Erkenn (2014) 79:597–645
DOI 10.1007/s10670-013-9562-0

ORIGINAL ARTICLE

Truth, Revenge, and Internalizability

Kevin Scharp

Received: 13 October 2010 / Accepted: 9 October 2013 / Published online: 13 December 2013

Abstract Although there has been a recent swell of interest in theories of truth that attempt solutions to the liar paradox and the other paradoxes affecting our concept of truth, many of these theories have been criticized for generating new paradoxes, called *revenge paradoxes*. The criticism is that the theories of truth in question are inadequate because they only work for languages lacking in the resources to generate revenge paradoxes. Theorists facing these objections offer a range of replies, and the matter seems now to be at a standoff. I aim, first, to bolster the revenge objections by considering a relation, internalizability, between languages and theories of truth. A theory of truth is *internalizable* for a language iff there is an extension of that language in which the theory is expressible and for which the theory provides an accurate and complete assignment of semantic values. There are good reasons to think that acceptable theories of truth are internalizable for any language. With this internalizability requirement in hand, I argue that properly formulated revenge objections are decisive and that the replies to them are inadequate. Second, I show that the internalizability requirement can be met by a certain theory of truth. The central claim of this theory is that truth is an inconsistent concept and should be replaced with a pair of consistent concepts that can then be used to provide a semantics for our truth predicates. This theory is compatible with classical logic, does not give rise to revenge paradoxes of any kind, and satisfies the internalizability requirement.

1 Introduction

In the last decade, the debate among analytic philosophers over the paradoxes associated with truth has experienced a resurgence of activity. Although the liar is

K. Scharp (✉)
The Ohio State University, Columbus, OH, USA
e-mail: Scharp.1@osu.edu

the most famous of these paradoxes—which I refer to as *aletheic paradoxes*—there are many others, including Curry's paradox and Yablo's paradox.[1] A host of new approaches to these paradoxes have appeared, and disputes about the adequacy of these approaches increasingly turns on the issue of revenge paradoxes. There is some disagreement about how exactly to characterize revenge paradoxes, but most agree that they are generated by reflecting on attempts to classify the sentences that figure in aletheic paradoxes. Revenge paradoxes often feature in criticisms of approaches to the aletheic paradoxes, and these criticisms usually turn on the expressive power of various languages. For example, one might argue that a particular approach to the aletheic paradoxes is inadequate because it avoids the paradoxes only in languages that do not contain certain linguistic expressions (i.e., those used to generate revenge paradoxes). Of course, theorists on the receiving end of these objections have offered a wide range of replies, with the result being something of a standoff.

I aim to strengthen the case against many contemporary approaches to the aletheic paradoxes by putting a new spin on revenge paradox objections. The criticism I develop focuses on a particular relation between theories of truth and the languages to which they are intended to apply: internalizability. Roughly, a theory of truth is internalizable for a language iff there is an extension of that language in which the theory is expressible and to which the theory is correctly applicable. With this notion in hand, I argue that a theory of truth that constitutes an acceptable approach to the aletheic paradoxes is internalizable for every language. Then I engage with the literature on revenge paradoxes and use the internalizability result to argue that many of the most prominent contemporary approaches to the aletheic paradoxes are unacceptable. Finally, I show that my favored approach to the aletheic paradoxes satisfies the internalizability requirement. Thus, the internalizability requirement serves as part of an argument for this approach.

2 Object Language and Metalanguage

The literature on revenge paradoxes and the motivation for the internalizability requirement are complex and probably best introduced by considering some reactions to Tarski's work. Tarski showed how to define what it is to be a true sentence of a certain type formal language in terms of satisfaction, which is akin to a generalized notion of reference.[2] He also proposed a condition for whether such a definition of truth is adequate, which has come to be known as Schema T: for each sentence p of the language for which truth is being defined (i.e., the *object language*), one should be able to derive from the truth definition that p is true iff q, where 'q' is replaced by the translation of p into the language in which the definition is given (i.e., the *metalanguage*). Another of his major contributions was to prove that (very roughly) if a classical language L has the capacity to describe its own

[1] See Friedman and Sheard (1987), McGee (1991: ch. 1), Sorensen (1998), Beall (1999), and Cook (2009) for discussion of other aletheic paradoxes.

[2] Tarski (1933); see also Tarski (1944) and Tarski and Vaught (1956).

syntax, then L does not contain a predicate that is true of all and only the true sentences of L (assuming truth obeys Schema T). One way to prove it is by reductio—assume the opposite, show that the language in question contains something like a liar sentence, and derive that it is true and not true. An example of a liar sentence is:

(1) (1) is not true.

It requires minimal logical resources to derive that (1) is true and (1) is not true as long as one can make use of Schema T and the fact that '(1)' is the name of '(1) is not true'.[3] This result led Tarski to remark that the metalanguage used to construct the truth definition must be "essentially richer" than the object language for which the definition is constructed.[4]

One familiar way to apply Tarski's work to natural languages is via the familiar idea of a hierarchy of artificial languages. What Kripke calls "the orthodox approach" is that natural language truth predicates are ambiguous and can express any one of the Tarskian truth predicates from the hierarchy on an occasion of use. Of course, Kripke goes on to present several devastating objections to the orthodox approach—objections that led an entire generation of truth theorists to follow Kripke's lead in seeking alternatives that avoid its shortcomings.

Kripke's own approach (for simplicity I focus on the internal Strong Kleene minimal fixed point version[5]) avoids one problem for the orthodox approach in that it defines what it is for a sentence to be true for a language that contains a univocal truth predicate. Of course, the language in question is not classical, and the truth predicate is specific to that language (i.e., it is 'true-in-L' for a language L), but the truth predicate does obey something like Schema T—for any sentence p of the language in question, p and 'p is true' have the same truth value (indeed, p and 'p is true' are intersubstitutable *salva veritate*).[6] Thus, there is no worry for Kripke's approach about how to apply it to languages like English that seem to contain their own univocal truth predicates. However, another problem remains. Kripke remarks:

> It seems likely that many who have worked on the truth-gap approach to the semantic paradoxes have hoped for a universal language, one in which everything that can be stated at all can be expressed. … Now the languages of the present approach contain their own truth predicates and even their own satisfaction predicates, and thus to this extent the hope has been realized. Nevertheless the present approach certainly does not claim to give a universal language, and I doubt that such a goal can be achieved. First, the induction defining the minimal fixed point is carried out in a set-theoretic metalangauge, not in the object language itself. Second, there are assertions we can make about the object language which we cannot make in the object language. For

[3] The resources required depend on the method of proof, but even many non-classical logics like intuitionistic logic and the logic of relevant implication (R) are strong enough.

[4] Tarski (1944: 352). See DeVidi and Solomon (1999) and Ray (2005) for discussion.

[5] I assume the reader has some familiarity with Kripke (1975); see Field (2008b: ch. 3) for discussion.

[6] Because of certain features of the conditional in Strong Kleene languages, not every instance of Schema T in the minimal fixed point language is true—some are gappy.

> example, Liar sentences are *not true* in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate. If we think of the minimal fixed point, say under the Kleene valuation, as giving a model of natural language, then the sense in which we can say, in natural language, that a Liar sentence is not true must be thought of as associated with some later stage in the development of natural language, one in which speakers reflect on the generation process leading to the minimal fixed point. It is not itself a part of that process. The necessity to ascend to a metalanguage may be one of the weaknesses of the present theory. The ghost of the Tarski hierarchy is still with us.[7]

When Kripke says that we cannot say in the object language that liar sentences are not true, he means that if p is a liar sentence of the object language, then the sentence ‘p is not true’ of the object language is not true (it is gappy). If one were to say in the object language that ‘p is not true’ is not true, then one could infer from this claim via the standard liar reasoning that ‘p is not true’ is in fact true. However, Kripke does seem to think that it is appropriate to say *in English* that a liar sentence of the object language is not true. Indeed, he says it himself. Thus, he thinks that as a model of English, the object language leaves something to be desired. We can, he suggests, think of the fragment of English without these resources as an object language and then construct a metalangauge for it in which it would be appropriate to say that liar sentences are not true. Following this route, we end up with a hierarchy of languages that is somewhat akin to the hierarchy utilized by the orthodox approach.[8]

There is a further reason for Kripke to “ascend to a metalanguage.” According to his approach, sentences of the object language are true, false, or gappy, and these categories do not overlap. However, his object language cannot have a predicate that expresses gaphood as long as the language is not trivial. For, if it did, it would also contain a sentence like the following:

(2) (2) is either false or gappy.

One can derive that (2) is true and either false or gappy in the following way. Assume for reductio that (2) is true. If so, then ‘(2) is either false or gappy’ is true by substitution. Then (2) is either false or gappy by Schema T. If (2) is either false or gappy, then (2) is not true. Thus, by reductio, (2) is not true. If (2) is not true, then (2) is either false or gappy. If so, then ‘(2) is either false or gappy’ is true by Schema T. Then (2) is true by substitution. Contradiction.

One can use this sort of result in an analog of Tarski’s theorem, except in this case it would prove the undefinability of gaphood instead of truth. Obviously, the derivation of a contradiction relies on certain classical inferences that are invalid in the object language. Still, given the passage above, Kripke seems to think that we

[7] Kripke (1975: 714).

[8] See Field (2008b: 214–224) for discussion.

can reason classically about his object language.[9] If that is correct, then it is hard to see how he could deny the validity of this derivation.

The derivation of a contradiction by reflecting on the status of (2) is an example of what has been called a *strengthened liar paradox* or a *revenge paradox*. There are many kinds of revenge paradoxes and they are often used to formulate objections to approaches to the aletheic paradoxes. They are the topic of Section Four.

Since Kripke is one of the fiercest opponents of the orthodox approach and its use of the distinction between object language and metalanguage, one might think that this revenge problem would lead him to reject his own approach. Instead, he offers the following footnote in his defense.

> Such semantical notions as ‘grounded’, ‘paradoxical’, etc. belong to the metalanguage. This situation seems to me to be intuitively acceptable; in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers.[10]

The idea here is that a hierarchy of languages where each contains a restricted truth predicate is not an acceptable model of natural language “in its pristine purity”, but a hierarchy of languages where each contains restricted predicates of the type he mentions *is* acceptable. I fail to see how this suggestion alleviates the worry about saying that liar sentences of the object language are not true; after all, ‘the liar sentence is not true’ contains no semantic vocabulary other than ‘true’. Moreover, I argue in Section Four that this sort of gambit is unacceptable even for the cases of the semantic vocabulary Kripke mentions.

Other theorists working on the aletheic paradoxes see the recourse to a metalanguage as a much more urgent problem. For example, William Reinhardt expresses this sentiment in the following passage:

> Let us suppose, as I believe is intuitively correct, that one of the primary features of [truth] is that it is one notion: in particular it does not split into some hierarchy of notions. … Let us explain that the truth predicate of our formal language (call the language L) is intended to be taken in the sense of our preexisting informal notion of truth. … Unless we are prepared to entertain splitting the notion of truth, we are forced to admit that the metalanguage is included in the object language. If the formal language is to provide an adequate explication of the informal language that we use, it must contain its own metalanguage. I take it that this is in fact a desideratum for success in formulating a theory of truth.[11]

[9] See also Kripke (1975: 700n18).

[10] Kripke (1975: 714n34).

[11] Reinhardt (1986: 227–228).

Reinhardt's mention of splitting the notion of truth is a reference to the claim made by proponents of the orthodox approach (described above) that natural language truth predicates are ambiguous and can express any one of the truth predicates in the Tarskian hierarchy of languages. He makes it a condition on any adequate theory of truth that constitutes an approach to the aletheic paradoxes that it not require an expressively richer metalanguage for its formulation. Reinhardt seems to think that this is the only way to treat the truth predicate of natural language as univocal. This does not seem like a good argument since Kripke's approach fails to meet the requirement but is compatible with the claim that natural language truth predicates are univocal, at least when found in their "pristine purity".

In a similar spirit, Vann McGee proposes the "integrity of language" requirement, which states, "[i]t must be possible to give the semantics of our language within the language itself."[12] McGee says of his requirement that it "is intended to hold open the possibility that the methods we develop can be applied to natural languages. If in developing the theory of truth for a language, we required the services of an essentially richer metalanguage, that possibility would be closed off. … [It] makes it reasonable to hope that our methods can be used to get a semantics of a natural language."[13] Given this explication, it is clear that McGee's integrity of language requirement is a condition on the acceptability of an approach to the aletheic paradoxes, not a condition on accounts of language. Any theory of truth that requires a metalanguage for its formulation that is expressively richer than its object languages fails McGee's requirement.[14]

In what follows, I argue that McGee's integrity of language requirement is too strong—we have no reason to think that our natural languages, as they are now, are semantically self-sufficient. However, in a different way, his requirement is too weak—there are theories of truth that do not require an expressively richer metalanguage but are unacceptable for reasons that are similar to those presented by McGee. Instead, one can formulate a condition that is similar to the intuition Reinhardt and McGee defend and use it show that a wide range of approaches to the aletheic paradoxes are unacceptable. That is the goal of the next two sections.

3 Internalizability

3.1 Theories of Truth

Let us start by being more precise about theories of truth. There are two broad categories mentioned in the literature on the aletheic paradoxes: axiomatic theories and semantic theories. Usually an axiomatic theory is a set of sentences of some formal language closed under logical entailment. However, in the truth literature, 'axiomatic theory' is more of a general term for any of the three major proof

[12] McGee (1991: 159).

[13] McGee (1991: 159).

[14] See McGee (1991: ix, 83–86; 1994: 628–629) for arguments for his requirement and Gupta (1997) and Eklund (2008) for criticism of these arguments.

theoretic structures—the other two categories being natural deduction systems and sequent calculi.[15] As such, a better term for this class of theories would be *proof theories of truth* ('axiomatic theory' is pretty entrenched, but I use 'proof theory' in what follows). Note that the logic according to which theorems are deduced is usually classical, but there are some non-classical proof theories of truth (e.g., Horsten's PKF[16]).

On the other hand, semantic theories of truth use set theoretic techniques (e.g., recursive definitions) to define a truth predicate for a particular artificial language. In fact, a semantic theory of truth defines a class of models for some artificial language that contains a truth predicate. Here, 'model' means a particular set theoretic object studied in the branch of mathematical logic known as *model theory*. As such, a better term for these theories would be *model theories of truth*. The term 'semantic theories' is relatively entrenched and trades on a sense of 'semantic' that goes back to Tarski. Nevertheless, I use the 'model theory' in what follows.

There is a third kind of theories of truth that do not fit neatly into either the proof theoretic or model theoretic categories—these theories determine a semantics for the truth predicate in the sense of contemporary linguistics and philosophy of language. A semantics in this sense is a theory that specifies the meanings of sentences containing a certain linguistic expression. The most familiar kind of semantics imbibes as input a sentence uttered and a context of utterance, and it excretes as output a truth value for that sentence in that context of utterance. The class of truth values assigned to the sentence across different contexts of utterance constitute the *truth conditions* for the sentence. The semantics might utilize information about the context of utterance, the proposition expressed by the sentence, or even standards supplied by an assessor of the sentence.[17] Instead of specifying truth conditions, a semantics might provide inferential roles or context change potentials. Providing a semantics for the truth predicate in this sense of 'semantics' constitutes a third kind of theory of truth. In what follows I refer to them as *semantics for truth*.

Three points should be emphasized. First, proof theories and model theories of truth need not be equivalent—some model theories of truth cannot be given a complete proof theoretic formulation. Moreover, finding a proof theory of truth that has as theorems all the sentences deemed valid in a certain class of models can be a non-trivial problem, especially when those models are non-classical.[18] Second, neither proof theories nor model theories of truth are theories of *truth*. Rather, they are theories of language-specific truth predicates, 'true-in-L', where 'L' is the name of some artificial language. When one presents either kind of theory, one claims that the artificial languages to which the theory applies serve as good models for natural

[15] For example, Halbach and Horsten (2006) contains a sequent-calculus theory of truth and Horsten (2011) defends a natural-deduction system. See Halbach (2011) for discussion.

[16] Horsten (2011).

[17] Sentences whose truth values change across contexts of utterance are often called context-dependent; see Kaplan (1989) for a classic treatment. Sentences whose truth values depend on judgments or standards from an interpreter are often called assessment-sensitive; see MacFarlane (2005a) for discussion.

[18] See Halbach (2011) and Burgess (2012) for discussion and details.

languages and that the language-specific truth predicates the theory describes serve as good models for the truth predicate of a natural language.[19] The use of language-specific truth predicates is ubiquitous in the literature on the aletheic paradoxes. Third, if we say that a theory of truth is just a definition of a (language-specific) truth predicate, then we can treat proof theories of truth, model theories of truth, and semantics for truth as instances of a single general schema. To do so, one needs to assume that the axioms and rules of a proof theory of truth are constitutive principles for or implicit definitions of the truth predicate they contain.[20] This take on theories of truth works well with some familiar examples: (i) Tarski's theory of truth defines a truth predicate for certain classical languages,[21] (ii) Kripke's theory (e.g., the Strong Kleene minimal fixed point) provides a recursive definition of a partially defined truth predicate for certain non-classical languages,[22] (iii) Gupta and Belnap offer a circularly defined truth predicate for a wide range of languages,[23] (iv) Field's theory defines a truth predicate too—for philosophical reasons, his definition is disquotational (i.e., it is just a set of T-sentences, or just an intersubstitutability principle), which we can treat as an axiomatic theory, but the biconditional in Field's T-sentences is illuminated by a model theory and an incomplete list of rules and axioms,[24] and (v) Beall's theory is like Field's in that they are both deflationary, but the biconditional in Beall's T-sentences has a model theory and a proper proof theory (it is the biconditional for the relevance logic BX).[25]

A theory of truth assigns semantic values to the sentences of the artificial languages to which it applies. Truth and falsity are the most familiar semantic values, but others are possible as well (e.g., gappy, glutty, indeterminate, ungrounded, and categorical). Given that a theory of truth is a definition of a truth predicate (or truth predicates—I shall omit this reminder from here on), it might not be obvious how a theory of truth assigns semantic values to sentences of its object languages. Consider a more familiar definition—the recently adopted definition of 'planet' by the International Astronomical Union in 2006: a celestial body that is (a) in orbit around the Sun, (b) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes a hydrostatic equilibrium, and (c) has cleared the neighborhood around its orbit.[26] This definition specifies conditions on what it takes to count as a planet. Anything that satisfies the definiens is classified by the definition as a planet; everything else is classified as not a planet. We can think

[19] Here, 'model' is used in an informal way familiar in science—a model of something is similar in certain ways that thing, and the model is used to illuminate, explain, predict, or control some aspect of the thing in question. Beall (2008b) is especially clear about the use of artificial languages as models (in this sense) of natural languages. See Frigg and Hartmann (2006) for an overview of the philosophical literature on models in science.

[20] The idea that a theory of X should take the form of a definition of an expression for X is prominent in Suppes (2002).

[21] Tarski (1933).

[22] Kripke (1975).

[23] Gupta and Belnap (1993).

[24] Field (2008b).

[25] Beall (2009).

[26] IAU (2006).

of the definition as assigning certain values (planetary values) to all the objects in the universe. The same goes for a definition of a truth predicate (i.e., a theory of truth). It classifies the sentences in its object languages as true, false, or perhaps something else. There are, of course, many kinds of definition and we need not think of a theory of truth as providing an explicit lexical definition in this way. Indeed, proof theories are most like implicit definitions, model theories are most like extensional definitions, and semantics are most like intensional definitions.

3.2 Internalizability Definition

Now that the preliminaries are out of the way, we can turn to what I take to be an important relation between a theory of truth and a language. We need a term for a theory of truth that assigns semantic values to all the sentences of an object language—I use ‘descriptively complete’:

> A theory of truth T is *descriptively complete* for a language L iff T assigns some semantic value to each sentence of L.[27]

Notice that descriptive completeness is a relation between a theory and a language. We can define various properties of descriptive completeness for theories by quantifying over languages.

We also need a term for a theory of truth that assigns the *right* semantic values—I use ‘descriptively correct’:

> A theory of truth T is *descriptively correct* for a language L iff every sentence of L to which T assigns a semantic value has the semantic value T assigns it.

Again, descriptive correctness is a relation between a theory and a language. However, a theory of truth does not apply directly to natural languages—instead it applies directly to artificial languages that then more or less model our natural language. That is at least how a lot of work in philosophy of language and philosophical logic gets done. A theory of truth classifies the sentences of an artificial language properly if we would judge the sentences of English represented by the sentences of the artificial language to be as the theory classifies the sentences of the artificial language. Again, there are lots of debates about how to deal with intuitions in philosophy and speaker’s intuitions in linguistics, but this kind of explanation has had many successes in both fields.[28]

There is a risk of circularity in the definition of descriptive correctness—it might seem that the key idea is that the theory’s assignments are *true*. Even if this were the case, I do not think that it would undermine the legitimacy of the definition or the use to which it is put. However, I do not think that the definition of descriptive

[27] There are some technical issues here about whether a theory like Kripke’s assigns some alternative semantic value (e.g., gappy) to certain sentences or whether it assigns no semantic value to them; for our purposes this debate does not matter. Descriptive completeness is akin to Jc Beall’s Exhaustive Characterization Project: “explain how, if at all, we can truly characterize—specify the ‘semantic status’ of—all sentences of our language (in our language).”(Beall 2007: 329–330).

[28] See Devitt (2006) and Ludlow (2011) for discussion of these issues.

correctness requires an explicit use of truth. All it requires is that if the theory says that something is X then that thing really is X.

Now we are ready to introduce the central concept:

> A theory of truth T is *internalizable* for a language L iff there is an extension L′ of L such that T is expressible in L′, T is descriptively complete for L′ and T is descriptively correct for L′.

Notice that internalizability is a relation between a theory of truth and a language. That is an important difference from the account given by McGee, which focuses on properties of languages. We will see that it is more effective for us to use internalizabilty considerations than for us to concentrate on properties of languages.

It could turn out that a particular theory is internalizable for one language but not for another. For example, if a theory of truth gives inconsistent results for certain sentences containing a specific word w, but the theory does not itself consist of any sentences that contain w, then it might be internalizable for a language that does not contain w and fail to be internalizable for some other language that does contain w. In this example, I am assuming that sentences of the language that contains w do not have contradictory semantic features, so the fact that the theory gives inconsistent results for some of them entails that it is not descriptively correct for that language. We will see in Section Four several actual examples of theories with this feature (e.g., Priest's theory, McGee's theory, Field's theory, and Beall's theory).

Aside from descriptive completeness and descriptive correctness, we need to say something about the expressibility relation. A theory T that belongs to one language L_0 is *expressible* in another language L_1 if and only if for every sentence q that composes T, there exists a sentence p of L_1 such that p is a translation of q.[29] This definition of 'expressible' relies on a notion of translation from one language into another. I assume that a sentence of one language is a translation of a sentence that belongs to another if and only if they have the same or relevantly similar meanings (or contents).[30] There is an additional difficulty here in the notion of content that is preserved in translation; I return to it below.

Much turns on the notion of expressibility in play in these debates, and it is easy find oneself equivocating on it. As such, I would like to be a bit more explicit about several different kinds of expressibility, what I take to be the relations between them, and how they bear on different kinds of languages. Let us distinguish between the following kinds of expressibility:

> A concept c is *extensionally expressible* in a language L iff for some expression w of L, w has the same extension as c.
>
> A concept c is *intensionally expressible* in a language L iff for some expression w of L, w has the same intension (content) as c.

[29] I use the word 'express' in two different ways. Words or sentences express concepts, while languages express sentences or theories. I define the latter in terms of translation and content. I discuss the former below.

[30] For the most part, I ignore the distinction between meaning and content, but the standard way of drawing it is that a context dependent expression has the same meaning in every context, but its content differs from context to context. The distinction for sentential meaning and sentential content is analogous.

> A concept c is *metaphysically expressible* in a language L iff for some expression w of L, w designates the same property as c.
>
> A concept c is *conceptually expressible* in a language L iff for some expression w of L, w has the same constitutive principles as c.

By 'extension', I mean the relevant set theoretic entity (e.g., a set of entities for a monadic predicate). By 'intension' I mean whatever an intensional semantic theory assigns to a linguistic expression. For example, the intension of a monadic predicate is a function from points of evaluation to extensions, where points of evaluation are n-tuples with at least one parameter being worlds, and perhaps another as times; they might have more exotic parameters like judges, standards, comparison classes, or delineations, depending on the kinds of linguistic expressions in the language. Other kinds of expressions are assigned different intensions; e.g., a logical connective like conjunction is assigned a truth function.[31] By 'property' I mean an entity that can be predicated of or attributed to an entity (e.g., whiteness). It makes the most sense in this setting to adopt an abundant conception of properties (i.e., there are maximally many).[32] A concept's constitutive principles are those that govern interpretation in the following sense: if a speaker utters a sentence that entails the negation of what the interpreter takes to be a constitutive principle for a concept expressed by a constituent of that sentence, then an interpreter should take this as strong but defeasible evidence that the speaker and the interpreter mean different things by that word. Violating a constitutive principle is an "interpretive red flag"—an indication of a potential problem in interpretation. This account of constitutive principles in no way commits me to the claim that they are true by virtue of their meanings alone. In fact, on an inconsistency approach to the aletheic paradoxes, which I accept, the T-schema is constitutive of our concept of truth, but the T-schema is not true in general—more on this view in Section Five. Note also that one might explicitly reject a constitutive principle for some concept and still possess that concept (e.g., Vann McGee on modus ponens and the conditional[33]) as long as the person recognizes that it is constitutive and has good reason to reject it. There is much more to be said about this issue, but I have discussed it at length elsewhere, and this should be good enough for my purposes here.[34]

Extensional expressibility in a language does not entail intensional expressibility in that language. Take the classic example of chordates and renates—these two terms have the same extension (since all and only animals with hearts have kidneys), but they have different intensions (since their extensions differ in other possible worlds). The concept of a renate is extensionally expressible in any language with a predicate whose extension is the set of things with kidneys, but it need not be intensionally expressible in such a language.

Metaphysical expressibility pertains mostly to predicates—it is difficult to see how it would be applied to other kinds of expressions. Depending on one's views on

[31] See Predelli (2005) for a recent overview.

[32] See Eklund (2008: 60) for discussion.

[33] See McGee (1985) and Williamson (2006) for discussion.

[34] See Scharp (2013a); this notion of constitutivity is similar to John Burgess's notion of pragmatic analyticity in Burgess (2004).

properties, one might think that if a concept is intensionally expressible in a language, then it is metaphysically expressible in that language, and vice versa. I take no stand on this issue and do not place much emphasis on metaphysical expressibility in what follows.

It might turn out that a concept is conceptually expressible in a language even though it is not extensionally expressible or metaphysically expressible in that language. Take our concept of truth according to an inconsistency view again. As long as a language has a predicate with the same constitutive principles as truth, truth is conceptually expressible in that language. However, if one thinks that truth is an inconsistent concept, then one might deny that it has an extension at all. One will certainly deny that it denotes the property of being true, since on an inconsistency view, there is no such thing. In each of these cases, one might accept that truth is intensionally expressible but deny that it is extensionally expressible if one thinks that truth predicates have a semantics that precludes a fixed extension (e.g., contextualists, revision theorists, or assessment sensitivity theorists). That is, in fact, the view I defend in Section Five. It follows from this example that intensional expressibility in a language does not entail extensional expressibility or metaphysical expressibility in that language.

There is a connection between the kinds of expression relations discussed so far and the notion of content at work in the appeal to translation for the definition of internalizability above. Specifically, we use intensions (e.g., functions from points of evaluation to truth values) as theoretical models of the contents of sentences when doing semantics, but we use constitutive principles as a practical guide to the contents of sentences when engaged in conversation. As such, 'content' is difficult to define. Instead of providing a proper definition, we can get along by saying that two expressions with the same content have the same constitutive principles and the same intensions. I will not speculate on sufficient conditions for sameness of content.

One reason for thinking we need these different notions of expressibility when assessing approaches to the aletheic paradoxes is that, for *artificial* languages, extensional expressibility and intensional expressibility are the most useful. Consider a first order predicate calculus that contains its own truth predicate. If we just consider a single model of the language, then we mean that every sentence of the language that is true in that model is in the extension of the truth predicate of that language in that model and nothing else is in its extension. When we say that it contains its own truth predicate we mean that the predicate in question extensionally expresses the concept of truth (or probably the concept of truth-in-L). If, instead, we consider a class of models, then we have something a bit closer to intensional expressibility since one can think of the different models as different possible worlds. Of course, in a language that is given an explicitly intensional semantics, it is easier to use intensional expressibility.

On the other hand, for *natural* languages, we do not have independent access to the semantics of any of our linguistic expressions. We cannot simply inspect them or rely on our stipulations as in the case of an artificial language. Instead, we have to look at how natural languages are used, which means focusing on conversations and the practice of interpretation. As such, for natural languages, conceptual expressibility is the gold standard, with intensional expressibility playing an important backup role. After all, there are linguistic tests for context dependence, for ambiguity, and for assessment sensitivity, but even their proponents admit that these tests are not definitive.[35] They merely provide evidence for one intensional account over another. If one uses a test for some semantic feature like ambiguity, one has to query natural language speakers and interpret the results, and interpretation is implicitly guided by constitutive principles and the idea of conceptual expressibility. For example, if one wants to figure out whether a word that is being used by another person means *true* (i.e., conceptually expresses the concept of truth), one has a conversation—one interprets that person. If the conversation goes smoothly under this assumption, then one's confidence that the word in question does express the concept of truth grows. By 'smoothly', I mean that the person does not violate any of the constitutive principles for the concept. If, however, the person explicitly rejects a constitutive principle or says something that implies that she does, then the participants in the conversation need to shift its focus to figure out whether they mean the same thing.

Another reason for thinking that constitutive expressibility is fundamental for natural languages is that the extensional, intensional, and metaphysical features of a natural language expression can depend in complex ways on discoveries about the natural world. For example, when we discovered that the speed of light is the same in all reference frames and first arrived at an adequate explanation of this fact in the form of special relativity, we also found out that whether two events are simultaneous is relative to a frame of reference—it can happen that there are two events that are simultaneous from one reference frame but not simultaneous from another. So we discovered that there is no such thing as absolute simultaneity. That is, there is no relation of simultaneity that holds of events independently of a reference frame. Those are discoveries about the world, but they have an impact on our language. Once we find out that there is no relation of simultaneity, we find out that: (i) 'simultaneity' does not designate any such property, (ii) 'simultaneous' does not have as its extension the (non-empty) set of ordered pairs of events that are simultaneous, and (iii) 'simultaneous' does not have as its intension a function that takes a physically possible world to the (non-empty) set of ordered pairs of events that are simultaneous in that world. These would have been the most obvious ways of explaining the semantics of 'simultaneous', but the discovery about the world rules them out. Instead, it has to have some other semantic features. It is possible that an error theory would be appropriate here, or perhaps the word is ambiguous and can mean *x and y are simultaneous with respect to reference frame f* for different choices of 'f', or maybe it is really a three place relation meaning *x and y are simultaneous in z*, or it could be context dependent so that it picks up a frame of reference from the context of utterance, or it might be assessment sensitive so that it picks up a frame of reference from the context of assessment. My money is on the last option, but nothing hinges on this choice.[36] The point is that the semantic

[35] See Zwicky and Saddock (1975), Cappelen and Lepore (2005), Stanley (2005), MacFarlane (2005a), and Cappelen and Hawthorne (2009) for discussion.

[36] See MacFarlane (2007b) and Pinilos (2011) for similar views.

features of natural language expressions depend on the way they are used (in particular, their constitutive principles) and how the world turns out to be. We cannot say the same about artificial languages. So, despite the fact that when we discovered that there is no such thing as absolute simultaneity, we also found out that simultaneity is not metaphysically expressible or extensionally expressible in natural language, we did *not* discover that simultaneity is not conceptually expressible in natural language.

Here is our somewhat awkward situation. Artificial languages are stipulated to have their semantic features, so extensional expressibility and intensional expressibility work well for them—we know the extension of a particular expression of an interpreted predicate calculus because we have stipulated that it has that extension, or we know the intension of a particular expression of an artificial language because we have stipulated that it has that intension. We cannot say the same for natural languages. For words that have established uses, we have to try to work out their intensions and extensions indirectly by figuring out their constitutive principles in conversations with native speakers. We can, of course, coin new expressions of natural language by stipulation, but all that tells us is their constitutive principles—we still have to go through the same process to figure out their extensional and intensional features. Simply stipulating that a certain word has certain semantic features is no guarantee that it actually has those semantic features. Constitutive principles are up to us, extensions are not.

In the definition of internalizability, I appeal to translation as the basis for the expression relation. Which expression relation is relevant? That will depend on which kind of language we are talking about. In most cases of artificial languages, all we have to go on is their stipulated extensional and intensional features. They do not have stipulated constitutive principles and no one uses them in conversation, so there is no matter of conceptual expressibility there. In natural language, we have speakers' intuitions about entailments, contradictions, felicity, and so forth. These guide us in latching onto constitutive principles and intensions. Thus, the notion of expressibility that is relevant depends on the language in question. This is, unfortunately, an unavoidable feature of our use of artificial languages to illuminate natural ones. When dealing with a relation like internalizability that applies across a range of kinds of languages, one should suppose the most appropriate sort of expressibilty relation.

3.3 Internalizability Requirement

In the present subsection, I present and defend a condition on theories of truth insofar as they constitute acceptable approaches to the aletheic paradoxes. The condition is the following:

> (INT) For any language L and theory of truth T, if T is acceptable, then T is internalizable for L.

In short, acceptable theories of truth are internalizable for any language. The argument for this claim rests on two assumptions:

> *Assumption One*: acceptable theories of truth are descriptively complete and descriptively correct for any language whose sentences have semantic values.
> *Assumption Two*: for any two languages L_1 and L_2, L_1 can be extended to a language L_3 and L_2 can be extended to a language L_4 such that L_3 and L_4 are intertranslatable.

The first assumption is that for a theory of truth to be acceptable, it should be able to assign semantic values to any sentences that have them. If a sentence of some arbitrary language is true or false (or something else), then an acceptable theory of truth ought to be able to specify that it is—provided it has enough auxiliary information. Assumption one does not require that a theory of truth adjudicate the semantic value of every claim; that would be preposterous. Instead, it should be able to specify the semantic value of any claim *given enough information about what the claim represents*. In short, it should be able to specify for each sentence that might be true, false or whatever, the conditions under which it would have those semantic values. A theory that fails to do so would leave us with, say, a true sentence but no explanation for why it is true or why we should think that when we call it true we are saying the same thing about it that we are saying about any other claim when we call it true.

Consider, for analogy, the definition of ‘planet’ given above. Notice, first, that this definition, all by itself, does not entail that any particular thing is a planet or is not a planet. It is only given additional information that Jupiter is in orbit around the Sun, has sufficient mass for persistent hydrostatic equilibrium, and has cleared the neighborhood around its orbit, that we can conclude that Jupiter is a planet. Only given that Pluto has not cleared the neighborhood around its orbit can we conclude that it is not a planet. That is analogous with the claim made above about acceptable theories of truth specifying the semantic values of sentences given enough additional information about them.

Imagine for a moment that we have independent reason to think that some particular thing is a planet, but the IAU’s new definition of planethood does not entail that it is a planet even given all relevant information about it. That would constitute good evidence that the definition was unacceptable. The same goes for theories of truth—they should account for true sentences and sentences with the other semantic values wherever they may roam. Failure to do so is tantamount to unacceptability.

Assumption two might seem like Donald Davidson’s notorious claim that any two languages are intertranslatable, but it is considerably weaker.[37] Instead, it is the claim that any two languages can be extended so that the extended languages are intertranslatable. The vocabulary to be added might be extensive, so there is no claim of intertranslatability even for something as expressively rich as natural languages. Just as in the definition of internalizability from the previous section, the standard for translation is sameness of content, but this might depend on the kind of languages in question.

[37] Davidson (1974); see also Davidson (1999).

I consider a technical objection to assumption two in just a bit, but consider by way of motivation for it a case of actual translation between two natural languages like English and German. If we want to translate a German sentence like ‘John fühlt sich Schadenfreude’ into English, we could use something like ‘John feels pleasure at someone’s misfortune’, but ‘pleasure at someone’s misfortune’ does not really capture the specific feeling denoted by ‘Schadenfreude’. That is, the former does not conceptually express the concept expressed by the latter. Instead, we might just add ‘schadenfreude’ to English as a loanword from German (or we might calque the word from German by adding something like ‘adversityjoy’ to English). Then we can translate the sentence into the extension of English as ‘John feels schadenfreude’.

Another example is a case of translation from English into a first order predicate calculus. It is easy to show by nonstandard models of arithmetic that no predicate of a first order language extensionally expresses the concept of well-ordering. So there is no translation of ‘the natural numbers are well ordered’ from English into a first order language. However, if we extend the first order language by adding the apparatus for second order quantification plus the needed arithmetic vocabulary, then we can translate ‘the natural numbers are well ordered’ into the resulting language. Assumption two says that we can always overcome a translation failure by adding resources to one or both languages.

The following is the argument for the central claim (INT). Assume that T is a theory of truth and that T is not internalizable for some language L. Because T is not internalizable for L, either: (i) T is not descriptively complete for L, (ii) T is not descriptively correct for L, or (iii) T is not expressible in L. On option (i), by assumption one, T is not acceptable. On option (ii), by assumption one, T is not acceptable. Consider option (iii), and assume that T *is* descriptively complete for L and that T is descriptively correct for L. Given that T is a theory, there is some language L′ in which T is expressible. By assumption two, L can be extended to a language L″ and L′ can be extended to a language L‴ such that L″ and L‴ are intertranslatable. If T is expressible in L′, T is expressible in L″; if T is expressible in L‴, then T is expressible in L″. Hence, T is expressible in L″. If T is expressible in L″, then either T is not descriptively complete for L″ or T is not descriptively correct for L″. Hence, there is a language for which either T is not descriptively complete or T is not descriptively correct. Therefore, by assumption one, T is not an acceptable theory of truth.

Here is an intuitive summary of the argument. T is not internalizable for L, but T has to be formulated in some language or other; if we consider the result of adding to L whatever expressive resources it takes to express T, then T will be descriptively incomplete for that extended language. For example, assume that T is the inner Strong Kleene minimal fixed point version of Kripke’s model theory of truth. This theory implies that a liar sentence like (1) is gappy. As I argued above and Kripke concedes, it also has contradictory implications for a sentence like (2) (i.e., ‘(2) is either false or gappy’). A proponent of such a theory might try to avoid this problem by restricting the theory to languages that do not contain gaphood predicates (there are, of course, other replies to this objection—like denying that there are fully general gaphood predicates—that will be taken up in the next section). Although the theory of truth in question might be descriptively complete for a language that does not contain a gaphood predicate, it is not both descriptively complete and descriptively correct for languages that have both a truth predicate and a gaphood predicate. Consequently, it is not an acceptable theory of truth. If this objection is cogent, then Kripke's "pristine purity" reply to the revenge objection his theory faces is unacceptable. However, before any anyone wedded to Kripke's theory will find this point decisive, a plethora of obvious objections need replies.

The following is an objection to assumption two voiced by Jc Beall. He argues that some linguistic expressions are incoherent with respect to certain languages; for example, an exhaustive and exclusive operator (i.e., one that obeys excluded middle and ex falso) cannot be added to a paracomplete or paraconsistent language that also contains a truth predicate and the means for representing its syntax without triviality.[38]

Beall defines 'trivial language' in the following way: "a trivial language (or theory) is one according to which everything is true. A non-trivial language is one that isn't trivial."[39] Unfortunately, this definition has a certain difficulty—languages do not make pronouncements about what is true. Perhaps, if one thinks of the expressions of a language as governed by certain constitutive principles, then one might think of a trivial language as having expressions with constitutive principles from which every sentence of the language follows. Still, one might deny that those sentences are true. Instead, a trivial language might be one whose truth predicate has an extension that contains every sentence of the language. That is probably what Beall has in mind here. However, that is still not a very good definition since, according to Beall's claim about EE operators, triviality is a feature of a the consequence relation of a language, not its truth predicate. Granted, if one can prove any sentence of a language from the empty set, that language contains its own truth predicate, and that truth predicate obeys T-In (i.e., if p then p is true), then the truth predicate's extension contains every sentence of the language. I prefer to keep triviality tied directly to a language's consequence relation by defining it in the following way: a language is *trivial* iff for every sentence p of the language and every set G of sentences of the language, p is a consequence of G. This definition says nothing about whether the language has a truth predicate or how that truth predicate might behave. It also avoids the awkward claim some things are true *according to a language*. Moreover, it fits well with Beall's point about EE operators above. That is, if we add an EE operator to a paracomplete language or paraconsistent language that contains its own truth predicate (which obeys Schema T) and has the means to represent its own syntax, then the extended language is trivial (in my sense).

Now that that matter is settled, we can continue with the objection, which might go in one of a couple of directions from here. One might argue that there are no trivial languages, so it is impossible to add an EE operator to one of the languages in question. I do not find this line plausible. One can construct an artificial language that is trivial and reason about it. It seems ad hoc in the extreme to claim that it is

[38] Beall (2008b: 6–7, 13).

[39] Beall (2007: 330n5).

not a language but some other formal structure that is similar in all other relevant respects is a language.

Instead, one might argue that sentences of a trivial language are not translatable with sentences of any non-trivial language. If the inferential role of a sentence determines (at least in part) its content, then one would think that no sentence with a trivial inferential role would be intertranslatable with a sentence that does not have a trivial inferential role.[40] That seems like a genuine worry.

As a reply, imagine we have a language as Beall describes it as a model for a natural language. It is a good model by any account. Then the natural language changes—the people in that linguistic practice start using a new expression. Let us say they use the new expression just as Beall describes an EE operator. What should we say about the new language? We might say that the new expression is really meaningless or that it has some other meaning—it does not mean what everyone thinks it means. I find this option rather implausible. Part of our job as linguists, philosophers of language, or philosophical logicians is to *explain* what we find in natural languages—look at how people use certain words and consider their judgments of synonymy, entailment, contradiction, etc. to construct and defend the best theories of natural languages we can. Pretending as if an expression with an established use is meaningless or has a meaning it clearly does not have is tantamount to a dereliction of duty.[41]

What should we say in a situation like this? We can use the distinction between expression relations to formulate a response. I think the best thing to do is to say is that the word in question means exactly what the participants in the linguistic practice think it means—that is, it has exactly the constitutive principles they think it has. We can, of course, specify artificial languages in whatever way suits our interests, but to say that a language like the one Beall describes cannot be extended to include an EE operator is not a good strategy. Instead, we do better to say that it can, and that operator will mean what people think it means in that it will have what they take to be its constitutive principles. Of course, those constitutive principles might be inconsistent given the other resources in the language. If that is the case, then one will need to appeal to a plausible theory of inconsistent concepts to figure out the semantic values of the expressions and sentences of the language. As I have said, my preferred approach here treats words that express inconsistent concepts as assessment sensitive.[42] I return to this issue in Section Five.

4 Revenge Paradoxes

So far, I have introduced what some have taken to be a problem for certain approaches to the aletheic paradoxes—they require expressively richer metalanguages and so it is unclear how they might apply to natural languages. In Section

[40] Given the emphasis linguists and linguistically oriented philosophers put on entailments, the antecedent seems plausible.

[41] See Dowty et al. (1980: 1–3) for a classic statement of this methodology.

[42] Scharp (2013a, b).

Three, I defined internalizabilty, which is a relation between theories of truth and languages, and argued that any theory of truth that constitutes an acceptable approach to the aletheic paradoxes is internalizable for every language. In this section, I apply the internalizability framework to a wide range of theories of truth and find many wanting. The main theme here is that theories of truth that generate revenge paradoxes are not internalizable for certain natural languages. Given the important role they play in the argument here, a discussion of revenge paradoxes is in order.

Just as there are many kinds of aletheic paradoxes, there are many kinds of revenge paradoxes. For example, the following is a liar sentence:

(1) (1) is not true.

Assume for a moment that we accept the inner theory of Kripke's Strong Kleene minimal fixed point. According to this approach, (1) is not in the extension of 'true' and not in the anti-extension of 'true'. We can characterize (1) by saying that it is gappy. If the object language has this expression as well, then it also has the following sentence:

(2) (2) is either false or gappy.

As we saw in Section One, this sentence seems to give rise to a revenge paradox for the approach under consideration.[43]

Notice that the revenge paradox generated by (2) is paradoxical only if we accept the above approach to the liar paradox. That is an important feature of revenge paradoxes: whether a sentence generates a revenge paradox is relative to a particular approach to the liar. (2) is a revenge paradox for the inner theory of Kripke's Strong Kleene minimal fixed point. It is not a revenge paradox for other approaches.

Other examples of sentences that generate revenge paradoxes are:

(3) (3) is either false or unstable (for revision approaches).[44]
(4) (4) is not true in any context(for contextual approaches).[45]
(5) (5) is either false or indeterminate(for paracomplete approaches).[46]
(6) (6) is just false(for paraconsistent approaches).[47]
(7) (7) is Bnot true[48](for paracomplete and paraconsistent approaches).

Revision theories say that (1) is unstable since its truth value never stabilizes in a revision sequence. The truth value of (3) never stabilizes in a revision sequence either, but if (3) is unstable then (3) is true since it says of itself that it is either false

[43] Notice that revenge paradoxes could be cast in the form of a sentence that figures in Curry's paradox or a sequence of sentences that figure in Yablo's paradox.

[44] See Gupta (1982) and Gupta and Belnap (1993) for revision theories.

[45] See Burge (1979), Simmons (1993), and Glanzberg (2004) for contextual theories. See Juhl (1997) for revenge considerations.

[46] See Kripke (1975), Soames (1999), and Field (2008a, b) for paracomplete theories. See Priest (2005, 2008), Rayo and Welch (2008), and Leitgeb (2008) for revenge considerations.

[47] See Priest (2006a, b) and Beall (2009) for paraconsistent theories. See Thomason (1986), Shapiro (2004), and Field (2008b) for revenge considerations.

[48] 'Bnot' expresses Boolean negation.

or unstable. So (3) is a revenge liar for approaches that appeal to revision sequences to define truth. Contextual approaches say that (1) is true in some contexts and false in others. That approach blocks the aletheic paradoxes since it stipulates that the context shifts in the midst of the reasoning. However, (4) poses a serious problem for contextual views because the claim that it is true in one context seems to imply that it is not true in any context, so the contextualist seems to have a problem. Paracomplete approaches typically say that (1) is indeterminate and they reject certain principles of classical logic involved in the liar reasoning. However, if the paracomplete approach calls (5) indeterminate, then it implies that (5) is true since (5) says of itself that it is indeterminate. Because sentences are not both indeterminate and true, (5) poses a problem for these approaches. Paraconsistent views say that (1) is both true and false and they offer a non-classical logic on which some contradictions are true (though they hold that not all contradictions are true). Paraconsistentists claim that (1) says of itself that it is not true, and it is not true, so it is true as well; (1) is both true and not true. However, there is a problem with saying that (6) is both true and false since (6) says of itself that it is false only. The paraconsistent view on (1) does not work for (6) since the claim that (6) is both true and false should be incompatible with what (6) says of itself—i.e., that it is only false. So the standard paraconsistent treatment of the liar seems to get the wrong answer for (6).

Sentence (7) contains an unusual term for negation. The ‘Bnot’ in (7) expresses Boolean negation. In multi-valued logics like paracomplete logic, Boolean negation takes indeterminacy to truth. So a theory that implies that (7) is indeterminate also implies that (7) is true. Thus, the revenge paradox generated by (7) is a variant of the revenge paradox generated by (5). In paraconsistent logics, Boolean negation takes gluts to truths, so a theory that implies that (7) is glutty also implies that (7) is just true. Hence, the revenge paradox generated by (7) is a variant of the revenge paradox generated by (6).

Consider how Graham Priest has long characterized the revenge paradox phenomenon:

> There is, in fact, a uniform method for constructing the revenge paradox—or extended paradox, as it is called sometimes. All semantic accounts have a bunch of Good Guys (the true, the stably true, the ultimately true, or whatever). These are the ones that we target when we assert. Then there’s the Rest. The extended liar is a sentence, produced by some diagonalizing construction, which says of itself just that it’s in the Rest. The diagonal construction, because of its ability to tear through any consistent boundary, may then play havoc. This shows, incidentally, that the extended paradox is not really a different paradox. The pristine liar is the result of the construction when the theoretical framework is the standard one (all sentences are true or false, not both, and not neither). ‘Extended paradoxes’ are simply the results of applying the construction in different theoretical frameworks.[49]

[49] Priest (2008: 226).

Priest's diagnosis is a good start. One can see the examples above fit well into the schema he provides. I am not one to care much about how to individuate aletheic paradoxes, and I do not think it matters whether revenge paradoxes are distinct from the liar or whether they are at root the same paradox. What matters is that they affect the concept of truth and that any acceptable approach to the liar has to incorporate some adequate approach to revenge paradoxes. However, it will pay to be a bit subtler about revenge paradoxes, and that is exactly what Jc Beall's analysis delivers.

Beall emphasizes that when one gives a formal theory of truth, one specifies an artificial language, L, that contains its own truth predicate 'true-in-L'. The theorist then shows that 'true-in-L' obeys various principles of the formal theory of truth, and the theorist can use L to show that the formal theory of truth is relatively consistent (often using classical logic and set theory in a metalanguage M). Finally, the theorist claims that natural languages are like L in relevant respects, so the theory of 'true-in-L' also applies to truth. Beall lays out three distinct revenge recipes to this sort of project:

1. Find some semantic notion X that is *used in M to classify sentences of L.*
   Show in M that X is not expressible in L unless L is inconsistent or trivial.
   Conclude that L is explanatorily inadequate since it does not explain how natural language, which contains X, is consistent.
2. Find some semantic notion X that is *expressible in M.*
   Show in M that X is not expressible in L unless L is inconsistent or trivial.
   Conclude that L is explanatorily inadequate since it does not explain how natural language, which contains X, is consistent.
3. Find some semantic notion X that is *expressible in natural language.*
   Argue that X is not expressible in L unless L is inconsistent or trivial.
   Conclude that L is explanatorily inadequate since it does not explain how natural language, which contains X, is consistent.[50]

The italics indicate the contrasts between the three recipes. In the first case, the concept X is used by the theory of truth to classify paradoxical sentences, whereas in the second case, the concept X is just expressible in the language of the theory—it need not be explicitly used by the theory. In the third case, the concept X is expressible in natural language and need not even be expressible in the language of the theory. In each case, the problem is that the theory in question does not apply to natural languages, so it does not really solve the aletheic paradoxes.

4.1 Revenge Objections

By using revenge paradoxes like those surveyed above, one can formulate objections to the approaches in question. Namely, the artificial language used to model natural language has an expressive limitation—i.e., it does not contain terms that feature in revenge paradoxes. Thus, the approach in question solves problems posed by the aletheic paradoxes only for expressively impoverished languages. That

[50] Beall (2008b: 11–12).

is, it does not solve these problems in general. In particular, it does not solve the problems posed by the aletheic paradoxes as they occur in natural languages.

Revenge objections are similar to the kind of objection we saw leveled by Reinhardt and McGee in Section Two. However, there are important differences. McGee's integrity of language requirement is that a semantics for a natural language ought to be expressible in that very language. If a theory of truth cannot be expressed in any of its object languages on pain of contradiction (or triviality) from revenge paradoxes, then it fails McGee's requirement. So McGee's requirement, if defensible, would bolster some revenge objections. However, many theories of truth, especially those proposed in the last decade, are expressible in some of their object languages, and so satisfy McGee's requirement.

For example, Field's paracomplete approach classifies liar sentences as indeterminate (i.e., not determinately true and not determinately false) and the object languages for his theory can have their own determinateness operators—of course, they will also have sentences like (5) in them as well. Field's theory implies that (5) is not determinately determinately true. In fact, his determinateness operator iterates non-trivially up into the transfinite. However, Field shows that there is no way to construct a revenge paradox using the resources in his object languages and, moreover, his object languages have the resources to classify every sentence in them as true, false, determinately true, determinately determinately true, etc. Thus, his theory satisfies McGee's requirement.[51] Nevertheless, as I argue, it faces revenge paradoxes. The same goes for Priest's theory, McGee's theory, and Beall's theory.[52] Each one is ingeniously constructed so that the theory does not rely on anything that might give rise to a revenge paradox. Each of these authors rightfully emphasizes that theories with this feature are to be preferred (other things being equal) over theories that do not (e.g., Gupta and Belnap's revision theory).

Why doesn't Field's theory have a problem with sentences like (5)? He interprets the 'indeterminate' in (5) as 'not determinately true and not determinately false', and his determinateness operator iterates non-trivially, so he can say that (5) is not determinately determinately true without thereby implying that (5) is not determinately true. If we use 'D' as a determinateness operator and superscripts for iterations, he can say that (5) is not D^2true, but he cannot say that (5) is not Dtrue. But what if we want to say that a given sentence is not determinately true in any way—it is not Dtrue, and it is not D^2true, and it is not D^3true, …? That, presumably, is how one would think of 'indeterminate' in (5) in the first place without knowing anything about Field's approach. It is impossible to say something like this in Field's object language. If it contained an idempotent determinateness operator ('idempotent' means that it iterates trivially), which we can use 'ID' to express, then Field's theory would entail that a sentence that says of itself that it is either false or neither IDtrue nor IDfalse is both IDtrue and not IDtrue. So, if it is non-trivial, Field's theory cannot apply to languages with idempotent determinateness operators. Field denies that this is a genuine problem, but we will have to come back to this issue in a moment.

[51] Field (2008b).

[52] See Priest (1979, 2006a, b), McGee (1991), and Beall (2009).

The significance of the revenge paradox phenomenon is most fundamentally that we seem to be unable to say something non-trivial and satisfying about the aletheic paradoxes. Something satisfying would be an account of what goes wrong in the reasoning and a way of defining truth and any other notions one would want to use in classifying sentences of languages capable of formulating these paradoxes that applies to any language that has a truth predicate regardless of the other notions expressible in it. So far, we do not know how to do this (or, rather, most theorists have yet to be convinced that it can be done—but in Section Five, I argue that it can). We can formulate satisfying accounts for a wide range of artificial languages, but in each case, there is always an Achilles' heel—some notion that, if it were expressible in a language to which the theory applies, the theory would be trivial: 'true in a context' for the contextualist, 'stable' for the revision theorist, 'idempotent-determinately' for the paracomplete theorist, 'just true' for the paraconsistent theorist, and the list goes on. In each case, even if the theory can be formulated in some of its object languages, and so does not officially use its Achilles' heel notion, it always seems odd to think that there is not and could never be such a notion—which is the position most often taken by proponents of these views—or that the expression in question is meaningless or unintelligible.

4.2 Revenge and Internalizability

Now that we have seen some revenge objections and some ways of countering them, we can link this theme with the other major topic of the paper—internalizability. Assume for a moment that we can formulate a legitimate revenge paradox for Field's theory using an idempotent determinateness operator. If so, then there is a language that contains a sentence like (5), where 'indeterminate' is the idempotent notion, not Field's notion. Field's theory entails a contradiction if it applies to a sentence like this. One way to avoid this result is to restrict his theory. That is, a proponent of Field's theory might deny that the theory applies to languages like this. Assume for a moment that we are engaging with just such a proponent.

The big question now is: does this restriction impact the ability of Field's theory to illuminate or explain our natural language and its concept of truth? If one were wedded to McGee's integrity of language requirement, then one would be impotent to mount any criticism of Field's theory in this regard. After all, his theory is expressible in some of its object languages, so there is no reason to think that we would need recourse to an expressively richer metalanguage if we were to treat English (or something very much like it) as an object language of Field's theory. Thus, even if one could get Field to accept that (5), properly understood, constitutes a revenge paradox, it would be impossible to use McGee's requirement to turn this fact into an objection.

Alternatively, one might argue about whether the English words 'determinately' or 'definitely' express an idempotent notion of determinateness. This turn in the debate is unlikely to be very fruitful. After all, I doubt whether native speakers have much in the way of intuitions about whether 'determinately determinately p' follows from 'determinately p'. Even if one could arrive at some robust data on this front, natural language usage is messy and complex, and there are bound to be competing considerations.

Instead of relying on McGee's requirement or fighting about the proper interpretation of English, we could use internalizability. Even if we assume that English contains no idempotent determinateness operator, Field's theory would fail to be internalizable for some languages if we could just extend English by adding one. That is another benefit of the internalizability condition over McGee's requirement. The internalizability requirement is not satisfied if one could extend English in a way that would preclude descriptive completeness or descriptive correctness for the extended language. Thus, even if Field's theory were expressible in and descriptively complete and descriptively correct for English as it is now, his theory would still be unacceptable given that it would not be descriptively complete and descriptively correct for the extension of English in question. The idea here is that a theory of truth that works for our natural language only by luck—because our language lacks the Achiles' heel vocabulary—is not a good theory. If we were to add that vocabulary, for whatever reason, the theory would be unacceptable. Stephen Yablo puts the point well: "if it is only by happenstance that the paradoxes are avoided, then although the immediate *semantical* challenge is met, *philosophically* we seem not further ahead."[53] This is exactly the worry that the internalizability requirement is designed to capture.

Assuming, still, that the revenge paradoxes are genuine, here is the situation. Field's theory is internalizable for the artificial language he constructs. There is no question about that. Perhaps it is internalizable for English as it is now—I doubt it, but suppose it is—that would happen if English contains no terms that figure in revenge paradoxes for Field's theory. Even so, it is not internalizable for simple extensions of English; e.g., by adding an idempotent determinateness operator. Thus, it is not internalizable for every natural language and so fails to be internalizable for every language. Consequently, it fails to satisfy the internalizabilty requirement. The same result holds for other theories that are expressible in and descriptively complete and descriptively correct for some of their object languages. Priest's theory, McGee's theory, and Beall's theory all have this feature.[54] If the revenge paradoxes that have been proposed for them are genuine, then they have the same status as Field's theory—they might be internalizable for English, but they fail to be internalizalbe for every natural language.

We can classify some familiar theories of truth into categories using the internalizability relation:

(i) Theories that are not internalizable for any language (e.g., Tarski's theory, Kripke's Strong Kleene theory, Gupta and Belnap's theory).[55]

[53] Yablo (1993: 393n12).

[54] Priest (2006a, b), McGee (1991), Beall (2009).

[55] Gupta and Belnap (1993: 253–256) admit that the crucial notion used by their theory—stability (sometimes called categoricality)—cannot be expressed in any of its object languages without triviality. They claim that it is also a circular concept and provide a theory for it that relies on a different notion of categoricality, which, of course, cannot be expressed in any of the object languages of this theory without triviality.

(ii) Theories that are internalizable for some languages but not for all languages (e.g., Priest's theory, McGee's theory, Field's theory, and Beall's theory).
(iii) Theories that are internalizable for every language.

In Section Five, I argue that the last category has at least one member. Again, I am assuming that the revenge paradoxes for these theories are genuine and that the revenge objections to them are cogent. Below, I consider some potential problems for this assumption.

In sum, the internalizability requirement has two major benefits over McGee's requirement. First, it lines up with revenge worries better—it casts doubt on certain theories of truth that generate revenge paradoxes even if those theories can be expressed in some of their own object languages. Focusing on internalizability helps us ignore a red herring—namely, that the key to an acceptable approach to the aletheic paradoxes is a theory of truth that is expressible in one of its object languages. This feature, in no way, guarantees that a theory of truth is "revenge-immune". Second, McGee's requirement focuses our attention on natural languages as they are now, whereas the internalizability condition emphasizes how they might be extended. A theory of truth that works now, but cannot handle legitimate extensions of the language is unacceptable.

4.3 Too Easy?

An advocate of a non-classical approach might respond to the points I have made by pressing the "too easy revenge" worry voiced by Jc Beall (and to some extent echoed by Lionel Shapiro). Beall offers a warning about the efficacy of these objections (I have changed the individual constants—L is the artificial language and E is the natural language being modeled):

> The weight of Rv1 or Rv2 depends on the sort of X at issue. … [I]f X is a classical, model-dependent notion constructed in a proper fragment of [E], then the charge of inadequacy is not easy to substantiate, even if the inexpressibility of X in [L] is easy to substantiate. In particular, if classical logic extends that of [L], then there is a clear sense in which you may 'properly' rely on a classical metalanguage in constructing [L] and, in particular, truth-in-[L]. In familiar non-classical proposals, for example, you endorse that [E], the real, target language, is non-classical but enjoys classical logic as a (proper) extension, in which case, notwithstanding particular details, there is nothing prima-facie suspect about relying on an entirely classical fragment of [E] to construct your model language and, in particular, classical model-dependent Xs. But, then, in such a context, it is hardly surprising that X, being an entirely classical notion, would bring about inconsistency or, worse, triviality, in the (classically constructed) non-classical [L].[56]

Beall claims that if L (in the above revenge recipes) is a non-classical language and X is a classical, model-dependent notion, then recipes 1 and 2 are not very

[56] Beall (2008b: 12); see also Shapiro (2011: 311–312) for a similar point made in defense of Field (2008a, b from the objection in Rayo and Welch (2008).

convincing since the theorist is simply using a classical metalanguage to construct a non-classical artificial language in an effort to argue that natural languages are non-classical.

> A would-be revenger, involved in too easy revenge, would have it easy but too easy. What is (generally) easy is showing that some classically constructed notion is inexpressible—or, at least, not consistently expressible—in a (classically constructed) non-classical 'model language'. What is too easy is the thought that showing as much is sufficient to undermine the adequacy of the given model language. The hard part is clearly establishing the relevance of such inexpressibility results, that is, clearly substantiating the alleged inadequacy.[57]

Beall calls these sorts of cases "too easy revenge", and dismisses objections based on them unless they are accompanied by additional considerations (e.g., the semantics for L are intended to model the semantics for E).

There are a couple of wrinkles in Beall's presentation that, when straightened out, cast doubt on his main point. First, it is not obvious what is meant by 'classical model-dependent notion'. I take it that a model-dependent concept is a mathematical concept that is defined in terms of a particular model or class of models (using 'model' in the mathematical sense of model theory). Although it is problematic in general, for our purposes, we can safely assume that any mathematical concept is expressible in a language that can express set theory (ZFC).

It is less clear what Beall means by 'classical notion'. It seems like a classical concept would be one that is expressible only in a language whose connectives obey classical logic (say, for definiteness, a classical first order predicate calculus with identity), but this definition applies only to logical connectives. For example, classical negation obeys the inference rules and theorems for negation in a classical language (e.g., double negation elimination, classical reductio, excluded middle, and ex falso), and classical disjunction obeys the inference rules and theorems for disjunction in a classical language (e.g., or-introduction, reasoning by cases, disjunctive syllogism, and excluded middle). For non-logical concepts, one might think that a classical concept expressed by a one-place predicate is one whose constitutive principles include all the classical theorems and inference rules pertaining to one-place predicates. For example, a classical concept of redness would have 'everything is either red or not red', 'nothing is both red and not red', and all the rest as constitutive principles, provided that the logical terms occurring in these principles are classical (in the sense specific to logical terms). Notice that a classical concept on this definition *is* expressible in a non-classical language without triviality. Imagine a first order language with the usual syntax and Strong Kleene connectives. There is nothing incoherent about this language containing a predicate expressing classical redness without triviality. In a non-classical logic, there are *fewer* rules or theorems than one finds in classical logic. Thus, if one cannot prove triviality for a classical language that expresses a certain concept, then one cannot prove triviality for a non-classical language that expresses that concept, as long as

[57] Beall (2008b: 11).

the two languages have all the same non-logical expressions. The inverse does not hold, however. For example, smooth infinitesimal analysis is a theory of infinitesimals (i.e., infinitely small non-zero quantities) that is consistent in intuitionistic logic but inconsistent in classical logic.[58] Assume that the theory of smooth infinitesimal analysis implicitly defines the concept of an infinitesimal; then the concept of an infinitesimal is a non-classical concept in the sense that the theory that implicitly defines it is trivial in classical logic.

Beall's claim that "it is hardly surprising that X, being an entirely classical notion, would bring about inconsistency or, worse, triviality, in the (classically constructed) non-classical [language]," does not make sense given these considerations. Other things being equal, classical notions in this sense are expressible in non-classical logics without inconsistency or triviality. It is *non-classical* notions that bring about inconsistency in *classical* logics. Either there is some hidden assumption in his reasoning, he means something else by 'classical notion', or he has mistaken the direction of the above conditional.

Instead, Beall might have in mind concepts expressed by terms that force sentences in which they occur to have classical truth values. For example, in a first order language with usual syntax and Strong Kleene connectives, the negation operator is called *choice negation* ('$\sim$'). If a sentence p is a truth-value gap, then $\ulcorner \sim \mathrm{p} \urcorner$ is a gap as well. We can define in such a language another connective called *exclusion negation* ($\neg$). If a sentence p is a truth value gap, then $\ulcorner \neg \mathrm{p} \urcorner$ is true. In fact, any sentence whose major operator is exclusion negation is either true or false. Thus, exclusion negation obeys excluded middle: $\mathrm{p} \vee \neg \mathrm{p}$. But exclusion negation is not classical negation since it is defined only for non-classical languages that have truth value gaps. Perhaps when Beall talks of classical concepts, he means something like exclusion negation. However, this reading does not make his claim ("it is hardly surprising that X, being an entirely classical notion, would bring about inconsistency or, worse, triviality, in the (classically constructed) non-classical [language]") any more sensible. For exclusion negation *is* expressible in certain Strong Kleene languages without triviality; in fact, it is *defined* above for such a language. For example, a Strong Kleene language with just the usual logical expressions, exclusion negation, and the basic arithmetic expressions used for Peano Arithmetic is not trivial. Again, I do not understand what Beall means by 'classical notion', at least, if his "too easy" reply is to make sense.

It seems to me that we should consider how to interpret Beall's talk of "an entirely classical fragment" of a non-classical language to get a better idea of what he might mean. Let C be the set of all the logical consequences among declarative sentences of the natural language E. C is a set whose members are order pairs of sets and sentences—for each pair, the sentence is the second entry, and it follows from the set of sentences that is the first entry. Since E is non-classical, only arguments that are valid in the non-classical logic in question are listed in C. Now consider the classical fragment of E. What is the set of logical consequences for this fragment? It is obviously a proper subset of C; otherwise, the fragment would not be a fragment of E. It is clearly the set of pairs where the first entry (the set) contains only

[58] See Bell (2008) for the theory and Hellman (2006) for some philosophical issues it raises.

sentences from the classical fragment and the second entry is a sentence from the classical fragment. Are there any pairs in this set where the second entry follows classically from the set that is the first entry, but not according to the non-classical logic in question? No. So something has gone wrong here.

Let us take one more crack at it. We have the language E and the set of logical consequences C as above, but now we deny that validity is closed under uniform substitution. It is possible that, for some particular sentence p and some particular set of sentences G, p is a logical consequence of G, but substituting uniformly in for the sentences of G to get G′ and for p to get p′, p′ is not a consequence of G′. In Beall's preferred approach, which is paraconsistent, he denies the validity of ex falso (i.e., q is a consequence of $p \wedge \sim p$). However, if he denies that validity is closed under uniform substitution, then he could accept that *for some particular sentences p and q*, q is a consequence of $p \wedge \sim p$. With this idea in place, we can define a subset F of the sentences of our language E such that for any classically valid argument with set G of premises and p as conclusion, any instance of that argument using only sentences of F is valid (according to the consequence relation C of E). Let the largest such F be the classical fragment of E.[59]

Given this definition of a classical fragment, one could use the classical fragment of a non-classical natural language to construct a non-classical artificial language to serve as a model for that natural language, which is exactly what Beall says. So that does vindicate one point Beall makes in his "too easy" reply. However, the revenge objector does not take issue with the way the non-classical theorist constructs the artificial language in question. The problem is that this artificial language is not a good model for what goes wrong with the reasoning in the aletheic paradoxes in natural language because this artificial language (however it gets constructed) is devoid of some crucial, relevant linguistic expression. Thus, even if this is the right way to understand what Beall means by 'classical fragment', it still does not make sense of what a classical concept is, and so does not address the central point of revenge objections. However, a major problem even with this reading of 'classical fragment' is that it depends on the claim that validity is not closed under uniform substitution. That assumption abandons the idea that logically valid arguments are those that are valid by virtue of their logical form. Indeed, it gives up on the idea that logical form has anything to do with logical consequence. Perhaps there is some *independent* reason to make this move, but if there is, Beall has not given us one. Moreover, this reading of Beall's "too easy" reply works only for non-classical approaches that have the feature (or bug) of uniform substitution failure. It is not in any way a general response to revenge worries. So, I do not see that this reading of Beall's reply is very helpful.

In sum, stating the relevance of revenge paradoxes for non-classical approaches is fairly straightforward. The theorist has constructed a theory of truth-in-L. Is that theory consistent or at least non-trivial? One would hope so. How does it avoid the reasoning in liar paradoxes? It tosses out as illegitimate some of the logical principles used in that reasoning. How does it avoid the reasoning in revenge paradoxes? It does not apply to sentences that have those expressive resources, for

[59] Thanks to Dave Ripley for this suggestion.

they are not in the model language. What if we extend the model language by adding these resources? Triviality. So the theory enjoys non-triviality only because it does not apply to languages that have certain linguistic expressions that figure in revenge paradoxes. It makes no difference whether this result is surprising or whether the linguistic expression in question is classical (whatever that means).

4.4 Expressibility Delimiters

Another sort of criticism of the points I have made could come from someone who endorses Lionel Shapiro's recent reply to revenge paradox objections. Shapiro's formulation involves three characters: the Puzzler, the Solver, and the Avenger. His claim is that the Avenger's arguments are based on a fundamental equivocation, and they are either sound but not paradoxical or unsound. Here is the setup.

The Puzzler offers two legitimate puzzles:

> Puzzle 1: There is a language L and a notion n such that without making reference to any features of L's syntax or semantics that distinguish L from English, we can prove that L does not express n, yet English appears to express n.
> Puzzle 2: There is a function **n**(**x**) from languages to notions such that we can prove that for any **L** satisfying certain conditions, **L** fails to express **n**(**L**), yet English, which satisfies these conditions, appears to express **n**(English).[60]

The Solver offers some solution to Puzzles 1 and 2. That solution probably involves a theory for truth, but according to Shapiro, it need not. The Avenger says that even if the Solver is right, there is some other notion that the language in question cannot express on pain of contradiction by liar reasoning.[61]

Shapiro's objections to the Avenger turn on his analysis of the Puzzler's arguments, the Solver's claims, and the Avenger's arguments, each of which depend on the notion of a *classical expressibility delimiter* (CED):

> The notion of being G is a CED for language L iff we have all instances of the following conditional:
> If a formula f(x) in L expresses the notion of being H, then for any name c in L, f(c) is G iff the referent of c is H, and ~f(x) is G iff the referent of c is H.

It is easy to show that if the notion of G is a CED for L, then L does not express the notion of being G, which Shapiro dubs the *Inexpressibility Schema*. The argument depends on some classical moves and a standard diagonalization strategy. It also presupposes what I have called extensional expressibility.

With the definition of a CED in hand, we get the reconstructed argument of the Puzzler:

[60] L. Shapiro (2011: 299). There is a third puzzle as well, but he thinks it is not legitimate:

Puzzle 3: For every language **L**, we can express (in some suitable metalanguage) a notion that is useful in semantic theorizing about **L**, but is not itself expressed by **L**.

Shapiro claims (2011: 312–313) that this is not a puzzle that most work on the aletheic paradoxes addresses.

[61] There is a second revenge type worry that Shapiro mentions but does not address that is based on what Beall calls the exhaustive characterization project; see L. Shapiro (2011: 298n2).

(P1) The notion of being a true sentence of L is a CED for L

(C1) L does not express the notion of being a true sentence of L (by the Inexpressibility Schema)

Because the Puzzler's argument makes no mention of the syntax or semantics of L, we get Puzzle 1, and because English seems to express the notion of truth, we get Puzzle 2.

Shapiro also considers how the Puzzler might defend the premise (P1), and he "see[s] no other justification"[62] than the following:

(T) Our grasp of the notions of a true sentence of a language and of a language's expressing a notion reveals that the notion of being a true sentence of language L counts as a CED for L

The Solver responds by denying (P1) and with it (T). One class of Solvers denies that truth is expressible in a classical metalanguage, and another accepts that it is expressible in a classical metalanguage but denies that it is a CED.[63]

For Shapiro all revenge strategies follow one of two forms, both of which press the Solver for some other CED. The First Avenger's argument is:

(P2) The notion of being a Good sentence of L is a CED for L

(C2) L does not express the notion of being a Good sentence of L (by the Inexpressibility Schema)

Again, for Shapiro, the First Avenger "must" appeal to the following to justify (P2):

(G) Our grasp of the notions of a Good sentence of a language and of a language's expressing a notion reveals that the notion of being a Good sentence of L counts as a CED for L

Because the First Avenger does not appeal to the semantics or syntax of L, we get Puzzle 1 back, and because English seems to express the concept of Goodness, we get Puzzle 2 back. However, Shapiro claims that the Solver can simply reject (P2) and with it (G), thus thwarting the First Avenger's strategy.[64]

The Second Avenger's argument pertains to any interpreted first order language L_M, such that its domain of discourse is a set, it contains its own (language-specific) truth predicate, and its predicates (except the truth predicate) can be translated into a classical fragment of our metalanguage. One can construct a classical model M for the truth-free fragment of L_M. Call the sentences of the truth-free fragment of the language that are designated in M, the *Solver-designated*$_M$ sentences.[65]

(P4) The notion of being Solver-designated$_M$ sentence of L_M is a CED for L_M

[62] L. Shapiro (2011: 304).

[63] Field (2008b) is an example of the former type of solver, and Maudlin (2004) is an example of the latter.

[64] L. Shapiro (2011: 305–306).

[65] Shapiro actually formulates the Second Avenger's argument for specific notions of designatedness (e.g., Kripke-designatedness). My formulation is meant to cover all the specific instances he mentions. See L. Shapiro (2011: 309–311).

(C4) L_M does not express the notion of being a Solver-designated$_M$ sentence of L_M (by the Inexpressibility Schema)

The Avenger justifies (P4) by appeal to facts about the model M that provides the semantics for L_M. Because the Second Avenger actually proves (P4), the Solver cannot reject it. However, by Shapiro's analysis, the Second Avenger fails to establish Puzzle 1, since the argument for (P4) depends on the semantics of the language in question. Moreover, Shapiro contends, "once we understand how the proposed proof of (C4) works, that should dispel any appearance that English expresses a notion that plays, with respect to English, the very role we exploited in proving that L_M does not express Solver-designatedness$_M$."[66] Thus, the Second Avenger fails to establish Puzzle 2 as well.

I disagree with just about everything in Shapiro's analysis, and as such I do not think that it does much to defend against revenge objections. Here are some of its major problems.

First, Shapiro's two puzzles fail to capture the most obvious problem posed by the aletheic paradoxes, which is that we can derive an intuitively unacceptable conclusion from intuitively acceptable premises via intuitively acceptable inferences. In fact, it seems to many people, myself included, that the premises, the inferences, and the negation of the conclusion are not just intuitive, but *constitutive* of the concepts involved. This is the central problem posed by the aletheic paradoxes and the central problem that just about anyone who offers an approach to them is trying to solve. Shapiro's puzzles are side issues. His entire way of framing the debate suffers from this defect. Revenge objections are not aimed at reinstating those puzzles—they are aimed at showing the approach in question is not an acceptable solution to the main problem.

Second, the fact that Shapiro sees no other justification for the Puzzler's premise (P1) is disturbing. I agree that if (T) is the best the Puzzler can do, then the puzzle is not very threatening. However, a much better case for (P1) comes from considerations about the way we use truth predicates. We treat 'true' as if it obeys the T-schema for a wide range of sentences including liar sentences. Thus, we accept all instances of the following conditional:

> If f(x) in L expresses the notion of being H then for any name c in L, f(c) is true iff the referent of c is H and ∼f(c) is true iff the referent of c is not H.

That is just how we use the word 'true' in English. Therefore, we treat it as if it is a CED. That is how we arrive at a justification for (P1) of the Puzzler's argument. It has nothing to do with conceptual analysis, or at least it need not. Of course, any Solver or theorist can deny that we treat 'true' in this way, but all that does is render his or her solution or theory irrelevant. So instead of showing the futility of revenge objections, Shapiro's analysis exposes a flaw in all the traditional approaches to the aletheic paradoxes that deny (P1), namely, that they deny that truth is a CED, but we treat truth as if it is a CED. Thus, Shapiro's attempts to discredit the use of CEDs in revenge arguments never get past the first stage.

[66] L. Shapiro (2011: 311).

Third, just as I do not think that Puzzle 1 or Puzzle 2 captures the central problem posed by the aletheic paradoxes, I do not think that that the central aim of revenge objections is to somehow reinstate these puzzles. Instead, the proper way of formulating the dispute is, as I indicated above, as a dispute about the acceptability of the Solver's explanation for what goes wrong in the paradoxical reasoning. As Beall indicates, the Solver responds to the Puzzler by doing three things: (i) proposing an artificial language that contains its own (language specific) truth predicate, (ii) showing that the paradoxical reasoning as represented in that language is unsound (or, at least, does not lead to triviality), and (iii) suggesting that the artificial language is a good model for natural languages. The Solver concludes that the aletheic paradoxes (as they pertain to natural languages) are unsound, and, moreover, indicates exactly what goes wrong in that reasoning.

The Avenger points out that the artificial language avoids triviality only because it is expressively impoverished—it lacks some notion that is either already in the natural language or could easily be added to it. If this notion were added to the artificial language proposed by the Solver, it would be trivial. Thus, the Avenger concludes that the artificial language proposed by the Solver is not a good model for our natural language (as it is or as it could easily be), and so the Solver's conclusion is implausible. Shapiro's analysis misses just about all of the central aspects of the debate.

Fourth, Shapiro's objection to the Avenger's first argument is just as unconvincing as his reconstruction of the Puzzler's argument. Shapiro thinks the Solver can just reject (P2) and (G) along with it, and thereby avoid reinstating the new puzzles. Again, his focus on the two puzzles he thinks are the main problems posed by the aletheic paradoxes allows him to mischaracterize revenge arguments. Frequently, the notion of Goodness that the Avenger uses to point out the expressive limitation is found in the Solver's own theory! When it is not, it is a notion that is either found elsewhere in the logical literature or one that can easily be constructed out of such notions (e.g., exclusion negation, Boolean negation, just true, idempotent determinacy, etc.). The fact that Shapiro cannot conceive of any other justification besides (G) for (P2) does not mean there is not one. There is no reason, other than the Solver's own claims, to think that these notions cannot be used in the way the Avenger suggests. Shapiro's major point here, "Once we allow the Solver to reject the claim that truth in L is a CED for L, it looks like dogmatism to insist without argument that some other notion must qualify as a CED,"[67] is strange. The Avenger does not insist without argument that some notion is a CED. Instead, the Avenger appeals to common usage. If the only reason that can be given for rejecting (P2) is that otherwise the Solver's theory is useless as a model for natural language and so the Solver's proposed approach to the aletheic paradoxes is unacceptable, then it is hard to see why anyone other than the Solver might find this convincing. Instead, to avoid making these sorts of ad hoc moves, the Solver would need to find *independent* evidence for rejecting (P2).

Fifth, Shapiro's criticism of the Avenger's second argument suffers from most of these problems. He points out that once one sees how to prove (C4), that should

[67] L. Shapiro (2011: 306).

remove any puzzlement we had at the fact that English cannot have a notion that plays the role of designatedness. But, again, that does not address the main problem, which is that the Solver's solution only works because the Solver's language does not contain the crucial notion of designatedness. If it did, then the language would be trivial. Since English either does have such a notion or it could easily be easily extended to include one, the Avenger's point stands—the artificial language used by the Solver is a bad model. It makes no difference whether we can understand *why* the bad model cannot contain the notion of designatedness in question. Understanding that does nothing to quell the Avenger's worries. The Avenger says that, say, the Strong Kleene minimal fixed point is a bad model for English because the Strong Kleene minimal fixed point does not contain Kripke-designatedness on pain of triviality, but English clearly does. After all, Shapiro's paper is written in English—or, perhaps an extension of ordinary English. Either way, the Avenger continues, the problem with the Solver's proposal is that it works only because it specifically excludes something that would prevent it from working. The Solver says that the paradoxical reasoning formulated in English is unsound because English is like the artificial language, and the reasoning is unsound there. The Avenger says that English contains the notion of designatedness or can easily be extended to include it (after all, Shapiro coins the term in his own paper, thereby extending English to include it), so it is not like the artificial language in one crucial respect. Thus, there is no reason to think that the Solver has hit on what is wrong with the reasoning in English.

Sixth, Shapiro's entire discussion, as he admits, takes classical logic for granted. I do not see that anyone who offers a non-classical approach to the aletheic paradoxes can find any solace in his criticisms of revenge objections. The non-classical theorist might think that one can use classical logic in various situations—those in which there is no threat of paradox. However, reasoning about classical expressibility delimiters is certainly not one of them. Thus, if one thinks that the problem with the aletheic paradoxical reasoning is that it depends on certain classical inferences, then one will find the same problems with Shaprio's argument for his Inexpressibility Schema, and thus, with his entire characterization of the debate in terms of CEDs.

Seventh, the puzzles he does consider, especially the second one, depend on considering English as it is now. This focus leads to arguments about whether English contains a CED, whether English contains a Goodness predicate, whether it contains a Solver-designatedness predicate, and so on. As I have argued, the emphasis for revenge debates should be on how we might *extend* English.

Finally, the whole notion of an expressibility delimiter is the wrong kind of tool for adjudicating these debates. It works fine for artificial languages, but for natural languages, we have no independent access to the semantic features of our expressions. We cannot inspect a truth predicate to see what its extension or intension might be; we cannot simply reflect on our logical concepts to determine whether they are classical or non-classical. Instead, we need to consider usage and that is a messy and complicated business. It is also one Shapiro neglects.

4.5 The Burden of Proof

Because the standard response to a revenge paradox is to deny the intelligibility of one or more concepts expressed by words composing the revenge paradoxical sentence in question, there is an issue of whether these terms are meaningful, whether they are coherent, and who has the burden of proof in these debates. Take, for example, Priest on Boolean negation. It is easy to show that if the artificial language Priest proposes as a model for our natural language contains Boolean negation, then it is trivial. It would contain a sentence that says of itself that it is Bnot true (again, 'Bnot' expresses Boolean negation) and one could then derive any sentence of the language via revenge type reasoning.

In response to this objection, Priest denies that there is such a thing as Boolean negation. He responds to any attempt to show that Boolean negation exists or is coherent by claiming that the argument in question begs the question against paraconsistent dialetheism.[68] That is, any attempt to show that Boolean negation exists or is coherent must appeal to Boolean negation, or at least, must presuppose that paraconsistent dialetheism is unacceptable. For Priest, the burden of proof is on the person pushing the revenge objection, while the theorist can sit back and play defense.

However, when criticizing others, Priest is all too happy to assume that the burden of proof is on the theorist. He writes when arguing the same kind of objection to Field: "There are notions which, for all the world, appear to us to be intelligible; these cannot, on pain of contradiction, be expressed in the object language. If we declare them meaningless, this is for no reason, in the last resort, other than that they lead to contradiction. As far as solutions to the paradoxes go, the result is, to put it mildly, disappointing."[69] It seems like Priest would not be satisfied if his target complains that any attempt to prove that the notions in question are meaningful or coherent begs the question against the target's favored theory. Instead, the target has the burden of proof to show that they are indeed meaningless or incoherent, and presumably this sort of argument would have to be *independent* of the favored theory in question. That is, one cannot just trot out the revenge paradoxes as evidence that the expressions are meaningless or incoherent. Moreover, we would also need an explanation of why they seemed meaningful or coherent in the first place. Of course, Boolean negation appears for all the world to be intelligible, so Priest's objection to Field seems unfair given Priest's defense of his own view.

Priest's double standard for revenge objections is obviously unreasonable, but it would be good to have a general position on who really has the burden of proof in these cases. Beall writes on this topic:

> The difficulty in successfully launching Rv3 might be put, in short, as follows. Theorist advances $\mathbf{L}_\mathrm{m}$ as a model of (relevant features of) $\mathbf{L}$, our real language. Rv3 Revenger alleges that $\mathbf{X}$ exists in $\mathbf{L}$, and shows that, on pain of triviality,

[68] See Priest (1990: 204–209, 2006b: ch. 5).

[69] Priest (2005: 46).

> **X** is inexpressible in $\mathbf{L}_m$. The difficulty in adjudicating the matter is that … Theorist may reasonably conclude that **X** is incoherent (given the features of our language that Theorist advances). Of course, if Revenger could establish that we need to recognize **X**, perhaps for some theoretical work or otherwise, then the debate might be settled; however, such arguments are not easy to come by.
>
> The burden, of course, lies not only on the Rv3 Revenger; it also lies with the given theorist. For example, typical paracomplete and paraconsistent theorists must reject the intelligibility of any EE device in our language. Inasmuch as such a notion is independently plausible—or, at least, independently intelligible—such theorists carry the burden of explaining why such a notion appears to be intelligible, despite its ultimate unintelligibility. Along these lines, the theorist might argue that we are making a common, reasonable, but ultimately fallacious generalization from 'normal cases' to all cases, or some such mistake. (E.g., some connective, if restricted to a proper fragment of our language, behaves in the EE way.) Alternatively, such theorists might argue that, contrary to initial appearances, the allegedly intelligible notion only appears to be a clear notion but, in fact, is rather unclear; once clarified, the alleged EE device (or whatever) is clearly not such a device. (E.g., one might argue that the alleged notion is a conflation of various notions, each one of which is intelligible but not one of which behaves in the alleged, problematic way.) Whatever the response, theorists do owe something to Rv3 revengers: an explanation as to why the given (and otherwise problematic) notion is unintelligible.[70]

I agree with Beall on this matter. Any theorist responding to revenge objections should be able to give *independent* evidence that the notions in question are meaningless or unintelligible and explain why they were taken to be intelligible in the first place. Moreover, anyone pushing a revenge objection should be able to provide evidence that the notions in question are meaningful and coherent, or why it does not matter whether they are coherent. I take the latter strategy to be fairly easy to accomplish, as I argue in just a moment.

In particular, I do not see that Priest's response to revenge objections concerning Boolean negation are successful. There is a long-standing problem in philosophy of logic about how one might justify one's basic logical notions and inference rules. For example, if one wants to justify the legitimacy of Modus Ponens, then one will have a hard time not using Modus Ponens in one's justification. The same goes for justifying the legitimacy of the classical conditional or classical negation.[71] Despite the fact that attempts to justify our basic logical rules and logical concepts are bound to be circular, we do not thereby conclude that they are illegitimate or that the linguistic expressions that purport to express them can be dismissed out of hand as unintelligible or meaningless. Instead, one would have to dig into this vexed issue and show somehow that Boolean negation is in an especially bad position—worse

[70] Beall (2008b: 13–14).

[71] See Dummett (1978), Boghossian (2000, 2003), Shapiro (2000), Williamson (2003), Wright (2004), Tennant (2005), Dogramaci (2010), and Kroedel (2012) for discussion.

than the standard circularity problem. Thus, Priest's strategy for replying to revenge objections should be rejected.

Here is what I take to be the upshot of these considerations on burden of proof. It is hopeless for a theorist to claim that the crucial linguistic expressions that figure in revenge paradoxes are meaningless. To say that is to abandon any reasonable stance in linguistics, which treats established usage as definitive of meaningfulness. Adopting an approach to the aletheic paradoxes that rejects out of hand a foundational assumption of the science of linguistics is just as ridiculous as being a young Earth creationist. I do not see how any self-respecting philosopher could advocate something like that.

Given that the expressions in question are meaningful, the question then becomes: are they intelligible? I think that the best way to understand this question is as asking whether they express consistent concepts. Consider the EE operator example again. An EE operator is one that is stipulated to obey excluded middle and ex falso. That is, these are its constitutive principles. I do not take the question of whether there really is such a concept seriously. We can stipulate that our words have certain meanings (in the form of constitutive principles) without any worry that they might not really have these meanings (i.e., constitutive principles). They might not have the extensional, intensional, or metaphysical features that we think they do, but there is no worry about their constitutive principles. As such, there is no worry about whether these concepts exist.

Thus, the best thing to say for a theorist in response to a revenge objection is that the concepts that figure in the revenge paradoxes are inconsistent. As such, there is no reason that the theory of truth being defended should be expected to apply to sentences with words that express them. The theorist concludes that we can simply dismiss these sentences out of hand, even if they do show up in English or could easily be added to it.

For example, Field claims that idempotent determinateness operators are unintelligible.[72] By that he does not mean that they are meaningless. Instead, he thinks they express concepts that are inconsistent. As such, he does not think that his theory of truth needs to say anything about sentences containing idempotent determinacy operators, any more than it ought to be able to say something about sentences containing 'tonk'.[73]

How plausible is this? Not very plausible in my opinion. First, philosophical moves do not get more ad hoc than this. There is no independent evidence that idempotent determinateness operators are problematic in any way. The only reason Field can give is: well, otherwise, my theory of truth would be unacceptable. Not very convincing. Second, Field's justification depends on his own view of truth. After all, if Field's object language had an idempotent determinateness operator but no truth predicate, then his theory would have no problem characterizing the semantic statuses of its sentences. So, even his rather weak case against idempotent determinateness operators rests on the questionable assumption that truth is a

[72] Field (2008a: 119).

[73] See Prior (1960) for a discussion of 'tonk'.

consistent concept. I have yet to see any independent reason for thinking that a revenge paradoxical notion is indeed inconsistent. Not one.

Finally, if Field were to engage with those advocating inconsistency views, there is good reason to think that he would not fare well. For, one can formulate revenge paradoxes for Field using resources other than an idempotent determinateness operator—for example, one can use exclusion negation, or an intuitionistic conditional, or a material conditional, or Boolean negation, or stronger relevant conditionals, or the phrase 'something other than true', or a whole host of other devices. In order to safeguard his theory against revenge worries, he has to say that all of these are unintelligible—that is, they express inconsistent concepts, and case against each one is just as flimsy as the case against idempotent determinateness operators. The inconsistency theorist says that truth is inconsistent. That is it. The claim that truth is an inconsistent concept, when properly developed, does away with all of the aletheic paradoxes and is not subject to any revenge paradoxes—as we will see in Section Five. So, in a debate about which concepts are inconsistent, it looks like Field's view is radically more complex than the inconsistency theorist's view. Thus, even if the theorist facing revenge paradoxes were to endorse some theory of inconsistent concepts, they would still be saddled with a vastly more complex and implausible view.

Field's response to these sorts of considerations is in the following passage:

> I've heard it argued that even if no *good* theory posits a Boolean negation, we haven't solved the paradoxes until we've given an account of how to apply the term 'true' to the sentences of someone who has a *bad* theory according to which the word 'not' obeys all of Boole's assumptions (or at least, to the sentences of someone for whom this bad theory plays such a central role in his linguistic practices that it determines what 'not' means for him.) But I don't think that this raises an interesting challenge. There is bound to be a certain arbitrariness in how we attribute truth-values to sentences that contain concepts we find defective [e.g., 'tonk' (Prior 1960) or 'Boche' (Dummett 1973: 454)]. We can't, for instance, regard both 'All Germans are Boche' and 'All Boche are prone to cruelty' as true, unless we share the racist beliefs of those who possess the concept; but a decision as to which (if either) should count as true, in the language of the racist for whom both beliefs enter into his practices in a "meaning-constituting" way, seems both pointless and unlikely to have a principled answer. Similarly for the use of 'not' by someone with a defective theory. Probably the simplest course is to translate his 'not' with our 'not', in which case his claims to the validity of excluded middle will come out false even though they were "meaning-constitutive"; but there's no reason to take a stand on this translation, or to think that there's a determinate fact of the matter as to whether it's correct.[74]

There is a lot going on in this passage. He describes those who use exclusion negation as having a bad theory of negation. By this, he means that someone who uses 'not' to express exclusion negation has false beliefs about negation. Field also

[74] Field (2008b: 309n.1–310n.1).

accepts Quine's criticism of analyticity and his argument for the indeterminacy of translation; the former attempts to undermine the idea that any sentences are true by virtue of their meanings alone, while the latter purports to show that meaning and translation are largely indeterminate.[75] Putting these points together: someone who uses 'not' to express exclusion negation has a bad theory of negation and we should attribute truth values to his or her sentences by simply translating them into Field's paracomplete language since there is no fact of the matter as to what they "really" mean. The effect is that Field advocates treating everyone who uses 'not' as meaning exactly what Field means when he uses 'not'.

Why is this a problem? For one, it seems desperate to rest one's response to revenge objections on something so controversial and radical as the indeterminacy of translation. In fact, to call it controversial is a bit optimistic. From my perspective, there is not much controversy—philosophers of language tend to reject it in overwhelming numbers. Moreover, most people do not have theories of negation—they utter sentences that contain 'not' in certain circumstances and they interpret others who utter sentences that contain 'not' in certain circumstances. The question for a philosopher of language or a linguist who works on semantics or pragmatics is: what is the best way to make sense of how English users behave? If it turns out that a theory on which 'not' expresses exclusion negation on some occasions provides the best explanation of how English users use 'not', then Field's approach has a major cost—namely, it cannot be applied to English.

Is there any reason to think that people use 'not' in this way? Yes. The intuitive evidence is the following. If I say that p is not true, and I mean to be saying that p is something other than true (which includes falsity, indeterminacy, or whatever), then I take myself to be saying something true, no matter how indeterminate p might be. But Field cannot accommodate this intuition. If Field is correct, then no matter how much I try, if p is indeterminate enough, my sentence will be indeterminate as well. To be a bit more careful about this point: Field will translate my claim that p is not true as 'p is not D^{σ}true' for some σ, but no matter how large σ is, if p's level of indeterminacy is higher, 'p is not D^{σ}true' is indeterminate (at some level or other). Moreover, according to Field, it is incoherent for me to say or believe that p is not true, where this claim or belief is true no matter what level of indeterminacy p has.

In addition, linguists claim that 'not' in English (at least sometimes) is properly interpreted as exclusion negation, and linguists use exclusion negation in their theories. Here are two examples.

Jay Atlas in *Philosophy without Ambiguity* (1989) argues that 'not' has a general sense and on particular occasions of use it can express either choice negation or exclusion negation. There is linguistic evidence that 'not' is univocal and invariant because it fails ambiguity tests and context-dependence tests; thus, it is neither ambiguous nor context-dependent. Nevertheless, on many occasions, it makes the most sense to interpret English speakers as meaning exclusion negation when they use 'not'.[76] A second example is that Laurence Horn in *A Natural History of Negation* (2001) surveys views on negation from Aristotle to present, the evidence

[75] See Quine (1951, 1960).

[76] Atlas (1989: ch. 3).

for choice negation readings of ‘not’ versus exclusion negation readings of ‘not’, and how these readings interact with other linguistic phenomena (presupposition, conversational implicature, scope, etc.). He too argues that ‘not’ is not ambiguous or context dependent. Rather, exclusion negation provides the semantics for natural language descriptive (non-metalinguistic) negation or predicate denial (in Aristotle’s sense), and what seems like choice negation is an artifact of pragmatic tendencies like that of reading topical/definite subjects as taking wide scope with respect to ordinary predicate denial.[77]

On the other hand, there is no evidence that English contains a transfinite hierarchy of determinate truth predicates. No scientists studying English have ever found any data to support such a view. The only support Field can offer is “well, that’s just a consequence of the best way I can think of to solve the liar paradox.” Perhaps this kind of armchair justification would be compelling if its consequences were obscure enough to have avoided any scientific inquiry, but that is not the case—linguists have plenty of data that suggest we often use ‘not’ to express exclusion negation.

I suppose Field could say that the evidence cited by Atlas and by Horn is compatible with English having some kind of expression that behaves like exclusion negation in their examples, but which fails to obey excluded middle in paradoxical settings. Of course, to be convincing, he would have to find some kind of *independent* evidence to support this claim. However, instead of developing this line of thought, it might be better to just stop banging on a square peg and recognize that the hole is round.

5 An Internalizable Theory of Truth

One might object that the standard I have proposed—that a theory of truth be internalizable for any language—is too hard to meet. And if no theory can meet it, then it hardly seems binding. If the objections I have raised cut down every theory of truth, then they effectively rule out none of them. However, there is at least one theory that satisfies the internalizability condition, which I demonstrate in this section. The theory and its motivations are complex, and I can only give the barest outline here.[78]

The theory in question is an inconsistency theory of truth—it takes truth to be an inconsistent concept. An inconsistent concept has inconsistent constitutive principles—the T-schema in the case of truth (taking classical logic and some way of representing syntax as our background). The theory is *not* eliminativist; we should continue using the truth predicates of natural languages in most situations because the concept of truth they express, though inconsistent, is useful and its inconsistency rarely inhibits its utility. Instead, we need to replace truth only for certain purposes with a team of concepts that will do its work without generating paradoxes. These replacement concepts play a pivotal role in the semantics for the truth predicate.

[77] Horn (2001).

[78] See Scharp (2011, 2013a, b) for details.

Because the constitutive principles for the concept of truth are inconsistent, there is no property of truth and no ordinary extension for the truth predicate. Instead, the truth predicate of a natural language is assessment-sensitive—it has the same content in every context of utterance, but its extension varies depending on the context of assessment. The theory of truth is the semantics for the truth predicate, which relies on the replacement concepts in a way described below.

5.1 The Prescriptive Theory

The two replacement concepts split truth's constitutive principles in the following way. One concept, which I call *ascending truth* is such that if p then $\langle p \rangle$ is ascending true. The other, which I call *descending truth* is such that if $\langle p \rangle$ is descending true, then p. The converse rules fail in general, but are valid for a certain class, which I call the *safe* sentences. That is, if $\langle p \rangle$ is safe and $\langle p \rangle$ is ascending true, then p; if $\langle p \rangle$ is safe and p, then $\langle p \rangle$ is descending true. There are a host of other principles for each concept, which, together constitute an axiomatic theory called ADT (for *A*scending and *D*escending *T*ruth). In the following, A(x) (for 'x is ascending true'), D(x) (for 'x is descending true'), and S(x) (for 'x is safe') are all monadic, univocal, and invariant.

D1 $D(p) \rightarrow p$
D2 $D({\sim}p) \rightarrow {\sim}D(p)$
D3 $D(p \wedge q) \rightarrow D(p) \wedge D(q)$
D4 $D(p) \vee D(q) \rightarrow D(p \vee q)$
D5 D(p) if p is a logical truth
D6 D(p) if p is a theorem of PA
D7 D(p) if p is an axiom of ADT

A1 $p \rightarrow A(p)$
A2 ${\sim}A(p) \rightarrow A({\sim}p)$
A3 $A(p) \vee A(q) \rightarrow A(p \vee q)$
A4 $A(p \wedge q) \rightarrow A(p) \wedge A(q)$
A5 ${\sim}A(p)$ if p is a contradiction
A6 ${\sim}A(p)$ if ${\sim}p$ is a theorem of PA

M1 $D(p) \leftrightarrow {\sim}A({\sim}p)$
M2 $S(p) \leftrightarrow (D(p) \vee {\sim}A(p))$
M3 $p \wedge S(p) \rightarrow D(p)$
M4 $A(p) \wedge S(p) \rightarrow p$

E1 If $\sigma = \tau$ and q results from replacing some occurrences of σ with τ in p, then $D(p) \leftrightarrow D(q)$
E2

If $\sigma = \tau$ and q results from replacing some occurrences of σ with τ in p, then $A(p) \leftrightarrow A(q)$

E3 If $\sigma = \tau$ and q results from replacing some occurrences of σ with τ in p, then $S(p) \leftrightarrow S(q)$

Some of these are redundant, but as I address below, there is a good reason for this feature. Further, these axioms are not meant to be *the* theory of ascending truth and descending truth; rather any theory of ascending truth and descending truth should have the preceding as a subtheory.

ADT plays the role of a *prescriptive* theory in this approach to the aletheic paradoxes—we need to change our linguistic practice and conceptual repertoire by adding the concepts it describes. It does not describe our language or thought as it is now. However, the concepts of ascending truth and descending truth play explanatory roles in the *descriptive* theory—the theory that purports to describe our language and thought insofar as they pertain to the concept of truth. That the descriptive theory does not rely on the concept of truth is a crucial feature of it. It is fine to use truth in most circumstances, but providing a semantics for an expressively rich language is not one of them. As such, the descriptive theory of truth utilizes the replacement concepts, not the concept of truth.[79]

5.2 The Descriptive Theory

In the past decade, a new kind of semantic theory has been proposed that promises to make sense of some puzzling natural language expressions. It treats certain natural language expressions (e.g., epistemic modals like 'might' and predicates of personal taste like 'tasty') as assessment-sensitive, which means a sentence containing one of these expressions is not context-dependent, so it expresses the same proposition in every context of use, but the proposition can have different truth values in different contexts of assessment.[80]

For example, consider 'tasty'. A contextualist about 'tasty' says that sentences containing it express different propositions in different contexts of utterance. For example, when Cletus asserts 'Possum is tasty', he expresses (roughly) the proposition that possum is tasty for Cletus, but when Brandine asserts 'Possum is tasty', she expresses the proposition that possum is tasty for Brandine.[81] On the other hand, an assessment-sensitivity theorist about 'tasty' says that sentences containing it express the same proposition in different contexts of use. So, Cletus and Brandine both express the same proposition—that possum is tasty. However, this proposition has a truth value only relative to a context of assessment. In the case of 'tasty', the context of assessment will contain a standard of taste—a *gustatory* standard. So, it could be that the proposition that possum is tasty is true from Cletus's context of assessment (according to his gustatory standard) and false from Brandine's context of assessment (according to her gustatory standard).

[79] See Scharp (2013a, b).

[80] See Lasersohn (2005, 2008, 2009) and MacFarlane (2005a, b, 2007a, b, 2008, forthcoming).

[81] See Glanzberg (2007) and Cappelen and Hawthorne (2009).

In what follows, I assume a standard intensional semantic framework (familiar from Kaplan) in which utterances (i.e., sentences uttered in a context of use) are assigned a logical form/index pair, where the *index* is an n-tuple of information from the context of use (e.g., speaker, place, time, etc.). The logical form is assigned a character, which, together with the index, determines a content. A *content* is a function from the set of points of evaluation to the set of truth values. A *point of evaluation* is an n-tuple of information that is needed to arrive at a truth value given a character and an index (e.g., world, time, standard). A truth value for the utterance is determined by the truth value of the content at the relevant point of evaluation. For utterances of sentences that contain assessment-sensitive terms, the theory assigns a truth value to the sentence in the context of use from the context of assessment. The relevant point of evaluation is one with the world and time from the context of use and the standard from the context of assessment. We can say that a theory of this sort has three parts—the *presemantic* part takes as input the sentence uttered and the context of use and outputs a logical form and an index, the *semantic* part takes as input the logical form and index and outputs a content, the *postsemantic* part takes as input the content and information from the context of use and context of assessment and outputs a truth value for the sentence in the context of use from the context of assessment.[82]

The presemantics, semantics, and postsemantics for truth predicates are as follows. Truth predicates are univocal and invariant one place predicates (in surface syntax and logical form). Truth is primarily attributed to sentences, understood as interpreted in a context of use. So the presemantics are rather uninteresting.

Truth predicates are invariant and have extensions relative to points of evaluation that are (at least) <world, aletheic value, aletheic standard> triples. The world parameter functions in the familiar way. The aletheic *value* parameter determines whether the point gives ascending truth conditions or descending truth conditions. That is, at a point of evaluation with ascending truth as its aletheic value parameter, the sentence in question is evaluated for its ascending truth value; likewise for those with descending truth as the aletheic value parameter. The aletheic *standard* parameter determines a reading for the truth predicates occurring in the sentence in question—an ascending standard treats ‘true’ in the sentence in question as ‘ascending true’, while a descending standard treats ‘true’ as ‘descending true’. Contents assigned to sentences containing truth predicates are functions from the set of points of evaluation to {0, 1}. The only difference between this semantic theory and what one might expect is the two nonstandard parameters in the points of evaluation.

The postsemantics employs two univocal, invariant, three-place predicates, ‘x is ascending true in context U from context A’ and ‘x is descending true in context U from context A’. Letting U be a context of use, A be a context of assessment, i_U be the index representing U, and i_A be the index representing A):

> A sentence p is *ascending true* at U from A iff the content assigned to the clause representing p with respect to i_U is assigned 1 at the point of evaluation <w, v, s>

[82] See Kaplan (1989), Lasersohn (2005), Predelli (2005), and MacFarlane (2005a, forthcoming) for details on this framework.

> where w is the world of i_U, v is *ascending* truth (i.e., the aletheic value parameter), and s is the aletheic standard from i_A.
>
> A sentence p is *descending true* at U from A iff the content assigned to the clause representing p with respect to i_U is assigned 1 at the point of evaluation <w, v, s> where w is the world of i_U, s is *descending* truth (i.e., the aletheic value parameter), and s is the aletheic standard from i_A.
>
> An argument $<\Gamma, \phi>$ is valid iff for every point of evaluation e, if for all $\gamma \in \Gamma$ iff if all the members of Γ are true-at-e, then ϕ is true-at-e.

There are several novel aspects to the postsemantics. First, it assigns semantic values to sentences in contexts of use from contexts of assessment; this feature reflects the fact that truth predicates are assessment sensitive. Second, it assigns two kinds of semantic values to sentences—ascending truth values and descending truth values; this feature reflects the fact that the descriptive theory of the truth predicate employs the replacement concepts, not the concept of truth.[83]

5.3 Examples

Let us begin with ADT and examples involving 'ascending true' and 'descending true'. Consider two sentences that are similar to liars and might be thought to make trouble:

(9) (9) is not ascending true.
(10) (10) is not descending true.

We can prove that (9) and (10) are ascending true and not descending true using the same kind of reasoning as in the liar paradox. However, this just shows that they are unsafe—there is no contradiction here. One cannot infer from the fact that (10) is not descending true that (10) is descending true, and one cannot infer from the fact that (9) is ascending true that (9) is not ascending true.

Instead, one might worry about revenge-type sentences like:

(11) (11) is either not ascending true or unsafe.
(12) (12) is either not descending true or unsafe.

Again, it is easy to show that (11) and (12) are unsafe. However, one cannot infer from this result that they are descending true or not ascending true since they are unsafe. In case the reader is not content to take my word for it (which is understandable given the fact that so many seemingly innocuous combinations of aletheic principles end up inconsistent), one can prove a relative consistency result for ADT using a generalization of neighborhood semantics.[84]

We can also classify "teller" sentences:

(13) (13) is ascending true.
(14) (14) is descending true.

[83] See Scharp (2013a, b).

[84] So this approach solves the other aletheic paradoxes too. See Scharp (2013a).

It turns out that both of these are safe—(13) is descending true and (14) is not ascending true.

Let us turn to examples with truth predicates. Consider a liar sentence like:

(1) (1) is not true.

Since (1) has a truth predicate, it is assessment sensitive. In fact, it is ascending true from the ascending standard and the descending standard, and it is not descending true from the ascending standard and the descending standard. One arrives at these results by considering that the ascending standard reads (1) as (9) and the descending standard reads (1) as (10). (9) and (10) are ascending true and not descending true. Thus, (1) has the following semantic values: (i) it is ascending true from contexts of assessment that employ the ascending standard, (ii) it is ascending true from contexts of assessment that employ the descending standard, (iii) it is not descending true from contexts of assessment that employ the ascending standard, and (iv) it is not descending true from contexts of assessment that employ the descending standard. A truth teller, on the other hand, like:

(15) (15) is true,

has a different batch of semantic values. Again, (15) is assessment-sensitive. It is descending true from the ascending standard and not ascending true from the descending standard. So, the theory classifies liars and truth-tellers differently.

Instead, one might try a different revenge strategy based on the observation that some unsafe sentences are derivable from ADT, like the descending liar, and some are not, like the ascending liar. Let U^+ (i.e., the *positive unsafe sentences*) be the class of unsafe sentences derivable from ADT and U^- (i.e., the *negative unsafe sentences*) be the class of unsafe sentences not derivable from ADT. Now consider:

(16) (16) is not descending true and (16) is not U^+.

One might think that the conjunction of descending truth and positive unsafey would act enough like truth to generate a revenge paradox. It is easy to show that (16) is unsafe. Assume (for reductio) it is descending true. Then '(16) is not descending true and (16) is not U^+' is descending true, which entails (16) is not descending true and (16) is not U^+. It follows that (16) is not descending true. Thus, by reductio, (16) is not descending true. Assume (for reductio) that (16) is not ascending true. Then '(16) is not descending true and (16) is not U^+' is not ascending true, which entails the negation of (16)—(16) is descending true or (16) is U^+. Of course, if (16) is descending true or U^+, then it is ascending true. Thus, by reductio, (16) is ascending true. So, we have shown that (16) is unsafe. Assume for reductio that (16) is U^+, which means that it is an unsafe theorem of ADT. Since we can take any axiom of ADT as an assumption, and (16) follows from some set of axioms of ADT, we can derive (16) itself—namely that (16) is not descending true and (16) is not U^+. It follows that (16) is not U^+. Thus, by reductio, (16) is not U^+. Therefore, we have shown that (16) is U^-. It is unsafe and it is not derivable from ADT. I leave it to the reader to try deriving something untoward from the assumption that (16) is U^- (knowing that one cannot because of a relative consistency result is no substitute for convincing oneself that it cannot be done by trial and error). In conclusion, (16) does not generate a revenge paradox. I encourage the reader to play around with other examples if these are not enough to give an intuitive understanding of how ADT avoids revenge.

Here is a different strategy for finding revenge. Consider the sentences:

(17) for all u, for all a, (17) is not ascending true in u from a
(18) for all u, for all a, (18) is not descending true in u from a

Let us drop the 'u' since nothing here is indexical. For all a, if p does not contain 'true', then p is ascending true from a iff p is descending true from a. The contexts of assessment do not affect anything but 'true'. It is easy to show that (17) is unsafe from every context of assessment. The same goes for (18). One still cannot derive anything contradictory about them. There are no paradoxes here and so still no revenge.

It might seem that ADT faces a self-refutation revenge paradox. After all, ADT implies that some of its own theorems are not descending true. For example '(10) is not descending true' is a theorem of ADT and ADT implies that (10) is not descending true. So ADT implies that some of its theorems are unsafe. Incidentally, this feature of ADT is why I wanted the redundancies in its formulation. Axiom schema D7 insures that all its axioms are descending true, but one is not thereby guaranteed that all its *theorems* are descending true. The more axioms one lists, the more are guaranteed to be descending true.

It is correct that ADT implies that some of its own theorems are not descending true (they are, of course, ascending true), but there are two points to be made here. First, if one has only one aletheic status to work with, e.g., truth, then this kind of consideration is decisive—the theory is self-refuting.[85] However, if one has two to work with, like ascending truth and descending truth, then we can formulate a criterion of adequacy for good (i.e., trustworthy) arguments that the theory, ADT, respects: namely that a valid argument will never take one from descending truths to something not ascending true. It might take you from descending truth to unsafety or from unsafety to something not ascending true. For example, the descending liar is provable in ADT, the axioms of ADT are descending true, but the descending liar is unsafe (so not descending true). Also, the ascending liar and its negation are unsafe (so ascending true), but their conjunction is a contradiction and so is not ascending true. Anyone who accepts that the new aletheic statuses are not preserved by valid arguments will be forced to distinguish between theorems of the theory that have top status, those that have middle status(es), and those that have bottom status. If valid arguments never take one from top to bottom, then the obvious condition on an acceptable theory is that all its axioms have top status. We know already that not all its *theorems* will have top status.[86] So the best we can hope for is top status

[85] See Maudlin (2004) for an example. I have not had space to devote to self-refutation problems like the one confronting Maudlin or why I think his response to them is inadequate. As such, Maudlin's theory avoids the central objection as it is presented here. See Scharp (2013a) for discussion.

[86] Field has argued, convincingly in my view, that no approach is compatible with the claim that all and only valid arguments are necessarily truth preserving. His argument carries over to the case of necessary descending truth preservation. See Field (2006).

axioms and no bottom status theorems. That is exactly the case with ADT: all the axioms of the theory are descending true.

The upshot is that there is no way of generating a revenge paradox for this approach. It assigns a semantic status to all sentences containing ‘true’ and there are no puzzling claims to the effect that certain linguistic resources are meaningless or unintelligible. Of course, the truth predicate expresses an inconsistent concept, but the whole point of the theory is to provide a consistent theory of it. The theory is compatible with classical logic, so there is no worry about logical resources that force classical reasoning like exclusion negation. As such, we have good reason to think that the theory of truth is internalizable for every language. Few English idiolects contain ‘ascending true’ or ‘descending true’, so most would have to be extended before the theory would be expressible in them. However, there is every reason to think that the theory is descriptively complete and descriptively correct for the extended idiolect.

6 Conclusion

I have introduced the internalizability relation and argued that an acceptable theory of truth is internalizable for every language. This result applies to a wide variety of theories and bolsters revenge objections to them. However, at least one theory of truth, the approach presented in Section Five on which truth is assessment sensitive and explained in terms of ascending and descending truth, meets the internalizability requirement.

This inconsistency approach is in a class by itself—endorsing it does not require one to pretend that some array of linguistic resources are unintelligible even though there is no independent reason to think that they are. Instead, one admits that the defect has all along been with our concept of truth. This insight, when properly formulated and developed, robs the concept of truth of its striking propensity for revenge.

Acknowledgments Versions of this paper have been given at The Ohio State University, Wayne State University, and the University of Georgia; thank you to those audiences for helpful feedback. For valuable conversations on the topics in this paper or constructive comments on earlier drafts, my appreciation goes to Bob Brandom, John McDowell, Anil Gupta, Hartry Field, Nuel Belnap, Joe Camp, Arthur Fine, Cristina Lafont, Neil Tennant, Stewart Shapiro, Robert Kraut, Seth Ard, Brad Cokelet, Graham Hubbs, John Morrison, Peter Guildenhuys, and Alison Duncan Kerr. Finally, I want to thank the referees for *Erkenntnis* for their comments, which greatly improved the paper.

References

Atlas, J. D. (1989). *Philosophy without ambiguity: A logico-linguistic essay*. Oxford: Oxford University Press.

Beall, Jc. (1999). Completing Sorensen’s menu: A non-modal Yabloesque Curry. *Mind, 108*, 737–739.

Beall, Jc. (2007). Truth and paradox: A philosophical sketch. In D. Jacquette (Ed.), *Philosophy of logic*. Dordrecht: Elsevier.

Beall, Jc (Ed.). (2008a). *The revenge of the Liar*. Oxford: Oxford University Press.

Beall, Jc. (2008b). Prolegomenon to future revenge. In Beall (2008a).

Beall, Jc. (2009). *Spandrels of truth*. Oxford: Oxford University Press.

Beall, Jc, & Armour-Garb, B. (Eds.). (2005). *Deflationism and paradox*. Oxford: Oxford University Press.

Bell, J. (2008). *A primer of infinitesimal analysis* (2nd ed.). Cambridge: Cambridge University Press.

Boghossian, P. (2000). Knowledge of logic. In P. Boghossian & C. Peacocke (Eds.), *New essays on the a priori*. Oxford: Oxford University Press.

Boghossian, P. (2003). Blind reasoning. *Proceedings of the Aristotelian Society, Supplementary Volume, 77*, 225–248.

Burge, T. (1979). Semantical paradox. *The Journal of Philosophy, 76*, 169–198.

Burgess, J. (2004). Quine, analyticity, and philosophy of mathematics. *The Philosophical Quarterly, 54*, 38–55.

Burgess, J. (2012). Friedman and the axiomatization of Kripke's theory of truth. In N. Tennant (Ed.), *Foundational adventures: Essays in honor of Harvey M. Friedman*. London: Templeton Press (Online) and College Publications. (2011).

Cappelen, H., & Lepore, E. (2005). *Insensitive Semantics*. Malden, MA: Wiley.

Cappelen, H., & Hawthorne, J. (2009). *Relativism and monadic truth*. Oxford: Oxford University Press.

Cook, R. (2009). Curry, yablo, and duality. *Analysis, 69*, 612–620.

Davidson, D. (1974). *"On the Very Idea of a Conceptual Scheme", in inquiries into truth and interpretation* (p. 1984). Oxford: Oxford University Press.

Davidson, D. (1999). Reply to Simon Evnine. In L. E. Hahn (Ed.), *The philosophy of Donald Davidson*. Peru, IL: Open Court.

De Vidi, D., & Solomon, G. (1999). Tarski on 'Essentially Richer' metalanguages. *Journal of Philosophical Logic, 28*, 1–28.

Devitt, M. (2006). *Ignorance of language*. Oxford: Oxford University Press.

Dogramaci, S. (2010). Knowledge of validity. *Noûs, 44*, 403–432.

Dowty, D., Wall, R., & Peters, S. (1980). *Introduction to montague semantics*. Dordrecht: D. Reidel.

Dummett, M. (1973). *Frege: Philosophy of language*. Cambridge, MA: Harvard University Press.

Dummett, M. (1978). The justification of deduction. In M. Dummett (Ed.), *Truth and other enigmas* (p. 1978). Cambridge: Harvard University Press.

Eklund, M. (2008). The liar paradox, expressibility, possible languages. In Beall (2008a).

Field, H. (2001). *Truth and the absence of fact*. Oxford: Oxford.

Field, H. (2006). Truth and the unprovability of consistency. *Mind, 115*, 567–605.

Field, H. (2008a). Solving the paradoxes, escaping revenge. In Beall (2008a).

Field, H. (2008b). *Saving truth from paradox*. Oxford: Oxford University Press.

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

Frigg, R., & Hartmann, S. (2006). Models in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2006 Edition). Online at http://plato.stanford.edu/entries/models-science/.

Glanzberg, M. (2004). A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic, 33*, 27–88.

Glanzberg, M. (2007). Context, content, and relativism. *Philosophical Studies, 136*, 1–29.

Gupta, A. (1982). Truth and paradox. *Journal of Philosophical Logic, 11*, 1–60.

Gupta, A. (1997). Definition and revision: A response to McGee and Martin. *Philosophical Issues, 8*, 419–443.

Gupta, A., & Belnap, N. (1993). *The revision theory of truth*. Cambridge: The MIT Press.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic, 71*, 677–712.

Hellman, G. (2006). Mathematical pluralism: The case of smooth infinitesimal analysis. *Journal of Philosophical Logic, 35*, 621–651.

Horn, L. (2001). *A natural history of negation* (2nd ed.). Stanford: CSLI Publications.

Horsten, L. (2011). *The Tarskian turn*. Cambridge, MA: MIT Press.

IAU. (2006). *IAU 2006 general assembly: Result of the IAU resolution votes*. Paris: IAU.

Juhl, C. F. (1997). A context-sensitive liar. *Analysis, 57*, 202–204.

Kaplan, D. (1989). Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almog, J. Perry, & H. K. Wettstein (Eds.), *Themes from Kaplan*. Oxford: Oxford University Press.

Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy, 72*, 690–716.

Kroedel, T. (2012). Implicit definition and the application of logic. *Philosophical Studies, 158*, 131–148.

Lasersohn, P. (2005). Context dependence, disagreement, and predicates of personal taste. *Linguistics and Philosophy, 28*, 643–686.

Lasersohn, P. (2008). Quantification and perspective in relativist semantics. *Philosophical Perspectives, 22*, 305–337.

Lasersohn, P. (2009). Relative truth, speaker commitment, and control of implicit arguments. *Synthese, 166*, 359–374.

Leitgeb, H. (2008). On the metatheory of Field's 'solving the paradoxes, escaping revenge. In Beall (2008a).

Ludlow, P. (2011). *The philosophy of generative linguistics*. Oxford: Oxford University Press.

MacFarlane, J. (2005a). Making sense of relative truth. *Proceedings of the Aristotelian Society, 105*, 321–339.

MacFarlane, J. (2005b). The assessment sensitivity of knowledge attributions. *Oxford Studies in Epistemology, 1*, 197–233.

MacFarlane, J. (2007a). Relativism and disagreement. *Philosophical Studies, 132*, 17–31.

MacFarlane, J. (2007b). The logic of confusion. *Philosophy and Phenomenological Research, 74*, 700–708.

MacFarlane, J. (2008). Truth in the garden of forking paths. In M. García-Carpintero & M. Kölbel (Eds.), *Relative Truth* . Oxford: Oxford University Press.

MacFarlane, J. (forthcoming). *Assessment sensitivity: Relative truth and its applications*.

Maudlin, T. (2004). *Truth and paradox: Solving the riddles*. Oxford: Oxford University Press.

McGee, V. (1985). A counterexample to modus ponens. *Journal of Philosophy, 82*, 462–471.

McGee, V. (1991). *Truth, vagueness, and paradox: An essay on the logic of truth*. Cambridge: Hackett Publishing Company.

McGee, V. (1994). Afterword: Truth and paradox. In R. M. Harnish (Ed.), *Basic topics in the philosophy of language* (pp. 615–633). London: Harvester Wheatsheaf.

Pinillos, A. (2011). Time-dilation, context, and relative truth. *Philosophy and Phenomenological Research, 82*, 65–92.

Predelli, S. (2005). *Contexts: Meaning, truth, and the use of language*. Oxford: Oxford University Press.

Priest, G. (1979). The logic of paradox. *Journal of Philosophical Logic, 8*, 219–241.

Priest, G. (1990). Boolean negation and all that. *Journal of Philosophical Logic, 19*, 201–215.

Priest, G. (2005). Spiking the field-artillery. In Beall and Armour-Garb (2005).

Priest, G. (2006a). *In contradiction: A study of the transconsistent* (2nd ed.). Oxford: Oxford University Press.

Priest, G. (2006b). *Doubt truth to be a liar*. Oxford: Oxford University Press.

Priest, G. (2008). Revenge, field, and ZF. In Beall (2007).

Prior, A. N. (1960). The runabout inference ticket. *Analysis, 21*, 38–39.

Quine, W. V. (1951). Two dogmas of empiricism. In *From a logical point of view*. Harvard University Press, Harvard (1953).

Quine, W. V. (1960). *Word and object*. Cambridge: MIT Press.

Ray, G. (2005). On the matter of essential richness. *Journal of Philosophical Logic, 34*, 433–457.

Rayo, A., & Welch, P. (2008). Field on revenge. In Beall (2008a).

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

Scharp, K. (2011). Xeno semantics for ascending and descending truth. In N. Tennant (Ed.), *Foundational adventures: essays in honor of Harvey M. Friedman*. London: Templeton Press (Online) and College Publications.

Scharp, K. (2013a). *Replacing truth*. Oxford: Oxford University Press.

Scharp, K. (2013b). Truth, the liar, and relativism. *Philosophical Review, 122*, 427–510.

Shapiro, S. (2000). The status of logic. In P. Boghossian & C. Peacocke (Eds.), *New essays on the a priori* (pp. 333–367). Oxford: Oxford University Press.

Shapiro, S. (2004). Simple truth, contradiction, and consistency. In G. Priest, J. C. Beall & B. Armour-Garb (Eds.), *The law of non-contradiction*. Oxford: Oxford University Press.

Simmons, K. (1993). *Universality and the liar: An essay on truth and the diagonal argument*. Cambridge: Cambridge University Press.

Shapiro, L. (2011). Expressibility and the Liar's revenge. *Australasian Journal of Philosophy, 89*, 297–314.

Soames, S. (1999). *Understanding truth*. Oxford: Oxford University Press.

Sorensen, R. (1998). Yablo's paradox and kindred infinite liars. *Mind, 107*, 137–155.

Stanley, J. (2005). *Knowledge and practical interests*. Oxford: Oxford University Press.

Suppes, P. (2002). *Representation and invariance of scientific structures*. Stanford: CSLI.

Tarksi, A. (1933). The concept of truth in formalized languages. In J. H. Woodger (tr.), & J. Corcoran (Eds.), *Logic, semantics, meta-mathematics*. Indianapolis: Hackett Publishing Company.

Tarksi, A. (1944). The semantic conception of truth. *Philosophy and Phenomenological Research, 4*, 341–376.

Tarski, A., & Vaught. (1956). Arithmetical extensions of relational systems. *Compositio Mathematica, 13*, 81–102.

Tennant, N. (2005). Rule-circularity and the justification of deduction. *The Philosophical Quarterly, 55*, 625–648.

Thomason, R. (1986). Paradoxes and semantic representation. In J. Y. Halpern (Ed.), *Reasoning about knowledge*. San Francisco: Morgan Kaufman.

Williamson, T. (2003). Understanding and inference. *The Aristotelian Society, Supplement, 77*, 249–293.

Williamson, T. (2006). Conceptual truth. *The Aristotelian Society, Supplement, 80*, 1–41.

Wright, C. (2004). Warrant for nothing (and foundations for free?). *Proceedings of the Aristotelian Society, 78*, 167–212.

Yablo, S. (1993). Hop, skip and jump: The agnostic conception of truth. *Philosophical Perspectives, 7*, 371–396.

Zwicky, A., & Sadock, J. (1975). Ambiguity tests and how to fail them. In Reimbald. (Ed.), *Syntax and semantics 4*. New York: Academic Press.