A THEORY OF IMPLICIT COMMITMENT FOR MATHEMATICAL THEORIES

MATEUSZ ŁEŁYK AND CARLO NICOLAI

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 so far 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 the reason why this is so. We argue that the theory has interesting 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 proposals for the analysis of implicit commitment based on truth or epistemic notions.

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 idealised, 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 *explicitly committed* to some sentence φ 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 φ if and only if there is a proof of φ 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 she will not be explicitly committed to the the various articulations of the soundness of S in the language $\mathcal{L}_S$, including 'S is consistent', or 'all theorems of S are true'. Several authors have proposed

[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.

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 (Turing, 1939; Feferman, 1962; Franzén, 2004; Cieśliński, 2017).[2] Underlying such projects there is what is has been referred to as the *implicit commitment thesis* by Walter Dean (2015):

> 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. Since there is no agreement on what are the basic properties of the notion of acceptance of a mathematical theory (see §2.1 for more details), in this paper we follow Fischer et al. (ming) and analyse the initial acceptance of S in terms of the more familiar notions of belief and justification.[3] Our task will then be twofold. Starting with justified belief in S, we will investigate what are the commitments implicit in this justified belief. This will be done axiomatically. 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 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.

The importance of ICT for logic and 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

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

> ...to obtain a more complete one by the adjunction as axioms of formulae, seen intuitively to be correct, but which the Gödel's theorem 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.

[3] To our knowledge, a clear distinction between acceptance and belief has only been drawn in the context of the epistemology of science and constructive empiricism in particular (van Fraassen, 1980).

techniques to theories of truth, the ordinal analysis of mathematical systems (Halbach, 2001; Franzén, 2004; Horsten and Leigh, 2017; Cieśliński, 2017; Fischer et al., 2017; Beklemishev and Pakhomov, 2019). However, there are are at least three problems with current analysis of implicit commitment for mathematical theories.

First, we lack an epistemological analysis of the process of reflection underlying ICT.[4] 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 iself*. 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 inherited to its implicit commitments. Our analysis also sheds light on the recent approach to the process of reflection in terms of epistemic entitlement (Horsten and Leigh, 2017). The theory will also give a definite verdict concerning Dean's non-uniformity objections to ICT: the notion of *epistemic stability*, on which Dean's critique is based, turns out to be much less appealing.

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 formal presentation of a (elementary) set of sentences of $\mathcal{L}_{\mathbb{N}}$ in the form of a specific axiomatization. For simplicity we will fix a proof-system for classical logic that will be common to all theories considered. Thus, we shall identify theories with Δ_0-formulae with one free variable that, provably in EA, define a set of sentences. From a technical point of view, such an approach is a common practice when talking about arithmetical reflection principles over a theory (Beklemishev, 2005).

[4] This witnessed e.g. by the following quote from Horsten (2018):

> What is still lacking, and what the subject sorely needs, is a careful phenomenological analysis of the process of reflecting on one's implicit commitments.

Crucially, it accords with our analysis of 'accepting a theory' in terms of justified belief. Primarily, justified belief in a theory is a relation between an agent and a collection of axioms presented in a concrete way. This also entails that, when an agent is justified in believing a specific theory, she does not need to be aware that two presentations single out the same theory, even if actually they do.

2.1. **Acceptance as Justified Belief.** Several recent works on implicit commitment involve the primitive notion of *acceptance* of a mathematical theory. There is no satisfactory epistemological analysis of acceptance. As we shall see later on is some more details, Cieśliński (2017) analyses the notion of acceptance of a theory in terms of the more amenable notion of belief: in particular, he articulates a theory of believability for the mathematical theory in question. Horsten and Leigh (2017) and Fischer et al. (ming) choose instead to analyse acceptance in terms of *justified* belief. In this paper we opt for the latter analysis. In fact, the relevant cases of application of ICT that we are interested in – including Dean's proposal of epistemically stable theories – are cases of *foundational* theories, which are rarely believed without justifications supporting them. Thus, in such cases, an agent has a justified belief in a mathematical theory, and our guiding questions are what the agent is implicitly commitment to, and whether such justified belief supports soundness assertions for the theory.

Finally, as it happens in these recent works on implicit commitment, we will set aside worries of logical omniscience. We will be exclusively concerned with highly idealized agents relating to abstract mathematical entities. In such contexts, justified belief in a set of axiom and a proof system can be taken to entail justify belief in their logical consequences. In particular, we will be only concerned with *classical* logical consequence.

2.2. **The principles of invariance and axiomatic reflection.** Our theory can be informally motivated by means of two basic principles. We call then principles of *invariance* and *axiomatic reflection*. When an agent is justified in believing a theory, she justifiedly believes a certain set of sentences and a body of inference rules. Our two principles for implicit commitments stem from these two sources.

The first one, the principle of invariance, derives from the agent's justified belief in certain inference rules (that is, the agent's ability to infer, from her justified belief in the premisses of any such a rule, her justified belief in the conclusion). Leaving aside issues of how this justified belief can arise, such an agent must treat sets of sentences, which are demonstrably equivalent, as being on a par. More precisely: if such an agent is able to establish that the consequences of two sets of sentences obtained from her inference rules are the same, then the two sets should be 'equally good' for her. Crucially, this obviously entails

that she 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 induction schemata ($\bigcup_n \mathrm{I}\Sigma_n$).[5] Suppose our agent is justified in believing a given proof-system for first-order classical logic as well as PA, and has at her disposal a simple procedure to transform each proof from the axiom set $\bigcup_n \mathrm{I}\Sigma_n$ to a proof in PA. Then, it's plausible to claim that she is implicitly committed to – and, as we shall argue, also justified in believing – the consequences of $\bigcup_n \mathrm{I}\Sigma_n$ as well. This is the informal reasoning behind a *principle of invariance* that underlies our theory:

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

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.

The second principle underlying our theory concerns the reflective capabilities of the agent justifiedly believing a theory. Suppose that our agent is justified in believing PA. Our such as PA can talk about their syntactic structure via arithmetization. In particular, since PA is a specific elementary formula, if it is indeed the case that the code of a formula $\varphi(n)$ is one of the axioms of PA, then both $\mathrm{PA} \vdash \mathrm{PA}(\ulcorner \varphi(\overline{n}) \urcorner)$ and $\mathrm{PA} \vdash \varphi(\overline{n})$ hold. Indeed, for a specific sentence $\varphi(\overline{n})$, 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 theory such as PA – or a Δ_0-formula τ more generally – , and has a 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 $\tau(\ulcorner \varphi(\dot{x}) \urcorner)$), then the implicit commitments of the theory τ include $\forall x \varphi(x)$.

The example generalizes to our principle of *axiomatic reflection*:

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

[5] More precisely, PA is given by the recursive equations for basic arithmetic operaions plus the full schema of first-order induction, and $\bigcup_n \mathrm{I}\Sigma_n$ is given by the union of the systems $\mathrm{I}\Sigma_n$, i.e. the extension of the same recursive equations with induction scheme restricted to sentences in a Σ_n-form.

The formal theory presented below will also make precise the way in which a sentence is 'uniformly and uncontroversially' recognized as an axioms: we will assume that such facts need to be established in a weak theory of formal syntax.

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 commitment of one's implicit commitments, and therefore part of her 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 τ and an associated proof-system, the implicit commitments stemming from justified belief in τ will contain the consequence of τ extended with formal counterparts of the principles of invariance and axiomatic reflection.

3. The Formal Theory

3.1. **Theories and Coding.** All 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 $\mathscr{L}_{\mathbb{N}}$. The arithmetical hierarchy for formulae of $\mathscr{L}_{\mathbb{N}}$ is then defined in the standard way (Beklemishev, 2005, p. 201).

The weakest theory that we shall consider is *Elementary Arithmetic* (EA), whose axioms 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 Δ_0-formulae of $\mathscr{L}_{\mathbb{N}}$.

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

EA enables us to develop in a natural way a theory of syntax for our formal language and theories. As in (Hájek and Pudlák, 1998, §1.1), we represent in arithmetical context, by means of the Ackermaniann membership relation, a standard set-theoretic development of syntactic notions and operations. For a given $\mathscr{L}_{\mathbb{N}}$-formula φ, $\ulcorner\varphi\urcorner$ denotes the canonical numeral naming its Gödelnumber. For an arbitrary formula $\varphi(z)$, $y = \ulcorner\varphi(\dot{x})\urcorner$ 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'. Once again, provably in EA, $y = \ulcorner\varphi(\dot{x})\urcorner$ is a provably total function of x and we shall allow ourselves to

use the functional notation, using $\ulcorner \varphi(\dot{x}) \urcorner$ as if it was a term of one free variable. Moreover, 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 Δ_0-formulae that, provably in EA, define sets of sentences. In other words, a Δ_0-formula $\tau(x)$ is a theory if and only if

$$\mathrm{EA} \vdash \forall x \big(\tau(x) \to \mathrm{Sent}_{\mathcal{L}_{\mathbb{N}}}(x)\big),$$

where $\mathrm{Sent}_{\mathcal{L}_{\mathbb{N}}}(x)$ is the definable predicate naturally expressing that x is a code of an arithmetical sentence. For τ, τ' theories, when working in the metatheory we will employ the following abbreviations:

$$x \in \tau :\leftrightarrow \tau(x);$$
$$\tau \subseteq \tau' :\leftrightarrow \forall x (x \in \tau \to x \in \tau').$$

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

$$\mathrm{Prov}_\tau(x) :\leftrightarrow \exists y\, \mathrm{Proof}_\tau(y, x),$$
$$\mathrm{Con}_\tau(x) :\leftrightarrow \forall y < x\, \neg \mathrm{Proof}_\tau(y, \ulcorner 0 = 1 \urcorner),$$
$$\mathrm{Con}_\tau :\leftrightarrow \forall y\, \neg \mathrm{Proof}_\tau(y, \ulcorner 0 = 1 \urcorner).$$

We say that τ is Σ_1-*complete, provably in* EA, if for every Σ_1 formula $\varphi(x)$

$$\mathrm{EA} \vdash \forall x \big(\varphi(x) \to \mathrm{Prov}_\tau(\ulcorner \varphi(\dot{x}) \urcorner)\big).$$

It can be shown that any extension of Robinson's arithmetic Q is Σ_1-complete provably in EA.

The *Principle of Invariance* introduced in the previous section made use of 'simple' proof transformations. The next definition fills in the details.

Definition 1. *Suppose that τ and τ' are two theories. We say that τ is* elementarily reducible *to τ', denoted $\tau \leq_{er} \tau'$, iff there exists an* EA*-provably total elementary function f such that*

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

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

Another important tool for our purposes is proof-theoretic reflection.[6] 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 τ is a theory. The *Uniform Reflection Principle* over τ, denoted $\mathrm{RFN}(\tau)$, is the following collection of $\mathcal{L}_{\mathbb{N}}$ sentences

$$\forall z\big(\mathrm{Prov}_{\tau}(\ulcorner \varphi(\dot{z}) \urcorner) \rightarrow \varphi(z)\big)$$

for all $\varphi(v) \in \mathcal{L}_{\mathbb{N}}$. Over EA, $\mathrm{RFN}(\tau)$ entails Con_{τ}, and hence its addition amounts to a proper extension of τ.

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 part of the implicit commitments of the theory (or more precisely, the implicit commitments of someone who justifiedly believes τ).[7] In the definition, we fix a 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 τ, τ' be Δ_0-formulae.*

(INVARIANCE) *if* $\tau' \leq_{er} \tau$, *then* $\mathcal{I}(\tau') \subseteq \mathcal{I}(\tau)$;

(REFLECTION) *if* $\mathrm{EA} \vdash \forall x\, \tau(\ulcorner \varphi(\dot{x}) \urcorner)$, *then* $\forall x\, \varphi(x) \in \mathcal{I}(\tau)$.

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 τ is included in its implicit commitments, since it is allowed for quantifiers in it to be vacuous.

INVARIANCE states that if τ'-proofs can be elementarily transformed into τ-proofs in a way that τ recognizes as correct, then the implicit commitments of τ will include all implicit commitments of τ'. For instance, one might consider PA, i.e. the Δ_0-presentation of Peano Arithmetic as $\mathrm{EA} + \mathrm{Ind}(\mathcal{L}_{\mathbb{N}})$, and the $\mathcal{L}_{\mathbb{N}}$-formula

$$\mathrm{PA}_{\mathrm{I}}(x) :\leftrightarrow \mathrm{PA}(x) \vee x = \ulcorner 0 = 0 \urcorner.$$

[6] For a thorough survey of reflection principles, see again Beklemishev (2005).

[7] Crucially, since we are aiming to capture necessary conditions for implicit commitment, our operator $\mathcal{I}$ may not *exhaust* the implicit commitments of the theory. In fact, at least in the non-iterated form, our operator will certainly not-exhaust the commiments of a theory even in the most conservative interpretation of our picture.

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

By contrast, if we consider the formula

$$\mathrm{PA_{II}}(x) :\leftrightarrow \mathrm{PA}(x) \vee (\exists y \leq x)(x = \ulcorner \mathrm{Con_{Z_2}}(\dot{y}) \urcorner),$$

things change. If EA trivially proves that $\mathrm{PA}(x)$ is contained in $\mathrm{PA_{II}}$, and therefore its commitments are included in those of $\mathrm{PA_{II}}$, the same does not hold for the converse. $\mathrm{PA_{II}}$ is not elementary reducible to PA, and so INVARIANCE does not apply.[8]

REFLECTION says that if our theory of formal syntax EA can establish the sentence: 'For every object x, $\varphi(\underline{x})$ is a member of τ' (where $\underline{x}$ is a name for x), then it follows from τ's commitments that every object satisfies $\varphi(x)$. REFLECTION 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'.[9] Moreover, as we shall see shorty, there is a precise technical sense in which REFLECTION for a theory U is not stronger than U itself.

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

$$\text{for every } n, \text{ EA proves that } \varphi(\underline{n}) \text{ is a member of } \tau.$$

With such a premise, REFLECTION would amount to

$$\text{if } \forall n\, \mathrm{EA} \vdash \tau(\ulcorner \varphi(\dot{n}) \urcorner) \text{ then } \mathcal{I}(\tau) \vdash \forall x \varphi(x). \tag{1}$$

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

$$\mathrm{PA_{III}}(x) := \mathrm{PA}(x) \vee \exists y \left(x = \ulcorner \mathrm{Con_{ZFC}}(\dot{y}) \urcorner \wedge \mathrm{Con_{ZFC}}(y)\right).$$

[8] To see that it is so, observe that for all τ and τ', which provably in EA are Σ_1-complete, EA proves that

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

This holds, since by a well-known fact (see (Beklemishev, 2005)), for a Σ_1-complete theory τ, Con_τ is equivalent to uniform reflection for Π_1-formulae over τ. And if we have the uniform Π_1 reflection over τ, then $\mathrm{Con}_{\tau'}$ follows immediately from $\forall y\, \mathrm{Prov}_\tau(\ulcorner \mathrm{Con}_{\tau'}(\dot{y}) \urcorner)$.
Of course, if one replaced EA with a theory that proves $\mathrm{Con}(\mathrm{Z_2})$ in the notion of reducibility, $\mathcal{I}(\mathrm{PA})$ and $\mathcal{I}(\mathrm{PA_{II}})$ would turn our to be equivalent. However, this would be a highly controversial choice.

[9] One could succinctly paraphrase the situation by saying that, if Uniform Reflection can be seen as an ω-rule with recursively enumerable premisses, REFLECTION is an ω-rule with *elementarily* recognizable premisses.

$\mathrm{PA}_{\mathrm{III}}$ is elementary reducible to PA.[10] However, if one assumes $\mathrm{Con}_{\mathrm{ZFC}}$, and that implicit commitments are closed under (1) and INVARIANCE, then we would conclude that the implicit commitments of PA include $\mathrm{Con}_{\mathrm{ZFC}}$.[11] Intuitively, this is because the specific formulation of (1) allow EA to access 'external', metatheoretic facts (such as $\mathrm{Con}_{\mathrm{ZFC}}$) to satisfy its premise. 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 $\mathrm{Con}_{\mathrm{ZFC}}$. As we will see shortly, $\mathcal{I}(\mathrm{PA}_{\mathrm{III}})$ will *not* entail $\mathrm{Con}_{\mathrm{ZFC}}$.

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 objects. We could then write $\tau^*(\ulcorner\varphi(v)\urcorner, x)$ for 'the result of applying the elementary predicate φ to the object x is a member of the "theory" τ^*'. REFLECTION would then be reformulated as:

$$\text{if EA} \vdash \forall x\, \tau^*(\ulcorner\varphi(v)\urcorner, x), \text{ then } \forall x \varphi(x) \in \mathcal{I}^*(\tau^*).$$

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 to the formulation of the theory. Specifically, we require that τ^* satisfies the following natural condition

$$\forall x\big(\exists y(\mathrm{name}(x) = y \wedge \tau^*(\ulcorner\varphi(v)\urcorner[y/v], \emptyset)) \rightarrow \tau^*(\ulcorner\varphi(v)\urcorner, x)\big),$$

where $\ulcorner\varphi(v)\urcorner[y/v]$ stands for result formally replacing v with the name y of x in the formula φ. 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 shall 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 one's theory of choice. Moreover, we also show that, once taken in isolation, each principle does not force logical strength as it admits interpretations that are conservative over the underlying theory.

PROPOSITION 1. *If* τ *extends* EA, *then* $\mathrm{RFN}(\tau) \subseteq \mathcal{I}(\tau)$.

[10] Notice that to establish this elementary reducibility we only appeal to provable Σ_1-completeness, and we do not rely on what's true in the standard model of PA.

[11] More generally, any such notion of implicit commitment will reflect all the Π^0_2-consequences of one's metatheory.

Proof. The proof is essentially Feferman's reasoning that, for a theory τ, closure of τ under the Kleene's rule is equivalent to the uniform reflection over τ (Beklemishev, 2005).

Given an arbitrary $\varphi(v) \in \mathcal{L}_{\mathbb{N}}$, we shall first show that $\mathrm{EA} \vdash \forall x \mathrm{Prov}_\tau\big(\ulcorner \theta(\dot{x}) \urcorner\big)$, where $\theta(x)$ is the so-called small reflection principle for φ and τ, defined as

$$\theta(x) := \forall y_1, y_2 \big(y_1 = (x)_1 \wedge y_2 = (x)_2 \wedge \mathrm{Proof}_\tau(y_1, \ulcorner \varphi(\dot{y_2}) \urcorner) \rightarrow \varphi(y_2)\big).^{12}$$

Working in EA, we fix an arbitrary x and let $y_1 = (x)_1$ and $y_2 = (x)_2$. If $\mathrm{Proof}_\tau(y_1, \ulcorner \varphi(\dot{y_2}) \urcorner)$, then $\mathrm{Prov}_\tau(\ulcorner \varphi(\dot{y_2}) \urcorner)$ and the claim follows by logical reasoning inside the provability predicate for τ. Similarly, if $\neg \mathrm{Proof}_\tau(y_1, \ulcorner \varphi(\dot{y_2}) \urcorner)$, then provable Σ_1-completeness entails that

$$(2) \qquad \mathrm{Prov}_\tau(\ulcorner \exists y_1, y_2 \big(y_1 = (\dot{x})_1 \wedge y_2 = (\dot{x})_2 \wedge \neg \mathrm{Proof}(y_1, \ulcorner \varphi(\dot{y_2}) \urcorner)\urcorner).$$

Therefore, in either case, the claim follows.

Now, for θ as above, we define

$$\tau'(x) := \mathrm{EA}(x) \vee \exists y \leq x \;\; x = \ulcorner \theta(\dot{y}) \urcorner.$$

Then the previous argument shows that $\tau' \leq_{er} \tau$. So by INVARIANCE

$$(3) \qquad \mathcal{J}(\tau') \subseteq \mathcal{J}(\tau).$$

However, by the definition of τ', we obtain that $\mathrm{EA} \vdash \forall x\, \tau'(\ulcorner \theta(\dot{x}) \urcorner)$. Then, by REFLECTION, we have $\forall x \theta(x) \in \mathcal{J}(\tau')$. Hence, by the definition of θ and the fact that $\mathrm{EA} \subseteq \tau' \subseteq \mathcal{J}(\tau')$, we can conclude that

$$\mathcal{J}(\tau') \vdash \forall x \forall y \, \big(\mathrm{Proof}_\tau(\dot{x}, \ulcorner \varphi(y) \urcorner) \rightarrow \varphi(y)\big),$$

which entails the uniform reflection axiom for φ. Therefore, by (3) and the closure of implicit commitments under derivability, we obtain that $\mathrm{RFN}(\tau) \subseteq \mathcal{J}(\tau)$. □

Proposition 1 gives us necessary conditions for implicit commitment. Given a theory τ, $\mathcal{J}(\tau)$ cannot be weaker than $\tau + \mathrm{RFN}(\tau)$. This claim can be made even sharper: the implicit commitments of a theory τ afforded by INVARIANCE and REFLECTION alone cannot surpass what is provable in $\tau + \mathrm{RFN}(\tau)$.[13]

[12] The parameters y_1 and y_2 are needed because, whereas we have introduced REFLECTION for τ 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.

[13] Let us call $\mathcal{J}_{\mathrm{RFN}}(\tau)$ the set of implicit commitments of τ given by (the deductive closure of) $\tau + \mathrm{RFN}(\tau)$. We show that $\mathcal{J}_{\mathrm{RFN}}(\tau)$ satisfies INVARIANCE and REFLECTION. For INVARIANCE, assume that $\tau' \leq_{er} \tau$. Then, since $\mathcal{J}_{\mathrm{RFN}}(\tau) \vdash \mathrm{RFN}(\tau')$, we have that $\mathcal{J}_{\mathrm{RFN}}(\tau) \subseteq \mathcal{J}_{\mathrm{RFN}}(\tau')$. To verify REFLECTION, fix a formula $\varphi(x)$ and assume that $\tau \vdash \forall x \tau(\ulcorner \varphi(\dot{x}) \urcorner)$. Then obviously $\tau \vdash \forall x \mathrm{Prov}_\tau(\ulcorner \varphi(\dot{x}) \urcorner)$ and consequently $\forall x \mathrm{Prov}_\tau(\ulcorner \varphi(\dot{x}) \urcorner) \in \mathcal{J}_{\mathrm{RFN}}(\tau)$. Now by $\mathrm{RFN}(\tau)$ we immediately obtain $\forall x \varphi(x) \in \mathcal{J}_{\mathrm{RFN}}(\tau)$.

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 $\mathcal{I}_{\mathrm{I}}$ satisfying INVARIANCE would not rule out trivial interpretations. For instance, one can let $\mathcal{I}_{\mathrm{I}}(\tau)$ to be τ itself. Since EA is arithmetically sound, the assumption $\tau' \leq_{er} \tau$ would immediately entail that $\mathcal{I}_{\mathrm{I}}(\tau') \subseteq \mathcal{I}_{\mathrm{I}}(\tau)$.

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 $\mathcal{I}_{\mathrm{II}}(\cdot)$ satisfying REFLECTION which is nonetheless conservative over PA. It suffices to let

$$\mathcal{I}_{\mathrm{II}}(\tau) = \{\forall x \varphi \mid \mathrm{EA} \vdash \forall x\, \tau(\ulcorner \varphi(\dot{x}) \urcorner)\}$$

Then $\mathcal{I}_{\mathrm{II}}(\mathrm{PA})$ is deductively equivalent to PA.[14]

4. Justified Belief

In §2.2 we started with the agent's justified belief in τ, and introduced informal principles to characterize (part of) the agent's commitments. Such principles were then made formally precise. Now we turn to the epistemological analysis of the agent's implicit commitments so characterized. If an agent is justified in believing a theory τ, what's her epistemic attitude towards sentences in $\mathcal{I}(\tau)$? We will argue that INVARIANCE and REFLECTION preserve justified belief in a sense that will be made precise shortly. Proposition 1 will then entail that justified belief in τ is preserved to all instances of Uniform Reflection for τ. Our framework, therefore, strengthens the Implicit Commitment Thesis in a significant way.

Let us start with INVARIANCE. It is uncontroversial that elementary reducibility preserves justified belief, in the sense that justified belief in τ transfers to any τ' that is elementary reducible to it. The mathematical theories τ under consideration contain a fair amount of formalized metamathematics, in particular we stipulate that $\mathrm{EA} \subseteq \tau$. Under the assumption that justified belief in τ entails justified belief in the logical consequences of τ, the proof-transformations that are required by the notion of elementary reducibility, if available, are clearly formalizable in τ and therefore amount to justified beliefs. This entails that, if the justification for τ' is derived from τ via elementary reducibility, and such

[14] However, other choices of τ may lead to much stronger $\mathcal{I}_{\mathrm{II}}(\tau)$. For instance, $\mathcal{I}_{\mathrm{II}}(\mathrm{PA}_{\mathrm{II}})$ will prove the consistency of Z_2.

justification can be transferred to some member φ of the commitments $\mathcal{J}(\tau')$, the sentence φ will be justified on the basis of justified belief in τ.

We now turn to REFLECTION. Philosophers and logicians have recently started to pay attention to the epistemology of proof-theoretic reflection principles (Horsten and Leigh, 2017; Fischer et al., ming; Cieśliński, 2017). The main focus has been on the difference between entitlement and justification in the context of reflection. Essentially, entitlement and justification differ in the kind of warrant that they require (Wright and Davies, 2004; Burge, 1996). Justification requires self-evidence, or a deductive or inductive rule acting on warranted premisses. Entitlement doesn't. For instance, perceptual beliefs such as '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 (Graham, 2020).

In the light of this, it seems clear that the justification for a reasonable τ does not immediately transfers to Uniform Reflection for τ. First, there is not a deductive *logical* rule that warrants it, as Uniform Reflection is unprovable in τ.[15] Also, inductive rules have little role to play in such abstract contexts such as mathematical theories. Is Uniform Reflection self-evident? One may argue that, given a reliable process of formalization in the background, justified belief in τ may be 'expressed' by Uniform Reflection, and this would guarantee its self-evidence. This conclusion should be resisted, even if one disregards the difficult choices of implementation details in the process of formalization. For instance, as argued in Dean (2015), one can reasonably consider as self-evident fragments of first-order arithmetic – based for instance on some combinatorial or other apriori evidence (Tait, 1981; Parsons, 2007) –, and yet consider Uniform Reflection as unwarranted given the equivalence of Uniform Reflection and full number-theoretic induction.[16]

For similar reasons, Horsten and Leigh (2017) and Fischer et al. (ming) argue that, although one cannot be immediately justified in believing in Uniform Reflection for τ, she can at least be entitled to it. The warrant for Uniform Reflection for a reasonable τ consists in fact in a cognitive project meeting the following additional two conditions (Wright, 2004): (i) We have no sufficient

[15] Although Uniform Reflection for τ can be derived from *non-logical* rules of inference such as Kleene's rule. However, Kleene's rule is equivalent to Uniform Reflection, and is therefore in need of justification in the same way as Uniform Reflection is.

[16] Another possibility may be to resort to ideological expansions of the theory, for instance by means of a truth predicate. In this scenario, the justified belief in τ would entail a justified belief in a theory of truth of τ, ideally one that proves $\mathrm{RFN}(\tau)$ (Shapiro, 1998; Franzén, 2004; Ketland, 2005). However, this strategy does not directly help us, because it only shifts the required preservation of justified belief from the reflection principle for τ to a theory of truth for τ. And it is far from clear that justified belief in τ warrants a justified belief in a non-conservative theory of truth for τ. Moreover, in the context of theories of truth the addition of Uniform Reflection is more prone to brute error than in the purely arithmetical context (Fischer et al., 2017).

reason to believe that Uniform Reflection for τ does not hold; (ii) Any attempt to justify such Uniform Reflection would involve further presuppositions in turn of no more secure a prior standing. The idea is that we have some form of procedural warrant for Uniform Reflection that is not appropriate for justification. We believe that claim (ii) is not correct. We will provide an alternative, deductive route to the justification of Uniform Reflection that is based on more basic principles.

A crucial feature of REFLECTION is that it is *deductively light*, in the sense that its range of applicability is well-delineated and is never confused with the deductive apparatus of τ. This feature of REFLECTION manifests itself in different ways. First, the possibility of conservatively interpreting REFLECTION over τ in τ itself shows that the deductive apparatus of τ is rich enough to represent the logical structure of REFLECTION.[17] Second, the premiss of REFLECTION involves an elementarily decidable property – being a member of the set τ –, whereas Uniform Reflection involves a significantly more complex property – indeed, a recursively enumerably complete property. And elementary properties are completely transparent to any choice of τ: there is always an elementary procedure for deciding whether or not some sentence is a member of τ or not, and therefore this procedure is always available in τ itself. Third, the addition of REFLECTION to τ is fully grounded in τ-provability and does not allow for any *new proofs* obtained from the combination of τ rules with REFLECTION. Since REFLECTION cannot be iterated, it only adds one layer of proofs to the pre-existing structure of τ-proofs. This is in stark contrast with standard procedures for extending theories such as τ with new axioms or new rules of inference.

These distinctive properties of REFLECTION entail that the possibility of error is substantially less likely when extending τ with REFLECTION in place of Uniform Reflection. This can for instance be seen if one considers the paradigmatic case of consistent but ω-inconsistent theories. Whereas the addition of Uniform Reflection can result in inconsistency when τ is ω-inconsistent, REFLECTION determines an inconsistency only if the ω-inconsistency is evident in the concrete presentation of τ itself: e.g. if τ has axioms $P(\overline{n})$ for all n, and $\neg\exists x P(x)$. For instance, one may be justified in believing in the theory FS from (Friedman and Sheard, 1987; Halbach, 1994) on the grounds of its arithmetical soundness. Now FS + RFN(FS) is inconsistent, whereas FS+REFLECTION for FS is not. Of course, we are not claiming that ω-inconsistent theories should have interesting implicit commitments – in fact, in our theory, canonical ω-inconsistency will result in inconsistent implicit commitments! –[18], but that the standard obstacles

[17] While conservativeness is not a *sufficient* reason for justification (Fischer et al., ming, §4.3), it is its combination with the other features that makes REFLECTION special.

[18] By 'canonical ω-inconsistency' in this context we mean an ω-inconsistency that can be formalized in EA.

to the transition from justified belief in τ to justified belief in $\mathrm{RFN}(\tau)$ are not immediately present when considering REFLECTION.

To sum up, in our reflection process we start with the justified belief in τ and in $\forall x\, \tau(\ulcorner \varphi(\dot{x}) \urcorner)$. Specifically, the justification for $\forall x\, \tau(\ulcorner \varphi(\dot{x}) \urcorner)$ is given by one's justified belief in our formal syntax theory EA, and it is a basic assumption of our framework that such justification is compatible with any of the particular justifications one might have for different choices of τ. The deductive lightness of REFLECTION just described enables one to justifiedly believe $\forall x \varphi$.

Consequently, given one's justification for τ, all reasoning steps in Proposition 1 can be seen to preserve such a justification. Closure of justified belief under logical context in our abstract and mathematical context then entails that Uniform Reflection is also justified on the basis of τ.

5. Epistemic Stability

Walter Dean 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 *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,[19] 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 puts the notion of epistemic stability into question. We argue that, if one sticks with the given definition of epistemic stability, there cannot be epistemically stable theories. We also suggest a possible way out for the advocate of epistemic stability, based on a weaker formulation of it.

On the most straightforward understanding of Dean's definition of epistemic stability, the agent is aware of a coherent rationale for 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 INVARIANCE and 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 recognizing whether or not a syntactic object is an axiom of the theory. Proposition 1 then tells us that all instances of

[19] 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 *Isaacson's Thesis*.

Uniform Reflection for the theory can be warranted by this rationale. Therefore, no theory can be epistemically stable 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). We do not take a stance on 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 *accept* PA without possessing a general notion of what an axiom of PA is. And this for us 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 $\mathscr{L}_{\mathbb{N}}$.

To avoid this conclusion, the advocate of epistemic stability could dispense with the notion of acceptance or belief altogether in the formulation of epistemic stability. An epistemically stable theory τ 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. To illustrate this, let's consider the example of a finitist as depicted by Tait (1981). As Tait admits, the characterization of primitive recursive functions as the finitist ones is out of reach for the finitist theorist. 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, although this reformulation of epistemic stability would not immediately contradict the Implicit Commitment Thesis, there are serious issues with it. The advocate of this form of epistemic stability would need to clarify in which way a theorist could coherently hold some foundational standpoint without having access to the very axioms that capture it. We feel that the burden of the proof is on the defender of the new version of epistemic stability, and rest content with pointing out the the original version of epistemic stability is not a viable option.

Finally, in §6.3 we will discuss another potential problem for epistemic stability, applying to any advocate of the epistemic stability of a theory τ who is happy to commit to minimal truth-theoretic extensions of τ.

6. Other Proposals

In this section we connect our proposal with other approaches in the liteature.

6.1. **Reflection as Entitlement.** Although we have argued that both INVARIANCE and REFLECTION preserve justified belief, our formal theory can shed light on positions that regard soundness statements for a base theory τ as warranted by entitlements and not directly by justifications. One such view is advanced in Horsten and Leigh (2017), where it is argued that whenever one is justified in believing a base theory τ, one is *entitled* to a uniform reflection principle for τ. Horsten and Leigh's picture apply this idea to the case in which τ is a theory of disquotational truth, and use reflection to obtain compositional principles for truth: 'the compositionality of truth is implicitly contained in disquotational axioms' (Horsten and Leigh, 2017, p. 209). It's not entirely clear what is the status of such compositional principles in Horsten and Leigh's picture – whether we are only entitled to them, or fully justified in believing them –, but they are certainly logical consequences of a complex net of justifications and entitlements.

A more articulated picture is given in Fischer et al. (ming). There it is argued that, given one's justified belief in a base theory τ 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. The proposal is reminiscent of Crispin Wright's accout of logical knowledge, in which the transition between entitlements to logical principles and their justification is tracked by the distinction between rule formulation of such principles and their object-linguistic formulation by means of the material conditional.

Both approaches are based on the idea that justified belief in τ supports an entitlement to Uniform Reflection for τ. Our formal theory of implicit commitment presented in §3.2 provides a further analysis of the structure of the entitlements involved in the reflection process. In §4 we argued that INVARIANCE preserves justified belief because all the transformations required by elementary reducibility are completely transparent to the theory τ one starts with. We believe that our line of reasoning is available to the proponent of the entitlement-only approach to Uniform Reflection as well. Hence our formal framework can be used to locate the source of the entitlement in a principle that is *properly weaker* than Uniform Reflection. This is a precise sense in which the analysis of implicit commitment via entitlement may be sharpened by our proposal.

6.2. **Believability.** Cezary Cieśliński 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 τ':

> S believes that for every theorem φ of τ there is a normally-good-enough reason to believe that φ. (Cieśliński, 2017, p. 251)

'Having a normally-good-enough reason' is abbreviated as 'being believable',[20] so we shall employ it as well in the context of Cieśliński's system.

Cieśliński defines a formal theory of believability for a system τ (denoted $Bel(\tau)$). It contains an additional predicate B for 'believable' and extends τ with the following axioms, where τB denotes the trivial extension of τ with no non-logical axioms for the predicate B:

REF $\forall\varphi \;\; \mathrm{Prov}_{\tau B}(\varphi) \to B(\varphi)$.
MP $\forall\varphi,\psi \;\; B(\varphi) \wedge B(\varphi \to \psi) \to B(\psi)$.
ωR $\forall\varphi \;\; B(\forall x B(\varphi(\dot{x}))) \to B(\forall x \varphi(x))$

Finally, the entire system is closed under the necessitation rule, NEC,

$$\frac{\varphi}{B(\ulcorner \varphi \urcorner)}.$$

Then, the implicit commitments of τ are defined as the internal theory of $Bel(\tau)$, i.e. the set

$$Int_{Bel(\tau)} = \{\varphi \in \mathcal{L}_B \mid Bel(\tau) \vdash B(\varphi)\}.$$

Crucially, Cieśliński shows that $Int_{Bel(\tau)}$ contains ω-iterations of Uniform Reflection over τ, and therefore a highly non-trivial set of commitments for τ.

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 'φ is believable' to 'I believe φ' is not immediate, for there might be some independent reasons which make us reject φ. In such a situation it might even be the case that both φ and $\neg\varphi$ are believable. This virtue makes the notion of commitment based on believability rather weak and a conditional one: *if* φ 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.

The next difference lies in the complexity of rules used in the formal model. Although ωR (and its predeccesor GEN from Cieśliński (2017)) inspired our REFLECTION, we think that our axiom is significantly simpler: Cieśliński assumes

[20] Although perhaps 'believable' is not the best choice, since it may be confused with the weaker 'it's possible to believe'.

that the set of believable sentences is closed under ωR. Together with NEC this yields closure under the following rule of reasoning

$$\frac{\forall x B(\varphi(\dot{x}))}{B(\ulcorner \forall x \varphi(x) \urcorner)}.$$

Our REFLECTION axiom rests on weaker assumptions: we might infer $\forall x \varphi(x)$ only if we establish (i.e. prove in τ) that for every x, $\varphi(\overline{x})$ is an *axiom* of the accepted theory. Moreover, $\tau(x)$ is an elementary formula, hence membership in τ can easily be decided.

We finally come to our main worry concerning Cieśliński's theory of believability. It concerns the *closure* under the rule NEC. Cieśliński argues that it is justified, since 'a proof of φ in $Bel(\tau)$ is provided, it is simply the proof itself, which is seen as good reason to accept φ' (Cieśliński, 2017, p. 254-55). We have no doubt that, if τ is accepted, *proofs in* τ amount to good reasons to accept their conclusions. But what grounds our acceptance of *proofs in* $Bel(\tau)$? This is left unjustified by Cieśliński. One might try to overcome the issue by showing that applying NEC to derivations outside of τ is not necessary to yield non-trivial implicit commitments. Let us define the system $Bel'(\tau)$ to be the restriction of $Bel(\tau)$ in which the rule NEC can be applied only once and only to consequences of τ. Unfortunately, in all cases of interest, this system does not produce any non-trivial implicit commitments. The following proposition shows that the strength of $Int_{Bel(\tau)}$ essentially lies in the interplay between the rule NEC and the the axioms for the believability predicate.

PROPOSITION 2. *Let τ be Σ_1-sound. Then $Int_{Bel'(\tau)}$ is conservative over τ.*

Proof. Fix τ. We shall show how to interpret $Bel'(\tau)$ in τ. Let us define the translation * by recursion in τ, putting

$$B(t)^* = \mathrm{Prov}_{\tau B}(t)$$

and letting * be identity on the arithmetical formulae and commute with connectives and quantifiers (hence * does not relativize quantifiers). Now, we check that if φ is an axiom of $Bel'(\tau)$, then $\tau \vdash \varphi^*$. For (REF) and (MP) this is obvious. Let us check ωR. Observe that

$$\big(\forall \varphi (B(\forall x B(\varphi(\dot{x}))) \to B(\forall x \varphi(x)))\big)^*$$

is equal to

$$\forall \varphi \big(\mathrm{Prov}_{\tau B}(\forall x B(\varphi(\dot{x}))) \to \mathrm{Prov}_{\tau B}(\forall x \varphi(x))\big).$$

Reasoning in τ, we see that for every φ the antecedent of the implication is false, since it is consistent with pure logic that the extension of B is empty. Hence (ωR)* is provable in τ as well. Finally, we deal with NEC: whenever we have a

$Bel'(\tau)$ proof

$$(\varphi_0, \ldots, \varphi_k, B(\ulcorner \varphi_k \urcorner))$$

ending with an application of the NEC rule, then we know (by the restriction on Bel'), that

$$(\varphi_0, \ldots, \varphi_k)$$

is a proof in τ. Hence $\tau \vdash \mathrm{Prov}_\tau(\ulcorner \varphi_k \urcorner)$, since Prov_τ weakly represents provability in τ. Consequently, $\tau \vdash \mathrm{Prov}_\tau(\ulcorner \varphi_k \urcorner)$, which witnesses that * is indeed an interpretation.

Now it follows that for every arithmetical φ, if $Int_{Bel'(\tau)} \vdash \varphi$, and then $Bel'(\tau) \vdash B(\ulcorner \varphi \urcorner)$. Hence, by the above considerations, $\tau \vdash \mathrm{Prov}_{\tau B}(\ulcorner \varphi \urcorner)$. By Σ_1-soundness of τ, it follows that $\tau B \vdash \varphi$. However, τB is trivially conservative over τ, so it follows that $\tau \vdash \varphi$. □

6.3. **The semantic core.** Nicolai and Piazza (2018) react to Dean's counterexamples to ICT. They argued that, even in epistemically stable theories τ, one can isolate a set of non-trivial implicit commitments, which are nonetheless conservative over τ. This is the basic idea of their *semantic core*, a minimal set of implicit commitments that is shared by anyone accepting τ. Formally, the semantic core consists in an extension of τ with a compositional theory of truth over τ and the sentence

(∗) All sentences in τ are true.

Let is call such an extension of τ $\mathrm{CT}^-[\tau]$. Crucially, in many natural cases $\mathrm{CT}^-[\tau]$ is a *conservative* extension of τ. 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*:

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

We shall now show that, modulo INVARIANCE, WICT implies its original, stronger version.

The theory $\mathrm{CT}^-[\mathrm{PA}]$ is conservative over PA if PA is formally represented as finitely many axioms of Q and the induction scheme (Kotlarski et al., 1981). However, consider

$$\mathrm{PA}_{\mathrm{RFN}} := \mathrm{EA} + \mathrm{RFN}(\mathrm{EA}).$$

Kreisel showed that $\mathrm{PA}_{\mathrm{RFN}}$ and PA are the same theory, and this fact can be formalized in EA. The two theories are then elementarily equivalent (Beklemishev, 2005). However, $\mathrm{CT}^-[\mathrm{PA}_{\mathrm{RFN}}]$ proves the global reflection principle over

EA,[21] i.e. the sentence

$$\forall\varphi\big(\mathrm{Prov}_{\mathrm{EA}}(\varphi) \to \mathrm{T}(\varphi)\big).$$

By a result Cieślńiski (2010), already the assertion that logical validites are true, that is

$$\mathrm{CT}^{-}(\mathrm{PA}) + \forall\varphi\big(\mathrm{Prov}_{\emptyset}(\varphi) \to \mathrm{T}(\varphi)\big)$$

implies the Global Reflection principle for PA, that is the sentence:

$$\forall\varphi\big(\mathrm{Prov}_{\mathrm{PA}}(\varphi) \to \mathrm{T}(\varphi)\big).$$

Therefore, $\mathrm{CT}^{-}[\mathrm{PA}_{\mathrm{RFN}}]$ is non-conservative over PA. In fact, it is much stronger: not only $\mathrm{CT}^{-}[\mathrm{PA}_{\mathrm{RFN}}]$ proves the uniform reflection over PA, but it also able to demonstrate each finite number of iterations of uniform reflection over PA (Kotlarski, 1986; Smoryński, 1977; Cieśliński, 2017; Łełyk, 2017).

Nicolai and Piazza (2018)'s proposal only applies to *schematic theories*, that is theories that τ can be presented by means of some first order formula $\varphi(P)$ with a free second order variable P such that $\tau(x)$ says:[22]

> There is a $\psi(y)$ such that x is the result of replacing $P(y)$ with $\psi(y)$ in φ.

Admittedly, schematic theories include a substantial variety of natural axiomatizations of foundationally relevant mathematical theories. However, what we said about INVARIANCE, we should not disregard non-schematic presentations of theories such as $\mathrm{PA}_{\mathrm{RFN}}$: they are also quite natural qua axiomatizations by refleciton and are elementarily equivalent to schematic presentations.

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 τ to articulate impilcit commitment, the fact that even WICT leads beyond what's provable in τ amounts to additional evidence of the problematic nature of the notion of epistemic stability.

[21] Indeed, reasoning internally in $\mathrm{CT}^{-}[\mathrm{PA}_{\mathrm{RFN}}]$ for every φ the sentence

$$\mathrm{Prov}_{\mathrm{EA}}(\varphi) \to \varphi$$

is an axiom of $\mathrm{PA}_{\mathrm{RFN}}$. Hence, within $\mathrm{CT}^{-}[\mathrm{PA}_{\mathrm{RFN}}]$ we have $\forall\varphi\mathrm{T}(\mathrm{Prov}_{\mathrm{EA}}(\dot{\varphi}) \dot{\to} \varphi)$ and by compositional axioms we deduce

$$\forall\varphi\big(\mathrm{Prov}_{\mathrm{EA}}(\dot{\varphi}) \to \mathrm{T}(\varphi)\big).$$

[22] Leigh (2015) proves in fact that, for schematic theories τ, $\mathrm{CT}^{-}[\tau]$ is always conservative over τ.

7. Conclusion

In this paper we have 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 τ in terms of two simple axioms, INVARIANCE and REFLECTION. Unlike standard formal soundness claims for τ, these axioms may be conservatively interpreted in τ. However, when combined, they necessarily entail the uniform reflection for τ.

The main consequence of this formal framework is that justified belief in τ can be preserved, via the principles of INVARIANCE and REFLECTION, to Uniform Reflection for τ. The framework has consequences for other aspects of the epistemology of proof-theoretic reflection. On the one hand, the recent analysis of it in terms of entitlement can be substantially refined by it. 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.

References

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

Beklemishev, L. D. and Pakhomov, F. N. (2019). Reflection algebras and conservation results for theories of iterated truth.

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

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

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

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 Czermak, J., editor, *Philosophy of Mathematics*, pages 1–147. Hölder-Pichler-Tempsky.

Fischer, M., Horsten, L., and Nicolai, C. (forthcoming). Hypatia's silence. *Noûs*.

Fischer, M., Nicolai, C., and 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.*, volume 16 of *Lecture Notes in Logic*. Association for Symbolic Logic, Urbana, IL; A K Peters, Ltd., Wellesley, MA.

Friedman, H. and Sheard, M. (1987). An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic*, 33(1):1–21.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatsh. Math. Phys.*, 38(1):173–198.

Graham, P. (2020). What is epistemic entitlement? reliable competence, reasons, inference, access. In Greco, J. and Kelp, C., editors, *Virtue-Theoretic Epistemology: New Methods and Approaches*. New York, USA: Cambridge University Press.

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

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic*, 35(1):311–27.

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

Horsten, L. (2018). Book review: Cezary Cieslinski, The Epistemic Lightness of Truth: Deflationism and its Logic. Cambridge University Press, 2017. *Notre Dame Philosophical Reviews*.

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

Isaacson, D. (1987). Arithmetical truth and hidden higher-order concepts. In *Logic Colloquium '85*, volume 122 of *Studies in Logic and the Foundations of Mathematics*, pages 147 – 169. Elsevier.

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

Kotlarski, H. (1986). *Mathematical Logic Quarterly*, 32:531–534.

Kotlarski, H., Krajewski, S., and 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.

Łełyk, M. (2017). Axiomatic truth theories, bounded induction and reflection principles.

Nicolai, C. and Piazza, M. (2018). The implicit commitment of arithmetical theories and its semantic core. *Erkenntnis*, pages 1–25.

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.

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

Smoryński, C. (1977). ω-consistency and reflection. In *Colloque International de Logique (Clermont-Ferrand, 1975)*, volume 249 of *Colloq. Internat. CNRS*, pages 167–181. CNRS, Paris.

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

Turing, A. M. (1939). Systems of Logic Based on Ordinals. *Proc. London Math. Soc. (2)*, 45(3):161–228.

van Fraassen, B. (1980). *The scientific image*. Oxford University Press, Oxford.

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

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

mlelyk@uw.edu.pl ; carlonicolai6@gmail.com ; carlo.nicolai@kcl.ac.uk