Scharp on Replacing Truth

Andrew Bacon*

October 5, 2016

Abstract

Kevin Scharp's 'Replacing Truth' is an ambitious and far reaching account of the semantic paradoxes. In this critical discussion we examine one the books central claims: to have provided a theory of truth that avoids the revenge paradoxes. In the first part we assess this claim, and in the second part we investigate some features of Scharp's preferred theory of truth, ADT, and compare it with existing theories such as the Kripke-Feferman theory. In the appendix a simple model of Scharp's theory is presented, and some potential consistent ways to strengthen the theory are suggested.

Kevin Scharp's 'Replacing Truth' is an ambitious and far reaching account of the semantic paradoxes.[1] The book is nominally a defence of an inconsistency account of truth. However Scharp brings a wide range of ideas from epistemology and the philosophy of language to bear on these issues. The result is a very distinctive view that combines Scharp's novel theory, involving two complementary truth-like predicates, with elements from the inconsistency theory and recent work on relativism. Scharp somehow manages to take this large collection of disparate ideas and bring them together into a single package. Although the scope of the book is wide, there are several central themes that recur throughout the book. I shall be focussing on a couple.

One of the main considerations Scharp offers in support of his theory is that his theory, unlike his rivals, escapes the revenge paradoxes. I examine this claim in section 1. Roughly, while many theorists are engaged in the project of giving a substantive classification of sentences which are 'paradoxical' (the feature that distinguishes the liar from ordinary sentences), historically this has not been a project that inconsistency theorists have concerned themselves with. Since it is exactly this project that is under threat from the revenge paradoxes, those who say nothing or little about it are not targets of a revenge paradox. However, solving the revenge paradoxes and merely not being the target of one, are not the same thing. In section 1 I argue that a substantive classification of sentences of this type is required of an explanatory theory of the liar, and moreover that

*Many thanks to Kevin Scharp and Jeremy Goodman for discussion.

[1] References to page numbers and sections will be to Scharp (2013a) unless otherwise noted.

Scharp doesn't achieve this with his own fairly schematic notion of 'safeness', which he uses to classify the paradoxical sentences.

Another central component of Scharp's book is his distinctive formal theory, ADT, that governs two predicates that are intended to replace the inconsistent concept of truth. In section 2 I contrast it with a better known theory, the Kripke-Feferman theory, and in section 3, argue that Scharp's system is exceedingly weak and uninformative by comparison. Unfortunately Scharp doesn't provide us with an underlying picture of descending and ascending truth that would guide us toward a better theory. These problems, I suggest, ultimately indicate that a theory based on an antecedently understood picture of truth, such as the Kripke-Feferman theory, might be more suited to Scharp's purposes.

1 Diagnoses and revenge

The naïve concept of truth at least appears to be governed by the T-schema:[2]

T. 'P' is true if and only P

By now we should all be familiar with problems that beset this naïve conception. If L is the sentence 'L is not true', then by substituting L into T and applying Leibniz's law we can infer that L is true if and only if it isn't. If we furthermore assume the classical laws of logic we can derive from this any conclusion we like.

Something has clearly gone wrong somewhere. But saying *where* it went wrong is only half the problem. It is surely also important that we say something about *why* it went wrong. Inconsistency theories have traditionally concerned themselves with answering the second question, leaving logicians to answer the first (see Chihara (1979), p590-591). According to the inconsistency theory the concept of truth is governed (in some sense or other) by inconsistent rules. Abstracting from the specifics, the idea is that these rules are so closely tied to the meaning of the word 'true' that it is no surprise that those employing the concept are disposed to find instances of T attractive. Of course this response is not satisfactory without a general story about where derivations like these go awry – some kind of formal theory of truth – and one of Scharp's goals is to 'get in the game' and provide a response to the first question that is congenial to the diagnosis in terms of inconsistency.[3]

It is worth pointing out that it is also unsatisfactory to only meet the first challenge. It would not do to point out the step in the derivation of absurdity that you reject and say nothing more. One also wants a picture of what's going on with liar-like sentences which predicts that you wouldn't expect T to hold in the problematic cases – some illuminating background picture that makes clear which part of the naïve picture breaks down and why. (We'll have more to say about this later.)

[2]Here, and elsewhere, the 16th letter of the alphabet is to be substituted for a declarative sentence of English both when it appears in quotation marks and without.

[3]See Scharp's response to Beall and Priest's objections to the inconsistency theory in Scharp (2013b).

Now, one the key themes of Scharp's book – one of the main arguments in favour of his own version of the inconsistency theory over his rivals – is that, unlike these rival theories, Scharp's account addresses and ultimately does not fall afoul of the revenge paradoxes. Before we assess this contention, then, it is worth getting clear on exactly how the revenge paradoxes end up entering into the story, either within an inconsistency account of truth or otherwise.[4]

The standard recipe for revenge goes something like this. In the process of addressing the second question – of providing a diagnosis of the paradoxes – one often attempts to identify some feature of the liar sentence that is shared by other problematic instances of T (instances involving the Curry sentence, liar pairs, Yablo's paradox, and so on), but not shared with the unproblematic instances of T (instances such as 'snow is white', '1+1 = 5' and so on). Extant examples of this strategy include the claim that the bad instances of T involve sentences that are ungrounded (Kripke (1975)), fail to express a proposition (e.g. Prior (1961)), are context sensitive in certain ways (e.g. Burge (1979)), have been introduced by impredicative definitions (Russell (1903)), are semantically indefinite (McGee (1990)) or are semantically unstable (e.g. Gupta and Belnap (1993)).[5] In each case one can construct a revenge sentence R which says of itself that it is either untrue or has the status in question. The assumption that the sentence doesn't have the problematic status leads directly to contradiction.[6] So we can prove that R doesn't have the status; and so, of course, R either doesn't have the status or is untrue. But then we have just proven R itself.[7] It seems that if one attempts to diagnose the paradoxes this way, one has to assert sentences that, like the liar, have the very status alleged to be problematic.

Note, however, that traditional inconsistency theories of truth tend not to spend much time on the project of characterising the problematic sentences: finding that feature of the liar that distinguishes it from those sentences to which disquotational reasoning straightforwardly applies. Rather they focus on explaining why we find T so seductive. It is therefore not straightforward to run a revenge paradox on the inconsistency theorist, because we have nothing playing the role that, say, groundedness plays in Kripke's theory.

The cost of this apparent victory is that this leaves us lacking an informative characterisation of the problematic instances of T. Although we have a story

[4] A word of caution: in the literature on the liar there is a cluster of very closely related problems that get labelled the 'revenge paradox' by different people. In the following I focus on revenge paradoxes that arise when one introduces vocabulary for the purposes of diagnosing the semantic paradoxes (see Bacon (2015)). Scharp doesn't focus explicitly on this aspect of the revenge paradoxes, and so his discussion of revenge is slightly more narrow than I think it should be. At any rate, the problems I'm highlighting here need addressing whether or not one calls them 'revenge'.

[5] The above theories all characterise the problematic instances by a restriction on the sentences appearing on the left-hand side of T. In Bacon (2015) I suggest that we should be using a classification that restricts on the right-hand side: a distinction that would be drawn using an operator rather than a predicate expressing a property of sentences.

[6] If it didn't have the status then we'd be able use T (since it's not liar like) and conclude that R is true if and only if it's either untrue or doesn't have the status. Under the assumption that it doesn't have the status this is of course impossible.

[7] See Bacon (2015) for discussion and a more rigorous version of the argument.

about how things went wrong, this theory says nothing predictive about *when* things will go wrong. The crucial question, then, is whether we really needed such a classification in the first place. Is it sufficient to simply state that T is at fault, but stop short of giving a characterisation of the instances of T are at fault?

In my view some kind of informative classification of this sort is required of an explanatory theory of the paradoxes. To see why, consider the following sentence:

(L^*) Either L^* isn't true or 'snow is white' isn't true and snow is white

Two further instances of T + Leibniz's law gives us:

1. 'Snow is white' is true if and only if snow is white.
2. L^* is true if and only if either L^* isn't true or 'snow is white' isn't true and snow is white

Neither of these two instances of T are inconsistent on their own, however from both together one can prove a contradiction. A predictive theory of the liar ought to tell us which instance is causing the problem.

Most of the theories listed earlier tell us which instance to reject. Unsurprisingly, they all deliver the verdict that it's 2. The grounding theorist, for example, can point out that L^* is ungrounded, whereas the sentence 'snow is white' is not: the latter bottoms out in non-semantic facts, whereas the former does not. The revision theorist can point out that L^* is semantically unstable in the sense of Gupta and Belnap (1993), so we should favour 1 over 2. And so on.

But which instance does the inconsistency theory predict we should reject? One could point out that L^* contains a word whose constitutive rules are inconsistent – the word 'true' – whereas 'snow is white' does not. This will clearly not do, however, because there are pairs of instance of T analogous to 1 and 2 where it is clear which instance to reject, but where both contain the word 'true'. Consider the sentence L^{**} = 'either L^{**} isn't true or "snow is white' is true' isn't true and 'snow is white' is true' and the corresponding two instances of T:

1′. "Snow is white' is true' is true if and only if 'snow is white' is true.

2′. L^{**} is true if and only if either L^{**} isn't true or "snow is white' is true' isn't true and 'snow is white' is true.

Where the grounding theory, for example, predicts we should relinquish 2′, the above proposal classifies both in the same way since both instances concern sentences containing the word 'true'. (Nor would it do to suggest that the bad instances of T be inconsistent: both 1′ and 2′ are consistent on their own; it is only together that they're inconsistent.)

So while it is certainly true that one cannot flatfoodedly run the revenge paradoxes against inconsistency theories of truth, it is primarily because it is

unclear how they resolve the question of *which* instance of T must be rejected; simply noting that the word 'true' has constitutive principles that are inconsistent does not answer this.

In this sense, the accounts of Chihara (1979) and other similar inconsistency theories, cannot really be said to avoid the revenge paradoxes. They are merely not engaging in the kind of project to which revenge style objections are usually directed.

How, then, does Scharp's claim that his own theory avoids the revenge paradoxes bear on this? On this point Scharp offers something more than the basic inconsistency theory delivers: he provides us with a formal theory, that contains something he calls a 'safety' predicate, which plays something like the role described above in classifying sentences into problematic and unproblematic. A safe sentence, according to Scharp, is a sentence to which one can apply disquotational reasoning. Moreover, Scharp's theory of safety is consistent: he shows that one cannot derive any contradictions from his theory of safeness and truth. In particular there are revenge sentences involving the safety predicate, but one cannot derive any contradictions from them. (What happens to the revenge argument we gave earlier? It turns out that the theory proves of some of its theorems that they are unsafe; however since safety is a technical notion that Scharp takes as primitive it is hard to assess whether this result is problematic or not.[8])

Perhaps, then, we can answer our challenge by arguing that one of L^{**} or "snow is white' is true' is unsafe. Although it's not implausible that Scharp could deliver this particular verdict, Scharp tells us very little about what a safe sentence is, other than that it is a sentence to which one can safely apply disquotational reasoning, and that it is governed by the other principles of Scharp's theory. Unfortunately the theory in question is quite weak; a point we shall return to in section 3.[9]

Contrast this with the grounding or revision theories of truth. Given a simple language, such as the language of arithmetic with a truth predicate, these theories provide explicit definitions of the grounded and semantically stable sentences of that language. We have necessary and sufficient criteria for working out when a sentence of this language is of the good or bad form. Scharp's theory of safety, on the other hand, does not supply criteria like this because the theory that governs it is radically incomplete. For example, although the axioms of Scharp's theory are provably safe, if you conjoin some of those axioms together the resulting conjunction is not provably safe. The theory on its own does not tell us whether these conjunctions are safe (see section 3 for more examples like this).

Now presumably if Scharp were to flesh out the notion of safeness in more

[8]The standard puzzle for this kind of theory is that it seems as though Scharp must go about asserting sentences which by his own lights are unsafe. Scharp says very little about the relation between safety and assertion; although see the short discussion in §8.6.

[9]A sentence is safe in Scharp's theory if it is either descending true or descending false. But the notion of being descending true is similarly fairly unconstrained – see the discussion in sections 2 and 3.

detail – give us a background picture, from which answers to these questions would just fall out – we would have the resources to determine whether conjunctions of Scharp's axioms are safe.[10] But the problem is that it is far from clear that once we have such a characterisation of the safe sentences we won't inadvertantly reinstate the revenge paradox. This is exactly what is so hard about the revenge problem: as soon as one tries to say something informative about the problematic sentences, one runs into revenge. It is for exactly this reason, as well, that revenge paradoxes can be directed at grounding and revision theories of truth. Of course, the less one says about the safety predicate, the harder it is to formulate the revenge objection. This much goes for any subject matter – the less one says about something, the harder it is to formulate objections to what you're saying – but it would be misleading to think of this as a way of *avoiding* certain kinds of objections.

In short, revenge is quite a delicate matter for the inconsistency theorist. Although the inconsistency theorist typically doesn't accept the kind of ideology needed to get a revenge paradox going, it is arguable that this ideology must be part of any explanatory theory. Scharp's introduction of a safety predicate seems like a step in the right direction in this regard, however it is radically underdetermined which sentences are safe. More information about what counts as safe is needed before we can really assess the claim that Scharp has avoided the revenge paradoxes.

2 Scharp's theory of ascending and descending truth

Let us now turn to Scharp's positive theory of truth. According to Scharp, the ordinary concept of truth is inconsistent: some of the principles constitutive of the concept are false (possibly even inconsistent).[11] The following are two such principles for the concept of truth:

T-In If p then 'p' is true.

T-Out If 'p' is true then p.

A centerpiece of Scharp's book is the introduction of a pair of replacement concepts for theorizing consistently about truth. To adopt Scharp's own analogy, this pair of concepts stand to truth as rest mass and relativistic mass stand to the pre-relativistic uses of the word 'mass' before we discovered that the principles constitutive of the latter concept were false.

In the case of the pre-relativistic concept of mass some principles constitutive of the original concept are satisfied by rest mass, whilst some are satisfied by

[10] Scharp does give us a model theory, but it has little intuitive rationale. He does not, at any rate, go to any length to explain what safety amounts to in terms of the model theory, or to say informally what is wrong with sentences that turn out unsafe in the model.

[11] A principle is constitutive of a concept, according to Scharp, if possessors of the concept automatically gain a kind of (defeasible) entitlement to those principles.

relativistic mass. Scharp's replacement concepts, which he calls 'descending' and 'ascending truth' respectively, behave similarly with each inheriting one of the two jointly inconsistent principles constitutive of the original concept of truth:

AT-In If p then 'p' is ascending true.

DT-Out If 'p' is descending true then p.

Neither concept alone can do the explanatory work that the inconsistent concept of truth was supposed to do, but just as with the case of mass, they are each supposed to occupy a consistent fragment of the role of truth.

Considerations of liars involving these new predicates straightforwardly raise the question of whether the two revised principles are consistent. One needs not only to account for the sentences which say of themselves that they are not ascending true, and descending true, respectively, but also more complicated sentences such as the sentence that says of itself that it is not ascendingly descending true, and so on. Although Scharp doesn't go this route, there is actually an extremely natural way to co-opt constructions familiar from more standard theories of truth to achieve these principles.[12]

Before we consider Scharp's solution to this problem it will be instructive to consider one of these theories – the Kripke-Feferman theory (henceforth, KF) – as a useful point of comparison in what follows.[13] KF is a consistent theory of truth which accepts all instances of of T-Out. An intuitive way to think about the theory is by considering a particularly natural model for the theory: given a classical model, M, of a non-semantic language $\mathcal{L}$, we can extend it to a model of $\mathcal{L}_{Tr}$ ($\mathcal{L}$ with a truth predicate) by simply setting the extension of the truth predicate to the set of $\mathcal{L}_{Tr}$ sentences that are grounded and true relative to M (for a precise definition of 'grounded' see Kripke (1975)).

According to this theory the liar sentence L is neither true nor false.[14] Thinking in terms of our model this is easily verified, since neither the liar or it's negation is grounded. Since, in particular, L is not true, one has to relinquish T-In, otherwise one could conclude that 'L is not true' is true, and hence that L is true after all. However, there is another predicate definable within KF that does satisfy the axiom T-In, the dual of the truth predicate: 'doesn't have a true negation' (it stands to truth as possibility does to necessity). Since the original truth predicate classifies the liar as neither true nor false, the dual predicate classifies the liar as both true and false. More importantly, every instance of

[12]There is, of course, a very simple way of seeing the consistency of these two principles: just let the descending truth predicate apply to nothing and the ascending truth predicate apply to everything. We are looking for a pair of predicates with more interesting laws than this.

[13]I am here working with the version of KF includes the claim that not all sentences are true, which is sometimes omitted from axiomatisations. A close variant of KF, VF, based on supervaluations is closer to Scharp's proposal, since it guarantees that the theorems of classical logic are true (Cantini (1990)). The difference between these theories will not really matter for my exposition, however, so I shall stick with KF since it is more familiar.

[14]By false I simply mean: has a true negation.

T-In is validated for the dual predicate: the schema 'if p then 'not p' is not true' is a theorem of KF since it follows from the 'not p' instance of T-Out by contraposition and the double negation laws – thus the dual of truth satisfies T-In within KF.

These observations make clear that one can get a rich and interesting two predicate theory of truth out of a single predicate theory such as KF by simply interpreting descending truth with the theory's truth predicate and interpreting ascending truth with its dual. Since we have such a simple and natural two predicate theory at our disposal it is worth examining the reasons Scharp doesn't simply adopt this kind of theory.

According to Scharp the basic theory KF is unsatisfactory because the theory proves that some of its own axioms aren't true: each instance of T-Out is part of the theory, but the claim that those instances of T-Out are true is not, and moreover cannot be consistently added to the theory. In the two predicate variant this results in 'a theory of descending truth that is not itself descending true' (p150). He calls this the 'self-refutation problem', and considers it to be a form of the revenge paradox (see the discussion of Maudlin on pp102-104). Thus Scharp rejects the two predicate variant of KF on the grounds that it succumbs to the self-refutation problem.

In fact this feature is not just an artefact of KF. One can show that any theory of descending truth (i.e. a theory that contains DT-Out) which also contains (i)-(iii) is inconsistent with the existence of a descending truth liar:

(i) The instances of DT-Out are descending true: $D(\ulcorner D(\ulcorner \phi \urcorner) \rightarrow \phi \urcorner)$

(ii) The logical axioms are descending true: $D(\ulcorner \phi \urcorner)$ whenever ϕ is a logical axiom.

(iii) Modus ponens preserves descending truth: $D(\ulcorner \phi \rightarrow \psi \urcorner) \rightarrow (D(\ulcorner \phi \urcorner) \rightarrow D(\ulcorner \phi \urcorner))$.

This result is known as Montague's paradox (see Montague (1963)). Since denying either (i) or (ii) would involve succumbing to the self-refutation problem, Scharp opts for the third option.[15] Modus ponens, and indeed many other rules of proof in Scharp's preferred theory ADT, do not preserve descending truth. The result, as Scharp puts it, is a theory of descending truth that is itself descending true.

However one has to be careful: although the *axioms* of ADT are descending true, many of its *theorems* aren't simply because the rules of proof can take descending truths to things that aren't descending true. Indeed there are theorems of ADT that ADT actually proves aren't descending true, such as the

[15] He cites Hartry Field's arguments to the effect that all theories of truth have rules that do not preserve truth (Field (2006)). Note that, *pace* Field, this is not entirely correct. There are, for example, theories based on revision and internal supervaluational constructions can be formulated in such a way that all of their rules provably preserve truth. These theories crucially must not be thought of as having the rule of inference 'from ϕ infer $Tr(\ulcorner \phi \urcorner)$' and its converse, although they can have the *rule of proof* 'if ϕ is provable, so is $Tr(\ulcorner \phi \urcorner)$' and it's converse, which can be captured in the object language using provability predicates: for the in and out rulse this includes $Pr(\ulcorner \phi \urcorner) \leftrightarrow Pr(\ulcorner Tr(\ulcorner \phi \urcorner \urcorner)$.

descending liar. If one takes the self-refutation problem seriously, as Scharp apparently does, I think more discussion of this feature is required.

The first thing to note is that although Scharp's theory counts all of its axioms as descending true, this feature of the theory is extremely sensitive to how it is axiomatised. Logically equivalent axiomatisations of ADT can disrupt this feature. For example, the rule of double negation introduction does not preserve descending truth. Suppose that A is any one of Scharp's axioms – for concreteness sake, suppose it is an instance of DT-In. It is impossible to prove that $\neg\neg A$ is descending true in ADT (the model I describe later, for example, does not count these sentences as descending true), so if I were to remove A from the theory and replace it with $\neg\neg A$, although I would have a logically equivalent axiomatisation, the resulting system no longer proves that its axioms are descending true. A similar point can be made about conjunction introduction: according to Scharp's theory one can have two descending true sentences whose conjunction is not descending true. So if I had decided to conjoin the axioms of Scharp's theory instead of merely listing them separately I would no longer have a theory that can prove that all of its axioms are descending true (again, this is demonstrated by the model below).

When Scharp presents his theory ADT he does not say which of the logical connectives he is treating as primitive and which (if any) as defined. For reasons similar to those discussed above, this actually makes a difference to the content of the theory. If you took every connective as primitive then the theory can prove $D(\ulcorner D(\ulcorner \phi \urcorner) \rightarrow \phi \urcorner)$ but not $D(\ulcorner \neg D(\ulcorner \phi \urcorner) \vee \phi \urcorner)$, whereas if you defined $A \rightarrow B$ as simply $\neg A \vee B$ then these would be identical and both would be part of the theory. These issues becomes significant later when we try to reaxiomatise Scharp's theory by treating different things as primitive.

Of course Scharp could augment his theory so that it included the claim that this or that particular logical equivalent of an axiom is also descending true (he would have to modify his consistency proof). But the point I'm making can be generalised quite significantly: one can show that descending truth cannot be preserved under logical equivalence, on pain of paradox. This follows from an argument Scharp discusses himself, attributed to Dana Scott (see §6.6.4), that establishes that descending truth cannot be consistently closed under the substitution of logical equivalents (and thus neither can any candidate extensions of Scharp's system). It follows that it is always possible to find reaxiomatisations of the theory which fall afoul of the self-refutation problem: logically equivalent ways of axiomatising the theory whose axioms are not descending true.

It is also unclear how stable Scharp's position on Montague's paradox is is. Scharp's attitude that it's OK to have rules of proof that do not preserve truth, but not OK to have untrue axioms requires one to draw a somewhat arbitrary line between rules of proof and axioms. Even in the case of propositional logic there is no standard way of haloing certain valid inferences as *the rules* of propositional logic and other valid principles *the axioms*. Natural deduction systems favour rules over axioms, for example, whereas Hilbert formalisations favour axioms over rules. In a Hilbert system one can derive the rules of the natural deduction systems, and in the natural deduction systems one can derive

the axioms of the Hilbert systems, but the choice to use one formalism rather than the other seems of little significance other than economy.

Indeed, if ones goal is simply to provide a theory of descending truth which proves that all of its axioms are descending true at the expense of having rules that preserve descending truth, then that goal can be achieved extremely easily. All one needs to do is reformulate the theory – be it KF, or what have you – so that it has no axioms at all, and has instead a bunch of rules of inference in their place. Here's a simplistic way to do that: replace each axiom, A, of the original system with the rule 'infer A from no premises'. I am certain that this maneuver would not satisfy Scharp, but I take it that this just goes to show that we need a clearer conception of what the self-refutation problem is, and what it would take to solve it. As such, I think, it is unclear whether Scharp's own proposal really meets the challenges he sets against his rivals.

3 The theory ADT

So what exactly is Scharp's theory of truth? Scharp presents his theory as a list of 20 axioms that govern both descending truth and ascending truth, and a notion of a sentence being 'safe' which effectively amounts to saying that the sentence is either descending true or descending false: in other words, it is not a descending truth gap (or equivalently, it is not an ascending truth glut). However there is some redundancy in Scharp's axiomatisation which obscures some notable features of the system. One thing we've noted already is that once we have either a descending or ascending truth predicate, we can simply define the other as its dual. To get a feel for the theory, then, it might make sense just to look at an axiomatisation purely in terms of one of the truth predicates, and treat the other predicate as defined. By getting a clearer picture of one of these predicates, we automatically can see what happens to the other.

Below is one such axiomatisation of ADT in terms of a descending truth predicate only. I assume here that the base language is that of arithmetic, and that in addition to the axioms below we have all of the axioms of Peano arithmetic. As noted above, it potentially matters which logical connectives I take as primitive; to avoid arbitrariness I shall take them all as primitive, although I shall treat ascending truth and the safety predicate as defined: $A(\ulcorner\phi\urcorner)$ as $\neg D(\ulcorner\neg\phi\urcorner)$ and $S(\ulcorner\phi\urcorner)$ as $D(\ulcorner\phi\urcorner) \vee D(\ulcorner\neg\phi\urcorner)$.

1. $D(\ulcorner\phi\urcorner) \rightarrow \phi$
2. $D(\ulcorner\neg\phi\urcorner) \rightarrow \neg D(\ulcorner\phi\urcorner)$
3. $D(\ulcorner\phi \wedge \psi\urcorner) \rightarrow D(\ulcorner\phi\urcorner) \wedge D(\ulcorner\psi\urcorner)$
4. $D(\ulcorner\neg(\phi \vee \psi)\urcorner)) \rightarrow D(\ulcorner\neg\phi\urcorner) \wedge D(\ulcorner\neg\psi\urcorner)$
5. $D(\ulcorner\phi\urcorner) \vee D(\ulcorner\psi\urcorner) \rightarrow D(\ulcorner\phi \vee \psi\urcorner)$
6. $D(\ulcorner\neg\phi\urcorner) \vee D(\ulcorner\neg\psi\urcorner) \rightarrow D(\ulcorner\neg(\phi \wedge \psi)\urcorner)$

7. $s = t \rightarrow D(\ulcorner \phi \urcorner) \leftrightarrow D(\ulcorner \phi[s/t] \urcorner)$.[16]

8. $D(\ulcorner \phi \urcorner)$ if ϕ is a theorem of PA, an instance of 1-7, or is one of the principles (i)-(ix) (see footnote.[17])

To show the consistency of ADT Scharp introduces a fairly elaborate construction involving revision sequences and a novel way of modeling truth predicates which aren't closed under deduction which he calls 'xeno models'. As far as I can see, these models have little intuitive rationale, and are primarily introduced as a technical device for showing that no contradictions follow from the axioms of ADT. While there is certainly nothing wrong with this approach to consistency arguments, I suspect that the complexity of this particular construction obscures some of the more quirky features of the theory.

Alternatively, the consistency of ADT can be demonstrated with a fairly humdrum construction, albeit one which I think reveals the structure of descending truth in an illuminating way. Consider the following three rules of inference:

> From ϕ infer $\phi \vee \psi$ and $\psi \vee \phi$
>
> From $\neg\phi$ infer $\neg(\phi \wedge \psi)$ and $\neg(\psi \wedge \phi)$.
>
> From ϕ infer ϕ' whenever $a = b$ is a true arithmetical identity statement and ϕ' is the result of substituting some occurrences of b with a in ϕ.

We can then think of the sentences that are descending true in Scharp's theory as those sentences that are either axioms of his theory (the principles (1)-(7), (i)-(ix) listed above), theorems of PA (including the theorems of logic) or things one can prove from those two sets using only the above rules. Indeed, it turns out that if we let M be a standard model of arithmetic that assigns the descending truth predicate the set of sentences just described then M is a model of ADT (see appendix).[18]

This model is also, in some sense, a *minimal* model of ADT: the descending truth predicate is as small as it can be in a way that's consistent with the

[16] This principle corresponds to Scharp's E principles, which appear to be stated as rules. However he treats them as axioms in the statement of his axiom D7.

[17] Principles (i)-(ix) are principles that are redundant in Scharp's axiomatisation, modulo definitions, except for the fact that he declares them descending true. They are, respectively: (i) $\phi \rightarrow \neg D(\ulcorner \neg\phi \urcorner)$, (ii) $\neg\neg D(\ulcorner \neg\phi \urcorner) \rightarrow \neg D(\ulcorner \neg\neg\phi \urcorner)$, (iii) $\neg D(\ulcorner \neg\phi \urcorner) \vee \neg D(\ulcorner \neg\psi \urcorner) \rightarrow \neg D(\ulcorner \neg(\phi \vee \psi) \urcorner)$, (iv) $D(\ulcorner \phi \urcorner) \leftrightarrow \neg\neg D(\ulcorner \neg\neg\phi \urcorner)$, (v) $D(\ulcorner \phi \urcorner) \vee D(\ulcorner \neg\phi \urcorner)) \leftrightarrow (D(\ulcorner \phi \urcorner) \vee \neg\neg D(\ulcorner \neg\phi \urcorner)$, (vi) $\phi \wedge (D(\ulcorner \phi \urcorner) \vee D(\ulcorner \neg\phi \urcorner)) \rightarrow D(\ulcorner \phi \urcorner)$, (vii) $\neg D(\ulcorner \neg\phi \urcorner) \wedge (D(\ulcorner \phi \urcorner) \vee D(\ulcorner \neg\phi \urcorner)) \rightarrow \phi$, (viii) $s = t \rightarrow (\neg D(\ulcorner \neg\phi \urcorner) \leftrightarrow \neg D(\ulcorner \neg\phi[s/t] \urcorner))$, (ix) $s = t \rightarrow (D(\ulcorner \phi \urcorner) \vee D(\ulcorner \neg\phi \urcorner)) \leftrightarrow (D(\ulcorner \phi[s/t] \urcorner) \vee D(\ulcorner \neg\phi[s/t] \urcorner))$ when $s = t$ is a true arithmetical identity. Note that all of these principles are obviously logically equivalent to or provable from axioms (1)-(7). However we have to include them manually since descending truth isn't closed under logical equivalence or consequence, and Scharp's theory treats these particular sentences as descending true, modulo definitions.

[18] Note here that I don't include conjunction elimination, despite the presence of axiom 3. This is because the set I've described is already closed under conjunction elimination: the only conjunctions in the extension of D are theorems of arithmetic, and the conjuncts in this case are also theorems of arithmetic. A similar point applies to axiom 4, and the rule it corresponds to.

axioms of ADT. What we can see quite clearly from this model is that the descending truth predicate is closed under disjunction introduction, another rule very similar in spirit to disjunction introduction (given deMorgan equivalence) and a principle that allows us to substitute coreferring terms. As we noted above, it cannot be closed under modus ponens on pain of paradox.

But there is a lot of space between the rules we can close descending truth under in Scharp's framework, and the rules we can't. For example, note also that conjunction introduction does not preserve descending truth in our model, nor does double negation elimination and introduction, the deMorgan laws or the contraposition laws. In fact, in the above model even though the axioms (1)-(7) are descending true, no conjunction of them is descending true. Not even the conjunction of an axiom with itself is descending true in this model, so it follows that although ADT can prove that $D(\ulcorner 0 = 0 \urcorner) \rightarrow 0 = 0$ is descending true, for example, it can't prove that the conjunction $(D(\ulcorner 0 = 0 \urcorner) \rightarrow 0 = 0) \wedge (D(\ulcorner 0 = 0 \urcorner) \rightarrow 0 = 0)$ is descending true. Scharp's framework goes a small way to respecting deMorgan equivalence, but not very far: we make sure that we are not only closed under disjunction introduction, but also negated conjunction introduction. However we fall far short of the full deMorgan equivalences.

A more striking feature is that descending truth is not provably closed under the associativity or commutivity laws for conjunction. This fact is not demonstrated by the model I give in the appendix, but can be seen by taking any three axioms, A, B and C, and adding the conjunction $((A \wedge B) \wedge C)$ and $(A \wedge B)$ to the extension of the descending truth predicate described in the appendix model and closing under the three laws of inference. In the resulting model $((A \wedge B) \wedge C)$ is descending true but $(A \wedge (B \wedge C))$ is not. One can similarly construct a model in which $(A \wedge B)$ is descending true but not $(B \wedge A)$.

Now it is clear that Scharp does not consider ADT to be the full story (see §6.4). As a theory it is surely incomplete: perhaps there are further truths about descending and ascending truth that the theory does not capture. Incompleteness is an inevitable feature of many theories, after all. For instance, there are many truths about the natural numbers that one cannot prove within the mathematically rich system of Peano arithmetic. If the issue was just a matter of incompleteness, then although the above observations reveal the extent of the incompleteness there is no particularly pressing worry for Scharp's overall project. Of course one needs to know that any desirable principles we are missing can be consistently added to ADT, and since Scharp has no model of these further principles we have no guarantee that they can be; this at most indicates that further work is still needed.

However I think that behind the issue of incompleteness lurks a more substantial, methodological problem with Scharp's approach. The notions of ascending truth, descending truth and safety are all technical notions: they are replacements for the inconsistent pretheoretic concept of truth, and they simply aren't concepts we had prior to theoretical investigation. In so far as we have a grasp on these notions at all, it is acquired solely by the things Scharp tells us about them. In this respect ascending and descending truth are quite unlike the arithmetical concepts: the former concepts are explained to us by the theo-

retical role that they play, whereas we clearly have an self-standing grasp of the arithmetical concepts independently of any axiomatisation of them.

The above observations, however, indicate that this theoretical role is actually quite thin. We are told, for example, that ascending and descending truth are supposed to obey T-in and T-out respectively. This is consistent with ascending truth applying to every sentence, and descending truth applying to no sentence. The more substantial theory, ADT, does a little more to narrow things down, but as we have seen it leaves much open. One might hope to beef up the theoretical role of ascending and descending truth by relating them to the norms of assertion and belief. Unfortunately, neither ascending nor descending truth hook up with assertion or belief in a straightforward way: clearly the theorems of ADT ought to be assertable and believed, but the theorems aren't all descending true, and dually there are ascending truths which aren't assertable – ascending truths can sometimes be inconsistent with one another. In short, we have very little to go on.

In light of this I think talk of the incompleteness of ADT or the idea of a better, more substantial theory of descending and ascending truth, characterising a larger range of principles governing the concepts, is a bit premature: we don't really know what we're aiming for. It is worth contrasting the situation here with the situation in KF, for example. KF sets out to axiomatise the notion of a grounded truth: a sentence whose truth is ultimately grounded by the non-semantic facts. The concept the theory is characterising is clearly in sight prior to the project of trying to axiomatise it, much like Peano arithmetic, and moreover, the theory has a set of intended models with an intuitive rationale behind them. As such we have a clear guide as to which inferences involving truth are to be accepted, and we have the resources to prove that a given theory is complete.[19] But what is guiding our choices when we come to strengthening ADT? There are pairs of natural axioms that can be consistently added to ADT on their own, but not together – how do we decide in these cases? Why, for example, does Scharp insist that descending truth is closed under disjunction introduction, but not conjunction introduction? It is hard to imagine motivating a background picture which favours one of these laws but not the other.[20]

It is perhaps interesting to note at this juncture that Scharp could have consistently added more principles to ADT if he had wanted. For example, if we added conjunction introduction and elimination and double negation introduction and elimination to the three rules listed above, then the consistency proof in the appendix extends completely straightforwardly; one *could* have these principles if one wanted.[21] There are, however, other rules I could have

[19] Indeed, you can prove the completeness of KF with respect to the class of all models that result from a fixed point of a Kripke construction. Perhaps of more interest is the fact that if we are just interested in the *minimal* fixed points of Kripke's construction then no recursive axiomatisation is possible: the model theory is all we have to go on.

[20] Conjunction introduction is one of the few rules, along with modus ponens, that Scharp is explicit about: descending truth is closed under neither.

[21] Once one has added conjunction introduction the resulting set is no longer closed under conjunction elimination, so we must add it manually to get axiom 3. The proof of consistency is basically the same, except that one has to consider a few more cases for the conjunction

added instead which would have precluded me from consistently adding these rules;[22] the moral is that we need to be told more about the concepts we are axiomatising before we start trying to make these choices.

4 Appendix

Let $C(S)$ be the smallest set containing the set S that is closed under the three rules of inference listed in section 3, and let $\mathcal{L}_{Tr}$ be the language of arithmetic with a truth predicate.

Let M be the model of this language that assigns the standard interpretation to the arithmetical vocabulary and assigns $X = C(\{\phi \mid \phi$ an instance of (1)-(7), (i)-(ix) or a theorem of PA$\})$ as the interpretation of D. Then M is a model of ADT.

Axiom 2: is validated since X is a classically consistent set (you can construct a standard model by setting the extension of D to be the empty set).

Axiom 3/4: if $\phi \wedge \psi \in X$ then it is because $\phi \wedge \psi$ is a theorem of PA (no other members of X are conjunctive). Thus both ϕ and ψ are theorems of PA and are thus in X. The argument for axiom 4 is basically the same.

Axiom 5/6: if $\phi \in X$ or $\psi \in X$ then $\phi \vee \psi \in X$ since X is closed under disjunction introduction. Same argument for (6).

Axiom 7: if $a = b$ is true in M it is a true arithmetical identity statement so ϕ is in X iff ϕ' is, since X is closed under the third rule.

Axiom 8: each of the axioms (1)-(7), (i)-(ix) and the theorems of PA are members of X by stipulation.

Axiom 1: we prove by induction on formula complexity that every member of X is true in M. Say that ϕ and ψ are s-equivalent if they can be gotten by making a substition of two terms that constitute a true arithmetical identity. Note that s-equivalents of theorems of PA are theorems of PA.

> If ϕ is atomic and $\phi \in X$ then ϕ is a theorem of PA (no other members of X are atomic). So ϕ is true in M since M is a standard model.
>
> $\phi = \phi \wedge \psi$. The only members of X that are conjunctive are the theorems of PA, so these instances of (1) are true in M.
>
> $\phi = \psi \vee \chi$. Suppose $\phi \in X$. Then either ϕ is a theorem of arithmetic (see above) or ϕ was inferred by disjunction introduction (or is s-equivalent to something inferred by disjunction introduction). So we know that $\psi \in X$ or $\chi \in X$, and by inductive hypothesis either ψ or χ is true in M. Either way $\psi \vee \chi$ is true in M. (If ϕ was merely s-equivalent to something inferred by disjunction introduction then either something s-equivalent to ψ or something s-equivalent to χ is in X, and we can reason analogously.)

and negation clauses of the inductive proof of axiom 1.

[22] The rule: from A and $\neg(A \wedge B)$ infer $\neg B$, for example, is not consistent with double negation elimination, for then $\neg(A \wedge \neg B)$ would behave exactly like the conditional in Montague's theorem.

$\phi = \neg\psi$. Suppose $\phi \in X$. Then either ϕ is a theorem of arithmetic (see above) or $\phi = \neg(\psi \wedge \chi)$ and was inferred by the second rule (or is s-equivalent to something inferred by that rule). So either $\neg\psi \in X$ or $\neg\chi \in X$, and by inductive hypothesis either $\neg\psi$ or $\neg\chi$ is true in M. Either way $\neg(\psi \wedge \chi)$ is true in M. (If ϕ is merely s-equivalent to something inferred by the second rule we can reason analogously.)

$\phi = \psi \rightarrow \chi$. Suppose $\phi \in X$. Then either ϕ is a theorem of arithmetic (see above), or an instance of the principles (1)-(7), (i)-(ix) or s-equivalent to one of them. If it is an instance or s-equivalent to an instance of (2)-(7) then ϕ is true in M by the above arguments. . Suppose, then, that it is an instance of (1): $\phi = D(\ulcorner\psi\urcorner) \rightarrow \psi$. We can see that this is true in M as follows: suppose $\psi \in X$. Since ψ has less complexity than ϕ we know that ψ is true in M, so $D(\ulcorner\psi\urcorner) \rightarrow \psi$ is true in M. (This works if ϕ is merely s-equivalent to an instance of (1).) Finally, M also validates principles (i)-(ix) since they are all provable from principles (1)-(7) and M is a classical model.

References

Bacon, A. (2015). Can the classical logician avoid the revenge paradoxes? *Philosophical Review 124*(3), 299–352.

Burge, T. (1979). Semantical paradox. *The Journal of Philosophy 76*(4), 169–198.

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to id1. *The Journal of Symbolic Logic 55*(01), 244–259.

Chihara, C. (1979). The semantic paradoxes: A diagnostic investigation. *The Philosophical Review 88*(4), 590–618.

Field, H. (2006). Truth and the unprovability of consistency. *Mind 115*(459), 567–606.

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

Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy 72*(19), 690–716.

McGee, V. (1990). *Truth, vagueness, and paradox: An essay on the logic of truth*. Hackett Publishing Company Inc.

Montague, R. (1963). Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. *Acta philosophica fennica 16*, 153–167.

Prior, A. (1961). On a family of paradoxes. *Notre Dame Journal of Formal Logic 2*(1), 16–32.

Russell, B. (1903). *The principles of mathematics*. WW Norton & Company.

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

Scharp, K. (2013b). Truth, the liar, and relativism. *Philosophical Review 122*(3), 427–510.