

Truth, Reflection, and Hierarchies*

Michael Glanzberg
Department of Linguistics and Philosophy
MIT
Cambridge, MA 02139
USA
glanzber@mit.edu

To appear in *Synthese*.

*Thanks to Jody Azzouni, Matti Eklund, Chris Gauker, Volker Halbach, Richard Heck, Peter Koellner, Vann McGee, Agustín Rayo, Carol Voeller, Ralph Wedgwood, David Wiggins, and Steve Yablo for comments on earlier drafts of this paper, and for many helpful conversations. A version of this material was presented at the Boston Colloquium for Philosophy of Science. Thanks to my audience there for valuable questions and comments.

It is often noted that Tarski's [1935] hierarchy of languages and metalanguages fragments the concept of truth. Instead of one concept, we have infinitely many, arranged in a hierarchy. Subsequent work along Tarskian lines [e.g. Parsons, 1974b; Burge, 1979; Barwise and Etchemendy, 1987] has proposed, in various ways, to base hierarchies on what speakers can express in a given context, rather than on multiple languages or concepts of truth *per se*. In my own work [2001; MS], I argue that a hierarchy of domains of propositions is in fact generated by the workings of linguistic context. I thus argue that natural language considerations lend some plausibility to a hierarchical approach. But the objection of fragmentation still stands. One way or another, hierarchical theories all require that speakers cannot in any one instance express the entirety of a unified concept of truth.

In this essay, I shall defend hierarchical approaches in general against the objection of fragmentation. I shall do so in two ways. First, I shall argue that the fragmentation required is familiar. We see just the same sort of fragmentation in the concept of *mathematical proof*. In the case of mathematical proof, we find the fragmentation banal, or at least unthreatening, and perhaps even a source of richness of the concept. It is not problematic there, and it is not problematic in the case of truth either. Second, in the course of this argument, I shall attempt to make clearer the source and nature of the fragmentation of these concepts. It has been observed both for proof [Kreisel, 1970] and for truth [Parsons, 1974a,b] that fragmentation arises because these concepts exhibit a kind of failure of closure under reflection. Very roughly, any sufficiently precise articulation of either concept allows for a kind of reflection upon the correctness of that articulation that leads to a distinct, stronger one. I shall here investigate this reflection. Building on the work of Kreisel and Parsons, I shall offer a more precise characterization of it in the setting of formal theories of truth, and I shall show that this helps us to understand the hierarchical nature of the concept of truth in a more general setting.

My arguments below will focus on a particularly strong and rarely noticed form of the Liar paradox, which I have dubbed the *Fortified Liar paradox*. As the Strengthened Liar paradox shows that a natural answer to the Liar paradox leads back to paradox, the Fortified Liar shows that a natural answer to the Strengthened Liar leads back to paradox as well. I shall argue that the Fortified Liar helps to make clear why the concept of truth is hierarchical, and helps explain the sense in which the hierarchy is generated by a process of reflection.

This essay will proceed as follows. In Section (1), I shall introduce the Strengthened Liar as a problem for theories of truth that see the truth predicate as *partial*. I shall develop this problem in the setting of formal theories of truth. Then in Section (2), I shall present the Fortified Liar, which shows that a natural response to the Strengthened Liar coming from partiality theories fails. The failure of this response will point the way towards a precise characterization of the process of reflection that makes truth hierarchical. Section (2) will involve a modestly technical discussion of *reflection principles* in proof theory. In Section (3) I shall show that the situation that emerges in Section (2) is similar to the one well-known for mathematical proof. I shall claim on this basis that the objection of fragmentation loses its force. In Section (4), I shall go on to argue that the conclusions reached in previous sections are robust, and not merely generated by the limitations of formal theories. Finally, I shall conclude in Section (5) with a brief comparison of the view I offer here with some important opposing views.

1 Partiality and the Strengthened Liar

Tarski's [1935] response to the Liar paradox is drastic. It bans any application of the truth predicate to itself (more properly, to sentences containing it, but I shall abuse notation and talk about predicates and concepts like truth applying to themselves). This appears to be unacceptable, in light of the many natural language examples where applying the truth predicate to itself is entirely harmless, or even quite useful. One of the leading ideas for avoiding such a drastic step is to conclude that the paradox shows us that the concept of truth is *partial*. There are sentences, such as the Liar sentence, which fall outside the domain of application of the predicate.

This idea is compatible with more liberal hierarchies than Tarski's. In other work [MS], I have used partiality techniques to allow for a reasonable amount of self-application of semantic concepts within levels of a hierarchy. Kripke [1975] ultimately admits to a hierarchy of truth predicates, as does Soames [1999]. But others, including Parsons [1984] (Terence Parsons), Reinhardt [1986], and McGee [1991], argue that the right sort of partiality, properly understood, avoids the need for any hierarchy. Showing where their strategy goes wrong will help make the nature of the hierarchy more clear.

The standard response to the partiality-based, anti-hierarchy view is the Strengthened Liar paradox. This is a species of reasoning, which observes

that any partiality theory must ensure that the Liar sentence does not come out true. But then, we have learned from the theory that the Liar sentence is not true. This conclusion is just the Liar sentence, so we are back in paradox.

Though this argument is intuitively compelling, it is also in some ways problematic. It requires reasoning about a given ‘solution’ to the paradox. Those opposed to hierarchies maintain that it is *mistaken* reasoning, arising from a misunderstanding of the way in which the truth predicate is partial.

I shall argue in Section (2) that the Strengthened Liar cannot be blocked. First, it will be useful to formulate the matter in a more formal setting. Let \mathcal{L} be the usual language of arithmetic, and \mathcal{L}^{Tr} be the language extended by adding a predicate $\lceil Tr \rceil$.¹ Let PA^{Tr} be the usual theory of Peano arithmetic with the induction schema extended to \mathcal{L}^{Tr} . To introduce some further customary notation, let the term $\lceil \phi \rceil$ be an appropriate Gödel code for $\lceil \phi \rceil$, and let the function \dot{x} have as value the $x + 1$ st formal numeral for input x .

As Tarski noted long ago, the obvious axiom to govern $\lceil Tr \rceil$ is the inconsistent T-schema:

$$(T) \quad \forall x (Tr(\lceil \phi \dot{x} \rceil) \leftrightarrow \phi x).$$

We need some consistent principles for $\lceil Tr \rceil$ to replace (T). According to the partiality view, we will find them by looking for principles which embody the partiality of the truth predicate.

There have been quite a few proposals for how to pursue this idea. For our purposes, an elegant and useful one is to replace the inconsistent (T) with a collection of rules of inference. Let TP (for *Truth Partial*) contain PA^{Tr} and be closed under the following:

$$(TrInf) \quad \frac{TP \vdash Tr(\lceil \phi \dot{x} \rceil)}{TP \vdash \phi x} \quad \frac{TP \vdash \neg Tr(\lceil \phi \dot{x} \rceil)}{TP \vdash \neg \phi x}$$

$$\frac{TP \vdash \phi x}{TP \vdash Tr(\lceil \phi \dot{x} \rceil)} \quad \frac{TP \vdash \neg \phi x}{TP \vdash \neg Tr(\lceil \phi \dot{x} \rceil)}.$$

(I have put these rules in parameterized form, allowing a free variable $\lceil x \rceil$ to occur in $\lceil \phi \rceil$. This is quite commonplace (see, for instance, Friedman

¹In my own more systematic work, I insist on applying the truth predicate to propositions, and argue for a hierarchy of domains of propositions expressible in contexts. I will here acquiesce in talking about a truth predicate applied to sentences, as it makes the hierarchical structure plain, and will facilitate comparisons with proof.

and Sheard [1987]), and seems to do little more than avoid an unnecessary restriction on formulas. I believe doing so captures the philosophical idea behind the rules accurately. Proof-theoretically, however, it has consequences which I exploit below.)

TP embodies the idea that $\lceil Tr \rceil$ is partial in the following way. For some sentences, we have $TP \vdash Tr(\ulcorner \phi \urcorner)$. These are true. For some, we have $TP \vdash \neg Tr(\ulcorner \phi \urcorner)$ (equivalently, $TP \vdash Tr(\ulcorner \neg \phi \urcorner)$). These are false. But for some sentences, we have neither.² Let $\lceil \lambda \rceil$ be the Liar sentence, i.e. a sentence such that:

$$TP \vdash \lambda \leftrightarrow \neg Tr(\ulcorner \lambda \urcorner).$$

The diagonal lemma guarantees that $\lceil \lambda \rceil$ exists. As TP is known to be consistent [Friedman and Sheard, 1987], we know that $TP \not\vdash Tr(\ulcorner \lambda \urcorner)$ and $TP \not\vdash \neg Tr(\ulcorner \lambda \urcorner)$. In this sense, we can observe that $\lceil \lambda \rceil$ is outside the domain of application of $\lceil Tr \rceil$ according to TP .³

We may use TP to give a more precise version of the Strengthened Liar reasoning. We have noted that $TP \not\vdash Tr(\ulcorner \lambda \urcorner)$. Furthermore, TP is designed exactly to insure this result, as that is how it can be consistent. The same holds for any consistent extension of TP . Hence, the issue is not mere incompleteness. We may thus conclude that according to TP , $\lceil \lambda \rceil$ is not true. Insofar as TP is the correct theory of truth, this shows $\lceil \neg Tr(\ulcorner \lambda \urcorner) \rceil$. We are now back in paradox.

Now, what is controversial about this reasoning is evident. It requires reasoning about our theory TP . (Indeed, if this reasoning could be carried

²Though TP describes a partial truth predicate, the background logic is entirely classical. The partiality of TP is embodied in what it does and does not prove, not some alternative logic. The specification of TP via the rules of (TrInf) is done for ease of exposition and for comparison with (T), but it is not the most concise formulation available. As Volker Halbach pointed out to me, the elimination rules of (TrInf) (the top line) appear to be redundant. Rules like (TrInf) figure prominently in McGee [1991], but in a somewhat different way, as a source of adequacy conditions for theories.

³ TP is a straightforward statement of a partiality theory, but it is extremely weak. A natural extension of it, adding some further basic principles about $\lceil Tr \rceil$ and a formalized statement of bivalence, is known to be a conservative extension of PA [Friedman and Sheard, 1987]. A further extension, adding conditionals stating that quantifiers commute with $\lceil Tr \rceil$, is also consistent but ω -inconsistent [Cantini, 1996]. Kripke fixed point techniques may be used to build natural models of TP . However, in this case, the weakness of the theory allows for the easy construction of an ω -model. $\langle \mathbb{N}, \{\ulcorner \phi \urcorner \mid TP \vdash Tr(\ulcorner \phi \urcorner)\} \rangle \models TP$ [Friedman and Sheard, 1988]. Many theories stronger than TP are known [see Friedman and Sheard, 1987; Feferman, 1991; Cantini, 1996].

out in TP , it would be inconsistent.) Those who see the argument as sound will point out that in spite of being about TP rather than in TP , the reasoning appears perfectly good. Insofar as TP is the theory that tells us about truth, and insofar as it is correct, we can learn from what it does with $\lceil \lambda \rceil$ that $\neg Tr(\lceil \lambda \rceil)$.

Defenders of hierarchical approaches see this reasoning as indicating a step in the hierarchy. Different theories explain the step differently, but they all agree that the reasoning is correct, and non-paradoxical, because the conclusion at the end of the Strengthened Liar reasoning is made one level higher up in the hierarchy than the Liar sentence itself.

2 The Fortified Liar and Reflection Principles

The anti-hierarchy view, obviously, does not accept this conclusion. It responds that there is a mistake in the Strengthened Liar reasoning, and no need for a hierarchy. I shall now explain why I do not think this response works. This will help clarify the nature of the hierarchy towards which the Strengthened Liar already points.

The mistake the anti-hierarchy view sees in the Strengthened Liar reasoning lies in going from $TP \not\vdash Tr(\lceil \lambda \rceil)$ to the conclusion that $\lceil \lambda \rceil$ is not true (a conclusion that only makes sense one level up in a hierarchy). The anti-hierarchy response runs something like this: “The reasoning misuses TP . The conclusion that some sentence is true is only licensed by TP when TP proves it to be true. Likewise, the conclusion that some sentence is not true is only licensed when TP proves it not to be true. By design, TP proves nothing about $\lceil \lambda \rceil$. To draw *any* conclusion from this is to misunderstand the partiality of $\lceil Tr \rceil$. Hence, the step in the Strengthened Liar of concluding $\lceil \neg Tr(\lceil \lambda \rceil) \rceil$ is a mistake.”

The suggestion here is not that we should conflate truth with provability. It is rather to insist that TP is already a correct theory of truth (nearly enough for current purposes), and to enjoin us not to step beyond what it says. A theory of truth provides basic principles that govern the concept of truth, and explains how inferences may be drawn from them. TP provides (TrInf) as the basic principle of truth, which together with the usual rules of first-order logic and the auxiliary PA^{Tr} , guides the ways we are to reason

with truth. (Actually, for most interesting claims, we will have to see TP as working together with other theories not about truth—physics, chemistry, and so on—to tell us what is true. But in the case of $\lceil \lambda \rceil$, and all the other sentences that will be at issue here, the syntax expressible in PA^{Tr} is enough, so we may suppress this point for current purposes.)

TP thus offers an analysis of the concept of truth, and at the same time an analysis of how that concept may be deployed in reasoning. Insofar as it is a correct analysis, it appears that correct conclusions about what is true are conclusions that may be proved in this theory. In response to the Strengthened Liar, the anti-hierarchy view claims that the analysis is correct. The crucial step in the Strengthened Liar inference goes beyond the analysis, and so it is a mistake. As the analysis is given as a formal theory TP , the proper way to formulate this response is as the claim that we may conclude a sentence is true when it is proved to be true by TP , and we may conclude a sentence is not true when it is proved not to be true by TP . TP 's silence is not enough.

In answer to this response to the Strengthened Liar, I shall offer what I call the *Fortified Liar*. The Fortified Liar plays the very response we have just seen back against the anti-hierarchy view, much as the Strengthened Liar plays back the partiality of the truth predicate.

The response to the Strengthened Liar insists on the correctness of TP by offering three principles: (i) if $\lceil Tr(\ulcorner \phi \urcorner) \rceil$ is provable in TP , $\lceil \phi \rceil$ is true; (ii) if $\lceil \neg Tr(\ulcorner \phi \urcorner) \rceil$ is provable in TP , then $\lceil \phi \rceil$ is not true (is false); (iii) that is all the theory tells us. As is customary when working with partial predicates, we may take the third principle as tacitly given with the first two. The first two may be formalized in \mathcal{L}^{Tr} , as:

$$\begin{aligned} (\text{Tr-RFN}) \quad & \text{Prov}_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner) \rightarrow Tr(\ulcorner \phi \dot{x} \urcorner) \\ (\neg\text{Tr-RFN}) \quad & \text{Prov}_{TP}(\ulcorner \neg Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner) \rightarrow \neg Tr(\ulcorner \phi \dot{x} \urcorner). \end{aligned}$$

(Here $\lceil \text{Prov}_{TP} \rceil$ is the canonical predicate expressing provability in TP .)

These express the correctness of TP as a theory of *truth*. They have a familiar form. They are special cases of the uniform reflection principle for TP :

$$(\text{RFN}_{TP}) \quad \text{Prov}_{TP}(\ulcorner \phi \dot{x} \urcorner) \rightarrow \phi x.$$

(‘Uniform’ to indicate the presence of parameters.) This principle expresses the soundness of TP , and thus captures the correctness of the theory.

(Tr-RFN) and (\neg Tr-RFN) do the same for formulas starting with $\lceil Tr \rceil$ or $\lceil \neg Tr \rceil$, i.e. for TP as a theory of truth.

It is the status of these principles that poses a problem. The problem arises even for (Tr-RFN) alone, so I shall discuss only it. In responding to the Strengthened Liar reasoning, the anti-hierarchy view offers us (Tr-RFN) as part of an explanation of where the reasoning goes wrong. But as such, this principle must be properly assertible. The norms of assertion require us to only assert what we take to be true. But by the very view being offered, the only ground for truth there can be is the provability of truth in TP . Hence, the explanation *requires* the provability of the truth of (Tr-RFN) in TP , for the explanation to be acceptable by its own lights. In order for the anti-hierarchy view to make the response offered, it must thus have:

$$(\text{Prov-Tr-RFN}) \quad TP \vdash Tr(\ulcorner Prov_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner) \urcorner \rightarrow Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner).$$

This the theory cannot have. One application of (TrInf) yields:

$$TP \vdash Prov_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner) \rightarrow Tr(\ulcorner \phi \dot{x} \urcorner).$$

If this holds, then TP is *inconsistent*. From Löb's theorem, it follows that $TP \vdash Tr(\ulcorner \lambda \urcorner)$, and we are back in paradox.⁴

This should hardly come as a surprise. (RFN $_{TP}$) is a strong consistency statement for TP . (Its restriction to Π_1 formulas is equivalent to $\lceil Con_{TP} \rceil$ (the canonical consistency sentence for TP) [see Smorynski, 1977].) The second incompleteness theorem makes clear that we cannot have $TP \vdash \text{RFN}_{TP}$. We have simply observed that the special case (Tr-RFN) already cannot be proved. On the other hand, we are now beginning to see what is fundamentally wrong with the anti-hierarchical response. It is a claim about the correctness of TP , and thus a kind of soundness claim. But by the lights of the anti-hierarchical response, I have argued, it is a soundness claim that must be provable. We generally should not expect to be able to have this, and we have seen we cannot in this case.

In reply to this argument, it might be suggested that I have yet again under-appreciated the partiality of TP . Formally, the partiality of TP is

⁴Reinhardt [1986] offers a version of the response under consideration, and explicitly states a principle like (Tr-RFN). He is well aware of the technical situation, and disavows (Prov-Tr-RFN). I shall discuss this further in Section (5).

expressed by formulating principles as inference rules. So, perhaps, the anti-hierarchy view might require not (Prov-Tr-RFN) but the rule:

$$\text{(Tr-RFN-Rule)} \quad \frac{TP \vdash Prov_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner)}{TP \vdash Tr(\ulcorner \phi \dot{x} \urcorner)}.$$

It turns out this rule is no weaker than (Prov-Tr-RFN), in that they both imply that TP contains the full (RFN_{TP}) , and so is inconsistent.

This follows from a basic result. Let $TP + \text{Tr-RFN}$ be the theory that results from adding (Tr-RFN) to the axioms of TP . Let TPR be the theory that results from adding the rule (Tr-RFN-Rule) to the rules of TP . (It is crucial that the new rules and axioms be for $\lceil Prov_{TP} \rceil$, not the provability predicate for the resulting stronger theory. As we will see below, the latter makes the theory inconsistent.) Then we have:

Proposition 1 $TP + \text{Tr-RFN} \equiv TP + RFN_{TP} \equiv TPR$.

Clearly, to prove this, it suffices to show that $TPR \vdash RFN_{TP}$.

The proof of this is a modification of the corresponding result of Feferman [1962]. It follows from some simple lemmas. The first is also proved in Feferman [1962].

Lemma 2 $PA \vdash Prov_{TP}(\ulcorner Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, y) \urcorner \rightarrow \phi \dot{x} \urcorner)$.

$\lceil Proof_{TP}(x, y) \rceil$ is the canonical predicate expressing that y is a proof of x . $PA \vdash Prov_{TP}(x) \leftrightarrow \exists y Proof_{TP}(x, y)$.

The second is an observation about provability in TP . A proof of $\lceil \phi x \rceil$ can be extended to a proof of $\lceil Tr(\ulcorner \phi \dot{x} \urcorner) \rceil$, and a proof of $\lceil Tr(\ulcorner \phi \dot{x} \urcorner) \rceil$ can likewise be extended to a proof of $\lceil \phi x \rceil$, by one application of the rules of (TrInf). Formalizing this gives the second lemma.

Lemma 3 *There are primitive recursive functions e and f such that:*

$$PA \vdash Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, y) \rightarrow Proof_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner, \dot{e}(y))$$

and

$$PA \vdash Proof_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner, y) \rightarrow Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, \dot{f}(y)).$$

Hence,

$$PA \vdash Prov_{TP}(\ulcorner \phi \dot{x} \urcorner) \leftrightarrow Prov_{TP}(\ulcorner Tr(\ulcorner \phi \dot{x} \urcorner) \urcorner).$$

(Here $\lceil e \rceil$ and $\lceil f \rceil$ are formulas that represent e and f .)

These lemmas combine to give us Proposition (1). They imply:

$$PA \vdash Prov_{TP}(\ulcorner Tr(\ulcorner Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, y) \rightarrow \phi \dot{x} \urcorner) \urcorner).$$

As TPR is closed under (Tr-RFN-Rule), we have:

$$TPR \vdash Tr(\ulcorner Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, y) \rightarrow \phi \dot{x} \urcorner).$$

From (TrInf) we then have:

$$TPR \vdash Proof_{TP}(\ulcorner \phi \dot{x} \urcorner, y) \rightarrow \phi x.$$

By logic, we then have:

$$TPR \vdash Prov_{TP}(\ulcorner \phi \dot{x} \urcorner) \rightarrow \phi x.$$

As an immediate corollary, we have:

Corollary 4 *If (Prov-Tr-RFN) or (Tr-RFN-Rule) holds for TP , then TP is inconsistent.*

Either of these implies $TP \vdash RFN_{TP}$, which implies $TP \vdash Con_{TP}$.

We now have the Fortified Liar. This is the reasoning which notes that in reply to the Strengthened Liar, the anti-hierarchy view offers (Tr-RFN). But for this claim to be acceptable, by the lights of the response being offered, the view must have (Tr-RFN-Rule), if not (Prov-Tr-RFN). Corollary (4) shows that this leads back to contradiction.⁵

⁵Allowing parameters in rules is crucial for Proposition (1), but not for the inconsistency of (Prov-Tr-RFN). It has sometimes been suggested to me that the anti-hierarchy view might then just reply by refusing to allow parameters in (TrInf). There are several reasons why I do not see this as a satisfactory response. First of all, as I already mentioned, I just do not see why the presence of parameters in these sorts of rules is philosophically suspect. But perhaps more importantly, I think it is clear that insofar as we are thinking of these rules as really capturing something about proof, the presence of parameters is above reproach. It is just not a problem to do a proof carrying along a parameter. Perhaps the anti-hierarchy view would want to replace this ordinary understanding of the kinds of rules we have been looking at with some other. But then, the difference between a rule like (Tr-RFN-Rule) and the conditional (Prov-Tr-RFN) becomes mysterious. In that case, the problems with (Prov-Tr-RFN) seem to me to be philosophically compelling alone. In proposing rule formulations like (TrInf) and (Tr-RFN-Rule), the anti-hierarchy view is

However, we learn more from the Fortified Liar than merely that the paradox may be reinstated. Proposition (1) shows us what the reasoning involved in the response to the Strengthened Liar entails. It leads to the full (RFN_{TP}) . To avoid a reinstated paradox, we must construe this reasoning as happening one step up in the hierarchy. Thus, this level includes (RFN_{TP}) .

This helps to make clear just what the step in the hierarchy involves. Typical steps in the hierarchy, as witnessed both by the Strengthened Liar and the Fortified Liar, involve a kind of reflection upon the theory of truth, which amounts to drawing conclusions from the correctness of the theory. In examining the Fortified Liar, we have seen that in fact such a step can involve the statement of the soundness of the theory, as expressed by (RFN_{TP}) . The Fortified Liar thus shows us that a good model of the next step in the hierarchy from TP is given by $TP + \text{RFN}_{TP}$.

With this model, we can indicate how to make sense of the Strengthened Liar as well. The Strengthened Liar draws a conclusion about the truth of $\lceil \lambda \rceil$ from $TP \not\vdash Tr(\ulcorner \lambda \urcorner)$. In $TP + \text{RFN}_{TP}$, we can state this reasoning directly. As $TP + \text{RFN}_{TP} \vdash Prov_{TP}(\ulcorner Tr(\ulcorner \lambda \urcorner) \urcorner) \rightarrow \neg Con_{TP}$, we have:

$$TP + \text{RFN}_{TP} \vdash \neg Prov_{TP}(\ulcorner Tr(\ulcorner \lambda \urcorner) \urcorner).$$

Now, the identification of the first-level truth predicate inside the second level is an extremely delicate matter (hence the appearance of paradox). In other work [MS], I develop the idea of an *internal truth predicate* to explain it. But very roughly, from the perspective of $TP + \text{RFN}_{TP}$, we are thinking of first-level truth as provable truth in TP , so we should think of first-level truth as corresponding to $\lceil Prov_{TP}(\ulcorner Tr(x) \urcorner) \rceil$. Under this rough identification, we may think of $\lceil \neg Prov_{TP}(\ulcorner Tr(\ulcorner \lambda \urcorner) \urcorner) \rceil$ as expressing the Liar sentence. Hence, the conclusion above leads to the truth of the Liar sentence.

already relying on some ideas of a proof-theoretic nature. It is not then acceptable to reject other natural ideas about proof, such as allowing parameters, when it becomes a problem.

That being said, there are some issues related to parameters that deserve more extensive discussion, but which I shall merely mention. First, as is already obvious from results like Proposition (1), parameters can have significant proof-theoretic strength. In particular, uniform reflection principles like (RFN_{TP}) are equivalent to versions of a formalized ω -rule. Second, in allowing parameters, I am talking as much about satisfaction as about truth *per se*. It is commonplace to see the differences between these as slight, but the proof-theoretic strength of parameters (and other differences that definability theory reveals) could raise questions about this. (Some further discussion of reflection principles in theories of truth, and of parameters, may be found in Halbach [2001].)

A slightly more accurate picture is given if we directly build a fixed point $\ulcorner \lambda' \urcorner$ such that:

$$TP + \text{RFN}_{TP} \vdash \lambda' \leftrightarrow \neg \text{Prov}_{TP}(\ulcorner \text{Tr}(\ulcorner \lambda' \urcorner) \urcorner).$$

By Lemma (3), $TP + \text{RFN}_{TP} \vdash \text{Prov}_{TP}(\ulcorner \text{Tr}(\ulcorner \lambda' \urcorner) \urcorner) \rightarrow \text{Prov}_{TP}(\ulcorner \lambda' \urcorner)$. From (RFN_{TP}) and the definition of $\ulcorner \lambda' \urcorner$, we have $TP + \text{RFN}_{TP} \vdash \text{Prov}_{TP}(\ulcorner \lambda' \urcorner) \rightarrow \lambda' \rightarrow \neg \text{Prov}_{TP}(\ulcorner \text{Tr}(\ulcorner \lambda' \urcorner) \urcorner)$. Hence, $TP + \text{RFN}_{TP} \vdash \neg \text{Prov}_{TP}(\ulcorner \text{Tr}(\ulcorner \lambda' \urcorner) \urcorner)$, i.e. $TP + \text{RFN}_{TP} \vdash \lambda'$. With closure under (TrInf) , we then have $TP + \text{RFN}_{TP} \vdash \text{Tr}(\ulcorner \lambda' \urcorner)$. We thus see that $TP + \text{RFN}_{TP}$ proves the truth of the (reconstructed) Liar sentence.

The identification of first-level truth with provable truth is highly specific to the proof-theoretic framework in which we are working. As I mentioned, I think a more refined model-theoretic construction can give us a more accurate picture of natural language occurrences of this sort of reasoning. (I shall discuss how this approach relates to the proof-theoretic framework in Section (4)). But $TP + \text{RFN}_{TP}$ does provide for just the sort of reasoning we use in the Strengthened Liar and in the Fortified Liar, and it accurately captures the kind of reflection upon the soundness of TP that generates this reasoning. It thus provides a clear expression of what sort of reflection takes us to the next level of the hierarchy. I propose we take the step from TP to $TP + \text{RFN}_{TP}$ as our model of the hierarchical nature of truth.⁶

A step in the hierarchy is generated by reasoning about the correctness of the previous level, as we see with the Strengthened Liar, and more clearly with the Fortified Liar. The result is the step from TP to $TP + \text{RFN}_{TP}$. This is a strikingly familiar pattern. It is just the pattern we have come to expect for mathematical proof in the aftermath of the collapse of Hilbert's program. To show that the hierarchy that has emerged is unproblematic, I shall investigate this similarity further in the next section.

3 Proof and Truth

Hilbert and his coworkers were confident that the concept of mathematical proof was closed under a process of reflection much like the one we saw in the

⁶Because the uniform reflection principle is equivalent to a formalized ω -rule, there is some connection between this proposal and ideas considered by Tarski [1935] for languages of infinite order.

last section. They argued as follows. A proof is a finite object. When properly formalized, proof thus falls within the domain of finitist mathematics. Any reflection on proof thus falls within the domain of a highly restricted sort of proof.⁷ As a result, the Hilbert school expected to be able to give consistency proofs for all of classical mathematics within finitist mathematics, thus protecting “the paradise that Cantor created for us” [Hilbert, 1926, p. 376] from the threat of paradox. Actually, Hilbert wanted more. He hoped to prove that classical mathematics is verifiably correct on its finitist fragment. Hence, he needed the finitistically provable soundness of classical mathematics for finitist statements. This amounts to the conservativity of classical over finitist mathematics.⁸

Of course, the incompleteness theorems doomed Hilbert’s program in its original form. The first incompleteness theorem shows that there will be no single system that formalizes all of classical mathematics, and provides theories that are not conservative over even PA . But the core of Hilbert’s program, securing the consistency of a reasonable system for classical mathematics, such as analysis or ZFC , in finitist mathematics, is really undermined by the second incompleteness theorem. Most any such system cannot be proved consistent in finitist mathematics.⁹

More specific to our interests, the second incompleteness theorem shows the argument that proof is closed under reflection to be mistaken. Like the argument we saw above, this one amounts to the provability of a reflection principle. Let F be some appropriate theory of finitist mathematics, and C a theory of classical mathematics.¹⁰ A proof in C is a finitist matter, so

⁷Such an argument is given in Hilbert [1926], and virtually repeated in Hilbert [1928]. A similar argument is given in Bernays [1930].

⁸Here we see Hilbert’s use of the notion of “ideal elements.” The interpretation of this notion is somewhat controversial. (See Mancosu [1998b] for a survey.)

⁹As Kreisel [1968] noted, this does not destroy all interest in Hilbert’s proof theory, but rather makes the choice of theories a crucial issue. Many proof theorists today describe their projects as “relativized” versions [Feferman, 1988a] or “partial realizations” [Simpson, 1988] of Hilbert’s program. It is well-known that in the paper that announced the incompleteness theorems [Gödel, 1931], Gödel claimed his results did not undermine Hilbert’s program. His unpublished work [Gödel, 1995] shows this was not his final view of the matter. It appears he wavered on this point because of questions about whether the needed consistency proofs could be carried out in intuitionistic mathematics, and whether this would go beyond finitist mathematics. (See Feferman [1988a] and Sieg [1988] for some discussion.)

¹⁰Kreisel [1960] argued that finitist mathematics corresponds to PA , but the consensus seems to be to follow Tait [1981] in identifying it with PRA . This is already enough to

$\lceil Prov_C \rceil$ is a predicate of finitist mathematics. Assuming the language of C extends that of F , the finitist correctness of $\lceil Prov_C \rceil$ can be expressed by the instances for each finitist sentence $\lceil \phi \rceil$ of:

$$(Rfn_C) \quad Prov_C(\lceil \phi \rceil) \rightarrow \phi.$$

This is a strong enough correctness statement to entail $\lceil Con_C \rceil$.

The closure argument concluded that because this schema is a finitist matter, if $\lceil Prov_C \rceil$ is correct we should have $F \vdash Rfn_C$ (more properly, each finitist instance of (Rfn_C)). In the special case of reflection on proof in F , we have have the local reflection principle for F :

$$(Rfn_F) \quad Prov_F(\lceil \phi \rceil) \rightarrow \phi.$$

(‘Local’ to indicate it is parameter-free.¹¹) Again, the closure argument concludes we should have $F \vdash Rfn_F$. The second incompleteness theorem tells us for any reasonable choice of F , we cannot have this. We thus see that proof in general, and even finitist proof in particular, cannot be closed under this sort of reflection.

The situation Hilbert was in is remarkably like that in which I argued the partiality theory of truth winds up. Both engage in a kind of reflection upon a given formalization, which is expressed by the appropriate reflection principle. Because each sees the system in question as closed under the sort of reasoning involved in the reflection, each requires that the appropriate reflection principle be contained in the original formalization. This turns out to be impossible.

Taking a cue from Kreisel, we may diagnose the problem in both cases as confusing what is *implicit* in a given formalization with what is explicitly part of it. We come to recognize what is implicit in a given formalization by reflection, which Kreisel [1970, p. 489] describes as asking, “What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, ‘accepted’) certain given concepts?” Let us call this *Kreiselian reflection*.

Reflection principles, like (Tr-RFN), (RFN_{TP}), (Tr-RFN-Rule), and (Rfn_F), state the results of Kreiselian reflection as they can be formulated

apply the incompleteness theorems, so the question does not really matter for our concerns here.

¹¹Local reflection principles generally agree with their uniform (parameterized) versions on Π_1 formulas, where both are equivalent to $\lceil Con \rceil$. Otherwise, local versions are weaker [see Feferman, 1962; Beklemishev, 1997].

in specific languages with specific resources. (I shall frequently identify the process of Kreiselian reflection with its result, and talk about the content of Kreiselian reflection.) The fully general form of Kreiselian reflection should express all the properties of a theory that follow from its correctness. It should thus express the soundness of the theory, in the strongest terms appropriate for the theory. I shall suggest that uniform reflection principles do just that.

For a theory like PA , the strongest statement of soundness is given by the full Tarskian truth theory for PA , which I shall call $Ta(PA)$. To determine the general form of Kreiselian reflection, we should investigate the relation between such a theory and uniform reflection.

$Ta(PA)$ is formulated in the language \mathcal{L}^{Ta} extending \mathcal{L} by a Tarskian truth predicate $\lceil Ta \rceil$. The axioms of $Ta(PA)$ are PA^{Ta} (PA with induction extended to $\lceil Ta \rceil$) together with the usual clauses of the inductive characterization of truth for PA . $Ta(PA)$ proves the properties of PA that follow from its correctness. It proves the soundness of PA in direct form, as the global reflection principle:

$$(GRFN) \quad \forall x(Prov_{PA}(x) \rightarrow Ta(x)).$$

(‘Global’ indicates this is non-schematic, unlike both local and uniform versions of reflection principles.) $Ta(PA) \vdash GRFN$ and $Ta(PA) \vdash Con_{PA}$. As we have a version of (T) for $\lceil Ta \rceil$, the uniform reflection principle for PA

$$(RFN_{PA}) \quad Prov_{PA}(\ulcorner \phi \dot{x} \urcorner) \rightarrow \phi x$$

emerges as a special case of (GRFN) together with (T).

Technically, $Ta(PA)$ is stronger than $PA + RFN_{PA}$. Feferman [1991] has observed that $Ta(PA)$ proves that each formula provable in $PA + RFN_{PA}$ is true, i.e. it proves global reflection for $PA + RFN_{PA}$, and hence $\lceil Con_{PA+RFN_{PA}} \rceil$. As both $Ta(PA)$ and $PA + RFN_{PA}$ state the soundness of PA , this may seem odd. It turns out that the difference between the two is a matter of some proof-theoretic subtlety. For instance, both can prove statements of the ω -consistency of PA , as a theory expressing the soundness of PA should. But $PA + RFN_{PA}$ proves only a uniform (schematic) form of ω -consistency, and cannot prove the global (non-schematic) form [Kreisel and Lévy, 1968]; whereas $Ta(PA)$ proves the full global form. A result of Smorynski [1977] shows that the global statement of ω -consistency for PA is equivalent to uniform reflection for $PA + RFN_{PA}$ restricted to Π_3 formulas

($\text{RFN}_{\Pi_3}(PA + \text{RFN}_{PA})$). This reflects the kind of added strength in $Ta(PA)$ Feferman’s observation indicates.

We can explain this difference if we recall that a theory like $Ta(PA)$ works like a weak second-order theory. As Parsons [1974a] noted, it essentially gives us a theory of predicative classes. A little more formally, it is well-known that $Ta(PA)$ is equivalent to the theory ACA of second-order arithmetic with comprehension restricted to arithmetic formulas. (Technically, $Ta(PA)$ and ACA are mutually interpretable, and prove the same arithmetic sentences.¹²)

Ordinary reflection principles, like (RFN_{PA}) , are first-order schemas that capture the content of virtually second-order theories like $Ta(PA)$. They express a basically second-order idea in a first-order setting. As such, reflection principles are accurate statements of soundness, even if they miss some of the proof-theoretic strength which can be gained by their virtually second-order analogs. The strongest of these for PA , the uniform reflection principle (RFN_{PA}) , thus provides the right statement of the soundness—the correctness—of PA in the language \mathcal{L} of PA . It expresses the correctness of PA as strongly as can be done in its own terms.

I suggest that this is just what we should expect of Kreiselian reflection. The idea of recognizing what is implicit in a formalization leads us to principles that may go beyond the formalization, but they should be principles expressible in the terms in which the formalization is given. I thus propose that we think of the uniform reflection principle for a given theory as giving the result of Kreiselian reflection upon the theory.

The results of Section (2) show that the hierarchical nature of truth flows from Kreiselian reflection. Steps in the hierarchy are induced by the reflection expressed by (RFN_{TP}) , which we saw to be equivalent to (Tr-RFN) . The reflection involved here is thus genuine Kreiselian reflection. Given that TP already has a truth predicate, it might have seemed that the reflection involved would be of the slightly stronger virtually second-order kind. But this would require building a theory of the strength of $Ta(TP)$ in \mathcal{L}^{Tr} . As TP makes its truth predicate $\lceil Tr \rceil$ self-applicative, this cannot be done by Tarski’s undefinability theorem. Kreiselian reflection on TP should be carried out in the terms of TP , so we should expect its results to be expressible in \mathcal{L}^{Tr} . This is just what is done by (RFN_{TP}) .¹³

¹²See, for instance, Halbach [1999]. ACA is essentially the first level of ramified analysis. The correspondence between iterations of $Ta(PA)$ and ramified analysis are investigated by Feferman [1991].

¹³It is, as far as I know, an open question whether $TP + \text{RFN}_{TP}$ (equivalently $PA +$

One of the fundamental lessons of the second incompleteness theorem is that for many concepts, the result of Kreiselian reflection upon a formalization of that concept is only *implicit* in the formalization. It cannot explicitly be in the formalization, as the resulting reflection principle cannot be proved by the formalization. This must be the case with any concept for which the natural formalizations are able to interpret weak theories of arithmetic. (Q suffices, as work on weak fragments of arithmetic has shown [see Hájek and Pudlák, 1993].) Let us call these *Kreiselian concepts*.

It must be stressed that for a Kreiselian concept, Kreiselian reflection does not amount to a concept shift. It does not ask us to replace one concept with another, or to reanalyze a concept in a different setting. It is not like, for instance, shifting from the theory of Euclidean spaces \mathbb{R}^n to the theory of manifolds. It is to insure this feature that I have insisted that we think of the results of Kreiselian reflection as expressible in the language of the original formalization. We start with a formalization of a concept, and reason about it. In particular, we reason about what follows from the correctness of the formalization. The results are new principles, but principles about the same subject-matter with which we started, expressible in the same terms. This provides a stronger formalization, but one that is still a formalization of the *same concept* with which we started.

The hallmark of a Kreiselian concept is that Kreiselian reflection upon a formalization of that concept leads to a stronger formalization, still of the same concept. Insofar as moving from formalization to formalization is moving to a *stronger* formalization, it may be tempting to say that we do have a kind of concept shift, in spite of my claim to the contrary. But this is to misplace emphasis upon individual formalizations, to the exclusion of the original Kreiselian concept of which they are formalizations. The shift is between individual formalizations, not in the original Kreiselian concept. As Kreisel himself reminded us [Kreisel, 1967], we do have to take the informal nature of Kreiselian concepts seriously. What is important, as I see it, is that each formalization is an articulation or precisification of the same Kreiselian concept. The shift does not amount to an articulation of a different concept, but rather a further articulation of the same concept. Hence, I maintain it is appropriate to hold that no concept shift occurs.¹⁴

RFN_{TP}) is arithmetically equivalent to $PA + \text{RFN}_{PA}$. (Thanks to Volker Halbach for pointing this question out to me.)

¹⁴This does raise the question of in what our grasp of the concept of truth consists; especially, a grasp which enables us to provide articulations of it. This is far too great

Though Hilbert thought otherwise, mathematical proof is a Kreiselian concept. The arguments of Section (2) show that the concept of truth, even self-applicative partial truth, is a Kreiselian concept as well. The self-applicative nature of truth, and the finite nature of proof, invite us to make the mistake of supposing these are not Kreiselian concepts. But it is a mistake.

In the case of mathematical proof, the Kreiselian nature of the concept did make trouble for Hilbert's ambitious program in the foundations of mathematics. Theorizing about Kreiselian concepts can be tricky. But beyond that, the Kreiselian nature of the concept is almost banal. As concepts go, mathematical proof is a clear one. For the most part, what makes a good proof in mathematics is pretty obvious. This is not to say there are no hard cases, nor that the Kreiselian nature of the concept does not open up some substantial theoretical issues.¹⁵ But one could hardly say that mathematical proof is ineffable, obscure, or otherwise philosophically suspect.

In particular, the Kreiselian nature of proof in no way leads us to conclude that the concept is fragmented in some pernicious way. It can be subdivided, of course, as when we distinguish analytic from elementary proofs in number theory. When it comes to formalizations, the Kreiselian nature of the concept requires that it be subdivided, and that there be a natural way to move from subdivision to subdivision. But if anything, this just points out the richness of mathematical proof.

I may now advance my primary claims. First, the sense in which there is a hierarchy of truth predicates is no more than that truth is a Kreiselian concept. In the proof-theoretic terms in which we have been examining the concept, any formal theory of truth leads to a stronger theory, still a theory of truth, by Kreiselian reflection. The hierarchy may be understood as the hierarchy of levels of this process. As the steps from level to level are generated by Kreiselian reflection upon a theory of partial truth, they are well-described by the model of the transition from TP to $TP + \text{RFN}_{TP}$. (I shall return in Section (4) to the question of how to understand this outside

an issue to be dealt with here, but let me merely say that I am inclined to follow such writers as Dummett [e.g. 1959, 1990] and Wiggins [e.g. 1980] in looking for an answer to it in the relation between truth and meaning or content. (I should also note that there is some overlap between the conclusions I have come to here and some of those reached by Dummett [1963], though there is also some significant disagreement.)

¹⁵The computer-based proof of the four-color theorem is often offered as an example of a hard case. (See the papers in Tymoczko [1986] for discussion.)

the proof-theoretic setting.)

Second, as truth is hierarchical precisely in being a Kreiselian concept, the hierarchy is no more problematic than what we find with any other Kreiselian concept, such as mathematical proof. As with proof, the hierarchical nature of truth does not imply that the concept of truth is ineffable, obscure, or otherwise philosophically suspect.

I thus conclude that the objection of fragmentation is without force. As we saw with proof, and as with any Kreiselian concept, there must be some subdivision of truth into formalizations related by Kreiselian reflection. But again as we saw with proof, this does not constitute a philosophical objection. Properly understood, the kind of fragmentation we find in the concept of truth is entirely acceptable.

4 Closure under Reflection

So far, we have investigated matters in the setting of formal theories of truth. However, such theories do not fully capture the kinds of natural language situations in which the truth predicate is ordinarily deployed, and in which complex Liar phenomena can arise. We must thus ask if the Kreiselian nature of the concept of truth is derived solely from features of formal theories, or if it will apply to more realistic theories as well. I shall argue in this section that it does apply more widely. Formal theories help make the nature of the hierarchy clear, but the basic point we have seen generalizes.

There is good reason to raise this question. It might be natural to suppose that the process of iterating Kreiselian reflection should reach a limit, at which point it would have ‘closed-off’ a concept under Kreiselian reflection. One might wonder if such a closing-off could not be achieved for a theory like *TP*. If it cannot, one might ask if the reason is specific to formal theories, and not indicative of the nature of truth itself. I need to explain why such a closing-off cannot occur, and why this is not specific to formal theories.

To investigate this matter, it will be useful to start by looking once more at the case of proof. Can we close-off the concept of proof under Kreiselian reflection? At least, can we close-off a more specific concept of proof, say, proof in arithmetic? It is known that there is a sense in which proof in arithmetic can be closed-off, but only a weak sense. I shall argue that, quite the reverse of what the question supposes, it is the availability of a weak closing-off that is specific to certain features of formalizations of proof.

These features do not extend to truth, which, I shall show, explains why the Kreiselian nature of truth is not just a matter of formal theories.

If we start with PA , and *iterate* the process of closing under (RFN_{PA}) transfinitely, we can reach closure in that we can reach a complete theory of arithmetic. However, the situation is somewhat delicate. Feferman [1962] showed that iteration through all of Kleene’s system of notations for recursive ordinals \mathcal{O} produces a complete theory. (\mathcal{O} is not a univalent notation system, but there is a path through \mathcal{O} recursive in \mathcal{O} which suffices.) But there are limits to this approach. Feferman and Spector [1962] showed that completeness can never be achieved along a Π_1^1 path through \mathcal{O} . As Feferman [1988b, pp. 143–144] himself later described the situation, “[This approach] is stalled as long as we don’t have a convincing answer to the question: *what is a natural path through \mathcal{O} ?* This should be closely related to the standing open question coming from proof theory: *what is a natural well-ordering?*”

What Feferman’s remark reflects, and the results show, is that the known techniques do not explain how to iterate reflection to closure in any informative way. To provide a path along which reflection is iterated is to explain the process of iteration that will be used. The known techniques only do this by relying upon \mathcal{O} , which is already far more complex than the set of truths of arithmetic. (The set of truths of arithmetic is Δ_1^1 , whereas \mathcal{O} is Π_1^1 -complete.) This amounts to solving the problem of how to iterate to closure by encoding it in a much harder problem. It does solve the problem, but by brute force, in a way that does not provide us with much useful information about how closure under reflection is reached.

It is worth noting that the issue here is not simply one of incompleteness. Of course, in light of the incompleteness theorems, we did not expect to be able to build a recursively enumerable complete theory. But in some cases, these sorts of limitations do not preclude informative results. Gentzen’s proof of the consistency of arithmetic stands out as an example. In spite of the second incompleteness theorem, this proof does shed substantial light on its subject-matter. The problem with the brute force method for iterating to closure, without any explanation of what makes a natural path, is precisely that it does not.¹⁶

When we turn our attention to formal theories of truth rather than proof

¹⁶It is striking that we do not get a complete theory even if we help ourselves to as complex a set as a Π_1^1 path through \mathcal{O} . An elegant result of Visser [1981] makes clear that even if we do help ourselves to such a path, the resulting theory is contained in a recursively enumerable one. Hence, the incompleteness theorem still applies.

in arithmetic, we find the situation to be even worse. Even highly aggressive iteration does not give us a way to avoid the Fortified Liar and reach closure. In proof-theoretic terms, we may think of each iteration of a Tarskian truth theory as a step in the progression of ramified analysis. (Feferman [1964] shows that if we start with a weak comprehension principle, iterating the uniform reflection principle can have a similar effect.) Feferman [1991] develops a single theory of partial truth which is equivalent to ramified analysis up to the ordinal Γ_0 . For such a theory, though, we may simply repeat the arguments of Section (2), which lead to further Kreiselian reflection. We have not iterated our way to closure under Kreiselian reflection.

In the case of truth, unlike that of arithmetic proof, we do not have at our disposal even a brute force way to iterate to closure. It is not clear what that would even be. In the case of arithmetic, we could set our sights on arithmetic completeness, and get it by brute force. But we have no such target for truth. We have no prior conception of what the complete theory of self-applicative truth should be. At best, we might strive for as much of schema (T) for self-applicative truth as we can get. But Tarski's theorem shows us that if this is the goal, it cannot be reached.

It might be objected that the problem here is still a vestige of proof-theoretic methods. Once we drop the proof-theoretic perspective, it might be proposed, we know very well for what to aim. Following Kripke [1975], it might be offered that we should aim for the fixed point property: $\langle \mathbb{N}, E \rangle \models \phi \leftrightarrow \langle \mathbb{N}, E \rangle \models Tr(\ulcorner \phi \urcorner)$. And we might further note that using techniques of inductive definitions, rather than proof-theoretic techniques, we know how to reach this by iteration of a kind of reflection construction in an elegant and informative way. It may look as if we have just what we need.

Indeed, it is known that the Kripke construction is closely related to that of iterating the process of constructing a Tarskian truth predicate [Kripke, 1975; McGee, 1991; Halbach, 1997]. Let us consider the proposal that it amounts, in model-theoretic terms, to a process of iterating Kreiselian reflection. Now, there are a great many questions that might be raised about this. In proof-theoretic terms, I insisted that Kreiselian reflection corresponds to adding (RFN_{PA}) rather than $Ta(PA)$. In more model-theoretic terms, this would not be appropriate, as (RFN_{PA}) involves a *proof* predicate. But it is plausible to suggest that the Kripke construction still involves Kreiselian reflection, in which the difference marked proof-theoretically by (RFN_{PA}) and $Ta(PA)$ is marked by the way in which the Kripke construction employs a truth predicate in a basically *partial* setting. Rather than pursuing this

in detail, I shall simply grant, for argument's sake, that we may reasonably see the Kripke construction as at least on par with a process of iterating Kreiselian reflection. With this granted, the basic question becomes clear: Does the fact that the process reaches a fixed point not show how to iterate Kreiselian reflection to closure? Is the result not a closing-off of a partial theory of truth like TP under such reflection?

It is not. Kripke himself noted that we do not really get the closure we are after. Though we have a great deal of closure, we may still reason as follows. The Liar sentence $\lceil \lambda \rceil$ cannot be in the minimal fixed point, or any other fixed point (so long as we require consistency). But then, insofar as a given fixed point is our theory of truth, we have observed that the Liar sentence is not true. We are back at the Strengthened Liar.

This exercise is exactly one of Kreiselian reflection (whatever the status of the steps in the Kripke construction themselves may be). It is reasoning about the correctness of the Kripke fixed point construction, understood as providing a theory of truth. Again, this reasoning takes place in a background of model theory and inductive definability. In this setting, as I mentioned, a formal description of Kreiselian reflection should not be put in terms of adding a proof-theoretic reflection principle, as we did above. Providing such a formal description would thus require a whole new set of technical apparatus, which I shall not pursue here.¹⁷ But informally, we may observe that the conclusion may be drawn once we have in place resources that allow us to show the existence of fixed-point models, and demonstrate certain basic properties of them. These amount to resources strong enough to prove the existence of certain inductively defined sets. In fact, they amount to definability resources that are already embodied in the Kripke construction itself. Hence, just as with the step from TP to $TP + \text{RFN}_{TP}$, we engage in the kind of reflection upon the correctness of the construction appropriate for the construction itself. This provides the resources for demonstrating that $\lceil \lambda \rceil$ is not in any fixed point. We have further, genuine Kreiselian reflection beyond a fixed point of the Kripke construction.

¹⁷In the setting of my [MS], a step in the hierarchy corresponds to the step from $HYP_{\mathfrak{M}}$ to $HYP_{\langle \mathfrak{M}, P \rangle}$ where P is inductive and non-hyper-elementary. I believe this does correspond to the kind of Kreiselian reflection we have seen with the step from TP to $TP + \text{RFN}_{TP}$. It is an expansion of resources appropriate for expressing soundness, yet in the same terms with which we began. It is one of using the same basic construction of HYP , but with an expanded ground structure that ultimately provides for longer effective iteration.

We are thus back in the situation we saw in proof-theoretic terms a moment ago. We are able to engage in Kreiselian reflection upon our model-theoretic construction—our model-theoretic theory of truth. Just as before, such reflection generates a substantial extension of the resources (the theory or model-theoretic construction) to which it is applied. It does so even for fixed point models. The conclusion that the Liar sentence is not true (and hence the second-level truth of the Liar sentence) may only be implicit in the theory embodied by the minimal fixed point model, but it becomes explicit once we allow for reflection on the model. We thus have the reflection involved in the Strengthened Liar.¹⁸ We can likewise produce a model-theoretic analog of the Fortified Liar by considering the model-theoretic analog of the reply to the Strengthened Liar, that what is true is what is in the fixed point.

If anything, we have had less success in reaching closure for truth by this technique than we saw in the case of proof in arithmetic. Unlike that case, we cannot buy the full closure of truth under reflection at any cost, no matter how much we are willing to pay. Why not? Let us look once more at the case of arithmetic proof. Achieving closure there relies upon the specificity of the subject-matter in question. We rely upon a prior account of truth in arithmetic, and in various ways code it into a proof system, to reach closure. Even this provides for only a very weak form of closure. We rely upon an account of truth with respect to the first-order language of arithmetic, not every truth about the natural numbers. The very constructions we use to achieve completeness for this language show us facts about the numbers that are not expressible in the language. Building a complete theory by iterating reflection leads us to at least a Π_1^1 set, or a path recursive in it, any of which is way beyond the arithmetic sets. We can thus engage in the appropriate definability-theoretic form of Kreiselian reflection even for these apparently closed-off theories.

The failure of the Kripke construction to really give us closure reminds us that we cannot even do as much as we did for arithmetic for truth. The

¹⁸In observing that the Liar sentence cannot be in the minimal fixed point, we are observing that it is *ungrounded* in Kripke’s sense. In Kripke’s own discussion, he says that this notion, and the conclusion that the Liar sentence is not true, “belong to the metalanguage” [Kripke, 1975, p. 80]. As is well-known, Kripke suggests that the metalanguage is produced by reflection “on the generation process leading to the minimal fixed point” [Kripke, 1975, p. 80]. There is clearly some affinity between what Kripke suggests and what I am claiming here; but I believe we understand the phenomenon at issue better if we see it as one of Kreiselian reflection.

concept of truth is entirely general, so we can always ask about the truth of a theory, or the correctness of a model, and thereby engage in Kreiselian reflection. There is no sense, not even the weak one we used for arithmetic, in which we can say we have exhausted our subject-matter and will look no further. The mere fact that a model might be a fixed point, or encode the entirety of some subject-matter like first-order arithmetic, makes no difference. We can still ask about its correctness, and thereby reach new truths. The complexity of the question may well increase. Reflection upon the minimal fixed point construction, for instance, requires us to ask about a Π_1^1 -complete set. But again, as truth is entirely general, this increase in complexity cannot bring Kreiselian reflection to a stop.

We may thus conclude that the Kreiselian nature of truth is not an artifact of proof-theoretic treatments. Proof-theoretic investigation is a fine way to make the phenomenon clear. But just as the concept of truth is entirely general, so is its susceptibility to Kreiselian reflection.

5 Conclusion

Before closing, let me return briefly to a couple of the more important anti-hierarchy positions on the matter we have been investigating.

At least two theorists, Reinhardt [1986] and McGee [1991], discuss the basic issue at stake here. Though he works with a somewhat different theory, Reinhardt states the corresponding version of (Tr-RFN), and considers the step of expanding his theory to include it. When it comes to the question of the status of his reflection principle, he describes it as “a formalist theory with respect to the non-significant sentences” [1986, p. 236].

McGee also works in a different framework from the one I have used here. First of all, he develops two notions: truth, which is at least formally bivalent, and definite truth. He provides an extremely elegant theory of definite truth based on \mathfrak{A} -logical provability (the generalization of ω -logic to an arbitrary structure). Formal issues like the ones I have discussed here appear in his framework as analogs of the second incompleteness theorem and Löb’s theorem for definite truth. Judging from his response to them, I believe his response to the puzzle of the Fortified Liar would be to note that in such situations, our theories cannot capture everything we might have pre-theoretically expected them to capture [cf. 1991, p. 280].

To the extent that they really repudiate hierarchies, I do not see how ei-

ther Reinhardt or McGee adequately accounts for the status of the principles generated by Kreiselian reflection. As I argued in Section (2), these principles must be true. I do agree with McGee that a formal theory often has to leave out some intuitively obvious principles, and that this is no grounds for rejecting the theory. But I do not see how it can constitute grounds for rejecting the truth of the principles either, especially those whose truth can be demonstrated. Recall, the problem here is not that by accident or by some pragmatic choice, the theory just does not include the principles generated by Kreiselian reflection. Rather, it is that the theory cannot include them. And thus, I maintain, we are left with no way to make sense of the truth of the principles, or the correctness of the theory, according to the view being proposed. Hence, though clearly McGee would reject (Prov-Tr-RFN),¹⁹ I do not see how this is a tenable position. The same applies to Reinhardt's position (at least, to the extent that I understand his idea of "formalistic" principles).

My objection to both Reinhardt and McGee is that on their own terms, I find the status they assign to the results of Kreiselian reflection mysterious. However, there is way I can make sense of their views on my terms. Reinhardt explicitly notes that the theory resulting from adding the relevant reflection principle is stronger than his original (and preferred) theory. McGee notes that given a partial interpretation, which fixes definite truth, it is possible to refine the interpretation. Thus, both seem to me to agree in some way to the Kreiselian nature of the relevant concepts. As I maintain that the sense in which truth is hierarchical just is its being a Kreiselian concept, I invite them to agree to the hierarchical nature of truth as well. Once this is agreed, I can interpret their proposals in my terms, in ways that seem quite reasonable. Reinhardt's category of merely formal corresponds quite closely to my second level, resulting directly from Kreiselian reflection. McGee's idea of refining an interpretation seems to include advancing in a hierarchy by Kreiselian reflection, but may be broader in including outright concept shifts as well.

I have argued that the hierarchical nature of truth consists in its being a Kreiselian concept. Steps in the hierarchy are taken by Kreiselian reflection. In proof-theoretic terms, this is accurately modeled by the step from TP to $TP + \text{RFN}_{TP}$. I have granted that this characterization is not fully

¹⁹I think this is fairly clear from his stated position, but he has also confirmed it to me in conversation.

general, as it does not apply to natural language, but I have also argued that the Kreiselian nature of truth is fully general. Some other work on the Liar paradox may be understood as investigating Kreiselian reflection in the natural language setting. I believe the idea from Parsons [1974b, p. 250] of reflection upon a “schema of interpretation” is best understood this way. I offer my own [MS] analysis of the way context shifts within Strengthened Liar inferences in natural language as a more refined theory. But I maintain that the fundamental sense in which truth is hierarchical is exactly that it is a Kreiselian concept.

References

- Barwise, J. and J. Etchemendy, 1987. *The Liar*. Oxford: Oxford University Press.
- Beklemishev, L., 1997. Notes on local reflection principles. *Theoria* **63**:139–146.
- Bernays, P., 1930. Die Philosophie der Mathematik und die Hilbertsche Beweistheorie. *Blätter für deutsche Philosophie* **4**:326–367. References are to the translation as “The Philosophy of Mathematics and Hilbert’s Proof Theory” by P. Mancosu in Mancosu [1998a].
- Burge, T., 1979. Semantical paradox. *Journal of Philosophy* **76**:169–198. Reprinted in Martin [1984].
- Cantini, A., 1996. *Logical Frameworks for Truth and Abstraction: An Axiomatic Study*. Amsterdam: Elsevier.
- Dummett, M., 1959. Truth. *Proceedings of the Aristotelian Society* **59**:141–162. Reprinted in Dummett [1978].
- , 1963. The philosophical significance of Gödel’s theorem. *Ratio* **5**:140–155. Reprinted in Dummett [1978].
- , 1978. *Truth and Other Enigmas*. Cambridge: Harvard University Press.

- , 1990. The source of the concept of truth. In G. Boolos, ed., *Meaning and Method: Essays in Honor of Hilary Putnam*. Cambridge: Cambridge University Press. Reprinted in Dummett [1993].
- , 1993. *The Seas of Language*. Oxford: Oxford University Press.
- Feferman, S., 1962. Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic* **27**:259–316.
- , 1964. Systems of predicative analysis. *Journal of Symbolic Logic* **29**:1–30. Reprinted in Hintikka [1969].
- , 1988a. Hilbert’s program relativized: Proof-theoretical and foundational reductions. *Journal of Symbolic Logic* **53**:364–384.
- , 1988b. Turing in the land of $O(z)$. In R. Herken, ed., *The Universal Turing Machine: A Half-Century Survey*, pp. 113–147. Oxford: Oxford University Press.
- , 1991. Reflecting on incompleteness. *Journal of Symbolic Logic* **56**:1–49.
- Feferman, S. and C. Spector, 1962. Incompleteness along paths in progressions of theories. *Journal of Symbolic Logic* **27**:383–390.
- Friedman, H. and M. Sheard, 1987. An axiomatic approach to self-referential truth. *Annals of Pure and Applied Logic* **33**:1–21.
- , 1988. The disjunction and existence properties for axiomatic systems of truth. *Annals of Pure and Applied Logic* **40**:1–10.
- Glanzberg, M., 2001. The Liar in context. *Philosophical Studies* **103**:217–251.
- , MS. A contextual-hierarchical approach to truth and the Liar paradox. Manuscript.
- Gödel, K., 1931. Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I. *Monatshefte für Mathematik und Physik* **38**:173–198. References are to the translation as “On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems I” by J. van Heijenoort in van Heijenoort [1967].

- , 1995. *Collected Works*, vol. III: Unpublished Essays and Lectures. Oxford: Oxford University press. Edited by S. Feferman, J. W. Dawson Jr., W. Goldfarb, C. Parsons, and R. M. Solovay.
- Hájek, P. and P. Pudlák, 1993. *Metamathematics of First-Order Arithmetic*. Berlin: Springer-Verlag.
- Halbach, V., 1997. Tarskian and Kripkean truth. *Journal of Philosophical Logic* **26**:69–80.
- , 1999. Conservative theories of classical truth. *Studia Logica* **62**:353–370.
- , 2001. Disquotational truth and analyticity. *Journal of Symbolic Logic* **66**:1959–1973.
- van Heijenoort, J., ed., 1967. *From Frege to Gödel: A Source Book in Mathematical Logic 1879–1931*. Cambridge: Harvard University Press.
- Hilbert, D., 1926. Über das Unendliche. *Mathematische Annalen* **95**:161–190. References are to the Translation as “On the Infinite” by S. Bauer-Mengelberg in van Heijenoort [1967].
- , 1928. Die Grundlagen der Mathematik. *Abhandlungen aus dem mathematischen Seminar der Hamburgischen Universität* **6**:65–85. References are to the translation as “The Foundations of Mathematics” by S. Bauer-Mengelberg and D. Føllesdal in van Heijenoort [1967].
- Hintikka, J., ed., 1969. *The Philosophy of Mathematics*. Oxford: Oxford University Press.
- Kreisel, G., 1960. Ordinal logics and the characterization of informal concepts of proof. In J. A. Todd, ed., *Proceedings of the International Congress of Mathematicians (at Edinburgh, 1958)*, pp. 289–299. Cambridge: Cambridge University Press.
- , 1967. Informal rigour and completeness proofs. In I. Lakatos, ed., *Problems in the Philosophy of Mathematics*, pp. 138–157. Amsterdam: North-Holland. Reprinted in Hintikka [1969].
- , 1968. A survey of proof theory. *Journal of Symbolic Logic* **33**:321–388.

- , 1970. Principles of proof and ordinals implicit in given concepts. In A. Kino, J. Myhill, and R. E. Vesley, eds., *Intuitionism and Proof Theory*, pp. 489–516. Amsterdam: North-Holland.
- Kreisel, G. and A. Lévy, 1968. Reflection principles and their use for establishing the complexity of axiomatic systems. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik* **14**:97–142.
- Kripke, S., 1975. Outline of a theory of truth. *Journal of Philosophy* **72**:690–716. Reprinted in Martin [1984].
- Mancosu, P., ed., 1998a. *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*. Oxford: Oxford University Press.
- Mancosu, P., 1998b. Hilbert and Bernays on metamathematics. In P. Mancosu, ed., *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*, pp. 149–188. Oxford: Oxford University Press.
- Martin, R. L., ed., 1984. *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press.
- McGee, V., 1991. *Truth, Vagueness, and Paradox*. Indianapolis: Hackett.
- Parsons, C., 1974a. Informal axiomatization, formalization, and the concept of truth. *Synthese* **27**:27–47. Reprinted in Parsons [1983].
- , 1974b. The Liar paradox. *Journal of Philosophical Logic* **3**:381–412. Reprinted in Parsons [1983].
- , 1983. *Mathematics in Philosophy*. Ithaca: Cornell University Press.
- Parsons, T., 1984. Assertion, denial, and the Liar paradox. *Journal of Philosophical Logic* **13**:137–152.
- Reinhardt, W. N., 1986. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic* **15**:219–251.
- Sieg, W., 1988. Hilbert’s program sixty years later. *Journal of Symbolic Logic* **53**:338–348.

- Simpson, S. G., 1988. Partial realizations of Hilbert’s program. *Journal of Symbolic Logic* **53**:349–363.
- Smorynski, C., 1977. The incompleteness theorems. In J. Barwise, ed., *Handbook of Mathematical Logic*, pp. 821–865. Amsterdam: North-Holland.
- Soames, S., 1999. *Understanding Truth*. Oxford: Oxford University Press.
- Tait, W. W., 1981. Finitism. *Journal of Philosophy* **78**:524–546.
- Tarski, A., 1935. Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia Philosophica* **1**:261–405. References are to the translation by J. H. Woodger as “The Concept of Truth in Formalized Languages” in Tarski [1983].
- , 1983. *Logic, Semantics, Metamathematics*. 2nd edn. Indianapolis: Hackett. Edited by J. Corcoran with translations by J. H. Woodger.
- Tymoczko, T., ed., 1986. *New Directions in the Philosophy of Mathematics*. Boston: Birkhäuser.
- Visser, A., 1981. An incompleteness result for paths through or within \mathcal{O} . *Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Mathematical Sciences* **43**:237–243.
- Wiggins, D., 1980. What would be a substantial theory of truth? In Z. van Straaten, ed., *Philosophical Subjects: Essays Presented to P. F. Strawson*, pp. 189–221. Oxford: Oxford University Press.