The Philosophical Quarterly Vol. 71, No. 4 *2021*
ISSN 0031-8094 doi: 10.1093/pq/pqaa083
Advance Access Publication 7th January 2021

ON REFLECTION

By Leon Horsten

This article gives an epistemological analysis of the reflection process by means of which you can come to know the consistency of a mathematical theory that you already accept. It is argued that this process can result in warranted belief in new mathematical principles without justifying them.

Keywords: reflection, reflection principle, consistency, epistemic entitlement, cognitive programme, implicit commitment.

> *Much justified mathematical belief is underwritten by non-demonstrative reasoning . . . Our belief in the consistency of arithmetic seems thoroughly warranted; in fact I think it constitutes knowledge. But no proof of it adds significantly to the ground for our belief.*
>
> (Burge 1998, p. 8)

I. INTRODUCTION

In the course of the debate about implicit commitment generated by acceptance of a theory, it has been claimed that in certain circumstances where you know a mathematical theory S and nothing more, you can come to know proof theoretic reflection principles for S without justifying them. Let us call this the *Implicit Commitment Thesis* (ICT). This thesis is a bold claim because if S is sufficiently strong (and consistent), then proof theoretic reflection principles for S are logically independent of S.

There is no agreement in the literature whether ICT is correct. But the Thesis is of considerable epistemological importance. ICT is relevant for the wider epistemological question to what extent, if any, you can come to know cognitive presuppositions of your cognitive projects without justifying them.

The only way in which progress can be made in this epistemological debate is by giving a detailed philosophical analysis of the cognitive process of reflection. Proposing an epistemological analysis of the process of reflection therefore constitutes the core of this article. Specifically, I will propose an analysis of the process of reflecting on the presupposition of *consistency* of the theory that you currently accept. Already this most basic reflective process will turn out to have a more complicated structure than might be expected, and I do not claim that you can come to know stronger reflection principles (such as uniform reflection principles or global reflection principles) through a reflection process that is similar to the one that is analysed in this article.

The Implicit Commitment Thesis that I will investigate is closely related to Feferman's views on the implicit commitments arising from accepting a theory. And this, in turn, is related to Wright's work on cognitive projects and their presuppositions. These connections are explored in the next two sections. Then, in Section IV, a detailed description of the process of reflection on consistency is given. In Section V, it is argued that this process vindicates ICT. In the final sections, wider ramifications and connections of my analysis of the process of reflection are discussed. The analysis of Section IV is compared to Cieśliński's recent account of the process of proof theoretic reflection, and to Wright's analysis of the presuppositions of our cognitive project of trying to understand the outside world. In the concluding Section I argue that my account is in harmony with Burge's general philosophical view of the philosophical process of reflection.

II. COGNITIVE PROJECTS AND THEIR PRESUPPOSITIONS

Wright has formulated an influential epistemological theory about what he calls *cognitive projects*. A cognitive project is "defined by a pair: a question, and a procedure one might competently execute in order to answer it" (Wright 2012, p. 466). One key objective of cognitive projects is of course to yield knowledge.

An example of a small-scale cognitive project is the ordered pair

$$\langle \text{What time is it?, Consult your watch} \rangle.$$

But by extension you can also conceive of a pair consisting of a cluster of questions and a battery of procedures for answering these questions as a (large-scale) cognitive project. One example of a very large-scale cognitive project can be taken to be the pair consisting of "Find out what you can about the external world" and {sense perception, logical reasoning, ampliative reasoning, ... }.

In his influential article *Warrant for Nothing* Wright (2004), Wright is concerned with this very large-scale cognitive project. In particular, he is concerned with our epistemic warrant for certain general propositions that he calls *cornerstone propositions*. Cornerstone propositions are propositions that play an organising role in our cognitive representation of reality. An example is the proposition:

There is an external world.

This proposition is a *presupposition* of the very large-scale project of trying to understand the world and our relation to it.

Wright is concerned with sceptical challenges that seek to undermine our epistemic warrant for our cornerstone propositions and thereby undermine our warrant for believing in ordinary propositions (such as “it is snowing outside”) in everyday circumstances.

One might think that scepticism about the outside world can in a Moorean fashion be refuted by what Wright calls a I-II-III argument Wright (2002):

I My visual perception suggests to me that I have hands.
II I have hands.
III There is an external world.

In Wright’s view, such an argument for an anti-sceptical conclusion is not rationally acceptable. We have a case of *warrant transmission failure*. The problem is that the argument from I to III is question-begging: one can only rationally accept the argument as a whole—and in particular the inference from I to II—on the condition that III holds. So it seems that we have to establish III in an independent way, and there seems no way to do this.

In response to this situation, Wright denies that III is in need of justification. Instead, we are entitled to rely on III without justification. Proposition III is then a *presupposition of cognitive project*: doubting III would rationally commit one to doubting the significance or competence of our cognitive project (Wright 2004, p. 193). Relying on presupposition III allows us to be *justified* in inferring II from I.

The notion of *entitlement of cognitive project* can then be defined along the following lines (Wright 2004, pp. 191–2):

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

In the light of this, Wright claims that we are entitled to *rely on* or *trust* (which appear to be near synonymous terms for Wright) cornerstone proposition III without having justification for it. According to Wright this does not, however, give us the epistemic right to *believe* III, for the very reasons that we have gone through above: to conclude III from I and II would be question-begging.

III. IMPLICIT COMMITMENT

Mathematics also contains cognitive projects. Mathematics itself can be seen as a large-scale cognitive project; subfields of mathematics can be seen as somewhat smaller-scale cognitive projects.

Typically, multiple theories are combined even in subfields of mathematics. Nonetheless, let us simplify matters—hopefully without affecting the strength of our argument—and identify a cognitive project with some first-order theory S. For example, we might (admittedly somewhat ridiculously) identify the cognitive project of number theory with first-order Peano Arithmetic (PA), or, somewhat pedantically, with discovering facts about the natural numbers on the basis of proof in PA.

A *proof theoretic reflection principle* for a mathematical theory S says that, or approximates saying that everything that S asserts, i.e., everything that is provable in the theory, is true (Kreisel and Levy 1968, p. 98). By Gödel's incompleteness theorems, proof theoretic reflection principles for S are under very general circumstances logically independent of S. A proof theoretic reflection principle for S can be said to be a *presupposition* of the cognitive project S.

Feferman actually formulated a claim that is somewhat stronger than what I have labelled ICT. He claimed that "[proof theoretic reflection is] the process of finding out what is implicit in accepting a basic system L_1, i.e., what one ought to accept, on the same fundamental grounds, when one accepts L_1" (Feferman 1988, p. 131). Let us call this thesis ICT$^+$. In this quote, 'ought to accept' should be interpreted as *is rationally required to accept*. The weaker claim ICT put forward in the introduction is obtained by replacing the expression 'ought to accept' in this quote by the expression '*can* come to know', and by replacing 'on the same rational grounds' by 'without justifying them'. I will argue in this article that ICT$^+$ is false, but its weaker cousin ICT is true.

In order to understand ICT$^+$ and ICT it is important to be clear about what is involved in *full* or *unconditional acceptance* of a theory S. Acceptance, as I will use the term, is a concept that has both a pragmatic and a doxastic component: see van Fraassen (1980). The pragmatic component concerns taking the theory as a guide to *action*. It involves committing to using S in your research. But it may also involve, for instance, building a bridge across a gorge based on a design which relies on theorems of S. The doxastic component of theory acceptance involves propositional mental representation and judgement: it involves *believing* the axioms of S. The notion of belief that is at work here is as full as the theory acceptance of which it is an aspect. So full acceptance entails full belief. The notion of full acceptance is qualitative in nature, but full acceptance of a theory S entails that when you are pressed to attach a 'degree of belief' to each of the axioms of S, you give them the maximal degree of belief. (Of course you may be sceptical about theories of 'degrees of belief'. If such scepticism is well-founded, then attempts to relate the qualitative notion of full acceptance to degrees of belief are futile.[1]) For a theory that has infinitely many axioms (such as PA), full acceptance of it has a dispositional component. For you unconditionally to accept PA, for instance, entails being disposed to believe any instance of the induction scheme that is presented to you.

This is the notion of full acceptance that I will be working with. There are other ways of employing the acceptance-belief distinction: see for instance Cohen (1992). In particular, it is important to observe that my use of the term 'acceptance' does *not* coincide with the way in which Wright uses the term in Wright (2004). Moreover, as far as I can see, there is not enough textual evidence to decide whether my use of the term coincides with Feferman's use of it in the quotation given earlier.

In a recent article, Dean sheds doubt on ICT$^+$ (Dean 2014, p. 35).[2] His reason for doubting ICT$^+$ is that many mathematical theories, such as PA, are *epistemically stable*: there appears to be nothing epistemologically blameworthy about someone who accepts such a theory and nothing more (Dean 2014, p. 53). I concur. But this leaves open the epistemological question whether ICT is correct. The correctness of ICT would still be surprising and significant because the claim that you can come to know a proposition that is logically independent from everything you know without justifying it, has an air of implausibility.

Kreisel has emphasised that in order to make progress on the kind of epistemological questions with which we are concerned, a *phenomenological description* of the reflection process is needed (Kreisel 1970, p. 489), even though, like Feferman, he himself did not provide one. Indeed, it is argued in Horsten and Leigh (2017) that a version of ICT—or even ICT$^+$?—for *theories of truth* is correct. But the authors do not give a sufficiently detailed account of the process of reflection involved, and for this reason their account falls short of being convincing.

In order to be as perspicuous and as explicit as possible, I will present a detailed epistemological analysis of the *simplest* reflection process: reflection on *consistency*. This analysis will then be used as a basis for evaluating ICT. The analysis that I will propose is based on a detailed description of the *structure* of the process of reflection. In a loose sense, this description can therefore be called phenomenological. But it is not an exercise in phenomenology in the specific sense of Husserl and his followers.

[1] Thanks to an anonymous referee for pointing this out.
[2] Dean calls this stronger thesis ICT (Dean 2014, p. 32).

IV. REFLECTION

IV.1. A cognitive project

I will be talking about *you* all the time in the fictional tale that I am about to tell. But nothing hinges on this. *You* might as well be a whole mathematical community for the purposes of the argument that follows.

Suppose you are a mathematician. As a mathematician, you accept and believe the axioms of PA. You do not accept them instrumentally or provisionally; you accept them *unconditionally*, without any reservations. Moreover, you unreservedly rely on the inference rules of classical logic when you construct *proofs* in your mathematical theory. You fully believe the *theorems* that you prove in PA.[3] This, as far as your mathematical work goes, is all *all* that you unconditionally believe and accept. In this situation, you are disposed unconditionally to believe all of (the classical closure of) PA and nothing more.[4]

In particular, the consistency of PA is not something you currently believe or are currently disposed to believing. Suppose that this disposition to believe, as far as mathematics is concerned, all of PA and nothing more, has somehow come to be hard-wired in you.

As a mathematician, you have an even deeper commitment to classical logic than to PA. If you were to derive a contradiction in PA, then you would reject some mathematical principles of PA rather than principles of classical logic.

Suppose that PA is in fact *true*.[5] Moreover, assume in addition that you as a matter of fact have epistemic *justification* for your belief in the axioms of PA. You may or may not know that you have, but you have. And suppose that your justification for PA does not justify *more* than PA. In particular, suppose that it does not justify the statement expressing the consistency of PA. (Otherwise our task would be too easy.)

That such a situation is possible (for a theory such as as Primitive Recursive Arithmetic, for instance) is argued for instance in Dean (2014), and I am assuming in article paper that this thesis is correct. Indeed, suppose we had a solid argument for the thesis that for every recursively axiomatised theory T in the language of arithmetic, for you fully and justifiedly to believe T, you would in addition have to have a justified belief in the formalised consistency statement for T. Then it would follow that for *no* recursively axiomatised theory T in the language of arithmetic, you could be justified in believing T and no more than that. I.e., then as far as arithmetic is concerned, your

[3] So I will from now on often identify PA with the closure of its axioms under classical logic.

[4] I am *not* assuming that at this stage, *you* believe *that* as far as mathematics is concerned, you believe the theorems that you prove in PA and no more than that. Thanks to an anonymous referee for asking me to point this out clearly.

[5] This is again a statement made 'from the outside': I am not assuming that *you* at this point believe that PA is true. Thanks again to an anonymous referee for asking me to be very explicit about this.

powers would outstrip those of any Turing machine. But—*pace* Lucas and Penrose—it is widely accepted that currently no such argument exists that carries conviction.

Insisting on restrictions on the *kind* of justification for the mathematical axioms of PA would limit the scope of my philosophical account of reflection, for there is no agreement about what justifies mathematical axioms that we think we know. So I impose no restrictions on the *kind* of justification that you have for the axioms of PA. Nonetheless, here is one example of a scenario that has been entertained (and criticised!) in the philosophical literature. The natural number structure is somehow given to you in intuition. You have justified the axioms of PA, and only them, by verifying that they hold in this "standard model"; you believe the axioms of PA on this basis.

The mathematical theory PA is then a fairly large scale *cognitive project* in Wright's sense of the word. It can be seen as an ordered pair:

$$\langle \text{questions expressible in the language of PA, proofs and refutations in PA} \rangle .$$

Nothing hinges on the maximal mathematical theory that you unreservedly accept and believe being PA; focussing on PA is mainly done for definiteness. The analysis of the process of reflection on consistency that I am about to propose is intended to have some generality: it is intended to apply to a variety of mathematical theories.

IV.2. The state of innocence

Your cognitive project *presupposes* the consistency of PA. Indeed, the consistency of PA is a cornerstone of your cognitive project.

You trust PA. This implies that you rely on its consistency even if you have never posed the question of the consistency of PA to yourself. So the formalised consistency statement for PA captures an aspect of your trust in PA. In Feferman's words (Feferman 1962, p. 261):

> 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 in the system of axioms A_k.

Perhaps you deeply distrust philosophy, all distinctively philosophical concepts and philosophical theories about them. In particular, you may not believe that there is a concept of truth that you may legitimately use in your reasoning. Nevertheless, if you were to discover that PA is inconsistent, then *as a mathematician* you would (rightly) feel compelled to revise your mathematical commitments.

The situation you are in satisfies Wright's conditions for entitlement of cognitive project (see page 5): you have no reason to think that PA is inconsistent, and an attempt to justify the consistency of PA would involve presuppositions in turn of no more secure prior standing. So you are *epistemically entitled to rely* on the consistency of PA. Call the situation that you are in at this point, i.e., before you start to reflect on your acceptance of PA, the *state of innocence*.

The next question is: how can you come to be entitled *to believe* in the consistency of PA? I will presently argue that you can come to be in this position by *reflecting* on what you are relying on in your cognitive project. Indeed, there are circumstances in which you can, by reflection, come to *know* the consistency of PA without justifying a statement that expresses the consistency of PA.

IV.3. Belief de se

In your pursuit of your cognitive project, you are guided by an algorithm *e* that produces all and only PA-provable statements.

We may assume that presently you do not know, or even believe, that you are guided by this algorithm. But by *reflection* on your cognitive situation, you can obtain beliefs about your cognitive situation. You can come to believe that in your mathematical work, you are disposed to accepting what is provable in PA. This reflective moment constitutes the first stage of the reflective process.

How does this happen? You consider all Peano Axioms except the mathematical induction scheme, and realise that you believe them. Concerning the scheme of mathematical induction, you notice that your acceptance of instances of mathematical induction does not depend on the particular formula for which it is instantiated, but that you are disposed to accept all arithmetical statements that have the *form* of a mathematical induction axiom. Similarly for the logical axiom schemes, and the logical schematic rules. Then, by mathematical induction (in a language that extends the language of arithmetic), you conclude that you are disposed to believing all proofs in PA.

Observe that this does not mean that you have thus come to believe that you are the algorithm *e* that was mentioned at the beginning of this subsection. You have come to believe that what you are disposed to believe is a *subset* of the arithmetical statements that are produced by *e*. I leave the question whether, and, if so, how, you can come to believe that you *are*, as far as your mathematical work goes, the Turing machine *e*, for later.

Note also that something fundamentally new has happened in this first reflective movement. Up until just now, self-awareness was not involved in the story. You were as a matter of fact explicitly accepting all of PA, but you did not know this.[6] Now, however, you do, and this involves acts of self-consciousness. This shows, incidentally, that the kind of reflection involved is somewhat similar to the examples of reflection that the classical rationalist philosophers were occupied with.

[6] A structurally similar characterisation of the 'state of innocence' of the finitist when she is working inside Primitive Recursive Arithmetic, is given in (Dean 2014, p. 53).

IV.4. Expressing your trust

In a second reflective act, you come to see that you have been, and are, relying on the consistency of your cognitive project. You make your implicit trust explicit. How? Not by ‘rational intuition’, presumably, but rather by *counterfactual reasoning*.

You have recognised that, as far as mathematics is concerned, you are disposed to believing what is provable in PA (stage 1 of the reflective process). You now realise that *if you were to derive a contradiction in PA*, your commitment to your cognitive project would collapse. Acquiring this counterfactual belief is a second moment of reflection. And this form of reflection is different from the first reflective moment. In particular, it is unlike the kinds of philosophical reflection studied by the classical rationalists.

At this juncture, there are two courses rationally open to you. *Either* you revise your commitment to your cognitive project, *or* you form a *belief* in the consistency of PA.

Suppose, for a moment, that at this juncture you do not form a belief in the consistency of PA, but instead remain agnostic about it. Then you can adopt a *instrumentalist* form of acceptance of PA that is difficult to distinguish from what I have called the state of innocence. You can resolve simply to continue with your mathematical practice unless and until you find a contradiction in PA. In other words, your acceptance can be an acceptance *as if* PA holds.

The reason why this kind of instrumental acceptance is hard to distinguish from the state of innocence is that you never *will* find a contradiction in PA, and even if you do, you will revise your practice in pretty much the same way as you would do if you had found the contradiction while in the state of innocence.

Nonetheless, your instrumental acceptance of PA is not the same as your full acceptance of PA in the state of innocence. Your instrumental acceptance is coloured by what you now take to be an epistemic possibility and which would undermine your cognitive project if it came to pass.

Concerning the doxastic aspect of your acceptance, it is admittedly *logically possible* for you not to change your unconditional belief in each of the axioms of PA while even at the end of the reflection process remaining agnostic about the consistency of PA. But it would be *irrational* to do so. It would be irrational even on a ‘liberal’ conception of rationality: recognising as an epistemic possibility a situation of which you know that it would undermine your belief in the conjunction of the axioms of PA, rationally compels you to have less than full belief in some of the axioms of PA. For those who are sympathetic to theories of degrees of belief, the problem can also be phrased in quantitative terms.[7] When coaxed to describe your mode of belief in quantitative terms, you give

[7] Thanks to Simon Goldstein for pressing me to express the situation in quantitative terms.

maximal credence to each axiom of PA (see page 7). Yet you recognise as an epistemic possibility a scenario in which PA would not hold. This is irrational.

There are situations in which it is perfectly rational not to form a consistency belief as a result of the reflection process, and to withdraw to less-than-full acceptance, such as the instrumental form of acceptance that was sketched earlier. Suppose that your starting theory is not PA but standard Zermelo-Fraenkel set theory with the axiom of Choice (ZFC), and you come to realise by means of the reflective process that you are relying on the consistency of ZFC, whereas you had not entertained the question of the consistency of ZFC before. You may, in that situation, not be sanguine that finding a contradiction in ZFC will never happen, even though you do not at present have even the vaguest inkling about how or where in set theory it might arise. (I know mathematicians who find themselves in this state.) In this situation, you may simply revise your unconditional acceptance of ZFC to a somewhat lower degree of acceptance. Your acceptance of ZFC becomes more cautious (or guarded, or provisional), even though this change does not leave a visible trace in your mathematical practice.

Suppose, however, that you maintain your *unreserved* commitment to your cognitive project through to the end of the reflective process. Then, if you are rational, you form an unqualified new belief: a full belief in the consistency of PA. (Observe that this does not mean that you *voluntarily decide* to believe that PA is consistent!) This concludes stage two of the reflective process.

IV.5. Arithmetisation

When you have arrived at this point, you have come to believe that PA is consistent. But you have not yet acquired a new *arithmetical* belief. Nevertheless, a belief in an arithmetised consistency statement for PA can be obtained by continuing your reflection process along the following lines.

Presently you come to realise that, given a simple coding scheme, provability in PA is expressed by an arithmetical predicate Bew_{PA}. This is also not a straightforward process—it took the mind of Gödel to think this through. Your reasoning goes roughly as follows.

You define some convenient computable coding $\ulcorner \dots \urcorner$ of terms and formulas of the language of PA ($\mathcal{L}_{\mathrm{PA}}$). You also construct a standard provability predicate Bew_{PA} for PA.

You want to convince yourself that:

$$\text{For all } \varphi \in \mathcal{L}_{\mathrm{PA}} : \mathrm{PA} \vdash \varphi \Leftrightarrow Bew_{\mathrm{PA}}(\ulcorner \varphi \urcorner).$$

You do this by *proving* this statement by mathematical induction (on the complexity of proofs). This statement relates syntax (symbols, terms, formulas) with numbers via your coding scheme. So, formally, this is an inductive argument in a language that does not only contain the familiar arithmetical vocabulary but also contains syntactic predicates and allows quantification over syntactic entities. This process of arithmetisation of PA constitutes the second stage of the reflective process.

It is of course possible that you reject meta-syntactic reasoning: if so, then you cannot carry out the proof of the statement. In this case, your reflective process will have ended at stage two. But I will assume that you accept the basic meta-syntactic reasoning required to prove the statement.

The point of spelling out what is involved in this argument in some detail is seeing that philosophical or semantic notions (such as rational belief, or truth) play no role in this reasoning. Moreover, and equally importantly, the theory in which this argument is carried out, is proof theoretically *conservative* over PA.[8]

As a particular instance of the new belief that you have acquired, you find that the consistency of PA is equivalent to the arithmetical statement $\neg Bew_{\mathrm{PA}}(\ulcorner \perp \urcorner)$.[9] You combine this belief with the outcome of stage two of the reflective process, i.e., with your belief that PA is consistent. Thus you come to believe a *new* arithmetical sentence, i.e., $\neg Bew_{\mathrm{PA}}(\ulcorner \perp \urcorner)$.

The assumption made at the beginning of this section, that you are using a *standard* provability predicate, is crucial.[10] For instance, suppose that you were to formalise provability in PA instead as

$$Bew_{\mathrm{PA}}(x) \wedge \mathrm{Con(ZFC)},$$

where Con(ZFC) is a standard way of formalising the consistency of ZFC in arithmetic. This new predicate would be co-extensive with the standard provability predicate Bew_{PA}. But *that* this nonstandard arithmetical provability predicate captures provability in PA can only be proved on the assumption Con(ZFC). However, if your process of reflecting on the consistency of PA required the consistency of set theory as an assumption, then it would of course not give you, at the end of stage 3 of the reflective process, a *new* entitled arithmetical belief. Similarly, there are nonstandard provability predicates Bew^*_{PA} such that already PA proves $\neg Bew^*_{\mathrm{PA}}(\perp)$.[11] If you use such a provability predicate, then again you do not arrive at a *new* entitled arithmetical belief.

This concludes my description of stage 3 of the reflective process, which is also the end of the reflective process as a whole. If it is at least roughly accurate, then which epistemological lessons can we draw from it?

[8] For a proof of this fact, see (Nicolai 2013, section IV.3).

[9] '⊥' stands for your favourite contradiction.

[10] Thanks to an anonymous referee for asking me to stress this point.

[11] For a discussion, see (Franzén 2004, section 12.2). Such provability predicates will not satisfy Löb's derivability conditions, which are seen as minimal requirements that any *standard* provability must satisfy.

V. COGNITIVE WORK

If you rely on a presupposition of your cognitive project, and are entitled to do so, then you are *entitled to articulate* what you are relying on in engaging unconditionally in your cognitive project. In this situation you are *entitled to believe* the presupposition of your cognitive project. You have 'warrant for nothing' in Wright's sense, but—*pace* Wright—not just warrant for trust, but warrant for belief. In fact, it is not completely accurate to describe the upshot as 'warrant *for nothing*'. You have earned your epistemic warrant for believing in the consistency of PA by doing cognitive work that carries cognitive risk. It is just that you have not given independent justification for the statement that expresses the consistency of PA: you did not need to.

In Burge's terminology, epistemic justification and epistemic entitlement stand to epistemic warrant as species to genus (Burge 2013b, p. 489).[12] In particular, having epistemically entitled belief does not entail having epistemically justified belief. At the end of the reflective process, you have acquired an epistemically entitled belief in the consistency of PA, but you have not justified the consistency of PA.

The assumption, in the fictional tale, that you are *justified* in your belief in the axioms of your starting theory PA to begin with (regardless of whether or not you are *aware* that you are so justified), is essential: your epistemic entitlement to believe in the consistency of PA, and therefore also your entitlement to believe $\neg Bew_{\mathrm{PA}}(\ulcorner \bot \urcorner)$, depends on your justification for believing the basic axioms of PA. Suppose for a moment instead that your starting theory is not the mathematical theory PA at all, but the teachings of a guru whom you have started to consult and base your beliefs on, simply because you assume that he is holy.[13] At some point you become aware of the fact that you are relying on the guru, while continuing to rely the guru in the same way as before. Then you form a belief in the reliability of the guru, but you are not *epistemically entitled* to this belief.

We do not have, in epistemology, anything like a clean definition of Gettier cases. But we know that Gettier can strike in the domain of mathematical beliefs as it can in other cognitive domains. You may derive, for instance, a true statement from a false mathematical axiom that you are justified in believing; then you are justified in believing the true statement, but you do not know it. So having true justified belief in PA does not entail that you *know* PA, and your true epistemic entitlement to believe $\neg Bew_{\mathrm{PA}}(\ulcorner \bot \urcorner)$ does not entail that you know this proposition.

[12] Graham (2020) gives an extended discussion of the relation between epistemic justification and epistemic entitlement according to Burge.

[13] This example was suggested to me by Cezary Cieśliński.

But *if you are not in a Gettier situation*, then you have more than justified true belief in the axioms of PA (and in theorems of PA): you *know* them. And then, after your process of reflection, you have acquired more than an epistemically entitled true belief in $\neg Bew_{\mathrm{PA}}(\ulcorner \bot \urcorner)$: you *know* this proposition. You have acquired knowledge of a cornerstone proposition of your cognitive project.

If your justification for the axioms of PA was a priori to start with, then the epistemic entitlement of your belief in the consistency of PA will likewise be a priori. The counterfactual reasoning that led you to believe that, in your cognitive project, you presuppose the consistency of your practice, is a priori. In the first reflective stage, you used an argument by mathematical induction. But if your justification for the arithmetical instances of mathematical induction was a priori, then so will, presumably, be your justification for applying mathematical induction to a formula involving the predicate 'I am disposed to believing *x*' (where *x* is an arithmetical sentence). Something similar can be said about the third reflective step. If your justification for the arithmetical instance of mathematical induction was a priori, then so will be, presumably, your justification for applying mathematical induction for a formula involving syntactic predicates.

In the course of your reflective process, you have done justificatory work. You might wonder if the theories in which these justificatory arguments are implicitly carried out, do not already entail the consistency of PA. If so, then your reflective process were circular.

But these worries are unfounded. The inductive argument in stage one is clearly carried out in a theory that is conservative over PA.[14] For the inductive argument in stage three this is slightly less straightforward, but we have seen that it is also carried out in conservative extension of PA.[15]

You can, of course, object to the mathematical induction argument in stage three, perhaps because you are loath to use predicates that are not fully arithmetical in the induction axiom. If this is your view, then your reflective process ends at the end of stage two at the latest, and you do not acquire a fundamentally new *arithmetical* belief. Nonetheless, if your reflective process carried you to the end of stage two, you have still acquired a fundamentally new entitled belief: the belief that PA is consistent. You can also object to the mathematical induction argument in stage one. In that case, you do not bring yourself to a belief in the consistency of PA by an reflective argument along the lines that I have sketched.

At the end of your reflective process, you have not *justified* $\neg Bew_{\mathrm{PA}}(\ulcorner \bot \urcorner)$. Where, exactly, in your reflective process, does the "drop" from justification

[14] More precisely, it is carried out in PA formulated in an extended language, with the induction axiom applying to the extended language.

[15] See Section IV.5.

to epistemic entitlement occur? It occurs at the moment in stage two where you form a belief in the consistency of PA *while* maintaining your unreserved acceptance of PA. This is a cognitive act for which you provide no justification. But, in the circumstances you are in, you are epistemically entitled to proceed in this way.

I take all this to be a vindication of ICT. But my epistemological analysis does not constitute evidence for Feferman's stronger thesis ICT^{+}. Throughout this article I have assumed (without argument) van Fraassen's 'liberal' conception of rationality (van Fraassen 1989, pp. 171–2):

> The difference [between Russell's traditional conception of rationality and the 'liberal' conception of rationality] is analogous to that between (or so Justice Oliver Wendell Holmes wrote) the Prussian and the English conception of law. In the former, everything is forbidden which is not explicitly permitted, and in the latter, everything permitted that is not explicitly forbidden. When Russell is still preoccupied with reasons and justification, he heeds the call of what we might analogously call the Prussian concept of rationality: what is rational to believe is exactly what one is rationally compelled to believe. I would opt instead for the dual: what is rational to believe includes everything that one is not rationally compelled to disbelieve. *Rationality is only bridled irrationality*.

In accordance with this view of rationality, I have insisted that it would not be irrational for you to refrain from following the reflective process of Section IV through to the end. I am *not* claiming that objecting to the inductive argument in stage three, or even to the more elementary inductive argument in stage one, would be irrational. At the point when you realise that you have been relying on the consistency of PA, you are rationally permitted to choose not unconditionally to rely on PA in the future and to revise your initial beliefs instead.

Moreover, going through the reflection process described in Section IV is not the *only* way of coming to know $\neg \mathit{Bew}_{\mathrm{PA}}(\ulcorner \bot \urcorner)$. To illustrate this, let us briefly back to the (admittedly somewhat naive) scenario that was briefly sketched in Section IV.1, where you have verified that the axioms of PA hold in the 'standard model'. Instead of going through the reflection process described above, you might go on to argue by mathematical induction, using a Tarskian compositional notion of truth, that all theorems of PA are *true*, and that therefore, since $\bot$ is not true, PA must be consistent. In this way, you might obtain *justified* (and not merely entitled) belief in $\neg \mathit{Bew}_{\mathrm{PA}}(\ulcorner \bot \urcorner)$. Or you might verify that the axioms of *second-order* number theory hold in the 'intended model', and *derive* $\neg \mathit{Bew}_{\mathrm{PA}}(\ulcorner \bot \urcorner)$ from them. Indeed, it is well-known that in this way the scope of mathematical knowledge can be extended in much more dramatic ways than by iterated consistency extensions.

Thus it is not part of the thesis that is defended in this article that the *only* successful way of arguing for the consistency of PA is based on the reflection process discussed in the previous section. But I do maintain that going through this reflection process is *one* way of coming to know $\neg Bew_{\mathrm{PA}}(\ulcorner \perp \urcorner)$. The interest of this particular reflection process is epistemological: it lies in the fact that you can acquire knowledge of new mathematical statements without justifying them, whereas adopting a new axiom, for instance, only gives you new knowledge if you have justification for it.

When you have come to the end of your reflective process, you have come to know a fundamentally new arithmetical statement: $\neg Bew_{\mathrm{PA}}(\ulcorner \perp \urcorner)$. But you have not come to know *that* it is fundamentally new, i.e., that it was not accessible to you in your state of innocence. This is because, in the first stage of the reflective process, you come to believe that what you are disposed to believe (as far as your mathematical work goes), includes PA, not that it *coincides* with PA.

In the state of innocence, you essentially "are" a Turing machine *e* that enumerates PA. You can acquire the first person knowledge that you are, insofar as mathematics is concerned, the machine *e*.[16] *How* can you come to know that as far as arithmetic is concerned, you are *e*? It does not happen by *intuition* or *direct introspection* (unless the meaning of those terms is stretched). All you have to go on is a finite set of *examples* of mathematical axioms that you believe, and theorems that you have come to believe by deriving them from the axioms using classical logic. Extrapolating from this finite collection of examples, you form the *hypothesis* that you are guided by *e*, and you come to believe this hypothesis. You are using some form of ampliative reasoning: we may call it *abduction*.

Your abductive argument is clearly fallible. We may suppose, however, that in this instance, you not only arrive at a true conclusion, but that in addition you are *justified* in believing this conclusion on the basis of your abductive considerations. That abductive arguments sometimes lead to justified beliefs is fairly widely accepted. But there is no consensus among epistemologists on *how* abductive arguments can generate justified beliefs. I have nothing to contribute to this large epistemological debate except to say that some reliabilist account is probably called for. This should not, however, be taken to imply that this reliabilist story must then account for *all* forms of knowledge. Indeed, it seems doubtful that the very same epistemological story that accounts for abductive reasoning will also account for your knowledge of the axioms of PA.

You might worry that it might not be possible for you to know that, as far as your mathematical work goes, you are a Turing machine. Lucas and Penrose have famously argued that it is not even possible for you to *be*, as far as your mathematical work goes, a Turing machine. But, as mentioned earlier, it is widely held that their arguments are unpersuasive. Reinhardt has argued that, as far as your mathematical work *and your knowledge of your own work goes*, for every Turing machine *e*, you cannot know that *e* enumerates what you know

[16] A brief discussion of this reflective is in (Franzén 2004, p. 216).

Reinhardt (1985). But here we are concerned only with your (true and justified) *arithmetical* beliefs. In this context, Reinhardt's considerations do not apply.

Despite all this, you may nonetheless be sceptical about the abductive argument given above. This would not make you irrational; it would just mean that you have not come to know *that* the new consistency beliefs that you have acquired in the reflective process described in Section IV are fundamentally new.

VI. SCOPE AND LIMITATIONS

In Section III I endorsed Dean's claim that finitism is an epistemically stable position in the foundations of mathematics. Moreover, it has been argued that accepting all of PA and no mathematics that goes beyond it is likewise a stable position: see Isaacson (1987). How are such claims compatible with the argument that was developed in Section IV? After all, in that section a reflection process is described by means of which someone who accepts all and only the principles of PA can come to know that PA is consistent. In addressing this question, I will now concentrate on finitism because this position has received a fair amount of attention in recent literature.[17]

Tait has argued—convincingly, in many scholars' view— that the extension of finitistically acceptable mathematics is captured by the system of Primitive Recursive Arithmetic: see Tait (1981). He also pointed out that the outer limits of finitism can only be seen from a vantage point that is external to finitism proper (Tait 1981, section IV).[18]

This does not make finitism internally unstable. Locally, the finitist can see of every proof principle of Primitive Recursive Arithmetic that it is justified. But she has no way of verifying that *only* the proof principles that are included in Primitive Recursive Arithmetic are legitimate. Indeed, such a claim involves a *general* concept of function, which the finitist does not have (Dean 2014, p. 53).

This means that the finitist does not accept all the steps of the reflection process that are described in Section IV. For instance, she will not accept the inductive argument in the first stage of the reflective process. After all, such an inductive argument is not a finitist proof![19] (The same holds, of course, for the inductive argument in stage three of the process.) In accordance with the 'British' conception of rationality that was discussed (and endorsed) earlier, I

[17] See for example (Parsons 2008, chapter 7), Dean (2014).

[18] See also (Dean 2014, section IV.1). Something similar can be said about Feferman-style predicativism.

[19] She could, however, come to believe that she is a "Primitive Recursive Arithmetic-machine" by abductive means. In that case, she could come to have an entitled belief in the consistency of Primitive Recursive Arithmetic. But, again, refusing to engage in the relevant abductive reasoning would not make her irrational.

maintain that thus refraining to accept some of the steps in the reflective process of Section IV does not make the finitist irrational. It is in this sense that, at least for all that what is said in this article, finitism is an epistemologically stable position.

What about the reflection process that can lead you to know a *strong* proof theoretic reflection principle, such as "everything that PA proves is true"? That reflection process is significantly more complex and requires a separate investigation. One key issue is that you may not possess the concept of *truth* for arithmetical sentences at the start of your reflective journey: it is a difficult question how you come to acquire it.

In (Cieśliński 2017, chapter 13), Cieśliński proposes an alternative account of the reflection process that leads to knowledge of the global reflection principle for a theory that one already accepts.[20] Central in Cieśliński's theory of implicit commitment is the notion of rational *believability*. The thought is that when a person reflects on the implicit commitments involved in her acceptance of a theory K, she comes to accept a theory of believability Bel(K) over K. Cieśliński explains how this process is structured, and he spells out Bel(K) as an *axiomatic* theory (Cieśliński 2017, p. 254). For instance, Bel(K) contains the principle

$$\forall \varphi \in \mathcal{L} : \mathsf{Bew}_{\mathrm{K}}(\ulcorner \varphi \urcorner) \rightarrow \mathsf{B}(\ulcorner \varphi \urcorner),$$

where B is the believability predicate, and $\mathcal{L}$ is the extended language containing the truth predicate and the believability predicate. He then shows that if K is a conservative disquotational truth theory, Bel(K) proves the believability of compositional truth laws and of reflection principles for K. From the believability of compositionality of truth and of global reflection, the cognitive agent then is entitled to infer to compositionality of truth and to global reflection *simpliciter*, *provided that there are no overriding reasons against doing so* (Cieśliński 2017, section 13.5).

There is a fundamental difference between Cieśliński's reflection process and mine. Global reflection, with which he is mainly concerned, involves the concept of truth. Moreover, Cieśliński's story about how you may rationally come to accept global reflection for a theory that you already justifiedly accept, is centred around the philosophical notion of rational believability. It may well be that the reflection processes that can lead you to accept reflection principles that are significantly stronger than consistency, inevitably require you to accept either new mathematical axioms or principles that involve philosophical notions. But if what was said in Section IV is right, then at least reflection on *consistency* does not involve the acceptance of a theory of rational believability,

[20] The account of the cognitive process involved in proof theoretic reflection of (Franzén 2004, chapter 14) is evaluated and found wanting in (Cieśliński 2017, section 13.1). A more philosophical account of acceptance of a mathematical theory and the process of reflection what you accept is found in Galinon (2014).

or even the notion of arithmetical truth. Indeed, one of the main claims of the present article is that reflecting on consistency does not require the acceptance of any *philosophical* notions whatsoever.

VII. REFLECTION AS A RATIONAL PROCESS

Reflection has a long history as a *philosophical* method for acquiring knowledge. Burge argues that classical rationalist philosophers attribute three cardinal properties to reflection (Burge 2013a, pp. 535–7):

(1) In reflection an individual brings to articulated consciousness steps or conclusions that are implicitly present, subliminally or unconsciously, in the individual's mind before reflection.
(2) Reflection can yield a priori knowledge of objective subject matters, beyond thoughts that the reflector is engaging in.
(3) Successful reflection requires skilful reasoning and is difficult: it is not a matter of one-off introspection or intuition.

Burge endorses theses 2. and 3., but rejects thesis 1. His main reason for disagreeing with thesis 1. is that reflection is more often than not applied in a situation where we do not have even an implicit, unclear, or confused idea or conception of a concept, but where we have not yet developed a concept at all. We may, for instance, merely have a small number of examples that we are inclined to see as similar in a way that we cannot describe. Or we may be disposed to classify a fairly well circumscribable number of examples as being similar in a significant way without this disposition being in any way conceptualised by us: think of this disposition as being hard-wired without an accompanying cognitive representation even at the sub-personal level.

In the present article, Burge's stance on theses 1., 2., and 3. can be taken to have been subjected to a *test* by applying it to a concrete example of reflection in the foundations of mathematics, viz., reflection on implicit commitments associated with the acceptance of mathematical theories. I claim that the process involved in proof theoretic reflection is in accordance with what Burge regarded as the cardinal properties of philosophical reflection. First, the reflective movements do not consist in drawing to the level of your consciousness representations that were already vaguely and subconsciously present. The beliefs that you form in reflection were not indistinctly, subliminally, or subconsciously, present in you mind in any way before the acts of reflection. Secondly, you have acquired knowledge not just about your mind or your commitment. You have acquired new knowledge about the world (of numbers): the statement $\neg Bew_S(\ulcorner \bot \urcorner)$ is, after all, a purely arithmetical proposition. Thirdly, reflection is a complicated process. I have concentrated on the *simplest* form of reflection, and the story already has a significant degree of epistemological complexity. It is even more complicated when we focus on stronger reflection principles.

We have seen how Wright concedes to the sceptic that we are not warranted to believe in the existence of the outside world without justifying this belief. Cognitive projects of a modest scale (such as finding out what time it is) face the same challenge as far as belief in their presuppositions is concerned. But here other cognitive projects may be able to help us out. A thorough inspection of the watch, for instance, might provide the basis for a warranted belief in its reliable functioning. The problem with the very large scale project of finding out about the external world is that it is hard to see what a cognitive project could look like that can provide a basis of warrant for belief in its presuppositions. It is far beyond the scope of this article to take a stance on this problem. But the process that was described in Section IV shows that the situation may not be entirely hopeless: the possibility that we may be able to obtain epistemic entitlement to believe without being able to obtain epistemic justification cannot be excluded from the outset.[21]

REFERENCES

Burge, T. (1997) 'Interlocution, Perception, and Memory', *Philosophical Studies*, 86: 21–47.

——— (1998) 'Computer Proofs, A Priori Knowledge, and Other Minds', *Philosophical Perspectives*, 12: 1–37.

——— (2013a) 'Reflection', in *Cognition Through Understanding*, 534–55. Oxford: OUP.

——— (2013b) 'Epistemic Warrant: Humans and Computers', in *Cognition Through Understanding*, 489–507. Oxford: OUP.

Carlson, T. (2000) 'Knowledge, Machines, and the Consistency of Reinhardt's Strong Mechanistic Thesis', *Annals of Pure and Applied Logic*, 105: 51–82.

Cieśliński, C. (2017) *The Epistemic Lightness of Truth*. Cambridge: CUP.

Cohen, L. (1992) *An Essay on Belief and Acceptance*. Oxford: OUP.

Dean, W. (2014) 'Arithmetical Reflection and the Provability of Soundness', *Philosophia Mathematica*, 23: 31–64.

Feferman, S. (1962) 'Transfinite Recursive Progressions of Axiomatic Theories', *Journal of Symbolic Logic*, 27: 259–316.

——— (1988) 'Turing in the Land of O(z)', in R. Herken (ed.) The Universal Turing Machine, 113–47. Oxford: OUP.

Franzén, T. (2004) 'Inexhaustibility. A Non-exhaustive Treatment', Wellesly: AK Peters.

Galinon, H. (2014) 'Acceptation, Cohérence et Responsabilité', in J. Dutant *et al.* (eds.) *Liber amicorum Pascal Engel*, 320–33. Genéve: Université de Genève.

Graham, P. (2020) 'What is Epistemic Entitlement? in J. Greco and C. Kelp (eds.) *Virtue-theoretic Epistemology: New Methods and Approaches*, 93–123. Cambridge: CUP.

[21] Versions of this article have been presented at a number of research seminars and conferences: Lingnan University (2018), University of Salzburg (2018), Oxford University (2018), Tokyo Forum for Analytical Philosophy (2018), Munich Center for Mathematical Philosophy (2019), UNICAMP (2019). I am grateful to the participants in those events for helpful comments and questions. In particular, this article has benefited from suggestions and objections by Rafael De Clercq, Cezary Cieśliński, Matteo Zicchetti, Hazel Brickhill, Kentaro Fujimoto, Carlo Nicolai, Henri Galinon, Giorgio Venturi, and three anonymous referees.

Horsten, L. and Leigh, G. (2017) 'Truth is Simple', *Mind*, 501, 195–232.

Isaacson, D. (1987) 'Arithmetical Truth and Hidden Higher-Order Concepts', In *Logic Colloquium '85*, edited by the Paris Logic Group, 147–69. Amsterdam: North-Holland.

Kreisel, G. (1970) 'Principles of Proof and Ordinals Implicit in Given Concepts', *Studies in Logic and the Foundations of Mathematics*, 60: 489–516.

Kreisel, G. and Levy, A. (1968) 'Reflection Principles and Their Use for Establishing the Complexity of Formal Systems', *Zeitschrift für Mathematische Logik und Grundlagenforschung der Mathematik*, 14: 97–142.

Nicolai, C. (2013) *Truth, Deflationism, and the Ontology of Expressions: An Axiomatic Study*. DPhil Thesis, Oxford University.

Parsons, C. (2008) *Mathematical Thought and Its Objects*. Cambridge: CUP.

Reinhardt, W. (1985) 'Absolute Versions of Incompleteness Theorems', *Noûs*, 19: 317–46.

Tait, W. (1981) 'Finitism', *Journal of Philosophy*, 78: 524–46.

van Fraassen, B. (1980) *The Scientific Image*. Oxford: OUP.

——— (1989) *Laws and Symmetry*. Oxford: OUP.

Wright, C. (2002) '(Anti-)sceptics Simple and Subtle: G.E. Moore and John McDowell', *Philosophy and Phenomenological Research*, 65: 330–48.

——— (2004) 'Warrant for Nothing (and Foundations for Free)?' *Proceedings of the Aristotelian Society Supplemental Volume LXXVIII*, 167–212.

——— (2012) 'Replies: Part IV', in A. Coliva (ed.) *Mind, Meaning, and Knowledge. Themes from the Philosophy of Crispin Wright*, 451–86. Oxford: OUP.

Universität Konstanz, Germany

E-mail: *leon.horsten@uni-konstanz.de*