Complexity and Hierarchy in Truth Predicates[*]

Michael Glanzberg
Northwestern University

Since seminal work of Tarski (e.g. Tarski, 1935), hierarchies have been much discussed in the literature on truth and paradox. Especially in recent years, this discussion has been decidedly negative. Tarski's hierarchy of languages is sometimes described as the "orthodox" response to the Liar paradox (e.g. Kripke, 1975), but it is an orthodoxy many authors have gone to great lengths to avoid. Frequently, the unpalatability of hierarchies is taken for granted, and the main task is taken to be developing theories of truth that avoid them. In this paper, I shall speak in favor of hierarchies. I shall argue that hierarchies are more well-motivated and can provide better and more workable theories than is often assumed. I shall not argue here that hierarchies are inevitable (though I have argued that elsewhere); rather, I shall argue that if we wind up with one, that is not by itself a bad result. Along the way, I shall sketch the sort of hierarchy I believe is plausible and defensible, which is different in important respects from the orthodox Tarskian one.

My defense of hierarchies will assume a particular view of the nature of truth that is fundamentally 'inflationary' and sees truth as a substantial semantic concept. My main thesis will be that if one adopts this view of truth, hierarchies arise naturally. In contrast, if you adopt a deflationist line, hierarchies are much less plausible, and certainly lack motivation. As a corollary of these claims, we will see an important way in which theorizing

[*] Versions of this material were presented at the Seventh Barcelona Workshop on Issues in the Theory of Reference, University of Barcelona, June 2011, the Truth at Work Conference, Institut d'Histoire et de Philosophie des Sciences et des Techniques, Université Paris-1 and Ecole Normal Supérieure, June 2011, and Birkbeck College, University of London, February 2012. Thanks to all the participants at those events for very helpful discussion. I am especially grateful to Eduardo Barrio, Jc Beall, Alexi Burgess, Marcus Giaquinto, Øystein Linnebo, and two anonymous referees for comments and discussion of previous versions of this paper.

about the nature of truth affects how we proceed with the task of addressing the paradoxes. We will also see along the way that the approach to truth I shall advocate makes truth a complex concept, and that in the presence of self-applicative truth and the Liar, truth becomes a very complex concept. As I shall show, this complexity helps motivate hierarchies. Complexity and hierarchy go together, if you adopt the right view of truth.

The plan for this paper is as follows. In section 1, I shall introduce the semantic view of truth I shall suppose throughout the paper, and contrast it with deflationist views. In section 2, I shall introduce the notion of reflection, which is a process by which we can make our implicit grasp of concepts like semantic truth explicit. I shall go on to argue that reflection is an engine that generates hierarchies, especially when combined with the complexity of truth according to the semantic view. Examples of reflection, and how they indicate that truth is a complex concept will be discussed in section 3. In section 4, I shall argue that approaching truth through reflection motivates hierarchies. I shall also sketch the form I think a good hierarchical theory of truth should take in this section. I shall continue my defense of hierarchies of this sort in section 5. I shall argue there that my favored form of hierarchy offers a plausible theory that works well, and is not vulnerable to some standard objections. The hierarchy is not without costs, but they are not nearly so high as it is often assumed. I shall conclude in section 6 by returning to the issue of the nature of truth with which the paper began. I shall argue there that the defense of hierarchies I offer on the basis of a semantic view of truth is not available to deflationists.

1 The Nature and Complexity of Truth

My discussion of hierarchies will be informed by some background ideas about the nature of truth, of the sort discussed in the more traditional literature on the metaphysics of truth but not typically applied to the paradoxes. I shall isolate two very broad approaches to truth: one deflationary, the other inflationary and motivated by the role of truth in semantics. I shall go on to argue that the inflationary view of truth can support and motivate hierarchies.[1]

[1] This way of thinking about the nature of truth and its role in solutions to the paradoxes comes from joint work with Jc Beall (e.g. Beall and Glanzberg, 2008), though Beall himself prefers a very different set of options to the ones I endorse here.

Let us begin by noting some general features of *deflationary* theories of truth. Though there are many such theories, different in important respects, they share the idea that in some ways truth is not a substantial property with an interesting underlying nature. I shall take as my representative deflationary position the *transparency* view of truth advocated by Beall (e.g. Beall, 2009) and Field (e.g. Field, 1994); not for least of reasons, they are explicit about the logical properties they attribute to truth, and are concerned with the paradoxes. In the spirit of disquotational view of truth, Beall and Field hold the main feature of truth to be the intersubstitutability of ϕ and $Tr(\ulcorner\phi\urcorner)$ in all non-opaque contexts. This in turn supports the logical role of truth in enabling such functions as simulating infinite conjunctions or disjunctions. Thus, the nature of truth is exhausted by its simple logical properties, which in turn endow it with a specific measure of logical utility. We can of course say these features are interesting, especially to logicians, but with the broader deflationist tradition, Beall and Field will insist there is no underlying metaphysical nature of truth which determines its logical properties. Most importantly for what follows, the rules of substitution are essentially all there is to the nature of truth, if indeed that is a nature at all. They are simple, obvious rules, and mastering them is complete mastery of the concept of truth.[2]

My main focus here will be on an opposing, inflationary, view of truth. To see how it works, we might start with important work of Field (1972). In that early work (he has drastically changed his view since), Field envisages a substantial theory of truth combining two components. One is the familiar Tarskian apparatus, which provides an inductive characterization of truth for a language. The other is an account of 'primitive denotation', which he imagines will be along the lines of the causal theory of reference and provide a reductive account of the basic relations of reference and satisfaction from which the Tarskian theory is constructed.

There are some features of Field's view that I shall ignore, especially, the specific role of the causal theory of reference in providing a reductive analysis of primitive denotation. Abstracting from such details, we get a picture of truth with several distinctive features. First, the truth of any sentence is determined by substantial facts about reference and satisfaction,

[2] The transparency view is a descendent of the disquotational view of truth associated with Leeds (1978) and Quine (1970). Of course, there are a number of other versions of deflationism, which are importantly different, but this sample view will suffice for our purposes here.

whatever underwrites those facts. These are basic *word-to-world* relations. Second, those facts are facts about the semantic properties of parts of a language, like the referents of singular terms and satisfaction for predicates. Third, truth is determined compositionally: the truth of a complex sentence is determined by the semantic properties of its parts, including the truth of embedded sentences, in a compositional way. Again, it is the language which determines how such composition is carried out.

One way to bring all these points together is to capture them under the idea that truth is a fundamental semantic property of sentences of a language.[3] We might start with the widespread idea that to understand a sentence is to a great extent to understand its truth conditions.[4] When we articulate a theory of truth, along the lines of Tarski plus Field, we are articulating the way this fundamental semantic property works for a particular language. Yet at the same time, as Davidson (1990) reminds us, we can see the fundamental properties of truth illustrated by the application to a specific language. We can see, for instance, the role of word-to-world relations and semantic composition in fixing truth. Truth, on this view, is a fundamental semantic property whose features we can see at work in particular languages, like the ones we speak.

A theory like this can be seen as a descendent of the traditional correspondence theory of truth, at least in a limited way. It bears out the traditional idea of truth as relying on substantial truth-making relations to the world; but it does not rely on a metaphysics of facts, or a structural correspondence relation between truth bearers and facts, as the traditional correspondence theories of the late 19th and early 20th centuries did.[5] The structural correspondence relation to a single fact is replaced by multiple relations of reference and satisfaction for parts of sentence. These are brought

[3] The usual provisos apply here, about sentences in contexts.

[4] This view of meaning is closely associated with Davidson (e.g. Davidson, 1967), but it is also part of the long tradition in philosophy of language from Frege to Carnap to Montague and beyond. Of course, like all philosophical views, it is controversial, and conceptual role or inferentialist approaches to meaning deny it. Indeed, deflationism of the sort described by Field (1986) also denies it. Though Davidson endorses the close connection between truth and meaning, he holds a very different view of the place of reference in semantics, as we see in Davidson (1977, 1990).

[5] I have in mind the correspondence theory in something like the form it appeared in work of Russell (e.g. Russell, 1912) and Moore (e.g. Moore, 1953). The sort of theory I advocate here perhaps has more in common with the sort of correspondence discussed by David (1994).

together by semantic properties of composition, rather than metaphysical ones of fact creation.

I do think this approach to truth is appealing, and captures a lot of what seemed right about the idea of correspondence. However, it is not my goal to defend this particular view of truth here. Instead, I shall try to argue that it offers motivation and support for hierarchies of truth predicates. For these purposes, one feature of the approach is especially important. In sharp contrast to deflationist approaches, this approach implies that truth has substantial internal structure. This structure is shown in the division of labor between facts about reference and the compositional determination of truth value. These combine to fix the truth values of sentences by way of the internal structure of the sentences and the nature of the things their constituents refer to. Whatever the logical behavior of truth is, it is determined by this internal structure of reference and satisfaction and semantic composition. This is a fundamentally different picture than the transparency view offers, according to which truth is at heart a simple property, fully captured by simple substitution rules, with effectively no internal structure and nothing more fundamental to determine them. Furthermore, Beall and I have argued (2008) that the approach I am endorsing does not yield full transparency. Moreover, whatever degree of transparency truth enjoys is not its basic feature, but a consequence of its more basic properties. Truth, on the approach I prefer, does have an underlying nature.

One consequence of this is that, according to the view I endorse, there is a sense in which truth is *complex*. The internal structure of truth shows possibly complex ways that the truth of sentences of a language are determined by reference and satisfaction. The underlying nature of truth reveals a specific kind of complexity, and it is not one the transparency view finds. As we will see in a moment, the mathematics of truth bears out this complexity in precise ways. Truth, seen from this view, is a complex property.

To fix some terminology, let us call the view that truth is a fundamental semantic property with complex internal structure the *semantic view* of truth.[6] I shall argue in what follows that the semantic view of truth provides motivation and justification for hierarchies. I shall also argue that it indi-

[6] The terminology is somewhat unfortunate, as Tarski already appropriated the term ‘semantic’ for his semantic conception of truth. Alas, it is not easy to say just what Tarski had in mind by that. Depending on what he did have in mind, my use of the term may or may not overlap with his. In previous versions of this work, I used the term ‘complex view’ of truth, but that proves confusing when we come to discuss complexity results below.

cates a special role for considerations of complexity in the response to the paradoxes, as we will see in the following sections.

2 Implicit Grasp and Reflection

Now that we have a basic view of truth in hand, what does it tell us about the paradoxes or hierarchies? In this section, I shall introduce the notion of what I call *reflection*. Reflection enters the picture when we ask if and how we understand the complex property of truth. In virtue of our grasp of our languages, I shall argue, we are in a position to engage in a form of reflection which reveals some of its features. I shall go on to argue in sections 3 and 4 that reflection, together with the complexity of truth (according to the semantic view) are the engines that generate hierarchies. In section 4 I shall explain what those hierarchies are like. But first, we need to see what reflection is, and how it works.

The notion of reflection may be introduced by asking what is involved in understanding the nature of the truth predicate? It follows from the semantic view that understanding truth is simultaneously incredibly easy and very hard. First, we might say, understanding truth is easy: in virtue of having competence with our languages, we already understand it. Or more precisely, you implicitly *grasp* it. Truth is, according to the view in question, a fundamental semantic property of your language. In virtue of understanding your language, you implicitly make use of the concept of truth. You thus have that concept in your cognitive repertoire. This is a form of implicit grasp, as you have and make use of the concept. It is implicit, as the concepts that figure into the basic functioning of your language need not be overly accessible to you. But regardless, you do, according to the semantic view, already possess substantial implicit grasp of the concept of truth.

On the other hand, coming to understand the nature of truth is, according to the semantic view, extremely hard. Truth has a substantial and complex underlying nature, including aspects of recursion, and complicated notions like reference. Coming to understand that can be, and experience shows is, an extremely difficult challenge. Indeed, according to the semantic view, it is hard in just the same way that coming to fully understand the semantics of a human language is hard. Anyone who has dipped so much as a toe into the field of semantics knows just how hard that is! Indeed, it may be harder, as we need to understand not merely the semantics of one language, but a fundamental concept that is common across the semantics of many languages.

How can understanding truth be both hard and easy at the same time? It is not really so mysterious. That is just what one would expect from implicit grasp. What is easy is the state of having such implicit grasp (well, easy in virtue of whatever enables us to learn our languages!). Great efforts have gone into explaining what that sort of state is, and I shall not delve into the issue here.[7] All that matters for us is that whatever this implicit grasp consists in, it is not *explicit*. When we see understanding truth as hard, we are asking for an explicit articulation of the concept, let us say, by offering a theory of it in the appropriate setting. For the semantic view, this setting will include the semantics of some language, and ultimately more than that; but regardless, we are asking for an explicit articulation of the complex nature of the concept. For concepts we grasp implicitly, it is making them explicit that can be hard.

When we encounter some concept with a complex underlying nature, that we also enjoy an implicit grasp of, we have at our disposal a unique way to study that concept. We can *reflect* on our own abilities that are underwritten by the implicit grasp, and thereby come to learn about the concept. We can reflect on our linguistic abilities, and thereby learn about the languages we speak, and the concept of truth which plays a fundamental semantic role in them. In doing so, we can begin to articulate the nature of the concept explicitly.[8]

Reflection as I understand it is an activity we can engage in, when we encounter concepts of which we have some implicit grasp. In the case of truth, we have assumed that our understanding of the terms and sentences of our languages includes an implicit understanding of their properties related to truth and reference. This understanding, though highly tacit, guides our linguistic uses, our comprehension of sentences, and other manifestations of our linguistic competence. This provides us with some evidence we can use to try

[7] See the large literature in the philosophy of language on tacit knowledge, including such contributions as Chomsky (1980), Davies (1987), and Higginbotham (1989).

[8] Notions of reflection have appeared in the literature on truth, though often with relatively little discussion. For instance, Kripke's famous remark about "some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point" (Kripke, 1975, p. 80) seems to be gesturing towards the sort of reflection I have in mind. I have discussed this idea in my (2006), and in a somewhat different form in my (2004c).

to make the nature of the concept explicit. The evidence is available through introspection of the contents of our sentences, and self-directed observation of our linguistic practices. It is thus generally evidence we can access by focusing our attention on ourselves, our thoughts, and our activities. Based on such evidence, we can begin to articulate the semantic properties of our languages, presumably in the form of some semantic theory. This, we assume, will include articulating a body of semantic facts about truth and reference, and how they combine compositionally. This, in turn, forms the basis of a theory of truth. Though the concept of truth is implicit in our languages, we can put ourselves in a position to offer an explicit articulation of it by reflection.

Reflection is a complicated and demanding process. In what follows, I shall focus mainly on articulating the semantic properties of our own languages or languages like them by reflection, or even highly simplified formal languages. Strictly speaking, this is only a part of the process of reflection that would have to go on to fully articulate the nature of the concept of truth, as that concept applies across languages. But it is already a complicated enough task, and it is the one that is of special interest when it comes to the paradoxes and hierarchies, so I shall focus on it.

With a task of this difficulty, there is no guarantee that we will in any one instance of reflection produce a particularly complete theory of the underlying concepts of our semantics. There is not even any guarantee we will get those properties right. We have a complex concept and only some indirect sources of evidence from which to try to characterize it. Even for a gifted linguist or logician, there may well be limits on how much can be accomplished in any one exercise of reflection. Indeed, as I shall argue, the paradoxes or various incompleteness phenomena show that in some cases, even the most gifted logicians will fail to capture the entirety of certain concepts in any one instance of reflection. In these cases, I claim, hierarchies ensue. But we can already see why that might be an unsurprising result. The difficulty in producing comprehensive theories in reflection is not one that stems only from incompleteness or paradox. It stems from the difficulties of reflection—the complexity of the task, and the limited resources we have to do it—as well.

3 Models of Reflection and Complexity

So far, we have seen that according to the semantic view of truth, we have implicit grasp of the complex concept of truth. That grasp is something we can try to make explicit by reflection, and in particular, we can try to make the semantic properties of the languages we speak or related ones explicit by reflection. In this section, I shall explore some examples of how that process might go, and what the results might be. I shall focus on formal languages, following the tradition in work on truth predicates. They will show us enough to see how complex the results of reflection must be. As we will see in section 4, this in turn will show us how reflection can generate hierarchies.

I shall begin with a very Tarskian case of a language with no semantic predicates. Though it will only occasionally matter, let us take the language $\mathcal{L}$ to be the language of arithmetic.[9] For a language like this, classic work of Tarski (1935) provides a very good illustration of what reflection should yield. Indeed, for this case, we can articulate what a fully successful exercise in reflection should yield, without having to face paradox problems.

The task of reflection is to make explicit the semantic properties of a language, especially, those properties relating to truth and reference. For a language like $\mathcal{L}$, that is in effect to write out the definition of truth in the way Tarski showed us. Hence, a Tarskian theory of truth for $\mathcal{L}$ is good representation of what successful reflection should look like for such a language.

We should be a little more specific about what will count as a 'Tarskian theory of truth', and we will see that there are several different ways to describe it formally. One way, oriented around model theory, is to see the Tarskian definition as the definition of truth in a model. Hence, to display the Tarskian truth theory, we need to provide a model $\mathfrak{M}$ of $\mathcal{L}$, and the definition of truth in a model $\mathfrak{M} \models \phi$ for sentences of $\mathcal{L}$. If we do this, we should not lose sight of the inductive nature of the definition of truth in a model and its route through satisfaction. Moreover, as reflection asks us to make the concepts at work in the semantics explicit, we should follow Tarski in displaying the truth predicate over the structure explicitly. This is in effect to define the Tarski truth predicate for $\mathcal{L}$ over $\mathfrak{M}$, which provides

[9] I will be moving back and forth between proof-theoretic and definability-theoretic perspectives. For proof theory, it will sometimes matter that $\mathcal{L}$ is the language of arithmetic, though we will rarely get into enough technical detail to see this. Definability theory often prefers to work with purely relational structures and replace functions with relations; but again, we will not get into enough details to see this.

an interpretation for a language extending $\mathcal{L}$ with a Tarskian truth predicate Tr.[10]

We can also take a more proof-theoretic approach, and ask for an axiomatic theory of truth. A good proof-theoretic representation of the way truth works in a language like $\mathcal{L}$ is provided by directly axiomatizing the compositional definition of truth Tarski provides. To do this, we add the truth predicate Tr to $\mathcal{L}$, and the following axioms:[11]

1. $\forall s \forall t (Tr(s \underset{\cdot}{=} t) \leftrightarrow s^\circ = t^\circ)$

2. $\forall x (Sent(x) \rightarrow (Tr(\underset{\cdot}{\neg} x) \leftrightarrow \neg Tr(x)))$

3. $\forall x \forall y (Sent(x \underset{\cdot}{\wedge} y) \rightarrow (Tr(x \underset{\cdot}{\wedge} y) \leftrightarrow Tr(x) \wedge Tr(y)))$

4. $\forall x \forall y (Sent(x \underset{\cdot}{\vee} y) \rightarrow (Tr(x \underset{\cdot}{\vee} y) \leftrightarrow Tr(x) \vee Tr(y)))$

5. $\forall v \forall x (Sent(\underset{\cdot}{\forall} v x) \rightarrow (Tr(\underset{\cdot}{\forall} v x) \leftrightarrow \forall t Tr(x(t/v))))$

6. $\forall v \forall x (Sent(\underset{\cdot}{\exists} v x) \rightarrow (Tr(\underset{\cdot}{\exists} v x) \leftrightarrow \exists t Tr(x(t/v))))$

We always think of axioms like these as added to some base theory. One good representative starts with PA, but with induction extended to the expanded language including the truth predicate Tr. The theory which adds the truth axioms to this base theory is known as the *compositional truth theory* or CT (following the terminology of Halbach 2011).[12]

Both the model-theoretic definition of truth over a structure and CT represent successful reflection for $\mathcal{L}$. Both provide essentially complete accounts

[10] When thinking about the semantic properties of a language like the ones we speak, we should probably focus on the intended interpretation, and so perhaps for $\mathcal{L}$ we should be working with $\mathbb{N}$ rather than an arbitrary model. Occasionally, we will need to know the model is reasonably nice, but for the most part, we will not be concerned with which structure it is. We should also note that the mathematical definition of truth is a mathematical representation of a concept with empirical applications (as Etchemendy (1988) and Soames (1984) reminded us).

[11] I follow the notational conventions of Halbach (2011). They are mostly standard. $^\circ$ is the evaluation function for terms, which is definable in PA. Recall that the language of PA has no predicates other than identity, and hence the form of the axioms below is specific to PA. Minor changes can accomodate other sorts of languages.

[12] In many cases, I shall talk about formal theories without going into full details of their expositions, but this case is central enough, and illustrative enough, that the details seem to be worth mentioning.

of truth for $\mathcal{L}$. The truth predicate we define model-theoretically provides an extensionally correct truth predicate for $\mathcal{L}$ over the base structure. Likewise, CT proves each instance of the T-schema, and so is extensionally adequate. But we have more than an extensionally correct truth predicate. Both approaches generate that predicate by describing correctly the semantic workings of the language, just as reflection asks. Both yield useful results. For instance, CT can prove the consistency of PA, while the model-theoretic definition of truth is the basis for pretty much everything else that happens in model theory. In a less mathematical vein, the way the theories illustrate the compositional determination of truth via satisfaction shows something important about how the semantics of a language can work. We thus have two good examples, for a highly idealized case, of how we may present a truth predicate, and how we can do so via reflection on the semantic properties of a language.[13]

In section 1, I discussed how the semantic view of truth makes truth a property with a complex underlying nature. Both our examples of reflection give us a way to make this mathematically more precise, as we can apply complexity measures to model-theoretic and proof-theoretic truth predicates. Let us begin with the proof-theoretic CT. We can measure its proof-theoretic strength in a few ways. It is stronger than PA, as it proves the global reflection principle for PA, and in fact, it is slightly stronger than $PA + RFN_{PA}$. Another measure is that CT is proof-theoretically equivalent to the second-order theory ACA, which is second-order PA with the full induction schema and a comprehension axiom for arithmetic formulas (i.e. ones with no second-order quantifiers).[14]

Definability theory also provides measures of the complexity of the model-theoretic truth predicate. Over reasonably nice models (including the standard model of arithmetic), the truth predicate is not elementary, but Δ^1_1.[15]

[13] There are some limits to what these sorts of models of reflection capture. For one thing, we may well learn about $\mathcal{L}$ and PA more explicitly than we learn our natural languages. Neither approach fully addresses the question of how reference and satisfaction are fixed for a language. This is illustrated by the fact that CT does not rule out nonstandard models. Generally, these are good theories of how truth works, but by no means complete theories of intentionality.

[14] For discussion of these sorts of results, see for instance Feferman (1991) or Halbach (2011). These results are proof-theoretically somewhat delicate; for instance, as is well-known, if we weaken the induction schema of CT we get back a conservative extension of PA.

[15] This result is quite general, and not really specific to arithmetic. Moschovakis (1974)

Both results are formal versions of a general point: truth is complex, and complex enough to be more complex than whatever we start with. If we start with PA, we get a stronger theory. If we start with a model, we get a predicate that is not elementary over that model. Reflection thus can, when successful, yield something markedly more complex than what we had when reflection began. We can capture this in terms of proof-theoretic strength of theories, or in definability terms for structures of certain sorts, but the idea of added complexity from reflection on truth stands out.[16]

The jump up in complexity is significant, but also limited. In proof-theoretic terms, for instance, we are able to prove facts about soundness we could not before, as well as some corresponding statements of arithmetic. Much more becomes elementary, in definability-theoretic terms. But at the same time, the compositional theory reveals a fairly modest jump in complexity. ACA is a weak second-order theory, building in a substantial amount of predicativity, while a Δ^1_1 predicate is just above elementary by standard definability-theoretic measures. We see in these examples the complexity that goes with the inner workings of truth, but only a limited amount of it.

In addition to the way it can yield complex results, this example shows one other feature that often goes with reflection. In many cases, reflection involves a change of topic or subject-matter of investigation. Suppose we start with the language of arithmetic, and a theory like PA. The subject-matter we investigate with these resources is clear enough: it is arithmetic. We do so in an interpreted language which has semantic properties, but those are not what the sentences of PA talk about or the subject-matter of arithmetic. When we engage in reflection on the semantics of a language, we make its semantic properties a topic of investigation. If it was not part of the subject-matter we were investigating before, it becomes so. Reflection can overtly change the topic. Once something becomes part of the subject-matter of investigation, we can start to build up explicit theories of it, and so, carry out the task of rendering something explicit that was previously merely implicit.

The extent to which reflection changes the topic is often a matter of

proves a general version only assuming what he calls an ‘acceptable structure’. Some assumptions are needed; for instance, the result fails for recursively saturated structures (cf. Barwise, 1975).

[16] This is what Horsten (2011) calls the power of the compositional theory of truth. As he notes, it is a surprising fact that we gain in arithmetic strength simply by adding semantic axioms.

degree, and it is not always trivial to tell how much takes place. To make this vivid, recall that PA can define lots of truth predicates, such as the Σ_n-truth predicates. It can also define proof predicates, and all sorts of other things that might not have transparently seemed to be part of the subject-matter of arithmetic. But the Tarskian example does show that in some cases, a fairly pronounced change of topic occurs. The complexity results in effect confirm this, by showing that we cannot take our original theory to implicitly characterize our new subject-matter. The way reflection can go with changes of topic will become important as we explore hierarchies in sections 4 and 5.[17]

So far, I have presented a view of truth which makes it a complex concept, and discussed the process of reflection which can make aspects of that concept explicit. Our example of a language like $\mathcal{L}$ of arithmetic is, of course, highly idealized. Though there is nothing wrong with working with idealized examples, there is one feature of this idealization which we must remove. By insisting that $\mathcal{L}$ contain no truth prediate, we avoid issues of paradox and self-applicative truth predicates that have been the focus of much of the logical work on truth. To address the issue of hierarchies, and to get a more useful model of reflection, we need to remove this restriction and work with languages with a self-applicative truth predicate.

In the case of a language with a self-applicative truth predicate, the basic task of reflection remains the same: to capture the semantic properties of the language. But the semantic properties include the word ‘true’, and in turn, the word ‘true’ is supposed to mean something which relates closely to the semantics of our language. We need the semantics and the word ‘true’ to properly relate. From the perspective I am taking, we should not assume they will be identical. The semantics of the language is part of its underlying workings, and those need not coincide with the meaning of any expression. But all the same, our insights into the nature of truth and into the meaning of the word ‘true’ are related, and we would clearly miss something about the meaning of the word ‘true’ if the two had nothing to do with each-other. So, even though the semantics of a language and its truth predicate need not be treated in exactly the same ways, they do interact.

One thing we have learned from the study of the Liar paradox is that this interaction is in fact quite complex. This complexity is the basis for a general

[17] I have argued (Glanzberg, 2002, 2004a, 2006) that in the kinds of cases at issue for the paradox, reflection invariably does change the topic.

strategy for dealing with self-applicative truth and paradox. The basic idea is that we can model self-applicative truth as the result of iterating a more Tarskian truth construction, where the iterations increasingly well capture the interactions between the truth predicate the semantic properties of the language it is in. When a self-applicative truth predicate figures into the language, these iterations can be very long indeed, but they do reach stages where a reasonable snapshot of the semantics of the language, including its truth predicate, is reached.

This is one way of thinking about the Kripke construction (Kripke, 1975), but I shall suggest, it is a common feature of a number of the leading approaches to the paradoxes. To flesh out the idea, I shall review a couple of ways of modeling it formally. However, the mathematics involved gets quite complex very quickly, and in many cases, the mathematical details will not really be important for the argument I am making here. So, I shall try to give a very rough indication of what the formal models might look like, but I shall often skip a great deal of substantial and interesting detail, and I shall occasionally simply cite results.

Let us start with the Kripke construction as an example. Recall the main features of the Kripke construction. We start with a language $\mathcal{L}^+$ that adds a truth predicate Tr to $\mathcal{L}$, and has no Tarskian syntactic restrictions on the application of Tr. We fix a model $\mathfrak{M}$ for $\mathcal{L}$, and our job is to extend it to a model for $\mathcal{L}^+$ by defining a truth predicate, as in the Tarskian case above. Unlike the Tarskian case, the truth prediate is interpreted partially, by an extension and anti-extension $\mathcal{I} = \langle E, A \rangle$. So a model of $\mathcal{L}^+$ looks like $\mathfrak{M}^+ = \langle \mathfrak{M}, \mathcal{I} \rangle$. We can think of each $\mathfrak{M}^+$ as summarizing an exercise in reflection for $\mathcal{L}^+$, and so reporting an account of its semantics. As before, we should remember that it is the structure and the definition of truth in it that provides our explanation of the semantics of the language. As we now are using partial predicates, this will involve a choice of valuation scheme—like the Strong Kleene or supervaluation scheme. Aside from the treatment of partial predicates the exercise goes much the way we learned from Tarski.

Part of the exercise in reflection will be the interpretation $\mathcal{I}$ of the truth predicate. Typically, this will not be all that good, as it will not come close enough to the semantics provided by $\mathfrak{M}^+$. Thus, our attempt at reflection represented by such structures may be only partially successful. But one of the insights of the Kripke construction (over and above the partial treatment of the truth predicate[18]) is that we can define a process for constructing a sequence of models of the form $\mathfrak{M}^+$ that provide better and better approximations of the semantics of $\mathcal{L}^+$ and the interpretation of Tr. If we iterate the process long enough, we can get something very good. Very good here means reaching a fixed point, where further iteration does not improve our model. In fixed point models, the semantics provided by $\mathfrak{M}^+$ and the interpretation $\mathcal{I}$ of Tr come very close to coinciding, as we get the fixed point property:

$$\mathfrak{M}^+ \models Tr(\ulcorner \phi \urcorner) \leftrightarrow \mathfrak{M}^+ \models \phi.$$

My notation here masks that $\mathfrak{M}^+$ is a partial model, and there will be may sentences, including Liar sentences, which are treated as gaps in fixed points. A classical version, known as the 'closed-off Kripke fixed point', simply replaces the partial interpretation with its extension. Doing so gives us:

$$\langle \mathfrak{M}, E \rangle \models ((Tr(\ulcorner \phi \urcorner) \vee Tr(\ulcorner \neg\phi \urcorner)) \rightarrow (Tr(\ulcorner \phi \urcorner) \leftrightarrow \phi)).$$

Either way, we see that at least for non-pathological sentences (those where we have $Tr(\ulcorner \phi \urcorner) \vee Tr(\ulcorner \neg\phi \urcorner)$), our semantics for $\mathcal{L}^+$ and the interpretation of Tr agree.

There are many different ways of developing this idea formally, and many more ways of interpreting it. I shall put in the setting of reflection. If we think of each stage in the process as the results of reflection on the semantics of a language, we can see the whole process as an extended iteration of reflection. Our process of reflection becomes an extended one, involving repeated reflection on the semantics of the language, relative to some hypothesis[19] about the interpretation of Tr in it, and repeated refinement of the results until we reach something that is a plausible interpretation of the whole language $\mathcal{L}^+$. To give this idea a name, call it the *long iteration* strategy for reflection on languages with self-applicative truth predicates.

The Kripkean implementation of the long iteration strategy comes close to simply a long iteration of the Tarskian sort of reflection we discussed above. As it is usually presented, it is not quite exactly that, as the use of a partial truth predicate at least as an intermediate step is not Tarskian. However, with some effort, the Kripke process and a transfinite Tarskian hierarchy of

[18] Which admittedly had precursors in the literature, such as van Fraassen (1968, 1970).

[19] The role of hypotheses in the process is highlighted by the revision theory of truth (Gupta and Belnap, 1993).

languages can be shown to be equivalent in important respects (Halbach, 1997).

I have emphasized that reflection is something we engage in; something we do. However, it is best not to think of long iteration as something we will carry out step by step. Reaching a fixed point often requires iterating well into the transfinite ordinals, and we cannot do that step by step. (The strong Kleene valuation reaches a fixed point at ω^1_{CK}, the first non-recursive ordinal.) Rather, we should think of the long iteration strategy as being used the very way that it is presented by Kripke and others. Typically, we are shown features of the process of building interpretations, like monotonicity, and then we see that a process is triggered which we can prove reaches a fixed point. This is a very complicated story, no doubt; but from the perspective on truth we are adopting here, such complication is no problem. We already observed that truth is a complex property, with a substantial underlying 'nature'. Even in the simple Tarskian case, we saw that reflection is a fairly complex endeavor, producing results of complexity measurably higher than we started with. What we learn here is that truth is *very complex*. We see this in the nature of the long iteration process. We also see it in the results. Whereas a Tarski truth predicate is Δ^1_1 over the ground model (with the right assumptions), the Kripke minimal fixed point is Π^1_1-complete. As I described it above, the Tarskian truth predicate provides only slightly more complexity, while the Kripke one provides a great deal more.

To illustrate the long iteration strategy, I have used the familiar Kripke construction as an example. This is not essential to my main point, which is that a more complicated form of reflection can provide an articulation of the semantics of a language containing its own truth predicate. Other approaches might provide slightly different results than the Kripkean one (e.g. Gaifman, 1992). Actually, long iteration is a feature of practically every modern approach to the Liar. It is clearly on display in the revision theory of truth (Gupta and Belnap, 1993), in recent work on paracomplete theories of Field (2008) and paraconsistent ones of Beall (2009), and as we will see in a moment, in influential proof-theoretic approaches too. Of course these theories differ in many respects, but they all make use of the general strategy of a long iteration process, where each stage shows some of the features we saw in the simple case of Tarskian reflection. Even if we have not yet fully understood all its details, I believe that long iteration shows us a fundamental aspect of the nature of (complex, semantic) truth.

In discussing reflection, I noted we can think of it in proof-theoretic rather than model-theoretic terms if we like. The same is true for the long iteration strategy, though the mathematical situation is not yet fully understood. We might think about long iteration simply in terms of iterating the theory CT. Indeed, if we take a fully Tarskian approach, with a hierarchy of truth predicates, we can do just that. The proof theory of hierarchies of Tarskian truth predicates and CT theories (as usual, up to appropriate proof-theoretic ordinals) has been explored by Halbach (1995, 2011). The correlation between CT and ACA continues, and levels of this hierarchy match-up to levels of ramified analysis. Yet just as with the Kripkean approach, this will not really be an adequate analysis of a language with self-applicative truth, and some care needs to be taken to build a consistent theory that does do a reasonable job of capturing self-applicative truth.

There are a number of axiomatic theories of self-applicative truth which have been explored in recent years. Perhaps the two most widely discussed are the Friedman-Sheard theory FS (Friedman and Sheard, 1987) and the Kripke-Feferman theory KF (Feferman, 1991).[20] I shall not attempt to go into much detail about either of these two theories, nor shall I advocate one over another as a theory of truth, but I shall discuss enough of their features to give some sense of how we can think of them as falling under the long iteration strategy.

First, just as I noted that the Kripke process relates closely to the Tarskian hierarchy of languages, we can find proof-theoretic connections between hierarchies of CT-theories and both FS and KF. KF defines each Tarskian CT-truth predicate up to the ordinal ϵ_0, and is arithmetically equivalent to ramified analysis up to that same ordinal.[21] Thus, KF encodes a long iteration, much like the Kripke minimal fixed point does. Indeed, KF models are precisely fixed point models (though not only minimal ones). Related but weaker properties hold for FS, which is arithmetically equivalent to the hierarchy of CT-theories up to ω. FS does not axiomatize anything like a fixed point property, but at least partially reflects the finite stages of the revision process of the revision theory of truth, and its notion of nearly stable truth. Thus, both FS and KF display features which connect them

[20] For extensive discussion of these and other theories, see Cantini (1996), Halbach (2011), and Horsten (2011). Feferman's work was circulated well before publication, and was reported in part by McGee (1991) and Reinhardt (1986).

[21] Feferman (1991) also presents an alternative version of KF which employs a different, and stronger, way of treating schemas. The result is equivalent to ramified analysis up to Γ_0.

to iteration of Tarski-like theories.

As with the Kripke construction, we see reflections of iterated Tarskian theories, but have to make some significant modifications to preserve consistency and generate reasonably good theories. FS in a way modifies CT the least, except for allowing a self-applicative truth predicate. FS extends the compositional axioms of CT to all sentences of $\mathcal{L}^{+}$, but keeps the axiom for atomic sentences only for $\mathcal{L}$. The theory thus contains CT, but proves far too little about iterated applications of the truth predicate. To make up for this, rule forms of the two directions of the T-schema (known as necessitation and co-necessitation) are added. The result is a very classical theory, and as we saw, one as strong as ramified analysis up to ω. Unfortunately, it is also known to be ω-inconsistent, though it is arithmetically sound. [22]

The other approach, KF, follows Kripke's lead in building in some partiality for the truth predicate. To do so, the compositional axioms need to be changed to reflect the characteristics of negation for partial predicates (or the theory can be formulated with both truth and falsity predicates). The resulting theory is formulated in a classical metalanguage, but reflects the partiality of Kripke's approach. Thus, like Kripke's theory, it builds on Tarskian ideas, but implements them with care about partiality in the truth predicate.

Both theories offer us proof-theoretic ways of thinking about the long iteration strategy. Both show us ways to start with the basic idea of reflection on the semantics of a language we saw with the Tarskian CT, and modify it in ways to make room for self-applicative truth. When we do, we get theories which capture the idea of iterating a Tarskian construction up to a suitable proof-theoretic ordinal. Thus, like the Kripke fixed point models, they capture the idea of a complex reflection on the semantics of a language with a self-applicative truth predicate, involving the core Tarskian insights, modified suitably, and iterated far enough to get a good theory. The naturalness of FS and KF help to substantiate the idea that the iteration was far enough to reach a good stopping place. Thus, though there are a great number of outstanding issues here, both technical and philosophical, I believe it is plausible enough to count these proof-theoretic options as falling within the long iteration strategy.

In either form, the long iteration strategy shows that self-applicative truth

[22] Many find ω-inconsistency a reason to reject FS. For an interesting discussion, see Barrio (2006).

is very complex. We already saw that the complexity of the Kripke minimal fixed point is quite great, and both FS and KF are much stronger than CT, going up quite high in the levels of ramified analysis (especially KF). As I said above, I suspect this complexity is a genuine feature of self-applicative truth, and one that bears out the idea from the semantic theory that truth has a complex underlying nature.[23]

The long iteration strategy, in either model-theoretic or proof-theoretic form, gives a way to think about the kind of reflection that would be involved in reflecting on languages with self-applicative truth predicates. It is a complex process, involving iteration of basic semantic insights about truth to properly relate them to the semantics of the truth predicate within the language.

In the Tarskian case, we saw that reflection involved a marked change in topic, as it adds an entirely distinct truth predicate to the language. In the self-applicative case, that does not happen. As I mentioned, the kind of change of topic or subject-matter we see in reflection comes in degrees, and one of the features of the long iteration strategy is that it involves only a much more modest change of topic. After all, our language already contains a self-applicative truth predicate, so we are already able to talk about the semantic properties of the language.

Even so, I believe we see some rather modest aspects of change of subject-matter in the long iteration strategy. Though we have a truth predicate in the language, the task of reflection is to describe the semantics of the whole language, and that is not quite the same as talking about the truth of some sentences using the truth predicate. In model-theoretic terms, we see this in reflection providing a model and definition of truth in the model for the whole language, not merely a truth predicate. In proof-theoretic terms, we see it in the added strength of our theories. Though we have a truth predicate in the language, we begin with PA formulated in that language. We switch

[23] This raises the question of whether the concept of truth is too complex to be grasped implicitly by all speakers, as the semantic view of truth requires. I do not have space to pursue this issue in depth, but let me quickly note that a great deal of work in cognitive science suggests we do have implicit grasp of complex concepts. A nice example is the concept of causation, which children have in some form starting as young as 6 months. The pressing question as I see it is not whether we can have implicit grasp of complex concepts, but how we can. The literature on perception of cause raises interesting questions about modularity and innateness of this concept. See, for instance Carey (2009) and Scholl and Tremoulet (2000).

to a stronger theory of truth, which tries to capture the semantic properties of the language compositionally. Though it is not a switch in topic marked by the introduction of a novel predicate, we see change in subject-matter either way. The difference between these two ways topics can be changed will become more important as we discuss hierarchies in sections 4 and 5.

We now have some idea what reflection on the concept of truth should look like in the presence of self-applicative truth predicates. We have seen that the long iteration strategy gives us a reasonable way to take the basic Tarskian insights into the nature of semantic truth and apply them in this complex setting. And, we have seen, the results are indeed very complex. My main contention all along is that reflection, and especially reflection of this very complex sort, can lead to hierarchies. It is now time to explain why.

4 Complex Truth and Hierarchies

Why might we expect hierarchies in our theory of truth? Here is the general idea. We began with the proposal that truth is a fundamental semantic property. This makes truth a complex property, but one whose nature can be studied by reflection. We saw that reflection genuinely indicates complexity, in mathematically measurable ways. When we take into account the self-applicative nature of truth predicates, and the interactions between the word 'true' and the underlying semantics of a language, we see that in fact truth is very complex. Again, this complexity can be measured mathematically in various ways, depending on the formal setting.

The complexity of truth should not come as a surprise, according to the semantic view. Reflection requires stepping out of the language you are speaking, and reflecting on its semantic properties as a whole. That the results turn out to be complex, measured against things you could do in the language, is not surprising (thought just what the degrees of complexity are might be surprising). What is special about the case of self-applicative truth is the additional complexity it creates. We handle that complexity, I suggested, by some form of the long iteration strategy, which provides a complex process of reflection suitable for the task of capturing the semantics of languages with self-applicative truth predicates.

If the long iteration strategy were to be fully successful, we would have a perfectly good theory of truth by lights of the semantic view of truth.[24]

[24] Or at least, almost. As I mentioned in section 1, we would have a perfect theory

Nothing like a hierarchy would ensue. On the other hand, as I stressed in section 2, there is no guarantee that any exercise of reflection will be fully successful, and no guarantee such an exercise will produce a complete or otherwise good theory. I argued that the long iteration strategy can produce reasonably good theories, but even so, for something as complex as the long iteration strategy, we might especially wonder if a complete theory of truth will be the result. We might wonder why such reflection should be able to return a fully correct theory of the semantics of a complex language all at once. If it does not, then the result is a hierarchy. We would have to restart the process of reflection, and generate a further truth predicate. The process could well be open-ended, as we have no guarantee we will ever reach a completely finished product. We would then have a hierarchy of accounts of the semantics of the language, each with a distinct truth predicate. We would indeed have a hierarchy.

Thus, I claim, if we see our theories of truth as the results of complex reflection on the semantics of a language, we should not be surprised if we encounter hierarchies, any more than we should be surprised that such complex tasks can yield incomplete results. Hierarchies, on this view, should not be surprising.

Actually, I believe something stronger. I maintain that reflection, even very good instances which generate plausible theories of truth, must be incomplete, and so, hierarchies are not just unsurprising, they are inevitable. This is, of course, a highly contentious claim. Many of the theories we reviewed in the last section are offered as non-hierarchical theories of self-applicative truth, and spirited defenses of them as such have been offered. This is so for the defenders of theories like KF or FS, such as again Halbach (2011) and Horsten (2011), and those who develop model-theoretic approaches in non-classical settings like the paraconsistent theories of Beall (2009) or Priest (2006) and the paracomplete theory of Field (2008) (all of whom, in one way or another, rely on the long iteration strategy).

My own view is that we cannot avoid hierarchies. I have argued this at length elsewhere, and I shall not try to mount a full defense of the claim here.[25] But it will help make clearer the nature of the hierarchy I think we

of truth as applied to one language, which has been the main concern of work on the paradoxes. We still might like to understand better the way truth works across languages.

[25] I discussed this in fairly general semantic terms in Glanzberg (2001), in model-theoretic terms in Glanzberg (2004a), and in proof-theoretic terms, focusing on FS, in Glanzberg (2004c).

are stuck with to roughly sketch why I think hierarchies are unavoidable. The main idea is that once we have our semantics in hand, it turns out we can engage in further reflection on how it works, and that leads us to more inclusive truth predicates. Each such reflection introduces a new level of the hierarchy, and the process is open-ended and does not terminate.

Of course, the Liar is doing a lot of work in showing that this further reflection really generates something new. Let us consider how this process of stepping back and seeing how the semantics that we generated in reflection works might go. Suppose we take our semantics to be a Kripke model $\mathfrak{M}^+$, understood as the result of a long iteration process of reflection. Above we noted that this is a pretty plausible semantics, in virtue of the fixed point property. Hence, it seemed that the long iteration strategy produced a successful exercise in reflection. But now we come to the Liar. Observing how the semantics works, we can observe that the liar sentence is not assigned the value true or false in this model (it gets the third or gap value), and it is not in the extension or anti-extension of Tr. But then, we can observe according to the semantics, the Liar is not true. But of course, this is just what the Liar says, so it appears we have used the semantics to show that the Liar sentence is in fact true. But then, we have found our semantics to be inadequate. The results of our initial exercise in reflection was not so successful after all. Likewise, our interpretation of Tr appears to be off, as we have convinced ourselves that the Liar is true, and so should be in the extension of Tr. If we take the closed-off model $\langle \mathfrak{M}, E \rangle$, we get similar results, though in a somewhat different way. The Liar is not in E, hence, the semantics simply says that the Liar sentence is true. But then we have an inadequate semantics, as our interpretation of Tr and the semantics fail to match-up adequately after all. Alternatively, we could note that though the Liar sentence is true, it is also true that $\neg Tr(\ulcorner L \urcorner)$ for the Liar sentence L, and so the semantics in effect denies the truth of the Liar sentence. There are other ways to illustrate the inadequacy of the semantics. For instance, it fails to define negation or some conditionals we need to describe the semantic properties of the Liar fully. In proof-theoretic terms, we get results like $KF \vdash L \wedge \neg Tr(\ulcorner L \urcorner)$ for Liar sentence L. Different ways of implementing the details here will capture the inadequacy differently, but I hope to have made clear why one way or another, our semantics and our account of Tr are not good enough.

Of course, this is just the Strengthened Liar. As I said, the standing of this and other 'revenge' arguments is hotly contested.[26] Also as I said, I have tried to defend a more carefully worked out version of it in other work, and I shall just accept it here. My main point now is to put it in the context of reflection. We have engaged in a process of reflection, which provides us with a seemingly plausible semantics for our language. But the exercise of the Strengthened Liar shows that we can observe how the semantics works, and come to see that it is not fully adequate after all. Just how that inadequacy manifests itself depends on the details, but one way or another, we can see it.

How are we to respond to this sort of problem? From the point of view we have taken here, where the task is one of reflection, the answer is easy. We found that our exercise in reflection yielded good, but we now see not good enough, results. As I have stressed, even without the force of paradoxes, this kind of situation is hardly surprising. And we know what to do when we find ourselves with reasonably good but not good enough theories. We should simply re-start the process of reflection to try to build a better theory. Building one will involve producing a wider truth predicate, and thereby a wider semantics, which will do better than the one we had. To see how this might work, let us again look at the closed-off Kripke model $\langle \mathfrak{M}, E \rangle$. This failed to accurately capture the properties of the Liar sentence, by failing to report the truth of the Liar sentence, which follows from the very semantics. We need to build a new truth predicate, which does better. But we also need to be careful, as simply throwing the Liar sentence into the extension of Tr gets us back into trouble. (Again, just what trouble depends on the details, but in the fully classical setting, we make $Tr(\ulcorner L \urcorner)$ true, which makes the Liar sentence false even though it is reported as true.) One way or another, we cannot simply adjust for the Liar sentence without getting a new paradox back. So, the basic task is clear—we need to re-start the process of reflection and build a better theory—but just how to do so in a productive way is a difficult task.

My own approach to this task is to rely on our observations about the semantics, as described in the earlier exercise of reflection, as we go forward. Thus, we start with the observation that he Liar comes out true *according to the semantics as we worked it out in the previous exercise of reflection.* What we need is an expanded semantics, which reports this, and so does a better job of making the semantic properties of the language explicit. Here

[26] See the papers in Beall (2008) for many different perspectives on revenge paradoxes.

we see the change of topic feature of reflection which was evident for the purely Tarskian case we explored in section 2. We are making the product of the previous round of reflection—the (reasonably good) approximation of the semantics—part of the subject-matter of our next round of reflection. In doing so, of course, we shift its role from the being the underlying mechanism of the language we are speaking—the real semantics—to something we are talking about. We can thus build accounts of the semantics of the language which take it into account, but depart from it. Of course, the basic approach to this further exercise of reflection will be the same: it will continue to use the long iteration strategy. But it will do so based on our new subject-matter, including the semantics produced by the previous round of reflection.

It is, unfortunately, somewhat complicated to capture this process formally. I developed a model-theoretic version of it in Glanzberg (2004a), but it relies on some fairly heavy use of definability theory. I shall here try to sketch the main idea with a minimum of formal apparatus. If we take $\langle \mathfrak{M}, E \rangle$ to represent the results of the first round of reflection, we want to repeat the long iteration strategy, taking it as part of the subject-matter. We thus want to re-do the Kripke construction, over the expanded structure $\langle \mathfrak{M}, E \rangle$. Of course, E is no longer interpreting the truth predicate, it is just another predicate. When we re-do the Kripke construction over this expanded model, we get an expanded truth predicate. The new interpretation of Tr includes facts like $Tr(\ulcorner \neg \underset{\cdot}{E}(\ulcorner L \urcorner) \urcorner)$. The new interpretation reports the facts about the semantics as it was before our new round of reflection. It is also much more complex than E, as we might expect. One reason the machinery gets complicated is it is not completely trivial to keep track of such iterated inductive definitions and their complexities, and so we wind up working with the machinery of next admissible ordinals.

The picture that emerges has some Tarskian aspects. As I loosely described it, we have the old truth predicate $\underset{\cdot}{E}$ and a new one Tr. This will help to fix ideas, but in fact, the model of Glanzberg (2004a) is slightly less Tarskian than that. The model does not simply add a new predicate to the language, and the main role of the prior interpretation E is to add complexity to the ground model. The expanded truth predicate allows us to reconstruct the semantics of the prior stage, and define the old truth predicate. So, $\underset{\cdot}{E}$ is definable, and we do not really have to outright change the vocabulary. Contextualism also enters the picture, as I take the ground structures to represent contextually salient elements, and work with domains of truth conditions relative to contexts. That helps to model the way in which we make the prior semantics a new topic.[27]

With all those details, we get a less Tarskian theory, but one that is still decidedly hierarchical. We can define the old truth predicate via the new one, and so, the old truth predicate is present in the language, though because of definability-theoretic strength rather than outright change of vocabulary. The response to the Strengthened Liar is likewise fundamentally hierarchical, as we say that the Liar sentence is true relative to the semantics as it was at the prior stage. It is true at the lower level of the hierarchy. As I mentioned, this is an effect of the topic-changing nature of reflection, though one that seems to be necessitated by the Strengthened Liar. I shall return to the comparisons between my preferred form of the hierarchy and Tarski's in section 5.[28]

Once we go down this road, an open-ended hierarchy ensues, for familiar reasons. The very reasons we found to be unsatisfied with the results of the first round of reflection can be applied again to the new results. They too will not be completely adequate, and we will trigger a new round of reflection. The process is open-ended. (Indeed, if it reaches a genuine top, we get paradoxes back.)

In thinking about reflection and the long iteration strategy, I grouped model-theoretic and proof-theoretic versions together as various ways to represent results of reflection. When it comes to capturing the kind of open-ended hierarchy generated by reflection we have been discussing proof-theoretically, there are few results available, but work of Fujimoto (2011) and Jäger *et al.* (1999) might be pressed into service. As with the model-theoretic approach, things can get quite complex very quickly, so I shall not explore this idea in any depth. One point is worth mentioning. The extant theo-

[27] I have defended the contextualist aspects of my view in Glanzberg (2001, 2004a, 2006). As I mentioned, contextualism helps with the development of the particular sort of hierarchical view I prefer, but generally, the step from any hierarchical view to contextualism is very small. One need only accept that reflection (or whatever else generates the hierarchy) takes place in real time as we work with and reason about our concepts, and so takes place within contexts. Contexts thus serve to index the stages of reflection. This is the core of the view I have defended, and I believe underlies other contextualist views such as those of Parsons (1974).

[28] A theory close in spirit to mine, but using very different resources, is developed by Barwise and Etchemendy (1987). Iterated Kripke constructions are also discussed by Field (2008), and briefly in the older discussions of Burge (1979) and the postscript to Parsons (1974).

ries explore iterating KF through appropriate proof-theoretic ordinals (and develop relations with theories of iterated inductive definitions), and hence might be plausible proof-theoretic representations of the kind of hierarchy that I claim we get. Interestingly, they also display the Tarskian features I remarked on in discussing the model-theoretic version. If anything, they do so more starkly. The known theories rely on a binary truth predicate, which in effect indexes the truth predicate to levels of a well-ordering (via some notation system). Thus, it seems that the only ways we know to develop the kind of iteration the Strengthened Liar reveals are at least somewhat Tarskian.

We now have at least a hint of what sort of hierarchy I think emerges for truth, and why. I also hope to have made clear why the hierarchy has some sound motivations and is not absurd on its face. As I have presented things, the hierarchy is a hierarchy of results of reflection. As truth, according to the semantic view, is a complex concept, we should not have particularly expected such reflection to yield fully complete theories in any one exercise. The extreme complexity of self-applicative truth only reinforces this expectation. That we find, after doing a good job of reflection, that we need to do more, is just not surprising or problematic. That is all there is to the hierarchy, and so, I claim, the mere fact that we get some sort of hierarchy is not problematic.

This is a partial defense of the hierarchy, and there remain some important questions. Just because some sort of hierarchy is not repugnant does not mean the one we have is plausible. The Tarskian aspects of the hierarchy I have noted might make us cautious about accepting it, even given the kind of motivation I have offered. I shall go on to discuss these issues in the next section.

There is one final point to make about the general idea that less-than-complete exercises of reflection are to be expected. Even if that is true, it does not address the fact that according to the kind of Strengthened Liar or 'revenge' reasoning I am relying on, such incompleteness is necessary. This shows that the hierarchy is not merely the result of our being fallible beings, with limited abilities to reflect on our own languages. Motivating and defending the hierarchy, as I am doing here, does not explain everything we might want to know about its source and nature.[29]

[29] There is something special about 'stepping back' and reflecting about the semantics of your own language that triggers hierarchies. I investigated some aspects of this in a

5 Varieties of Stratification

So far, I have argued that if we adopt the semantic view of truth, we naturally get a hierarchy (at least in the face of the Strengthened Liar), and the hierarchy is generally well-motivated. But as I mentioned, the very general kind of motivation I have offered does not show the specific hierarchy we get is plausible. To further defend the hierarchical approach I have been developing, I shall now explore in more detail how well it functions, and how vulnerable it is to objections. I shall argue it is quite successful on these counts, though it is not entirely without costs.

To do this, I shall begin by reviewing some common objections to hierarchies. I shall then discuss a somewhat broader notion of stratification, which is the core feature of hierarchies, but also includes some instances which are not usually labeled 'hierarchies'. I shall show that different forms of stratification make the objections more or less compelling. I shall argue that though the truth hierarchy is not at the completely innocuous end of stratification, it is close enough to evade the objections. Thus, I hope to offer a more detailed defense of the hierarchy, beyond very general motivations.

One preliminary point is needed. I shall in may cases contrast the sort of hierarchy I proposed in section 4 with an orthodox Tarskian one. I take it that the orthodox Tarskian hierarchy is well-known, and I shall not review its structure in detail.[30] But there are a couple of features of it that will become salient. First, the orthodox Tarskian hierarchy introduces new truth predicates at each level. The distinct predicates are indexed by the hierarchy. There is a syntactic restriction that each predicate can only apply to sentences of lower levels of the hierarchy, so each truth predicate can only apply to sentences containing only lower-level truth predicates. The orthodox hierarchy is thus syntactically driven, in the way it separates levels and restricts truth predicates.

Many have found hierarchies of truth predicates so obviously objectionable as to be dismissed without discussion. One reason behind this, I suspect, is what I shall call the *one concept* objection. There certainly seems to be one unified concept of truth, while a hierarchical theory might seem to present us with many distinct concepts of 'truth at some level'. Surely any such

number of papers Glanzberg (e.g. 2006). A very different view is presented in Gauker (2006).

[30] See McGee (1991) for a nice presentation of the details, and Halbach (1995) for an in-depth exploration.

analysis must be wrong, the objection goes.

Another set of objections to hierarchies is that they are unnatural or unworkable in some way or another. I shall group these together as the *clumsiness* objection. There are a couple of salient instances of this sort of objection. Perhaps the most important is Kripke's well-known 'Nixon-Dean' objection. Recall, Kripke presented an example where Nixon says 'Everything Dean says about Watergate is false' while Dean says 'Everything Nixon says about Watergate is false'.[31] Kripke rightly points out that the orthodox Tarskian hierarchy simply cannot capture these statements, at any level. The problem is with the fixed syntactic indexing of levels, which precludes reasonable assignments of levels to truth predicates in Nixon's or Dean's sentences. There are other versions of the clumsiness objection. For instance, Halbach (2011) and McGee (1991) complain that we cannot express in a hierarchy certain generalizations that we find theoretically interesting, especially semantic ones. Each version points out things we would like to do or say with truth predicates which seem problematic on hierarchical accounts. The theory is, it seems, too clumsy to work the way it should.

Another objection is the *weakness* objection, which says that whether or not they are clumsy or artificial, the theories we get from hierarchical approaches are just too weak. For instance, they might well be too weak to serve mathematical purposes, as discussed by Halbach (2011).

I shall discuss the one concept objection at some length in a moment. First, I want to address the clumsiness and weakness objections. As careful commentators such as Field (2008) and Halbach (2011) have noted, these objections may have some force against the orthodox Tarskian hierarchy, but they apply to more liberal hierarchies, like the one I sketched, only to a very limited extent. Actually, the weakness objection turns out not to be specifically one against hierarchical theories. We have already seen that in terms of pure mathematical power, versions of the Tarski hierarchy and various other non-hierarchical theories (like the Kripke construction or KF) turn out to be equivalent. So, there is no special weakness issue even for the orthodox Tarskian hierarchy. There are, of course, a number of more specific issues. The Tarski hierarchy itself has levels that are weaker in strength than some untyped theories. We see this, for instance, from the fact that CT is much weaker than KF. But this does not pose a problem for the hierarchy I en-

[31] For those who find this example dated, John Dean was White House Counsel to Richard Nixon, and went on to testify against Nixon at the Senate Watergate Hearings.

dorse, or any hierarchy that uses the long iteration strategy to build much stronger theories at each level. If anything, the kind of hierarchy I propose indicates much stronger theories than the familiar non-hierarchical ones. All of those (with the notable exception of the revision theory) mathematically correspond to specific levels of the Tarski hierarchy or ramified analysis, while the higher levels of my hierarchy will go beyond them. This is shown, for instance, in the impredicative nature of the iterated KF theory of Fujimoto (2011) and Jäger *et al.* (1999). So, there is nothing that makes hierarchies themselves a source of mathematical weakness. Rather, I believe the real worry behind the weakness objection is a general one about the state of our current theories of truth, which for the most part turn out to be mathematically weaker than we might require for some applications. The weakness problem is thus, as I see it, not a problem for hierarchies. It is a general problem for all of us working in theories of truth.

When we turn to the clumsiness objections, I believe that the objections are good as applied to the orthodox Tarskian hierarchy, but not to mine. Let us begin with the Nixon-Dean objection, which many (myself included) take to be decisive against the orthodox Tarskian approach. Though I think it is good against the Tarskian hierarchy, it does not apply to mine. The kind of hierarchy I endorse is very coarsely stratified, and so, it makes plenty of room for Nixon-Dean cases. This is made especially clear if we take my model-theoretic version, which builds on the Kripke construction. Thus, my hierarchy has no more trouble with the Nixon-Dean case than Kripke's own theory, and it solves it just the same way he does. Within the fixed point, we have lots of room for the kinds of complicated self-applications of truth the case presents.

Not only is the stratification I propose very course-grained, it does not rely on syntactically defined levels. Thus, the Nixon and Dean sentences are well-formed. Assuming there are no Liar cycles in the Nixon-Dean discourse, they will turn out to be grounded, and so get truth values according to the fixed point. Truth does find its own 'level' (as Kripke says), in the long iteration strategy, and so, the kind of failing Kripke illustrates for the orthodox Tarskian hierarchy are not problems for hierarchies based on long iteration.

There is one way in which we might see a variant Nixon-Dean problem, but I am not sure it is really a problem. If Nixon and Dean were to go in for some strengthened Liar discourse, we might have to hold there is a shift of levels within their discourse. There are two reasons I am not sure this is a problem. First, setting up this sort of discourse would be very unnatural compared to the original Nixon-Dean case, which though a little contrived, is an extension of a situation that occurs all the time. (In the Watergate scandal, lots of people were called liars!) So, the degree of clumsiness we might encounter would be at worst very small. But second, in the kind of case we are imagining, I am not sure the hierarchy would not be the right answer anyway. If Nixon and Dean were to create a complex Strengthened Liar, and then walk through the Strengthened Liar reasoning together, the hierarchy gives a very natural explanation of what is happening. Hierarchies may well show some clumsiness, but they also show theoretical naturalness in some cases too. So, as objections to theories go, the clumsiness one has relatively little force against my proposal.

The same can be said for the generalization version of the clumsiness objection. Because the hierarchy is not syntactically based, there is no barrier to forming all the sentence involving truth we might want, including generalizations about truth. The question becomes whether the right ones come out true in our model, or are provable by our theory. Here again, hierarchies can achieve reasonable but limited success. Lots of generalizations about truth come out true, or are provable, in the sorts of hierarchies I discussed in section 4. For instance, the closure of truth under modus ponens is true in the minimal fixed point and provable in KF. Now, the kinds of theories we have been discussing all fail to prove every generalization we might find plausible, and some of them prove things we might find implausible, as discussed at length by Field (2008). But notice, whatever problems these limitations really pose, they are problems about the theories we might adopt at various levels, not problems for the kind of hierarchy I offer. We can observe again that, from the perspective of reflection, our theories turning out to be defective in various ways, including missing or misdescribing some generalizations, is not a surprise. It is what our discussion of reflection showed we should expect. The response from the reflection viewpoint is simply to go in for further reflection and try to correct the defects. That is just what hierarchies do, on my view. The current non-hierarchical theories of truth all fail to secure some generalizations, or generate seemingly incorrect ones. Hierarchies give us a natural way to fix this problem. Thus, if anything, in this respect the hierarchical view can secure more generalizations over the long run than the non-hierarchical theories that have been developed to date.

As before, there is still a sense in which the objection applies. There is an absolute complete picture of truth, be it a theory or a semantics, which the hierarchical view says we cannot ever achieve. The points I made above all really point out that if we make the hierarchy coarse-grained and liberal in other respects, each level does as well as many other (non-hierarchical) theories of truth on offer. Of course, those other theories are presumably not the one complete theory of truth or semantics either. When compared to what we actually have, the hierarchy looks just fine. The objection is that the hierarchy makes a further objectionable step, by declaring that this situation ultimately cannot be remedied, by denying there can be one complete final theory of truth. Proponents of the non-hierarchical theories will presumably say their offerings may be incomplete or inaccurate, but they are merely the best that can be had so far, and they are striving for one single absolute theory of truth they have not yet achieved. Hierarchical approaches, including mine, say there is no such thing. My main point all along has been that from the point of view of reflection, this is not so objectionable after all. But it must be admitted that it is a limitation, and it might be disappointing. If I am right, then it is not really much more than that.[32]

On the weakness and clumsiness objections, I conclude that hierarchies do not fare particularly badly, and might, from the reflection perspective, actually fare reasonably well. They are not perfect, but the role of reflection helps to show why we can live with the problems.[33] Now, let us consider the one concept objection. I am going to argue that as a general point about hierarchies, the one concept objection fails. Not everything that counts as a hierarchy falls prey to the one concept objection. Even so, some do, including the orthodox Tarskian hierarchy. I shall suggest that my own proposal does reasonably well by lights of the one concept problem, but is not without drawbacks. Along the way, we will see how the ways topics can change in reflection can raise questions about one versus many concepts.

To discuss the general one concept problem, I shall introduce a general notion of stratification. A concept is *stratified* if we cannot provide a single theory or definition for it. Instead, we provide a family of related theories or

[32] In the background here is the issue of absolute generality, as hierarchies one way or another deny the possibility of expressing absolute generality. The attitude I am taking here about generalizations is much the same as the one I took about absolute generality in Glanzberg (2004b, 2006). The other papers in Rayo and Uzquiano (2006), and will give a good indication of the state of that debate.

[33] The careful discussions of hierarchies in Field (2008) and Halbach (2011) come to conclusions about these objections somewhat similar to mine, but disagree sharply on how well we can live with the problems we do find.

definitions, each of which is systematically connected to others. In effect, a concept is stratified if when we try to analyze it, we wind up with a hierarchy. However, we will see examples of stratification which look rather different than the kinds of hierarchies we have been discussing so far, so a different term seems in order.

The orthodox Tarskian hierarchy of languages, and the kind of hierarchy I have endorsed, are both examples of stratified theories of the concept of truth. The Tarskian hierarchy offers a family of truth predicates, though they are systematically related. (For instance, except for indexing of levels, we have the same definition or theory at each level.) My own suggestion here is also a family of definitions or theories of truth (thought of as the result of reflection on the semantics of the language). Again, they show the same systematic relations, and we move from level to level by stepping back and reflecting on the results at the previous stage.

The hierarchies we have considered, both my own variant and the Tarskian original, present truth as a stratified concept. But this is not the only way stratification can arise. For comparison purposes, let me mention two other cases. First, the case of mathematical proof, which I have discussed at length in Glanzberg (2004c). Recall that the incompleteness phenomena show us that concepts of mathematical proof are typically stratified. Suppose we begin with some formal theory for some reasonable piece of mathematics, say PA as a theory of arithmetic. This gives us an articulation of a concept of mathematical proof, in the specific domain of arithmetic. Furthermore, we might think of it as the result of reflection upon our mathematical practices, restricted to arithmetic. We know from the incompleteness theorems that neither PA nor any other reasonably good, recursively axiomatizable, theory of arithmetic will be a complete theory of arithmetic. Insofar as we have an implicit concept of proof in arithmetic and articulate it by reflection as a theory like PA, we have not completely articulated our concept. As we have come to expect, the task of reflection does not always yield complete theories. But for proof, we can say more about what is left out. We also know from the incompleteness theorems that statements of consistency or soundness are not provable in the theory. What is left out, at least, is the fact that the theory is *correct*.

Kreisel (1970) observed that this sort of fact is really implicit in our accepting or using the theory. He writes(Kreisel, 1970, p. 489), "What principles of proof do we recognize as valid once we have understood (or, as one sometimes says, 'accepted') certain given concepts?" His answer to the question is that statements of soundness are so recognized. Transposing Kreisel's proposal into the setting we are working in here, we start with an implicit concept of proof for arithmetic, and reflect upon it and produce an articulation, yielding a theory like PA. But we do not merely reflect as an exercise in theory construction. We make something that was implicit explicit. Insofar as we have a practice of doing arithmetic, we make its features explicit by articulating PA, and we also step into a practice of doing arithmetic formally, in PA. We thus use the theory, in a practice much like the one we had before but now more formally articulated. As Kreisel notes, doing so implicitly commits us to the soundness—the correctness—of our articulation. We are thus implicitly committed to the soundness of PA. That is not explicit, as it is not a consequence of PA, but it is implicit.

This further implicit content is available to reflection, which produces a further theory to capture something like PA together with its soundness. For theories like PA, in fact, we can make the soundness of the theory explicit as a *reflection principle*:[34]

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

Our new articulation is the theory PA plus the reflection principle RFN_{PA} for PA. This is a stronger theory.

We know not only that the result is a stronger theory, but also that the process we just started is open-ended. The new theory $PA + RFN_{PA}$ is (of course) incomplete, and we can again engage in just the same kind of reflection and produce a theory which adds an additional reflection principle for the new theory. The result is an open-ended family of theories, each a stronger articulation of the concept of proof in arithmetic with which we began.[35]

There are a number of ways one can extract morals from this case.[36] For our purposes, let me highlight what I take to be the important aspects of the case. Reflection generates a natural stratified analysis of the concept of

[34] The connection between 'reflection' and 'reflection principle' is probably no accident, and I am certainly exploiting it. This is the 'uniform reflection principle'. For more on reflection principles, see Feferman (1962), and Kreisel and Lévy (1968).

[35] It is natural to ask if we can get a single complete theory by iterating this process to a suitable end-point. Results in this area are quite subtle, but substantially negative. See the papers of Feferman (1962), Feferman and Spector (1962), and Visser (1981).

[36] I discussed the connection with Hilbert's program, and some specific issues about revenge paradoxes, at greater length in Glanzberg (2004c).

proof (mathematical proof, e.g. for arithmetic). Further results show that such analyses are unavoidable, and so, the concept really is stratified; but the role of reflection helps us to see how that situation can emerge and is natural. The kind of stratification we see here, where a single concept is only analyzable by a related family of theories, is one way stratification can occur. To give it a name, let us call it *Kreiselian stratification*, to highlight Kreisel's insight about what is implicit in accepting a theory.

The case of Kreiselian stratification is important for turning aside the one concept objection. Concepts of mathematical proof are stratified, and the Kreiselian view makes them in important respects similar to the case of truth. Not only are they stratified, but the stratification is evident in a process of reflectively articulating stronger and stronger theories. Moreover, as with the truth case, some deep underlying phenomenon, in this case Gödelian incompleteness and other results, guarantee that stratification is unavoidable.

Though we find proof to be stratified, it does not appear to be vulnerable to the one concept objection. What we have is a family of *theories*, all of which are recognizably articulations the concept of proof in arithmetic. They differ in strength, not subject-matter. Now admittedly, I am not offering a worked-out criterion for when we have distinct concepts versus distinct theories of the same concept. We can, of course, identify specific concepts going with our theories, like proof in PA, proof in $PA + RFN_{PA}$, etc. But these all seem clearly to be sub-concepts which simply map to theories. At the risk of appealing to a brute intuition, I think we can count this as a case of one concept with multiple theories. If that is right, then the mere presence of stratification does not indicate multiple concepts; at least, Kreiselian stratification does not. So, the one concept objection does not succeed as a general objection to stratification.

Once one goes looking for stratification like the Kreiselian variety, it is not that hard to find. Set theory, for instance, provides another example. If we start with a good set theory (say ZFC), we can build up stronger and stronger theories. One way to do this is to add more and more large cardinal axioms, which in many cases also flow from what are called reflection principles. These are different from proof-theoretic reflection principles, and express the idea that the universe of sets is maximally large. We might even tell a story about how this is implicit in the concept of set. I shall not explore this is detail, but simply note the appeal of the same intuition as in the Kreiselian case, that we have stratification but one concept.[37]

The one concept objection fails in general, as not all instances of stratification are unacceptable multiplication of concepts. At the same time, it appears the orthodox Tarskian hierarchy looks rather bad by lights of the one concept objection. After all, each new language with a new truth predicate seems to offer a new concept, and the sense in which they are all related is one which is highly implicit in the resulting hierarchy. On further inspection, however, it turns out the difference between the Kreiselian and Tarskian cases is not all that easy to specify. In the Kreiselian case, we also have distinct predicates. There are distinct predicates $Prov_{PA}$, $Prov_{PA+RFN_{PA}}$, etc., and these figure crucially in the theories we write down. So, what is the important difference? One is that for arithmetic, all these predicates are definable, and we do not need to jump to a new language altogether, even if we define and use new predicates which were not definable in the old theory.

It appears one crucial aspect of the difference between Kreiselian and Tarskian cases is that in the Tarskian one, we have to jump to a new language. That does not ‘feel’ like just articulating a new theory, which generally does not involve shifts in the signature of the language. The difference is sometimes hard to specify, as with increases of strength of theory, we can get ability to define things in the old language that we could not before, which is a lot like expanding the language. I believe that the phenomenon we are observing is the same one we talked about in terms of changing the topic or subject-matter in reflection. Reflection tends to do this in some way, as witnessed by the availability of new predicates, definable or not. But some exercises of reflection require a much more pronounced change of topic, perhaps going with a wholesale change of language. I do not have a full analysis of this difference, but I think we can use it, and the general difference between the Kreiselian and Tarskian cases, to try to measure the force of the one concept objection against the kind of hierarchy I sketched in section 4.

Let us now return to this sort of hierarchy. I shall argue that it is much less vulnerable to the one concept objection than the orthodox Tarskian one. But unfortunately, I doubt it will be possible to see the truth hierarchy, even in the form I prefer, as purely Kreiselian. There are a number of reasons for this. In the proof-theoretic variant, we might have hoped that we could get something

[37] Actually, I think a good story can be told here. Insofar as accounts like the iterative conception of set help to identify our concept of set, then we can observe how the multiple theories all express the iterative conception. Thus, I believe we can explain why we really have one concept.

very Kreiselian, by capturing the levels in the hierarchy by iterating a genuine proof-theoretic reflection principle. In some cases, this is simply not possible. It is known, for instance, that $FS + RFN_{FS}$ is *inconsistent* (Halbach, 1994, 2011). This is due to the ω-inconsistency of FS, and so the result is very specific to FS. It is not known to apply to KF, for instance. But all the same, it shows that adding proof-theoretic reflection principles is not an automatically available route. As I mentioned in section 4, the developments of iterated KF theories take a very different route, relying on a parameterized truth predicate.

The proof-theoretic situation shows that technically, we cannot take genuine Kreiselian structure for granted, and it might not be available. But beyond the technicalities, it also illustrates a genuinely non-Kreiselian feature of the hierarchy. The kind of reflection involved in moving between levels of the hierarchy is more complicated than merely accepting the correctness of the theory we had. If it were not, then adding a proof-theoretic reflection principle would not be able to make the trouble it does for FS.

We see this in the semantic versions of the hierarchy as well, and in the general motivation for the hierarchy I offered. We get the hierarchy because even when we are done with long iteration, we can step back and observe features of our semantics; especially, we can observe Strengthened Liar effects, and other inadequacies of the semantics. That leads us to take the semantics as we developed it as part of the subject-matter for further reflection. I discussed, in very general terms, how that might look in a model theory, and we also glanced at proof-theoretic versions. These steps, which make the semantics itself a topic of investigation, seem to be unlike what we saw in the Kreiselian story. They are not accepting the correctness of our theory, they are rather noting inadequacies of it, and modifying it.

Not only is this unlike the Kreiselian case, it has some important features in common with the Tarskian one. We do not literally have to expand the language, but we do something near to it. In the model-theoretic variant, we treat the prior truth predicate as a distinct predicate, not representing the semantics of the language. Even if the signature of the language does not change, this seems to be allowing it to be a distinct topic from the semantics of the language.

Thus, in terms of shifting the topic, the hierarchy I have proposed is not fully Kreiselian, and shows some Tarskian tendencies. But it still is not fully Tarskian stratification. Most importantly, at any level, we have only one genuine truth predicate. Though we can define other predicates that capture truth in the semantics of prior levels, we never have multiple genuine truth predicates at any level. It thus appears that the hierarchy I have proposed falls in-between Kreiselian and Tarskian hierarchies in the kind of stratification it proposes.[38]

It is worth mentioning that this is a point of contrast between the steps of reflection where we genuinely move up in level of the hierarchy, and the kind of internal steps of reflection that are part of the long iteration strategy. The latter are much more Kreiselian. In the steps of long iteration, we continually refine the one semantic predicate we are working with, and build better and better theories. (The monotonicity of the Kripke jump confirms this, for instance.) But the long iteration strategy does not generate the hierarchy. It is the less Kreiselian steps of reflection that mark the significant levels of the truth hierarchy.

Where does this leave my proposal with respect to the one concept objection? If Kreiselian stratification is completely immune to the objection, and Tarskian stratification vulnerable, then my own proposal falls somewhere in-between. The fact that it does not use multiple truth predicates is a reason to think that it models one concept of truth, and the processes we use to generate the levels is uniformly one of reflecting on the semantic properties of the language. Insofar as the semantic view says that is where the nature of truth is to be found, nothing in the apparatus or the way it is deployed really suggests it offers multiple concepts of truth. To that extent, it is not as badly off as the traditional Tarskian approach.

Where it is vulnerable is that even so, it does recognize distinct semantics at different stages. We may not have multiple concepts of truth, but we do have multiple representations of the semantic properties of the language we speak. We have, as it were, distinct concepts of 'truth as it appeared at a give stage of reflection'. Those are multiple concepts. My own view is that when we think about the semantic view of truth and the process of reflection, we should not be too unhappy about having those multiple concepts, as they arise from the kind of constrained processes of reflection we can engage in. They do not, strictly speaking, show truth proper to be anything other than one concept—it is our attempts at reflection which fragment, not truth. But

[38] I thus depart from the position I took in Glanzberg (2004c). I do still hold most of what I said there; particularly, that the comparison with Kreiselian stratification helps to show why the hierarchical nature of truth is unobjectionable. But, in that paper, I was more optimistic about how close the analogy between the Kreiselian case and truth could be drawn.

all the same, this shows a way in which the hierarchical nature of my proposal is substantial.

I conclude that a hierarchical theory of the sort I have sketched is not without costs, but it also has benefits. On objections like the weakness and clumsiness ones, it actually fares quite well. It is no more vulnerable to them than any other theories currently on the market, and when it comes to Strengthened Liar sorts of cases, it actually looks better and more natural than other non-hierarchical options. On the one concept objection, it does show some non-trivial Tarski-like effects of stratification, but not so much as to undermine its status as the results of reflection on one single concept of truth. Though I do grant this is a cost, it also has the benefit of allowing the kind of explanation of Strengthened Liar cases I think is appealing and natural. When combined with the kind of motivation for the hierarchy the semantic view of truth provides, it seems to me the benefits outweigh the costs.

The main cost of the hierarchy is its failing to provide a single complete theory of truth. Again, I grant this is a cost, as such a theory would be very nice to have. But as the discussion of the Kreiselian stratification shows, in many cases we find not achieving that goal not to be such a high cost after all. More importantly, when we measure the hierarchical approach not against our desires for complete theories—a desire that often cannot be fulfilled—but rather against other real options, hierarchies come out looking surprisingly good. I know of no approach to the paradoxes which does not have some costs that are hard to swallow (otherwise we would hardly call them paradoxes!), but I claim that hierarchies have fewer costs, and enjoy more solid motivations, than it is often supposed. We should accept the hierarchical theory.

6 Comparison with Deflationism

To conclude, I shall return briefly to the issue of the nature of truth with which I began. I have argued that if we assume a semantic view of truth, hierarchies are a defensible approach to the paradoxes. Most importantly, as I have argued throughout this paper, the semantic view of truth provides a solid motivation for hierarchies. The semantic view of truth makes the process of reflection substantial, and allows for a complex concept of truth. That, in turn, was the engine that produces hierarchies according to my proposal.

Moreover, I have argued that the kind of hierarchy reflection generates is one that is not vulnerable to a number of objections to the orthodox Tarskian hierarchy, and generally, hierarchies of my preferred sort provide useful and workable theories.

All this assumes the semantic view. If you adopt a deflationist view, then the results are very different. I shall not make any claim about whether deflationists can accept hierarchies in general, but virtually everything I proposed here about how hierarchies are generated, what they describe, and why they are plausible, is unavailable to a deflationist. My defense of the hierarchy is not one a deflationist can use.

First and foremost, if you adopt a deflationist view, then there is no underlying nature of truth, which we might implicitly grasp, and which we might make explicit via reflection. Truth is not a substantial semantic concept. As I noted in section 1, it is a simple one with transparent logical properties which are all there is to the nature of the concept. If there is no room for reflection, then the story I told about how hierarchies are generated does not get off the ground. Moreover, the defense of hierarchies I provided does not either. I repeatedly noted how for a very complex concept like truth, we should not expect reflection to generate complete theories, and hence hierarchies are a natural result. But for many deflationists, like the transparency theorists I mentioned in section 1, truth is a simple property, and the intersubstitutability of $Tr(\ulcorner\phi\urcorner)$ and ϕ in non-opaque contexts virtually exhausts what we need to say about it. The defense of hierarchies via complexity is thus not available.[39] There is little reason to think such a concept should show even Kreiselian stratification, much less the sort of I described for truth. Thus, the defense against the one concept objection is not available either. If you are a deflationist, none of the important parts of the defense of hierarchies I have offered here will be available.[40]

[39] Again, there are formal results to back this up. A natural deflationist analog to CT simply uses the Tarski biconditionals rather than the CT axioms. This theory is a conservative extension of PA, as Halbach (2011) discusses.

[40] Though I do not have the space to pursue the matter here, let me mention briefly that I suspect this is why Field (2008) endorses a hierarchy of determinacy operators but not a hierarchy of truth predicates. There are some important issues here, but in very crude terms, if you adopt the semantic perspective, not only can you offer the defense of truth hierarchies I have, you will also find determinacy operators to look a lot like they try to capture a notion of truth. If you adopt Field's own transparency view of truth, on the other hand, thy are clearly distinct conceptually, and the hierarchy seems unnatural and unmotivated for truth.

If one starts with deflationist views of truth, then perhaps the hierarchy just looks unacceptable. At least, the defense of it I offered here is not available. But, as I have tried to show, if one starts with the semantic view, then the hierarchy is a reasonably workable, natural, and well-motivated approach to truth and paradox.

References

Barrio, E. A., 2006. Theories of truth without standard models and Yablo's sequences. *Studia Logica* **82**:1–17.

Barwise, J., 1975. *Admissible Sets and Structures*. Berlin: Springer-Verlag.

Barwise, J. and J. Etchemendy, 1987. *The Liar*. Oxford: Oxford University Press.

Beall, Jc. (ed.), 2008. *Revenge of the Liar*. Oxford: Oxford University Press.

Beall, Jc., 2009. *Spandrels of Truth*. Oxford: Oxford University Press.

Beall, Jc. and M. Glanzberg, 2008. Where the paths meet: Remarks on truth and paradox. In P. A. French and H. K. Wettstein (eds.), *Midwest Studies in Philosophy Volume XXXII: Truth and its Deformities*. Boston: Wiley-Blackwell.

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.

Carey, S., 2009. *The Origin of Concepts*. Oxford: Oxford University Press.

Chomsky, N., 1980. *Rules and Representations*. New York: Columbia University Press.

David, M., 1994. *Correspondence and Disquotation*. Oxford: Oxford University Press.

Davidson, D., 1967. Truth and meaning. *Synthese* **17**:304–323. Reprinted in Davidson (2001).

———, 1977. Reality without reference. *Dialectica* **31**:247–253. Reprinted in Davidson (2001).

———, 1990. The structure and content of truth. *Journal of Philosophy* **87**:279–328. Reprinted in revised form in Davidson (2005).

———, 2001. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 2nd edn.

———, 2005. *Truth and Predication*. Cambridge: Harvard University Press.

Davies, M., 1987. Tacit knowledge and semantic theory: Can a five percent difference matter? *Mind* **96**:441–462.

Etchemendy, J., 1988. Tarski on truth and logical consequence. *Journal of Symbolic Logic* **53**:51–79.

Feferman, S., 1962. Transfinite recursive progressions of axiomatic theories. *Journal of Symbolic Logic* **27**:259–316.

———, 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.

Field, H., 1972. Tarski's theory of truth. *Journal of Philosophy* **69**:347–375.

———, 1986. The deflationary conception of truth. In C. Wright and G. MacDonald (eds.), *Fact, Science and Value*, pp. 55–117. Oxford: Basil Blackwell.

———, 1994. Deflationist views of meaning and content. *Mind* **103**:249–285.

———, 2008. *Saving Truth from Paradox*. Oxford: Oxford University Press.

van Fraassen, B. C., 1968. Presupposition, implication, and self-reference. *Journal of Philosophy* **65**:136–152.

———, 1970. Truth and paradoxical consequence. In R. L. Martin (ed.), *Paradox of the Liar*, pp. 13–23. Atascadero: Ridgeview.

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

Fujimoto, K., 2011. Autonomous progression and transfinite iteration of self-applicable truth. *Journal of Symbolic Logic* **76**:914–945.

Gaifman, H., 1992. Pointers to truth. *Journal of Philosophy* **89**:223–261.

Gauker, C., 2006. Against stepping back: A critique of contextualist approaches to the semantic paradoxes. *Journal of Philosophical Logic* **35**:393–422.

Glanzberg, M., 2001. The Liar in context. *Philosophical Studies* **103**:217–251.

———, 2002. Topic and discourse. *Mind and Language* **17**:333–375.

———, 2004a. A contextual-hierarchical approach to truth and the Liar paradox. *Journal of Philosophical Logic* **33**:27–88.

———, 2004b. Quantification and realism. *Philosophy and Phenomenological Research* **69**:541–572.

———, 2004c. Truth, reflection, and hierarchies. *Synthese* **142**:289–315.

———, 2006. Context and unrestricted quantification. In A. Rayo and G. Uzquiano (eds.), *Absolute Generality*, pp. 45–74. Oxford: Oxford University Press.

Gupta, A. and N. Belnap, 1993. *The Revision Theory of Truth*. Cambridge: MIT Press.

Halbach, V., 1994. A system of complete and consistent truth. *Notre Dame Journal of Formal Logic* **35**:311–327.

———, 1995. Tarski-hierarchies. *Erkenntnis* **43**:339–367.

———, 1997. Tarskian and Kripean truth. *Journal of Philosophical Logic* **26**:69–80.

———, 2011. *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

Higginbotham, J., 1989. Knowledge of reference. In A. George (ed.), *Reflections on Chomsky*, pp. 153–174. Oxford: Basil Blackwell.

Horsten, L., 2011. *The Tarskian Turn.* Cambridge: MIT Press.

Jäger, G., R. Kahle, A. Setzer, and T. Strahm, 1999. The proof-theoretic analysis of transfinitely iterated fixed point theories. *Journal of Symbolic Logic* **64**:53–67.

Kreisel, G., 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).

Leeds, S., 1978. Theories of reference and truth. *Erkenntnis* **13**:111–129.

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.

Moore, G. E., 1953. *Some Main Problems of Philosophy.* London: George Allen and Unwin.

Moschovakis, Y. N., 1974. *Elementary Induction on Abstract Structures.* Amsterdam: North-Holland.

Parsons, C., 1974. The Liar paradox. *Journal of Philosophical Logic* **3**:381–412. Reprinted in Parsons (1983).

———, 1983. *Mathematics in Philosophy.* Ithaca: Cornell University Press.

Priest, G., 2006. *In Contradiction.* Oxford: Oxford University Press, 2nd edn.

Quine, W. V. O., 1970. *Philosophy of Logic.* Cambridge: Harvard University Press.

Rayo, A. and G. Uzquiano (eds.), 2006. *Absolute Generality.* Oxford: Oxford University Press.

Reinhardt, W. N., 1986. Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic* **15**:219–251.

Russell, B., 1912. *The Problems of Philosophy.* London: Oxford University Press.

Scholl, B. J. and P. D. Tremoulet, 2000. Perceptual causality and animacy. *Trends in Cognitive Science* **4**:299–309.

Soames, S., 1984. What is a theory of truth? *Journal of Philosophy* **81**:411–429.

Tarski, A., 1935. Der Wahrheitsbegriff in den formalizierten 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). Original Polish version published in 1933.

———, 1983. *Logic, Semantics, Metamathematics.* Indianapolis: Hackett, 2nd edn. Edited by J. Corcoran with translations by J. H. Woodger.

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.