out the arithmetization of syntax within PA. Therefore, PA may be epistemically stable, and the agent who justifiably believes in PA may not have the resources to accept REFLECTION, and therefore they may not generate the reasoning required to conclude Uniform Reflection. We already argued that this would be certainly plausible if one justified arithmetization on the basis of the similarities between the structures of syntactic objects and numbers (or, equivalently, by means of essentially second-order resources). In our more direct approach to the development of syntax, which echoes other authors such as Ciesiński (2017), PA (or even EA) is sufficient to make sense of arithmetization as a procedure that links a direct axiomatization of syntax and its representation within the arithmetical system. It is fair to say that certain advocates of epistemically stable theories, such as Isaacson, do hold that arithmetization is a procedure exceeding the resources of the agent accepting PA. However, it’s not clear to us why such a mathematical agent wouldn’t be able to resort to our direct approach to arithmetization that licenses the adoption of INVARIANCE and REFLECTION.18

Secondly, one might hold that epistemic stability is formulated in terms of acceptance, and not of belief. On this view, accepting ρ is different from having full belief in ρ. Since our discussion targets (justified) belief in mathematical theories, it will not affect the theorist that merely accepts them. It seems clear that without a clear, workable notion of acceptance, it is difficult to even consider such a rejoinder. If one starts specifying the features of such a notion of acceptance, there are at least two

---

18 We thank the anonymous referee for raising an objection that led to this discussion.
options. The first is the direction we took, following Horsten (2021), in the discussion above: although acceptance is not belief, it entails full belief. Our analysis can then be employed, without essential modifications, to show that full belief in $\tau$ is preserved to Uniform Reflection for $\tau$. Moreover, since foundational mathematical theories are typically accepted (and therefore believed) with justification—this is clearly the case for the examples considered in Dean (2015) and Horsten (2021)—justified belief in $\tau$ is preserved to its uniform reflection principle.

Another reaction for the advocate of epistemic stability may be to resort to an understanding of ‘accepting a theory’ based on a disposition to assent to axioms of a theory (e.g., any instance of induction of PA) when presented with one. An epistemically stable theory $\tau$ would then be one for which there exists a coherent rationale for working in it which does not entail statements that cannot be derived from its axioms. Crucially, this coherent rationale would not be available to the mathematical agent engaged in this dispositional acceptance. For instance, for a finitist as depicted by Tait (1981), the characterization of primitive recursive functions as the finitist ones is out of reach. Consequently, so is the formulation of PRA as a formal system. However, the simple combinatorial intuition of a finite sequence may be sufficient to locally ground all finitist proofs (Tait, 1981, p. 529ff). In other words, one may regard PRA as sound and complete with respect to finitistic reasoning, and at the same time deny that any actual finitist believes in PRA in its full form. In fact, any justification of PRA that involves a global understanding of its syntactic structure qua formal system, by our results, would support also PRA-unprovable claims.

The example of PRA makes it clear that the local version of epistemic stability does not contradict the Implicit Commitment Thesis and it is not ruled out by our formal results and associated discussion. However, the mathematical agent associated with this form of epistemic stability would be prevented from articulating the simple, elementary procedure—a and as such expressible by a suitable $\Delta_0$-formula—of listing the infinite set of axioms of their dispositionally accepted formal system that capture it. The details of why the mathematical agent can lack such an explicit procedure may depend on their specific foundational standpoints. For our purposes, it suffices to point out that this is the only version of epistemic stability that survives our analysis, and this all comes at a cost.

### 6 Other proposals

In this section we employ our proposal to analyse other approaches to implicit commitment.

#### 6.1 Reflection as entitlement

Fischer et al. (2019) and Horsten and Leigh (2017) argue that, given one’s justified belief in $\tau$, one is entitled to the Uniform Reflection Principle for $\tau$. This is because

---

19 We again thank an anonymous referee for suggesting this formulation.

20 Or even P-time, if $S^1_2$ is chosen as base theory.
the warrant for Uniform Reflection for a reasonable \( \tau \) consists in a cognitive project meeting the following additional two conditions (Wright, 2004): (i) We have no sufficient reason to believe that Uniform Reflection for \( \tau \) does not hold; (ii) Any attempt to justify such Uniform Reflection would involve further presuppositions in turn of no more secure a prior standing.

We believe that claim (ii) is not correct. In fact, we provided an alternative, (meta-)deductive route to the justification of Uniform Reflection that is based on the principles of INVARIANCE and REFLECTION. However, our formal theory can shed light on positions that regard soundness statements for a base theory \( \tau \) as warranted by entitlements and not directly by justifications. One such view is advanced in Horsten and Leigh (2017), where \( \tau \) is taken to be a theory of disquotational truth, and Uniform Reflection is used to obtain compositional principles for truth: ‘the compositionality of truth is implicitly contained in disquotational axioms’ (Horsten & Leigh, 2017, p. 209). A more ambitious picture is given in Fischer et al. (2019). There it is argued that, given one’s justified belief in a base theory \( \tau \) in a suitable nonclassical logic, the interplay between two kinds of entitlements (one to full disquotational truth, and the other to Uniform Reflection) can yield justified belief of some new mathematical theorems.

Both approaches are based on the idea that justified belief in \( \tau \) supports an entitlement to Uniform Reflection for \( \tau \). Our formal theory of implicit commitment presented in Sect. 3.2 provides a further analysis of the structure of the entitlements involved in the reflection process. Our claim that justified belief in \( \tau \) is preserved via elementary reducibility is so basic that it should be accepted by the entitlement theorist as well. However, such a theorist may disagree with the meta-inferential picture of justification transfer via REFLECTION outlined in the previous section. If that is indeed the case, the entitlement theorist can replace their cognitive project based on Uniform Reflection with a logically weaker commitment to REFLECTION. If the entitlement theorist is willing to accept that Uniform Reflection satisfies Wright’s constraints (i) and (ii), they should certainly grant that REFLECTION does so as well. Entitlements are then to be located in the cognitive project on which REFLECTION is based: the idea that the formalized elementary notion of ‘being an axiom of \( \tau \)’ should be trustworthy and yield a correct representation of the intended domain of quantification.

### 6.2 Believability

Cieśliński (2017) analyses the notion of acceptance and provides a formal model for it. After critically evaluating some alternatives, he finally adopts the following reading of ‘agent \( S \) accepts \( \tau \)’:

\[
\text{\( S \) believes that for every theorem } \varphi \text{ of } \tau \text{ there is a normally-good-enough reason to believe that } \varphi. \quad (\text{Cieśliński, 2017, p. 251})
\]

‘Having a normally-good-enough reason’ is abbreviated as ‘being believable’,\(^{21}\) so we shall employ it as well in the context of Cieśliński’s system.

\(^{21}\) Although perhaps ‘believable’ is not the best choice, since it may be confused with the weaker ‘it’s possible to believe’. 
Cieśliński defines a formal theory of believability for a system $\tau$ (denoted $\text{Bel}(\tau)$). It contains an additional predicate $B$ for ‘believable’ and extends $\tau$ with the axioms expressing that all theorems of $\tau B$ are believable ($\text{ref}$), the set of believable sentences is closed under modus ponens ($\text{mp}$) and the $\omega$-rule of the form $\forall \varphi(v)(B(\forall x B(\varphi(x))) \rightarrow B(\forall x \varphi(x)))$. Finally, the entire system is closed under the necessitation rule, $\varphi \rightarrow B(\varphi)$. ($\text{nec}$)

The implicit commitments of $\tau$ are defined as the internal theory of $\text{Bel}(\tau)$, i.e. the set $\text{Int Bel}(\tau) = \{\varphi \in L_B \mid \text{Bel}(\tau) \vdash B(\varphi)\}$. Crucially, Cieśliński shows that $\text{Int Bel}(\tau)$ contains $\omega$-iterations of Uniform Reflection over $\tau$, and therefore a highly non-trivial set of commitments for $\tau$.

The first difference between Cieśliński’s approach and ours is that he investigates the notion of believability (as explained above), while we stick to the more familiar one of justified belief. The difference between the two is substantial: for Cieśliński the transition from ‘$\varphi$ is believable’ to ‘I believe $\varphi$’ is not immediate, for there might be some independent reasons which make us reject $\varphi$. In such a situation it might even be the case that both $\varphi$ and $\neg \varphi$ are believable. This virtue makes the notion of commitment based on believability rather weak and a conditional one: if $\varphi$ is believable, then we are committed to it unless $\neg \varphi$ is believable as well. As a consequence, we cannot conclude that ICT holds unconditionally, even with respect to arithmetical theories. Admittedly, this does not do much harm to the main claim of Cieśliński’s. However, it makes his conceptual analysis not applicable to our project of vindicating ICT.

Another objection to the adequacy of Cieśliński’s system concerns the closure under the rule $\text{nec}$. Cieśliński argues that it is justified, since ‘a proof of $\varphi$ in $\text{Bel}(\tau)$ is provided, it is simply the proof itself, which is seen as good reason to accept $\varphi$’ (Cieśliński, 2017, pp. 254–255). We have no doubt that, if $\tau$ is accepted, proofs in $\tau$ amount to good reasons to accept their conclusions. In fact, this was one of the reasons supporting INvariance. But what grounds our acceptance of conclusions of proofs in $\text{Bel}(\tau)$? This is left unjustified by Cieśliński, however, as we shall show, the main argument in Cieśliński (2018) crucially depends on that. It is fairly easy to observe that the strength of $\text{Int Bel}(\tau)$ essentially lies in the interplay between the rule $\text{nec}$ and the axioms for the believability predicate. Indeed, let $\text{Bel}(\tau) |$ be the restriction of $\text{Bel}(\tau)$ in which the rule $\text{nec}$ can be applied only once and only to consequences of $\tau B$.

**Proposition 4** Let $\tau$ be $\Sigma_1$-sound. Then $\text{Int Bel}(\tau) |$ is conservative over $\tau$.

The above observation—whose proof is in the Appendix—puts into question Cieśliński’s justification of compositionality from disquotation alone.

### 6.3 The semantic core

Nicolai and Piazza (2019) accept Dean’s counterexamples to ICT. However, it is argued that, even in epistemically stable theories $\tau$, one can isolate a set of non-trivial implicit commitments, which are called the semantic core of $\tau$. Formally, the semantic core consists of an extension of $\tau$ with a compositional theory of truth over $\tau$ and the

---

$\tau B$ denotes the trivial extension of $\tau$ in the language of $\tau$ with predicate $B$ added and no non-logical axioms for formulae of the extended language.
sentence

All sentences in \( \tau \) are true. \((*)\)

Let us call \( \text{CT}^-[\tau] \) such an extension of \( \tau \). Crucially, in many natural cases \( \text{CT}^-[\tau] \) is a conservative extension of \( \tau \). For example the theory \( \text{CT}^-[\text{PA}] \) is conservative over \( \text{PA} \) if \( \text{PA} \) is formally represented as finitely many axioms of \( Q \) and the induction scheme (Kotlarski et al., 1981). Building on this, Nicolai and Piazza argue that epistemic stability is compatible with non-trivial commitments. In other words, they offer the following Weak Implicit Commitments Thesis:

\[ \text{wict} : \text{In accepting a formal system } \tau \text{ one is also committed to additional resources not available in the starting theory } \tau \text{—i.e. } \text{CT}^-[\tau]—\text{but whose acceptance is implicit in the acceptance of } \tau. \]

In the framework of the current paper, Nicolai and Piazza (2019) define the following operator

\[ I_c(\tau) = \text{CT}^-[\tau], \]

claiming that it is a well behaved notion of implicit commitments which additionally supports \text{wict}. However, let us observe that \( I_c \) satisfies \text{reflection}. This follows trivially, since, by compositional conditions, for every arithmetical formula \( \varphi(x) \) the sentences \( \forall x T(\varphi(\dot{x})) \) and \( \forall x \varphi(x) \) are equivalent. Hence, since e.g. \( I_c(\text{PA}) \not\vDash \text{RFN}(\text{PA}) \), then, by Proposition 1, the notion must fail to satisfy \text{invariance}. This, as we argued, is a highly unwelcome consequence.

Let us give a concrete example of failure of \text{invariance} for \( I_c \). Assume that \( \text{PA} \) represents the usual axiomatization of Peano Arithmetic by induction scheme. Consider

\[ \text{PARFN} := \text{EA} + \text{RFN}(\text{EA}). \]

Kreisel showed that \( \text{PARFN} \) and \( \text{PA} \) are the same theory, and this fact can be formalized in \( \text{EA} \). The two theories are then elementarily equivalent (Beklemishev, 2005). However, \( \text{CT}^-[\text{PARFN}] \) proves that all theorems of \( \text{EA} \) are true i.e. the sentence \( \forall \varphi (\text{Prov}_{\text{EA}}(\varphi) \rightarrow T(\varphi)) \). Quite surprisingly, by a result of Ciesliński (2010), over pure compositional axioms for the truth predicate this assertion already implies that all theorems of \( \text{PA} \) are true. Therefore, \( \text{CT}^-[\text{PARFN}] \) is non-conservative over \( \text{PA} \).

In fact, arithmetically, this is a pretty strong theory: not only \( \text{CT}^-[\text{PARFN}] \) proves the uniform reflection over \( \text{PA} \), but it is also able to demonstrate each finite number of iterations of uniform reflection over \( \text{PA} \) (Ciesliński, 2017; Kotlarski, 1986; Łełyk, 2017; Smoryński, 1977).

Nicolai and Piazza (2019)’s proposal only applies to schematic theories (see Leigh (2015), Theorem 2), as defined at the end of Sect. 3.1. Admittedly, schematic theories include a substantial variety of natural axiomatizations of foundationally relevant

\[^{23} \text{For the definition of the theory } \text{CT}^- \text{ consult Nicolai and Piazza (2019).} \]
mathematical theories. However, by what we said about \textit{INVARIENCE}, we should not disregard non-schematic presentations of theories such as PA$_{RFN}$.

That being said, we believe that the semantic core captures a very natural notion of implicit commitment involving a truth predicate. Our observations can therefore be also seen as a reductio argument for the notion of epistemic stability. If one accepts truth-theoretic extensions of $\tau$ to articulate implicit commitment, the fact that even WICT leads beyond what’s provable in $\tau$ amounts to additional evidence of the problematic nature of the notion of epistemic stability.

7 Conclusion

We proposed a theory of (the necessary conditions for) implicit commitment for formal mathematical theories. The theory aims to justify the move from justified belief in a theory to its own soundness assertions, and to provide a better framework to study the epistemology involved in the process.

From the formal point of view, the theory provides an analysis of one’s minimal commitments implicit in the acceptance of a theory $\tau$ in terms of two simple axioms, \textit{INVARIENCE} and \textit{REFLECTION}. We showed that each of these principles can be conservatively interpreted in $\tau$, that they are computationally simple, and that they supervene on $\tau$-provability. However, when combined, they entail the uniform reflection for $\tau$.

The main consequence of this formal framework is that justified belief in $\tau$ can be preserved, via the principles of \textit{INVARIENCE} and \textit{REFLECTION}, to Uniform Reflection for $\tau$. The framework has consequences for other aspects of the epistemology of proof-theoretic reflection. On the one hand, the recent analysis of reflection in terms of entitlement can be refined by our framework. Moreover, our theory puts into question the notion of epistemic stability, which has been recently employed to support foundational standpoints that aim to isolate epistemologically privileged portions of the mathematical universe. Finally, our theory provides new insights on alternative proposals to implicit commitment based on epistemic and semantic notions.

Acknowledgements We would like to thank Martin Fischer, Kentaro Fujimoto, Leon Horsten, Lavinia Picollo, Johannes Stern, and Matteo Zicchetti for their feedback on the paper. Two anonymous referees provided insightful comments that improved the paper. Special thanks go to Cezary Cieśliński for detailed feedback on different versions of this work. The first author was supported by the National Science Centre, Poland (NCN), Grant No. 2019/34/A/HS1/00399.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit \url{http://creativecommons.org/licenses/by/4.0/}. 
Appendix: Proofs

Most of the proofs below make crucial use of the EA-provable $\Sigma_1$-completeness for theories $\tau$, i.e. the fact that for every $\tau$ we consider (in particular extending EA) and for every $\Sigma_1$ formula $\varphi(x)$

$$EA \vdash \forall x \left( \varphi(x) \rightarrow \text{Prov}_\tau (\ulcorner \varphi(x) \urcorner) \right).$$

Proofs for Section 3

Proof of Observation 1

First observe that for all $\tau$ and $\tau'$, EA proves that

$$\forall y \text{Prov}_\tau (\ulcorner \text{Con}_{\tau'}(\ulcorner y \urcorner) \urcorner) \rightarrow (\text{Con}_\tau \rightarrow \text{Con}_{\tau'}).$$

This holds, since by a well-known fact (Beklemishev, 2005), for a $\Sigma_1$-complete theory $\tau$, $\text{Con}_\tau$ implies the uniform reflection for $\Pi_1$-formulae over $\tau$. Finally, $\text{Con}_{\tau'}(x)$ is a $\Pi_1$-formula.

Now, given that $EA \nvdash \text{Con}_{\text{PA}} \rightarrow \text{Con}_{\text{ZFC}}$, we obtain

$$EA \nvdash \forall x \text{Prov}_{\text{PA}} (\ulcorner \text{Con}_{\text{ZFC}}(\ulcorner x \urcorner) \urcorner).$$

As a consequence PA II is not elementarily reducible to PA. \hfill \Box

Proof of Observation 2

The proof is a simple application of EA-provable $\Sigma_1$-completeness of PA. Obviously it is sufficient to show that

$$EA \vdash \forall x (\text{PA}_{III}(x) \rightarrow \text{Prov}_{\text{PA}}(x)).$$

Working in PA fix an arbitrary $x$ and assume that $\text{PA}_{III}(x)$ holds. If $\text{PA}(x)$ holds, then we are done. Otherwise it has to be the case that for some $y$, $x = \ulcorner \text{Con}_{\text{ZFC}}(y) \urcorner$ and $\text{Con}_{\text{ZFC}}(y)$ holds. In such a case, $\text{Prov}_{\text{PA}}(x)$, because $\text{Con}_{\text{ZFC}}(y)$ is a true $\Sigma_1$-sentence. This concludes our proof. \hfill \Box

Proof of Proposition 2

Let us call $I_{\text{RFN}}(\tau)$ the set of implicit commitments of $\tau$ given by (the deductive closure of) $\tau + \text{RFN}(\tau)$. We show that $I_{\text{RFN}}(\tau)$ satisfies INVARIANCE and REFLECTION. For INVARIANCE, assume that $\tau' \leq_{\tau} \tau$. It follows that $\tau + \text{RFN}(\tau) \vdash \text{RFN}(\tau')$, hence we have that $I_{\text{RFN}}(\tau) \vdash \text{RFN}(\tau')$ and consequently $I_{\text{RFN}}(\tau') \subseteq I_{\text{RFN}}(\tau)$. To verify REFLECTION, fix a formula $\varphi(x)$ and assume that $EA \vdash \forall x \left( \ulcorner \varphi(x) \urcorner \right)$. Then obviously $EA \vdash \forall x \text{Prov}_\tau (\ulcorner \varphi(x) \urcorner)$ and consequently $\tau + \text{RFN}(\tau) \vdash \forall x \text{Prov}_\tau (\ulcorner \varphi(x) \urcorner)$. However, in $\text{RFN}(\tau)$ this implies $\forall x \varphi(x)$, hence we obtain $\forall x \varphi(x) \in I_{\text{RFN}}(\tau)$. \hfill \Box

Proof of Proposition 3

In the proof we make a detour through axiomatic theories of truth. Let $I_c$ be the semantic core operator defined in Sect. 6. As we argued, by the properties of the truth predicate, whenever $EA \vdash \forall x \left( \ulcorner \varphi(x) \urcorner \right)$, then $\text{CT}^- [\tau] \vdash$
\(\forall x \varphi(x)\). Consequently, for every \(\tau\), \(\mathcal{I}_c(\tau) \vdash \mathcal{I}_{\Pi}(\tau)\). The thesis follows, since, if \(\tau\) is schematic, \(CT^-[\tau]\) is conservative over \(\tau\), by theorem of Visser and Enayat (2015) and Leigh (2015).

\(\Box\)

Proofs for Section 6

Proof of Proposition 4

Firstly, define \(Bel^-(\tau)\) to be the restriction of \(Bel(\tau)\) in which the use of NEC rule is disallowed. We observe that the use of NEC in \(Bel(\tau)\)| can be simulated in \(Bel^-(\tau)\): whenever we have a \(Bel'(\tau)\) proof

\[
(\varphi_0, \ldots, \varphi_k, B(\neg \varphi_k))
\]

ending with an application of the NEC rule, then we know (by the restriction on \(Bel(\tau)\)|), that \((\varphi_0, \ldots, \varphi_k)\) is a proof in \(\tau B\). Hence \(\tau B \vdash Prov_{\tau B}(\neg \varphi_k)\) and by the REF axiom we have \(B(\neg \varphi_k)\) in pure \(Bel^-(\tau)\).

Now it is sufficient to prove that \(Int_{Bel^-(\tau)}\) is conservative over \(\tau B\). We shall show how to interpret \(Bel^-(\tau)\) in \(\tau B\) in such a way that whenever the former theory proves \(B(\neg \varphi)\), the latter proves \(\varphi\). Let us define the translation \(*\) by recursion in \(\tau\), putting \(B(t)^* = Prov_{\tau B}(t)\) and letting \(*\) be identity on the arithmetical formulae and commute with connectives and quantifiers (hence \(*\) does not relativize quantifiers).

Now, we check that if \(\varphi\) is an axiom of \(Bel^-(\tau)\), then \(\tau \vdash \varphi^*\). For REF and MP this is obvious. \((\omega R)^*\) is equal to

\[
\forall \varphi (Prov_{\tau B}(\neg \forall x B(\varphi(x))) \rightarrow Prov_{\tau B}(\neg \forall x \varphi(x))).
\]

Reasoning in \(\tau\), we see that for every \(\varphi\) the antecedent of the implication is false, since it is consistent with pure logic that the extension of \(B\) is empty. Hence \((\omega R)^*\) is provable in \(\tau B\) as well. This shows that if \(\varphi \in Int_{Bel^-(\tau)}\), then \(\tau B \vdash Prov_{\tau B}(\neg \varphi)\). Since \(\tau B\) is \(\Sigma_1\)-sound, the latter condition is equivalent to the provability of \(\varphi\) in \(\tau B\). Finally, \(\tau B\) is a trivial extension of \(\tau\), so for any arithmetical sentence \(\varphi, \varphi\) is provable in \(\tau B\) if and only if it is provable in \(\tau\).

\(\Box\)

References

Beklemishev, L. (2005). Reflection principles and provability algebras in formal arithmetic. Russian Mathematical Surveys, 60(2), 197–268.

Beklemishev, L. D., & Pakhomov, F. N. (2019). Reflection algebras and conservation results for theories of iterated truth. Annals of Pure and Applied Logic, 173(5), 103093

Burge, T. (1996). Our entitlement to self-knowledge. Proceedings of the Aristotelian Society, 96(1), 91–116.

Buss, S. (1986). Bounded arithmetic. Bibliopolis.

Cieśliński, C. (2010). Truth, conservativeness, and provability. Mind, 119(474), 409–422.

Cieśliński, C. (2017). The epistemic lightness of truth: Deflationism and its logic. Cambridge University Press.

Cieśliński, C. (2018). Minimalism and the generalisation problem: On Horwich’s second solution. Synthese, 195(3), 1077–1101.

Dean, W. (2015). Arithmetical reflection and the provability of soundness. Philosophia Mathematica, 23(1), 31–64.
Feferman, S. (1962). Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic*, 27(3), 259–316.

Feferman, S. (1993). What rests on what? The proof-theoretic analysis of mathematics. In J. Czermak (Ed.), *Philosophy of Mathematics* (pp. 1–147). Hölder-Pichler-Tempsky.

Feferman, S. (2010). Conceptual structuralism and the continuum. *Slides for the Talk PhilMath Intersem, Université de Paris*, 7(6), 08.

Ferreira, F., & Ferreira, G. (2012). Interpretability in Robinson’s Q. *Bulletin of Symbolic Logic*, 19(3), 289–317.

Fischer, M. (2021). Another look at reflection. *Erkenntnis*. https://doi.org/10.1007/s10670-020-00363-9

Fischer, M., Horsten, L., & Nicolai, C. (2019). Hypatia’s silence. *Noûs*, 55(1), 62–85. Online first.

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

Franzén, T. (2004). *Inexhaustibility. A non-exhaustive treatment*. Lecture Notes in Logic. Association for Symbolic Logic (Vol. 16). A K Peters, Ltd.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik*, 38(1), 173–198.

Gödel, K. (1947). What is Cantor’s continuum problem? *The American Mathematical Monthly*, 54(9), 515–525.

Graham, P. (2020). What is epistemic entitlement? Reliable competence, reasons, inference, access. In J. Greco & C. Kelp (Eds.), *Virtue-theoretic epistemology: New methods and approaches*. Cambridge University Press.

Hájek, P., & Pudlák, P. (1998). *Metamathematics of first-order arithmetic*. Springer.

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

Horsten, L. (2021). On reflection. *The Philosophical Quarterly*, 71, pqa083.

Horsten, L., & Leigh, G. E. (2017). Truth is simple. *Mind*, 126(501), 195–232.

Isaacs, D. (1987). Arithmetical truth and hidden higher-order concepts. In *Logic Colloquium ’85. Studies in logic and the foundations of mathematics* (Vol. 122, pp. 147–169). Elsevier.

Ketland, J. (2005). Deflationism and the Gödel phenomena: Reply to Tennant. *Mind*, 114(453), 75–88.

Kotlarski, H. (1986). Bounded induction and satisfaction classes. *Mathematical Logic Quarterly*, 32, 531–534.

Kotlarski, H., Krajewski, S., & Lachlan, A. H. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin*, 24(1), 283–93.

Krämer, S. (2014). Implicit commitment in theory choice. *Synthese*, 191(10), 2147–2165.

Leigh, G. (2015). Conservativity for theories of compositional truth via cut elimination. *The Journal of Symbolic Logic*, 80, 845–865.

Leigh, G. E., & Nicolai, C. (2013). Axiomatic truth, syntax and metatheoretic reasoning. *Review of Symbolic Logic*, 6(4), 613–636.

Łełyk, M. (2017). Axiomatic truth theories, bounded induction and reflection principles. https://depotuw. ceon.pl/handle/item/2266.

Maddy, P. (2011). *Defending the axioms: On the philosophical foundations of set theory*. Oxford University Press.

Mount, B. M., & Waxman, D. (2021). Stable and unstable theories of truth and syntax. *Mind*, 130(518), 439–473.

Nelson, E. (1986). *Predicative arithmetic*. Princeton University Press.

Nicolai, C., & Piazza, M. (2019). The implicit commitment of arithmetical theories and its semantic core. *Erkenntnis*, 84(4), 913–937.

Parsons, C. (2007). *Mathematical thought and its objects*. Cambridge University Press.

Peacock, H. (2011). Two kinds of ontological commitment. *Philosophical Quarterly*, 61(242), 79–104.

Pettigrew, R. (2009). On interpretations of bounded arithmetic and bounded set theory. *Notre Dame Journal of Formal Logic*, 50(2), 141–151.

Shapiro, S. (1998). Proof and truth. *Journal of Philosophy*, 95(10), 493–521.

Smyrliski, C. (1977). o-consistency and reflection. In *Colloque International de Logique (Clermont-Ferrand, 1975) Colloq. Internat. CNRS* (Vol. 249, pp. 167–181). CNRS.

Tait, W. W. (1981). Finitism. *Journal of Philosophy*, 78(9), 524–546.

Turing, A. M. (1939). Systems of logic based on ordinals. *Proceedings of the London Mathematical Society* (2), 45(3), 161–228.

van Fraassen, B. (1980). *The scientific image*. Oxford University Press.
Visser, A., & Enayat, A. (2015). New constructions of satisfaction classes. In K. Fujimoto, J. M. Fernández, H. Galinon, & T. Achourioti (Eds.), *Unifying the philosophy of truth*. Springer Verlag.

Wright, C. (2003). Some reflections on the acquisition of warrant by inference. In S. Nuccetelli (Ed.), *New essays on semantic externalism and self-knowledge* (pp. 57–78). MIT Press.

Wright, C. (2004). Intuition, entitlement and the epistemology of logical laws. *Dialectica*, 58(1), 155–175.

Wright, C., & Davies, M. (2004). On epistemic entitlement. *Aristotelian Society Supplementary*, 78, 167–245.

**Publisher’s Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.