Inquiry, 2018
https://doi.org/10.1080/0020174X.2017.1420012

Replies to Bacon, Eklund, and Greenough on *Replacing Truth*

Kevin Scharp

University of St Andrews

Arche Philosophical Research Centre

Centre for Exoplanet Science

My deep appreciation goes to Andrew Bacon, Matti Eklund, and Patrick Greenough for their stimulating and challenging comments on my book, *Replacing Truth*.[1] The area in which we all work is immensely richer for their contributions, and this exchange has already improved my understanding of the issues facing us. Each set of comments is self-contained, and they differ significantly from one another; after a brief summary of the book, I respond to them independently.

Replacing Truth Summary

A quick reminder of the major views presented and defended in *Replacing Truth*:

- A theory of inconsistent concepts, which takes concepts to be determined by their constitutive principles; an inconsistent concept has some false constitutive principles.
- That truth is an inconsistent concept; the principles of truth that lead to the liar and other paradoxes are constitutive of truth, and the logical principles invoked in the reasoning of the paradox are constitutive of the logical concepts in question.

[1] Bacon (2017), Eklund (2017), Greenough (2017), and Scharp (2013).

- That the inconsistency in the concept of truth blocks us from doing certain things with it, like understanding our own natural languages; hence, truth should be replaced for certain purposes (e.g., doing semantics for expressively rich languages).
- A *prescriptive* theory, which suggests ascending truth and descending truth as replacements for truth.
    - An axiomatic theory of ascending truth and descending truth, ADT.
    - Xeno semantics for ADT and relative consistency proof for ADT.
    - A measurement-theoretic (i.e., metrological) account of the nature of ascending truth and descending truth in the spirit of Davidson's unified theory of rational phenomena.
    - Theories of relations between ascending and descending truth and other concepts like proof, objectivity, belief, assertion, knowledge, validity, and predication; the most important of these connections is to meaning—ascending truth and descending truth can be used as the basis for a successor to truth-conditional semantics, which I call *AD semantics*.
- A *descriptive* theory of truth that is based on the replacement concepts.
    - An AD semantic theory for truth predicates, which entails that they are assessment-sensitive because they express the inconsistent concept of truth.
    - A pragmatic theory for truth predicates, based on Roberts' scorekeeping pragmatics.
    - That, despite its defects, the concept of truth can be legitimately used in most situations; the replacement concepts are only needed when the distinction between ascending truth and descending truth isn't negligible.

Overall, *Replacing Truth* outlines a conceptual engineering project – our defective concept of truth is replaced for certain purposes with a team of concepts that can do some of the jobs we thought truth could do.

Bacon

Andrew Bacon focuses on the formal aspects of ascending truth and descending truth, which are the concepts I suggest as replacements for the concept of truth. He offers a number of objections and suggestions, which I formulate below in bold.

***Objection 1*: I advertise my view as being immune to revenge paradoxes. However, only theorists offering classifications of paradoxical sentences are subject to revenge paradoxes, and I don't provide a substantive classification of paradoxical sentences (safety isn't substantive enough). Thus, avoiding revenge paradoxes isn't a genuine feature of my view.**

It will be helpful to break this into two separate points:

(i) Simply being an inconsistency theorist doesn't tell us which instances of the T-schema to reject. And without specifying that, we can't get revenge paradoxes.

(ii) It isn't clear which sentences are safe, and any attempt to give a background picture that makes safety more substantive might reinstitute revenge paradoxes.

Reply 1: The revenge paradox phenomenon is complex, and I disagree with most of how Bacon characterizes it. Revenge paradoxes need not result from attempts to characterize paradoxicality or attempts to say which instances of the T-schema to reject. Theorists who advocate non-classical solutions typically accept all instances of the T-schema, but they are certainly subject to revenge

paradoxes. Moreover, there are plenty of paradoxes associated with truth that do not depend on the T-schema at all, and attempts to solve them also give rise to revenge paradoxes.[2] Nor is characterizing paradoxicality necessary for revenge. According to Hartry Field's terminology, a classical gap theorist claims that liar sentences are neither true nor false, but freely admits that lots of things are neither true nor false, not just paradoxical sentences.[3] Still, the classical gap theorist is subject to a familiar revenge paradox.

A nice place to start when trying to understand revenge paradoxes and the problems they cause for approaches to the liar paradox is Jc Beall's aptly named introductory discussion to the 2008 collection he edited on the topic, "A Prolegomena to Any Future Revenge." Beall's treatment is also the basis for my discussion in *Replacing Truth*. Although a decade old, it holds up well, and the contemporary literature could benefit from focusing more on it. 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 some mathematical theory like 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 for 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] See Friedman and Sheard (1987) for some examples and see Scharp (forthcoming) for discussion.

[3] Field (2008).

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.[4]

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 certain 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 problems posed by the aletheic paradoxes.

Bacon's formulation – concept X is used to characterize paradoxical sentences of L – is a special case of only the first version. I argue in *Replacing Truth* and a follow up paper, "Truth, Revenge, and Internalizability,"[5] that most of the action is in *version three* because the best contemporary theories avoid versions one and two. For example, Hartry Field's theory of truth[6] does not require an object language / metalanguage distinction at all, and so it is not subject to version one or two type revenge paradoxes. Still, Field's theory is subject to revenge paradoxes (see *Replacing Truth*, pp. 106-110).

I suggested in *Replacing Truth* that although Beall's classification is powerful, tremendously illuminating, and helps serve as a foundational guide to disputes about revenge, it does not help

[4] Beall (2008: 11–12).
[5] Scharp (2014).
[6] See Field (2008); see Scharp (2013: 108-110) for discussion.

much with understanding *why* revenge paradoxes happen. Moreover, it focuses on revenge paradoxes for approaches to the liar that advocate weakening classical logic, but there are other kinds of revenge paradoxes that affect classical approaches as well. For example, certain classical theories of truth entail that some sentences composing the theory itself are not true. So there is a great variety of revenge paradoxes for different kinds of approaches. Here is the general explanation of revenge I favor, and is essentially the one I gave in the book.

We can split the T-Schema,

(T-Schema) <p> is true if and only if p,

which we know to be inconsistent in classical logic, into two parts:

(T-In) If p, then <p> is true.

(T-Out) If <p> is true, then p.

There is tremendous pressure to accept something from the inconsistent T-schema, and T-Out is far more popular than T-In. If you accept T-In, then it is very hard to avoid saying the liar sentence,

(Liar) (Liar) is not true,

is true. And no one really thinks the liar sentence is just plain true, even those who seriously restrict the logic.[7] Some say that it is a glut, which is a sentence that is both true and false. Others that it is meaningless or indeterminate or true in some contexts but not others or even just false. From this point of view, T-Out is by far the most plausible of the two principles, and it is hands down the favorite among theorists who choose between them. Those that accept (T-In) are usually the theorists that keep the entire T-Schema by seriously restricting the logic (although Greenough is an exception – see below).

[7] Another reason to think it is more popular is that non-contradiction is held more strongly than excluded middle. Including (T-Out) in a consistent theory of truth makes the inner logic of the theory violate excluded middle (i.e., there are sentences p such that ¬T<p> and ¬T<¬p>), but including (T-In) instead makes the inner logic violate non-contradiction (i.e., there are sentence p such that T<p> and T<¬p>). The concept one gets from including (T-Out) over (T-In) is more selective (i.e., fewer things fall under it).

Once one accepts (T-Out), one runs into a problem. The instance of (T-Out) for the liar is the following:

(T-Out-Liar) If (Liar) is true, then (Liar) is not true.

But (T-Out-Liar) is equivalent to (Liar) itself. (Liar) is equivalent to '(Liar) is not true or (Liar) is not true', which is equivalent to 'if (Liar) is true, then (Liar) is not true'. Therefore, with (T-Out) and the principle that truth is preserved by logical consequence (and this is one of the most popular definitions of logical consequence), any theory of truth has to say the same thing about (Liar) as it says about an instance of (T-Out). That is, any theory that includes (T-Out) and respects truth preservation as a condition on logical consequence is forced to say the same thing about itself as it says about (Liar). And since theorists are loath to call (Liar) true, any theorist in this predicament is going to accept a theory of truth that entails that it is something other than just plain true. In the rest of philosophy, this is called *self-refutation*.

The other option is to restrict the logic enough to break the equivalence between (Liar) and the instance of (T-Out), but any theory of truth that is non-trivial only in non-classical logics faces the question of what to say about languages that have stronger logical resources. These theories are inconsistent if they apply to these languages.

Overall it looks like the most reasonable options are: (i) keeping classical logic and endorsing a self-refuting theory of truth, or (ii) saying that logic is seriously non-classical and endorsing an inconsistent theory of truth.

However, there is a third option. It involves upholding classical logic but avoiding self-refutation by denying that truth is preserved by logical consequence. And moreover, supporters of classical logic are forced to endorse this option anyway! If they keep classical logic, they should reject that logical consequence is truth-preserving. We have an independent argument due to Field that no theory of truth that is compatible with classical logic is compatible with the claim that logical

consequence is truth preservation.[8] Field also argued that even theories of truth requiring nonclassical logics are not compatible with truth preservation, but a couple of examples have been disputed.[9]

The classical logic supporter has two options: (i) say that (T-Out) is not true – either because it is not part of the right theory of truth or because it is part of the right theory of truth and that theory entails that it is not true, or (ii) say that truth is not preserved by logical consequence. Call these the *preservation & self refutation* option and the *preservation failure & self-consistency* option.

I picked the latter for two reasons. First, the point already made – classical approaches to the liar are inconsistent with the idea that truth is preserved by logical consequence. So if there is already no hope or little hope of satisfying the truth preservation option, then the no self-refutation option is a sensible choice. Second, self-refutation is just about universally taken to be a deal-breaker across philosophy. It is only out of desperation with the liar that it is being considered for truth at all. I think the no self-refutation norm is a good one, and I have tried to abide.

But maybe Bacon is right and it will turn out that the best choice is a self-refuting theory. After all, the replacement concepts are only designed to do a very specific job in truth-conditional semantics, and it might be that they can do that job even though the theories they satisfy are self-refuting.

I think that won't be the case, however. Here is why. Consider first just the case of truth, a self-refuting theory as the basis for truth-conditional semantics will give the wrong truth conditions for the sentences that make up the theory. These sentences should be counted as true in the conditions in question – those conditions where truth satisfies the theory composed of those sentences. So

[8] Field (2006).

[9] Bacon does this – see Bacon (2017: ?).

basing a truth-conditional semantic theory on a conception of truth that is self-refuting is a bad idea – it results in a truth-conditional semantic theory that makes bad predictions.

Now consider the case of ascending truth and descending truth. The theory of ascending truth and descending truth should provide the right ascending truth and descending truth conditions for all the sentences of the fragment in question. Including the sentences of the theory of ascending truth and descending truth. These sentences should be counted as descending true in the conditions in question – those conditions where descending truth and ascending truth satisfy the theory composed of those sentences. These sentences composing the theory of truth should not have the same status as a liar sentence. The theory of ascending truth and descending truth does a much better job of representing ascending truth and descending truth than a liar sentence does of representing its own aletheic value. So basing a semantic theory on a conception of descending truth or ascending truth that is self-refuting is a bad idea. It results in a theory that makes bad predictions.

Another problem with self-refuting theories involves how to evaluate them. We cannot say that they are true without thereby saying something untrue. Normally, truth is the standard by which a theory is judged, but self-refuting theories resist this norm. Instead, we would need some new term, like 'right' (which I used above) to describe the self-refuting theory that gets things right. Surely there is a distinction among self-refuting theories between those that get things wrong (normally called untrue) and those, or maybe the one, that get(s) things right (normally called true). Now we have a new term, 'right', to mark this difference that, according to these theories, truth is not cut out to mark. But now it seems like we have just reinstated the liar paradox all over again in terms of 'right'. Surely the same problems arise for the theory that says only of itself that it is not right. So it seems like endorsing a self-refuting theory merely moves the bump around under the rug without actually solving anything.

Now on to the second point in the objection. Again it is:

(ii) It isn't clear which sentences are safe and any attempt to give a background picture that makes safety more substantive might reinstitute revenge paradoxes.

I agree about being more specific with respect to safety. Safety is defined in terms of ascending truth and descending truth (to be unsafe is to be ascending true and not descending true). The theory of ascending truth and descending truth, ADT, is only a *proto-theory* in the sense that it makes no claim to being thorough. I claim only that an adequate theory of ascending truth and descending truth with have ADT as a part. (Bacon himself has shown how to extend it in various ways, but that is a topic for the next objection.)

However, being more specific about safety is *not* going to give rise to any revenge paradoxes. Revenge paradoxes come in two sorts – inconsistency problems, which plague non-classical solutions, and self-refutation problems, which plague classical solutions. They are caused by the fact that principles one wants in one's theory of truth are logically equivalent to liar sentences. I respond to these paradoxes by denying that logical equivalents have the same descending truth value and the same ascending truth value. No amount of being more specific about safety is going to change that.

I agree wholeheartedly with Bacon that a formal theory of ascending truth, descending truth, and safety isn't going to be enough. People need more to go on than a formal theory in order to get a good sense of these replacement concepts and to assess my replacement strategy. Bacon laments that the replacement concepts don't hook up with other familiar concepts like belief and assertion, but I think he is wrong about this. I spent an entire chapter (Chapter Eight) outlining connections between the replacement concepts and other familiar concepts like knowledge, assertion, belief, and inquiry, but I focused on meaning. It is explaining meaning via something like truth-conditional semantics that serves as the primary application for ascending truth and descending truth. More on this later.

Bacon proposes a couple of examples pertaining to safety that he thinks illustrate a problem. He presents the following sentences and notes that sentences 1 and 2 together form an inconsistent set:

(L*) Either L* is not true, or 'snow is white' is not true and snow is white.

(1) 'Snow is white' is true if and only if snow is white.

(2) L* is true if and only if either L* is not true, or 'snow is white' is not true and snow is white.

Bacon claims, rightly, that other approaches to the aletheic paradoxes diagnose the problem as being caused by sentence (2). However, he suggests that my approach does not specify which sentence causes the problem. Bacon is wrong about this. In an effort to help the reader understand safety, I stipulated explicitly that all grounded sentences are safe (p. 170). Hence, my approach agrees with the other approaches Bacon mentions in diagnosing the problem with sentence (2) and not sentence (1). I also provided an intended model for the language in which theory ADT is formulated, and in this model, sentence (2) is unsafe, whereas sentence (1) is safe. (*Replacing Truth*, pp. 180-183).

One final point: Bacon suggests that grounding is supposed to be a better analysis than safety because it has an intuitive picture associated with it – the picture where all truth values for sentences with truth predicates are determined ultimately by truth values for sentences without truth predicates. It is a kind of supervenience or grounding picture. It is a helpful heuristic for thinking about which sentences might be paradoxical. But we know that groundedness is a very poor analysis of paradoxicality and even a poor characterization of paradoxicality. The reason is that there will always be ungrounded sentences that are not paradoxical; sentences like 'no sentence is both true and false' are ungrounded and non-paradoxical. This sentence has no truth value in Kripke's minimal fixed point language, which gives truth values to all the grounded sentences of the language.[10] Any non-paradoxical and intuitively true sentence that is not true in Kripke's minimal fixed point language is a counterexample to explaining paradoxicality in terms of groundedness.

[10] See Kripke (1975); for commentary see Field (2008).

Kripke did show that there are other fixed point languages where some ungrounded sentences have truth values; maybe one of these is a good model for paradoxicality? No. It does not matter how many ungrounded non-paradoxical sentences get truth values in some fixed point language, there will be others that do not.

(Proof. Let M_1 and M_2 be two maximal fixed points (i.e., they cannot assign truth values to any more sentences without being inconsistent). Let sentence ϕ be 'ϕ is true' and let sentence ψ be 'ψ is false'.[11] Let ρ be '$\phi \vee \psi$' and let σ be '$\sim \phi \vee \psi$'. ψ is paradoxical, so it is indeterminate in every fixed point. ϕ is ungrounded but not paradoxical—there are no constraints on the truth value M_1 and M_2 assign it. Let M_1 assign ϕ truth and M_2 assign ϕ falsity. Assume a strong Kleene scheme for the logical connectives; thus, a disjunction is true if any disjunct is true, false if both disjuncts are false, and gappy otherwise; a conjunction is false if any conjunct is false, true if both conjuncts are true, and gappy otherwise. Thus, in M_1 ρ is true and σ is indeterminate, and in M_2 ρ is indeterminate and σ is true. Therefore, either ρ is indeterminate or σ is indeterminate in any fixed point, and neither is paradoxical because if one is indeterminate in a fixed point then the other has a truth value in it.[12] Therefore, there is no fixed point in which only paradoxical sentences are indeterminate.)

The upshot is that being ungrounded is a poor choice for analyzing paradoxicality. All paradoxical sentences are ungrounded, but many ungrounded sentences are non-paradoxical.

***Suggestion 1*: I really need to develop my objection based on self-refutation revenge paradoxes because it isn't clear exactly what the problem is or what the options are.**

[11] The existence of these sentences is insured by the Gödel Diagonal Lemma.

[12] Here I am using Kripke's definition of paradoxicality (i.e., a sentence is *paradoxical* if and only if it does not have a truth value in any fixed point); Kripke (1975: 708). This definition of paradoxicality is sufficient to make my point, but it applies only to sentences belonging to the sorts of languages Kripke considers. Thus, it does not serve as an adequate definition of paradoxicality in general.

Response 1: I agree completely. There is not much literature on the issue to work with, so any inquiry is mostly a bushwhack. Here is how I have thought about it so far.

If a theory T is self-refuting then it is inconsistent with some truths from the theory of persons that are instantiated with T. The phrase "a theory of persons" was popularized among philosophers by David Lewis, and by it he meant a theory that allows us to explain the meanings of a person's words and the contents of that person's mental states, given certain kinds of facts about that person and her environment.[13] Given certain data, a theory of persons ought to specify all the propositional attitudes of a person and a semantics for that person's language. Along with a theory of persons must also be included theory that allows the explanation of intentional actions and of the pragmatic features of the person's utterances (a kind of intentional action).[14] Two cornerstones of any reasonable theory of persons is that to believe something is to take it to be true and to assert something is to present it as true. Thus, a self-refuting theory – one that entails that it is not true – is inconsistent with our best theory of persons. It cannot handle a situation in which John believes or asserting the theory, John is not irrational or deceptive in doing so, and the theory is self-refuting.[15]

Now the question is, what happens to this story when we replace the concept of truth with ascending truth and descending truth? Before we had only two truth values to use in characterizing assertion, but now we have four. And the four interact in complex ways. Nothing is descending true and not ascending true, but some things are ascending true and not descending true. The other possible combinations are descending true and ascending true, and not descending true and not ascending true. That is three out of four combinations as possibilities. Descending true (and ascending true) is the best (i.e., what one ideally aims for in belief and assertion, but is not always

[13] See Lewis (1974).

[14] This is a core commitment of Davidson; see Davidson (1986; 1990)

[15] For classic treatments of self-refutation, see Fitch (1946) and Mackie (1964).

attainable), ascending true but not descending true – also known as unsafe – is intermediate, and not ascending true (and not descending true) is worst.

One might hope to endorse theories that are entirely descending true – their axioms are descending true and descending truth is inherited by all the consequences of their axioms. However, given that we have three statuses to work with, there is another option – a theory whose axioms are descending true and none of whose theorems are not ascending true. A theory like this might have some descending true theorems, but some of its theorems might have the intermediate status – ascending true but not descending true (i.e., unsafe). There are two conditions that have to be met before a theory like this would be acceptable – it would have to be that some unsafe sentences are rationally assertible and believable, and it would have to be that one could show that simply deriving consequences of the axioms of the theory will never take one from something that is descending true to something that is not ascending true. And that is exactly what we have in the case of ADT. Therefore, once one moves from two values (true and not true) to three (descending true, unsafe, not ascending true), one has more options for acceptable theories. As usual, if it is consistent with a theory that all its theorems have the best value, then the theory is not self-refuting. However, failing to meet this condition need not result in a theory that is self-refuting. As long as it is consistent with a theory that none of its theorems have the bottom status and all of its theorems are rationally believable and assertible, then that theory is not self-refuting. And this is exactly the case with ADT.

Before moving on, I want to register how odd it is that Bacon would be so dismissive of the self-refutation problem given his own view on revenge paradoxes. He essentially agrees with me that self-refutation is the kind of revenge paradox facing those who opt for classical logic. Recall, he thinks that revenge happens when someone tries to characterize paradoxicality, and recall that this is one kind of revenge paradox, but only one kind.[16] He proves that any theory that contains a

[16] See Bacon (2016).

sufficient condition for the sentential restricted T-schema entails that some of its own theorems don't meet the condition. For example, if we restrict the T-schema so that it only holds for sentences that are unparadoxical, then the theory consisting of the claim 'if a sentence p is unparadoxical, then p is true iff p' entials that some of its own theorems are paradoxical. So there is no way to restrict the T-schema to the unparadoxical sentences without this claim itself having paradoxical consequences. Notice that this theorem Bacon proves shows that certain classical theories of truth are self-refuting in that some of their own theorems have the same status as the liar. So there is no unparadoxical sufficient condition for the T-schema to hold. Bacon afterward says that we could turn this self-refutation problem into a genuine inconsistency if we add some additional premises like every theorem of theory is provably unparadoxical.[17] Of course, the classical theorist is free to reject these additional premises. However, Bacon advertised his objection as applying to all classical approaches, so the real objection he presents is the one based on self-refutation, not inconsistency. Hence, Bacon's own diagnosis, as limited as it is, focuses on self-refutation in classical settings just as mine does.

***Objection 2*: The formal theory of ascending and descending truth, called ADT, is very weak compared to more familiar formal theories like KF, and I don't provide an underlying picture as a guide to strengthening ADT. There are many principles we could add to ADT, but there is no guidance about how to choose between them.**

Reply 2: As part of developing this objection, Bacon did something very nice. He found a new model for ADT. And not just any model, but a *minimal* model. Bacon pokes fun at the rather baroque relative consistency proof I gave in *Replacing Truth*, and rightly so; his is far more elegant. There are

[17] Bacon (2016: ?).

three reasons I gave the proof that occurs in *Replacing Truth*: (i) I do not really have the mathematical chops to come up with elegant proofs of novel results in contemporary mathematics (however, I'm glad I have colleagues like Bacon who do), (ii) I wanted to provide not just a model but a semantic theory for a language with ascending truth and descending truth that brings out the modal/aletheic analogy (ascending truth and descending truth can be expressed by *dual* logical expressions, just like possibility and necessity), and (iii) I wanted the semantic theory to illuminate how the aletheic values of sentences in the language with occurrences of 'ascending true' and 'descending true' might depend on the aletheic values of the rest of the sentences in the language. I do make substantive appeals to the results of the semantic theory used in my relative consistency proof in several places in the book (notably in justifying my views on the ascending truth teller and the descending truth teller), but I tried to stick to what is *provable* from ADT rather than what is true in the intended model. Bacon's minimal model answers to a different set of concerns and in particular doesn't address points (ii) and (iii). Anyway, finding this minimal model for ADT is going above and beyond the call of duty, and I appreciate it.

With his minimal model, Bacon argues that ADT is not a strong axiomatic theory and it could be extended by adding more principles to it to make it stronger. But it isn't clear how to decide which principles to add. And Bacon complains that I don't give any guidance on this front. I want to emphasize that I do not and did not take ADT to be *the* theory of ascending truth and descending truth. Rather I took it to be a proto-theory in the sense that any adequate theory should have ADT as a subtheory. Bacon knows this; I am just repeating it for the reader.

So what should guide the search for principles to add to ADT? My answer is easy – whatever is needed for ascending truth and descending truth to serve effectively in something like a truth-conditional semantics. I explained how to construct semantic theories that are like truth-conditional ones but instead utilize ascending truth and descending truth. These semantic theories attribute

ascending truth conditions and descending truth conditions to sentences of a language. I called this particular proposal *AD semantics*. It promises to deliver the same results as truth-conditional semantics on the unproblematic cases and it delivers consistent and reasonable verdicts on the problem cases – like liar sentences – on which truth-conditional semantics falters. Overall, the primary application for ascending truth and descending truth is semantics. That is the motivation to which I appeal over and over. Truth plays an explanatory role in natural language semantics, but it cannot perform this role as well as one might want because of the inconsistency in our concept of truth. Swapping out truth for ascending truth and descending truth frees the resulting semantic theory from these limitations.

What about the principles Bacon shows could be consistently added to ADT? Bacon claims ADT could have conjunction introduction and elimination for descending truth, and that it could have double negation introduction and elimination for descending truth. Double negation elimination and conjunction elimination are already part of the theory. As for adding the other two – one question is what this choice rules out. It might be that there are other principles that make descending truth and ascending truth work better in natural language semantics but are incompatible with these additional principles Bacon identifies. Right now, we do not understand the logical relationships among all the principles we might add to ADT well enough to say definitely that such and such principles should be added but so and so other principles should not be added. Until we have a complete catalogue of all the minimally inconsistent subsets and the maximally consistent subsets of potential principles, we won't be able to make these decisions.

***Suggestion 2*: Rather than ADT as a formal theory of ascending truth and descending truth, it would be better to use KF to define them.**

Response 2: After banging my head against the liar for several years, I came to the conclusion that I want an approach with "nothing ad hoc". I do not want to apologize for an unreasonable logic or apologize for pretending that linguistics isn't a science, or apologize for endorsing a self-refuting theory of truth, or apologize for pretending that exclusion negation is just a fiction. One result of this decision is that I avoid the "well, that's just the lesson of the liar!" locution, which is the phrase truth theorists enlist to apologize for ad hoc elements of their approaches. In other areas of philosophy, "biting the bullet" is much more common, and Wilfrid Sellars used the memorable phrase "grab the thistle," which I prefer; in Scotland, thistles are much more prominent than bullets.

The fact is that the liar has no lesson. It is one symptom of a colossal inconsistency in our concept of truth. It does not tell us anything deep about the nature of logic or assertion or reason or meaning or language or self-reference. Anyone who disagrees with me on this point would be, from my perspective, like a person who finds the face of Jesus in the cracks on the underside of a bridge. The cracks tell you that the bridge might not work very well in certain situations, but that is about it.

Just to be clear, I don't think I have reached the "nothing ad hoc" goal, but it still serves as a regulative ideal. And adopting a self-refuting theory like KF would take me in the opposite direction. Instead, I have endorsed what I think are the most *independently* supported views that, together, provide the most plausible approach to all the paradoxes affecting truth. There are concepts. Some of these concepts are inconsistent. Some inconsistent concepts should be replaced for certain purposes when these inconsistencies prevent us from doing what we want to do. The replacement concepts should do whatever it is we want to do better, and a certain semantic theory is appropriate for words that express inconsistent concepts. Inconsistent or self-refuting theories of these concepts are unacceptable.

These considerations tell against self-refuting theories in general, but in the case of KF there is an additional reason – KF is based on the idea that paradoxicality is ungroundedness. That is, KF is

an axiomatization of the truth predicate defined for Kripke's Strong Kleene minimal fixed point. This is the account of truth that results from taking all and only ungrounded sentences to be paradoxical. However, we have already seen that this guiding principle is false – there are plenty of ungrounded sentences that are not paradoxical. Therefore, regardless of one's views on self-refuting theories, choosing KF as a basis for one of the replacement concepts is a bad idea.

***Objection 3*: My way of handling Montague's paradox is to deny that descending truth or ascending truth is preserved under logical consequence. I make this choice so that I can have all the axioms of ADT turn out to be descending true (and thus ascending true). However, one result of this choice is that there will always be logically equivalent ways of axiomatizing ADT such that the resulting theory proves some of its axioms are not descending true. The problems are:**

(i) My choice for solving Montague's paradox forces an unnatural distinction between inference rules and axioms.

(ii) We can replace all axioms with inference rules. Why doesn't this avoid self-refutation?

Reply: I love this objection. It brings out a genuine conflict between the way logicians think about proof theory and one of the major logical results relevant to approaches to the aletheic paradoxes.

Once one distinguishes ascending truth and descending truth based on the two directions of the T-schema, one immediately runs into a problem for descending truth. Montague's theorem entails that four principles anyone would really really want for descending truth are inconsistent. They are:

(i) Logical tautologies are descending true

(ii) Logical consequence preserves descending truth

(iii) T-Out (i.e., if <p> is descending true, then p)

(iv) T-Out is descending true (i.e., 'if <p> is descending true, then p' is descending true)

Rejecting any one of these is going to be devastating and force a rethinking of the concepts involved. Obviously, T-Out defines descending truth, so we can't reject it. So we either say our core principle of descending truth isn't descending true, or we say logical truths aren't descending true, or we say that valid logical reasoning can take one from descending truths to conclusions that aren't descending true. There is no reason to think that we have to say logical truths aren't descending true or that our theory of descending truth isn't true. But there is good independent reason to think already that we cannot have logical consequence preserve descending truth. So even if we picked one of the other two options, we would be stuck with this one anyway. So it is the obvious choice. The funny thing is that few make this choice. Almost all the work on formal theories of truth concentrates on theories that, interpreted as theories of descending truth, deny that (T-Out) is descending true.

Overall, it is Montague's theorem that forces the distinction between axioms and inference rules, but it could be reformulated in terms of inference rules with no premises and inference rules with premises. And then we would again have the choice: do we start with something strong but deny that inference preserves it, or do we let inference preserve the strongest status even though what we start with now doesn't have it? Does this way of framing things depend on assumptions about the nature of proofs and arguments? Yes, absolutely, and when considering other types of proof systems, the question is: how do you formulate Montague's theorem with its distinction between assuming that the axioms of the theory are strong and assuming that the inference rules preserving what is strong?

Matti Eklund focuses on the idea that truth is an inconsistent concept and the recommendation that it be replaced for certain purposes. He offers a number of objections and makes several important points, which I formulate below in bold.

Objection 1: **I argue that truth is an inconsistent concept, and I appeal to constitutive principles for concepts in order to explain what an inconsistent concept is. But my account of constitutive principles is unclear and unmotivated in various ways. In particular, there are the following issues.**

(i) What's the difference between constitutivity and obviousness? For example, 'there are trees' and 'many trees have leaves' don't seem to be constitutive for 'tree', but denying them does seem to be a reason to think that someone doesn't mean what I mean by 'tree'. Why don't I just appeal to charity to distinguish between cases of different meanings vs. different beliefs?

(ii) I claim that anyone who possesses a concept is entitled to that concept's constitutive principles. But it isn't clear what grounds this entitlement. That is, by virtue of what is a given thinker entitled to believe the principles in question?

(iii) I suggest both that possessing a concept (or being competent with a word) are a matter of accepting enough constitutive principles, but that does not seem consistent with my view that the relationship between concept possession and constitutive principles is entitlement, not acceptance. If competence is a matter of what principles one is entitled to accept, how do questions about what one actually accepts enter in?

(iv) I seem to think that constitutivity comes in degrees, but if that is the case, then isn't my view just like Quine's view, which does not appeal to anything like constitutivity?

(v) The evidence I give for constitutivity of T-In and T-Out for the concept of truth focuses on the grounds for believing these principles, but officially, I'm supposed to be focused on entitlement to them. Why?

I think it makes sense to present the theory of constitutive principles from scratch because the discussion was obscure at best in *Replacing Truth*. The reason is that I did not fully appreciate that my theory of constitutive principles is not primarily about conceptual competence. And that marks a serious break from what others have been saying about the topic.

I am going to lay out a measurement system for constitutivity. See *Replacing Truth* for measurement systems for length and truth. If you don't care much about philosophical methodology, then this won't matter for you. Essentially, I am going to present a particular kind of formal theory and then an interpretation for that theory. The result should have plenty of empirical consequences that are testable (and testing is currently underway).

We need two languages. One for us to use for our theory (L) and one spoken by the people our theory is about (N). Language L is first order. Language N is a natural language.

We will need some definitions and we make use of set theory throughout. Let L contain all the individual constants and predicates necessary to define the following sets.

S is the set of names of sentences of N.

W is the set of names of words of N.

P is the set of people.

D is the set of degrees [0,1]

C is a subset sentences of L such that for all c in C, c= 'Constit(S_0, w_0, p_0, d_0)' where s_0, w_0, p_0, and d_0, are names of a sentence, a word, a person, and a degree, respectively.

C is the set of constitutivity claims.

This is the basis for our relational system.

The physical structure is a human linguistic community, and in particular, conversations among humans using natural language N. For these conversations, there are two major phases or modes they can be in. The first is regular (or transparent) mode, where the conversation goes along exactly as you would expect a well-behaved conversation to progress. The participants take turns making claims about the topic of the conversation (in formal pragmatic theories, this is usually called the Question Under Discussion).[18] But the second mode for a conversation is *meaning reflection*, where one or more participants questions whether everyone means the same thing by one or more of the words or sentences involved in the conversation. Participants treat utterances of certain sentences as default reasons to go into meaning reflection. These are the *prevented* sentences. For each case where a participant reacts to a prevented sentence or is disposed to react to a prevented sentence, we can identify a constitutive sentence. That is, we can identify a S_0 (i.e., the negation of the prevented sentence uttered), a w_0 (i.e., the word of the sentence in question whose meaning is questioned), a p_0 (i.e., the person initiating meaning reflection), and a d_0 (i.e., the strength of the reaction—how hard will it be to put the claim on the record?).

The system captures three important features:

1. constitutivity is relative to a word,

2. constitutivity is relative to a person,

3. constitutivity is gradable – it comes in degrees.

[18] See Roberts (1996, 2004, 2010) for a detailed model of conversational dynamics, but she says nothing at all about meaning reflection.

A sentence might be constitutive for one of the words contained in it but not for others. A sentence might be constitutive for one person but not for another. Sentences differ on how constitutive they are. For example, if you live in the Amazon jungle and say there is a penguin outside, then that might reasonably initiate meaning reflection, but it would be relatively easy to overcome. Asserting a contradiction, however, would be far more difficult to overcome. The strength of the constitutive principle is exactly how much harder it is to justify putting the claim on the record (the epistemic incline) than it would be for an ordinary permitted claim.

Now we need constitutivity thresholds. Intuitively, they are just degrees, members of D. But they mark off a boundary above which the sentence really is constitutive for that word for that person, and below which it isn't. The thresholds allow us to ignore the degrees and only look the sentences that are held with sufficient strength. Let C_d be the set of sentences of the form Constit (s, w, p, d_0) where d_0 is above d. C_d is the set of *constitutive sentences* for threshold d. Now, for each word w of L, collect all the members C_d that have w as a constituent. Call this $C_d(w)$ – the principles constitutive *for w* (relative to threshold d).

So far, all we have is a bunch of information about which people take which sentences to be constitutive of which words to which degree. How do we get from here to conclusions about which principles *really are* constitutive for which words? And how do we get *concepts* into the picture?

$C_d(w)$ is going to be the basis for the concept expressed by w. How? In order to define concepts in terms of sets of constitutive principles, we use an abstraction principle. For example, where x and y are lines, the direction of x = the direction of y iff x and y are parallel. This well-worn example features an abstraction principle for defining directions. The right-hand-side of the principle must be an equivalence relation, and being parallel fits the bill (every line is parallel to itself, if x is parallel to y, then y is parallel to x, and if x is parallel to y and y is parallel to z, then x is parallel to z).

We can utilize the same sort of principle to define concepts in terms of constitutive principles for words. If w_1 and w_2 are words, then,

(Concept) The concept expressed by w_1=the concept expressed by w_2 iff the set of constitutive principles for w_1 is identical to the set of constitutive principles for w_2 – that is, iff $C_d(w_1)=C_d(w_2)$.

We treat this as a definition of 'concept'. All you need for an abstraction principle is an equivalence relation on the right hand side, but we have an equality, which is stronger.

Now we have concepts and we know exactly how we know things about them – by knowing things about their constitutive principles. This is the key to individuating concepts. Different constitutive principles? Different concept.

All this still only gets us individual conceptual schemes, not shared conceptual schemes. The way in which language is social is a vexed problem, but we can identify three senses. First, there is the fact that each speaker might disagree with the rest about which sentences are constitutive. Nevertheless, each speaker assumes agreement until presented with evidence to the contrary. So although there are in the theory only idiolects, the practice of natural language users insures that divergences are small enough to not matter very often. The evidence for this claim is that meaning-reflection is relatively rare compared to normal conversation. The second source of sociality is that each person uses the sentences and words of her own idiolect to keep track of the concepts expressed by other people's words in their own idiolects. If there is too much divergence, then communication is impeded. Given the necessity and ubiquity of quick and effective communication in human societies, the pressure to keep one's conceptual scheme close to those of others is intense. Think of the social conceptual scheme like a flock of birds. It isn't just that the individual schemes (the birds) are close to each other, it is that each individual is constantly altering course and changing position in relation to the other members of the flock. It is this dynamic feedback-regulated process

that constitutes the flock. The same goes for a social conceptual scheme. The third source of sociality is that the semantic theory for N is based on a body of evidence, and that evidence comes from a huge number of individual people. Typically, linguists chalk up divergences to idiolect differences rather than to ambiguity in a shared word. The semantic theory for the natural language also individuates semantic units for us, and we use these as input for our theory of concepts and constitutive principles. This practice insures that the idiolects will, by and large, share their semantic features. If we think of meanings as given by semantic theories (extensions in contexts or intensions), then meanings and concepts are distinct.

Another crucial element in keeping individual conceptual schemes very similar to one another is the role of dictionaries in meaning reflection. Single language dictionaries serve as authorities in disputes about the meanings of words. They can be challenged of course and dictionaries differ from one another to some extent, but showing that someone has contradicted the dictionary definition of a word carries significant weight in everyday conversations about whether two people mean the same thing by some word. Emphasizing the importance of meaning reflection also highlights the crucial role dictionaries play in our linguistic practice.

The connection between semantic theory and theory of concepts is pretty tight. Because concepts are determined by constitutive principles and constitutive principles are among the most strongly held of our beliefs, they will show up quite strongly in the data for semantic theories, which are based on native speakers' judgments about acceptability, unacceptability, entailments, and other features of natural language. Other things being equal, semantic theory will respect constitutive principles, not because semanticists set out to do so but because of the connection between constitutive principles and acceptability judgments. Of course, in the case of an inconsistent concept, the semantic theory cannot respect the constitutive principles unless the semantic theory is itself inconsistent. As a result, semantic theories for words that express inconsistent concepts can't

possibly do justice to the data at face value. Instead, we should take these words to display a certain kind of subjectivity that gets modeled as semantic relativism – the idea that a word might express the same content in every context of utterance but still have a variable extension. A sentence might express the same proposition in every context but still differ in truth value. I have suggested assessment-sensitivity for 'true', which is a particular kind of semantic relativism that employs contexts of assessment in addition to contexts of utterance. Words that express inconsistent concepts have a kind of judgment-dependence where the word's semantic features change depending on how the word is treated. Assessing them in one way gives one answer but assessing them in a different way gives another. In many cases, these words that express inconsistent concepts would not normally be good candidates for semantic relativism. It is only because they express inconsistent concepts (or, in other words, the native speaker judgments that serve as data for a semantic theory are inconsistent), that semantic relativism is appropriate in these cases. So claims about inconsistent concepts have a share in the explanatory power of semantic theories themselves.

It is constitutive sentences for words that we take to be basic and indicated by meaning-reflection. The constitutive principles are always in the background but only come out to play in meaning-reflection. Like a plucky primatologist crouched in the Madagascar jungle at night in the hopes of seeing an Aye-Aye, we have to look at cases of meaning-reflection to catch constitutive principles out in the open.

In meaning reflection, the epistemic deck is stacked in various ways – in favor of some claims and against others. That is, the work you have to do to get a claim on the record differs depending on the claim. The closer it is to a constitutive principle, the easier it is to get on the record. The closer it is to the negation of a constitutive principle, the harder it is to get on the record. That is about it. Constitutive principles skew the space of reasons in this way. They have little to do with belief, nothing to do with truth, and quite a bit to do with justification. Constitutive sentences are

justified by virtue of their meanings rather than *true* by virtue of their meanings. And their negations are unjustified by virtue of their meanings. The brand of justification involved is entitlement – default and defeasible.

As far as I can tell, there is no term for the dual of entitlement (something whose negation is entitled), but *permission* is a fine term. If p is entitled, then it is not the case that not p is permitted. And the antonym for entitlement could be *prevention*. If p is entitled, then not p is prevented. Prevention and permission are default and defeasible as well. A single constitutive principle for a word gives rise to a constellation of entitlement, prevention, and permission. It is prevention and permission that guide the dance between normal conversation and meaning reflection. As long as the utterances in the conversation are permitted, the conversation goes along normally as described in numerous works on formal pragmatics. However, if an utterance of a prevented sentence occurs, then the conversation changes into meaning-reflection. In meaning reflection, the prevented sentence is facing an uphill battle for legitimacy. The standards are harsh, but they are, of course, attainable. If justified in the right way – for example, Williamson's Aristotelian conspiracy theorist – then the other participants in the conversation put the claim on the record and attribute the belief in question to the participant and take that participant's word to mean the same thing as everyone else's word.[19] The constitutive principles have little to do with assessing the competence of various speakers with various words; instead, they tilt the epistemic table one way or another – for some claims and against others. Just as it is difficult to get a *prevented* claim *on* the record, it is difficult to keep an *entitled* claim *off* the record.

It is the warpings and tiltings of epistemic standards for various claims that is the contribution of constitutive principles. They act like the rails of a railroad track – guiding the direction of the conversation without dictating it. The record can violate constitutive principles just as the train can

[19] The example is discussed in Williamson (2008: 86-92).

violate the track. But the rails make it more difficult for the train to leave the track and easier for it to stay on the track. Likewise, constitutive principles make it more difficult to take up certain positions and easier to take up other positions in a conversation. Constitutivity guides the conversation away from certain positions (prevented ones) and toward others (permitted ones). Any position is accessible in principle – even the position that not all vixens are vixens. You just have to work harder to get there.

Note well: it is *not* that you have to work harder to *justify* your claim if pressed on it. No, you have to work harder *to even make* this claim—to be credited with legitimately making this claim. You have to convince everyone that you do mean the same thing as everyone else and that you have a good reason for violating what you acknowledge is taken to be a constitutive principle. Your variance from the norm needs a good reason. Without that, you can't get the proposition in question on the record as one of your beliefs. If you don't do enough to justify your divergence from the norm, then your conversational participants will refuse to acknowledge your assertion of that sentence as an assertion of something with the content in question. Instead, they will treat you as if you mean something *else* by the word(s) in question and put *that* on the record (or put nothing at all). So you have to work hard to earn the right to mean something prevented.

Overall, the constitutive principles for some concept partition the space of sentences and the space of propositions into entitled, permitted, and prevented. And so it predisposes participants in the conversation to treat sentences and propositions in these ways.

That is the basic theory of constitutivity. Everything is based on meaning reflection. Constitutive principles have little or nothing to do with competence or concept possession. If you were in a conversation where someone asserted 'not all vixens are vixens', you would probably wonder whether you mean the same thing by some of those words, but you would not infer that this person does not possess the concept of vixen. After all, if your interlocutor means something else by

'vixen', then this sentence isn't about vixens. You can't infer anything about the concepts possessed by this person. Instead, constitutive principles only explain why people begin to question whether people are using the same words with the same meanings. Violating a constitutive principle is a reason for thinking that someone doesn't mean what you do by one of the words in the conversation.

However, we can add a theory of concept possession to our theory of constitutive principles and concepts in the following way. The idea is that in order to possess a concept, a person has to accept some of its constitutive principles. Any particular constitutive principle can be rejected but the concept possessor has to accept *some* of them. This is a compelling idea for some people, and I will just take it as an assumption to illustrate the theory of competence/possession. In the end, I think it is probably false.

Every person associates with each word of the language a competence threshold. For our purposes we can take these to be degrees (i.e., numbers between 0 and 1). And we stipulate that a person possesses a concept C if and only if the average constitutivity degree of the constitutive principles for C that the person accepts is above the possession threshold. That is, in order to possess a concept, the person has to accept enough principles constitutive of that concept that are of sufficient importance. Formally we introduce the theory by defining a two-place possession predicate, which holds of people and concepts. To formulate the definition, it helps to have a 'constitutive strength' function that simply outputs the degree from the constitutivity claim (Constit(…)). Use |c| for c in C. That is, if c is a constitutive principle sentence (statement of constitutivity?) for some word (c says that some sentence is constitutive for some word for some person to some degree), then |c| is the degree of c.

(Possession) Possess(p, c) iff there exist c in C_w such that for all s in the c_i, p accepts s from c and the ratio of the sum of $|c_i|$ for the c_is to the total number of cis accepted accepted is greater than or equal to the possession threshold for c.

I want to reiterate that the theory of concept possession is optional. It is tacked on to the theory of constitutive principles as an *application*. Constitutive principles, as I have outlined them here, are tied to meaning-reflection. And meaning reflection is distinct from concept possession. I can judge that you don't mean what I mean by 'boot' without having any idea what you might mean by it and without having any idea whether you possess the concept of sturdy footwear. Making a judgment that you do mean the same thing by what I mean does carry with it the judgment that you are competent with the word in question and you possess the concept in question. It is generally assumed that if someone, even a stranger, begins speaking to you in what sounds like a language you know, then you will by and large just assume that she means whatever you mean by the words she chooses, and you are more than willing to unconsciously grant her competence with these words and possession of the associated concepts. So far, I think this is exactly right. According to the theory of possession presented here, it follows that we just assume that people who seem to make sense accept a decent number of the principles we take to be constitutive of our concepts. For every word w uttered by a person p, the audience just presupposes that Possess(p, c). Only if meaning-reflection is initiated does the person question these assumptions.

The fact that I distinguish a theory of constitutive principles for concepts from a theory of concept possession represents a major shift from conventional thinking about the matter, which takes competence and concept possession to be *the* primary basis for thinking about constitutive principles. In *Replacing Truth*, I formulated the theory primarily in terms of concept possession/competence. I said that someone who possesses a concept is quasi-entitled to that

concept's constitutive principles. In doing so, I was following the tradition of inconsistency theorists, but I now see that this is a mistake.

I still think that people are quasi-entitled to the constitutive principles of the concepts they possess, but this isn't a *definition* of constitutivity. It is a consequence of the theory of meaning reflection outlined above.

A theory of constitutive principles and concepts like the one presented has obvious applications across philosophy – from assessing the claim that same sex marriage requires a redefinition of the word, to evaluating Hume's principle in the Neo-Logicist reduction of mathematics to logic. I use it primarily to make sense of inconsistent concepts, and the claim that truth is an inconsistent concept plays a central role in the book. Eklund questions this, and so it makes sense to go through the role it plays.

In the case of truth, I argued that ⟨T-In⟩ and ⟨T-Out⟩ are constitutive. There are a great many other constitutive principles for truth as well, but I say little about them in the book. They are inconsistent in classical logic and in the logically motivated non-classical logics (intuitionistic logic and the logic of relevant implication). Some of them are consistent in some of the more radical non-classical logics.

What is the problem with our concept of truth being inconsistent? We can use our nascent theory of constitutive principles to illuminate the situation. One problem is that we can just assume ⟨T-In⟩ and ⟨T-Out⟩ freely – any possessor of truth is quasi-entitled to them. But then one can use them to reason to a contradiction. And contradictions are prevented. Pretty strongly prevented as well – about as strong as anything. And one shouldn't be able to reason from something entitled or permitted to something prevented. So having T-In and T-Out causes an anomaly in epistemic space. Epistemic space can be thought of as a ranking of different positions in terms of how strongly they are justified or not. One has to have a very strong justification for a prevented position in order to

get it on the record. The inconsistent concept opens up a wormhole between the permitted, which don't require strong reasons to get on the record, and the prevented, which do require strong reasons. Reasoners are loath to use the wormhole but the fact that it is there is unsettling and decreases the overall integrity of the epistemic space. If there aren't many of these wormholes, then reasoners can just avoid them, but if they proliferate, then the whole issue of legitimacy in epistemic space becomes suspect.

Now (finally!) we can address Eklund's objections/questions.

(i) What's the difference between constitutivity and obviousness? Why don't I just appeal to charity to distinguish between cases of different meanings vs. different beliefs?

Now we can see the difference between constitutivity and obviousness. Something is obvious for a person if it is easy for that person to see that it is true. And many obvious claims are going to end up being constitutive. But being constitutive is tied to a very specific phenomenon – meaning reflection. Something is constitutive for a person iff when it is denied by someone in a conversation with that person, that person initiates meaning reflection. That is, the person starts to wonder whether everyone in the conversation means the same thing by the words being used. Obviousness has no such connection with meaning reflection. If I am in a conversation with someone who denies something I take to be obvious, that does not necessarily make me wonder whether we mean something different by one of the words in question. For a biologist, evolutionary theory might be obvious, but the biologist doesn't wonder whether any evangelical Christian who denies evolution means something else by 'human' or 'dog'. Likewise, it isn't at all obvious that Earth is roughly a sphere, but it is constitutive of 'Earth' that Earth is roughly a sphere (when one hears about the Flat Earth Society, one it is usually an interpretative red flag).

Moreover, obviousness seems to be factive. It would be odd to say that something was obvious and false. But constitutivity is explicitly *not* factive. Constitutive principles can turn out to be false, and in the case of truth or any other inconsistent concept, that is exactly what happens.

Finally, I'm not sure how to appeal to charity to distinguish belief change from meaning change because it is difficult to discern the principles governing appeals to charity. What exactly does charity predict in a given situation? It isn't clear. And if it isn't clear what charity predicts in any case, then it isn't clear what charity explains. On the other hand, the theory of constitutivity presented here has a rich and complex constellation of predictions and so has some explanatory power. It might be wrong, but at least it offers a substantive explanation.

(ii) By virtue of what is a given thinker entitled to believe the principles in question?

Grounding for epistemic properties like justification or entitlement is not a topic that has received much attention in epistemology, and I don't know what conclusions this nascent research program will reach. Regardless of how it goes, entitlements for constitutive principles do not seem to be especially problematic, so I am happy to be neutral about this topic.

For what it is worth, I am tempted by the theory that the quasi-entitlements, quasi-permissions, and quasi-prohibitions are instituted by our own attitudes. Ultimately, we are entitled to 'bachelors are unmarried' and prohibited from 'bachelors are married' because that is what we take each other to be.[20] It might not be obvious that we are taking these attitudes toward each other, but we are.

[20] See Brandom (1994) for an example.

(iii) If competence is a matter of what principles one is entitled to accept, how do questions about what one actually accepts enter in?

I have said little about competence. I have talked about, which constitutivity is based on meaning reflection. One can then use the theory of constitutivity to explain concept possession, but this is an optional application rather than the heart of the theory. Many people, myself included, are tempted to see competence with a word and possession of the associated concept as essentially the same thing. But I am open to deferring to linguists about word competence. Either way, it should be clear that constitutivity is not primarily tied to competence.

(iv) I seem to think that constitutivity comes in degrees, but if that is the case, then isn't my view just like Quine's view, which does not appeal to anything like constitutivity?

I hope it is obvious how the view differs from Quine's now. Quine appealed only to entrenchment, which does come in degrees, but constitutivity is far more subtle and theoretically fruitful than entrenchment. The fact that constitutivity comes in degrees makes the theory of constitutivity more complex, but it does not make it in any way Quinean.

(v) The evidence I give for constitutivity of T-In and T-Out for the concept of truth focuses on the grounds for believing these principles, but officially, I'm supposed to be focused on entitlement to them. Why?

If you look back at *Replacing Truth* in the chapter in which I argue that (T-In) and (T-Out) are constitutive of truth, you will see that I appealed to meaning reflection. I argued that denying either of these principles is a pro tanto reason to think that that person does not mean what you mean by 'true' (see pp. 62). Moreover, I argued that we also take the expressive role of the truth predicate to be constitutive of the concept of truth and without (T-In) and (T-Out) it would not have this expressive role. Denying (T-In) or (T-Out) constitutes an interpretive red flag; it institutes meaning

reflection. It is from this claim that that I infer they are quasi-entitled. Part of the theory of constitutive principles is that constitutive principles are quasi-entitled and their negations are quasi-prohibited. The three-part distinction between entitled/permitted/prohibited is part of a theory of conversational development (often studied under the heading of formal pragmatics). It could be paired with Stalnaker's theory of normal conversational development (which explains conversational development in terms of the context set) or Lewis's theory of normal conversational development (which explains development in terms of scorekeeping), but I prefer Craige Roberts' theory, which is a sophisticated scorekeeping theory.[21] She pioneered the use of the Question Under Discussion (QUD) in pragmatics and semantics. The QUD is what the conversation is about, and it is determined by the individual goals of each of the members of the conversation. In terms of Roberts' pragmatic theory, meaning reflection occurs when one of the participants puts something like "figuring out whether so and so means what I mean by such and such word" on their own list of goals for the conversation. Having this question added to the Questions Under Discussion is a more advanced stage of meaning-reflection. In the early stage, one or more participants question whether everyone means the same thing by a word, and in the advanced stage, the conversation becomes about whether everyone means the same thing by a word.

***Objection 2*: Consider theorist who agrees with me about what is true, but not about what is constitutive – say they have no views on constitutivity. Going through problems with other solutions to the liar won't show what is wrong with this theorist. And this theorist avoids revenge too. So what does constitutivity add?**

This is a great exercise. Let us look at the theorist who agrees with me about everything except my

[21] See Stalnaker (1970), Lewis (1979) and Roberts (1996); see Scharp (2013: 49-53) for discussion.

claims about constitutive principles and inconsistent concepts.

What about the replacement strategy? This theorist could, of course, suggest that we add two new words, 'ascending true' and 'descending true', to our language, and could offer ADT as a theory of them. But this entire prescriptive project would have nothing to do with the liar paradox or other paradoxes affecting truth. It would simply result in some new terms that behave sort of like 'true' and don't give rise to paradoxes similar to the liar. All by itself, this tells us nothing about the liar paradox, which is a paradox that involves 'true'. Contrast this with my proposal, which takes replacement to be exactly what is in order as part of an approach to the liar paradox. My claims about inconsistent concepts tie the prescriptive proposal to the problem of the liar.

Now what about the descriptive project – a semantic theory for 'true' that is based on the idea that words expressing inconsistent concepts have relativist semantic features. In particular, I offer an assessment-sensitivity semantic theory for 'true'. Could Eklund's imagined theorist advocate such a thing? Yes, absolutely. But the problem is that, if we stick to the usual kinds of evidence from linguistics, then we would never think that an assessment-sensitivity semantic theory would be right for 'true'. The standard test for semantic relativism is faultless disagreement – where two people are asserting contradictory claims but both seem right (or at least neither seems to be wrong). This sort of phenomenon is indicative of subjectivity. The word 'true' fails this test because it doesn't display faultless disagreement.

"That's true!"

"No it is not true!"

"Well, maybe we're both right."

No. No matter how much certain people might wish there were alternative facts, this is just a rhetorical ploy. Truth is exactly where one would *not* find faultless disagreement. So an assessment-sensitivity semantic theory for 'true' makes horrible predictions when based solely on the usual

linguistic data. No one should accept such a view. Moreover, assessment-sensitivity views are most often based on retraction. If someone says 'that is true' and then the standards for truth change, would the person reflecting back retract the claim? Notice how incoherent this question is. Standards for truth change? What would that be? There is no reason think that anything like that would ever happen. So the normal ways of justifying assessment-sensitivity seem completely ridiculous in the case of 'true'. Therefore, Eklund's imaginary theorist would be endorsing a semantic theory for 'true' that is completely unmotivated. It is my claim that truth expresses an inconsistent concept that justifies most of my descriptive theory.

Eklund could, I suppose, follow up with: "If assessment-sensitivity is prima facie such a bad semantic theory for 'true', then why does saying truth is an inconsistent concept help?" And, "Doesn't this attitude of promoting semantic theories that are not supported by the linguistic evidence constitute telling linguists what to do?"

A semantic theory isn't well supported from linguistics until it has been subject to all sorts of experiments – maybe these are based on fieldwork or on linguists' intuitions. My semantic theory for 'true' is not well supported from linguistics because linguists haven't spent much time working on 'true'.[22] Natural languages are huge, and much of the effort has been spent on large structural issues like quantifiers and tense. However, one can provide a measure of theoretical support for a semantic theory. One way of doing this is by thinking about faultless-disagreement and retraction. But there are others as well. By arguing for my semantic theory for 'true' on the basis of inconsistent concepts, I am effectively saying something like: when linguists get around to investigating 'true', they will find that people's judgments about 'true' – even core judgments that they are certain about – are inconsistent. No amount of trying to find some obvious parameter that is shifting around will succeed in eliminating the inconsistency. I suggest that in these cases, semantic relativist treatments

[22] See Moltmann (2015) for some results, however.

are the best. Of course, if linguists in the future decide to go through their standard testing of this assessment-sensitivity theory and they decide that it is unacceptable, then I would be happy to defer to their expertise.

Eklund pushes again: Why couldn't some theorist just point out the inconsistent judgments and advocate the same semantic theory on the basis of *that*, rather than on the basis of saying that truth is an inconsistent concept? One problem is that I don't know of anyone who has collected those data and I don't know any linguist who has thought about using tools from linguistics to find an underlying consistency in what seem like a batch of inconsistent judgments. So I suppose that my talk of inconstant concepts justifies my semantic theory for 'true' *in the absence of* this sort of detailed investigation by linguists. Eklund's imaginary theorist would have absolutely nothing to recommend the semantic theory and a lot to go against it. At least I have a promissory note for the theory and good reasons to dismiss the prima facie evidence against it.

***Objection 3*: One can agree that some constitutive principles involved are inconsistent with each other without accepting that truth is the responsible concept. And one can agree that truth is responsible without thinking that there are inconsistent concepts.**

I agree with both these points. I make the following claims: (i) truth is an inconsistent concept, (ii) truth is responsible for the paradoxes, and (iii) the concept of truth should be replaced. However, although they are related, these are independent of one another and are justified independently.

Truth is an inconsistent concept because the principles constitutive of truth are inconsistent with obvious facts like liar sentences exist and are meaningful. I justify for this claim in Chapter Three by arguing that ⟨T-In⟩ and ⟨T-Out⟩ are constitutive of the concept of truth.

Truth is responsible for the paradox because it makes the most sense to say that the liar paradox is

a symptom of the inconsistency in the concept of truth. I argue this point in Chapter Four by arguing that blaming truth is a far better explanation than blaming a litany of logical locutions.

Truth should be replaced for certain purposes because the inconsistency in the concept of truth is an impediment to using the concept of truth for certain purposes, like doing semantics for an expressively rich language. I argue for this claim in Chapter Five.

***Objection 4*: I argue that truth is an inconsistent concept by appealing to truth conditional semantics, but I don't really think truth-conditional semantics is acceptable. I effectively distinguish between strict truth conditional semantics, which uses truth, and rough truth conditional semantics, which uses something like truth. Now there are two versions of the meaning argument in Chapter Four. Neither version is a good argument. In particular, the consistency of truth and acceptability of rough truth conditional semantics don't seem to have anything to do with one another.**

I love this objection. It highlights exactly how hard it is to stay self-referentially consistent when engaging in conceptual engineering. I went over and over the project trying to make sure I had avoided every conceivable problem of this sort, but I missed this one! I suppose that is what I get when I practically dare the reader to find them.

Here is my meaning argument from Chapter Four of *Replacing Truth*:

(i) If truth is a consistent concept, then there are meaningful sentences that cannot be treated by truth-conditional semantics.

(ii) If there are sentences that cannot be treated by truth-conditional semantics, then truth-conditional semantics is unacceptable.

(iii) Truth-conditional semantics is acceptable.

(iv) So, truth is an inconsistent concept.

Once we distinguish between strict and rough truth-conditional semantics, we end up with two versions of this argument: one with 'strict' all the way through and one with 'rough' all the way through. The problem is that (iii) is false on the strict interpretation and (i) is false on the rough interpretation. Ouch.

Out of the context of my discussion, (i) probably looks very implausible all on its own. My argument for (i) is that, without implementing some approach to the liar and other paradoxes, a truth-conditional semantic theory for 'true' will be inconsistent. But any view that takes truth to be a consistent concept will inevitably engender revenge paradoxes and so any view that takes truth to be a consistent concept will have to be restricted so as to avoid applying to languages or sentences that cause revenge paradoxes. Here I am just following the almost ubiquitous refrain one hears from traditional theorists in response to revenge paradox objections – "oh, my theory doesn't apply to anything like that." Therefore, if some theory of truth T entails that truth is a consistent concept, then T will have to be restricted and so a truth-conditional semantic theory that implements T will have to be restricted in the same way (else: inconsistent). It is this convoluted conditional that is the basis for (i). So the justification for (i) treats it as implicitly about *strict* truth-conditional semantic theories. And remember (iii) is false on the strict interpretation – according to me at least. So that's bad.

But is it? This argument is meant to convince someone who thinks that truth is a consistent concept (or has no opinion on the matter) that truth is an inconsistent concept. So it makes sense to assess the argument against that background. And against that background, one should accept truth-conditional semantic theories (unless one has some other reason to reject them – but it would need to be an evidence-based reason, not deflationism or any other purely speculative position). They are endorsed by linguists who do natural language semantics in huge numbers; it dominates the

landscape (but remember that this impression is based on my own experience and testimony from linguists rather than any kind of poll).

Yes, there are other kinds of semantic theories that are also popular, including dynamic semantics, but that has nothing to do with whether truth-conditional semantic theories have a tremendous amount of explanatory power, both individually and when taken as a corporate body. The same goes for Newtonian mechanics even though Newtonian mechanics is false. Despite being false, it sets an explanatory bar that other more complex theories have to meet when one simplifies them by adding in various idealizing assumptions. Relativistic mechanics agrees with Newtonian mechanics (up to negligible discrepancy) in everyday situations. It is only on the cases that Newtonian mechanics got wrong that the theories diverge. If your theory doesn't agree with Newtonian mechanics on everyday predictions, then it is wrong. Dynamic semantics might eventually be to truth-conditional semantics as relativistic mechanics is to Newtonian mechanics, but we aren't there yet. And even if that day comes, it will have no impact whatsoever on my argument. Truth-conditional semantics will still have the same explanatory power (or probably greater because of the intervening research). Therefore, a person who thinks that truth is a consistent concept or hasn't really thought about it before should accept truth-conditional semantics. It is either right or it is close enough to right to serve as a standard in a wide variety of situations. Note that one should be reading my use of 'truth-conditional semantics' in this paragraph and the above paragraph as *strict*, not rough.

What about the meaning argument? Is it any good? I think the answer is *yes*. The intended audience for the argument should believe (i) – certainly after some prodding as above. And the intended audience for the argument should believe (iii). Again, maybe after being informed that linguistics is a science (some people still haven't heard, as preposterous as that sounds)[23] and the

[23] See Burgess (2015).

scientists who study these things utilize this theory in huge numbers. It has tremendous organizational and explanatory power for the entire study of natural language semantics.

Overall, the meaning argument has premises that should be accepted by people who are in its target audience. *I* don't accept them, but that is not the point. It should be fine to use an argument whose premises you don't accept if you think your audience does. For example an intuitionist logician might criticize a classical logician by using a classical reductio argument, even though intuitionists don't accept them – because the intuitionist knows the classical logician *does*. This isn't illegitimate. It's an internal critique. Overall I argue: If you think that truth is a consistent concept, then you both should and shouldn't accept truth-conditional semantics. But that is absurd. So truth is an inconsistent concept.

I can imagine Eklund objecting: if being superseded by a more complex and better empirically supported theory doesn't make truth-conditional semantics unacceptable, then why would restricting it to avoid paradox make it unacceptable? The answer is that the restriction to avoid paradox has to be massive. Remember Kripke's lesson about empirical or contingent paradoxicality – "many, probably most, of our ordinary assertions about truth and falsity are liable, if the empirical facts are extremely unfavorable, to exhibit paradoxical features."[24] So, in order to insure consistency in a truth-conditional semantic theory for 'true', one would have to restrict it severely. Moreover, implementing a restriction on a theory is different in kind from having a theory that makes some bad predictions. There is a big difference between having a mechanics that is restricted so that it says nothing at all about the procession of the perihelion of Mercury, and a mechanics that makes a prediction about it that is wrong by a significant margin. The former is obviously wrong as a general theory. It doesn't even try to be a general theory. And even though the second one gets it wrong, it still gets it pretty close – Newtonian mechanics is within 99% of the right value.

[24] Kripke (1975: 691).

Overall, it is the strict interpretation I was going for, and it makes sense to assume that my audience believes (i), (ii), and (iii) – again, perhaps after some prodding. Maybe it is better so say that, given what the audience believes about truth, (i), (ii), and (iii) are reasonable things for them to believe as well.

***Objection 5*: I don't have an argument for assessment sensitivity view of 'true' over a mere indeterminacy view of 'true'. In particular, the inconsistency view can't support the assessment-sensitivity semantic theory because inconsistency isn't necessary to support indeterminacy. Overall, inconsistency view plays little role in the book.**

I disagree with most of this one. The indeterminacy view Eklund mentions is often called local supervaluationism. It is the view that 'true' is indeterminate because there are multiple ways of interpreting it, and arguments involving sentences with 'true' should have the following logical standard: with premises G and conclusion p is valid iff for each interpretation I, if all the members of G are true in I, then p is true in I.[25]

The assessment-sensitivity semantics for 'true' that I propose in *Replacing Truth* entails local supervaluationism. But not vice versa. And I do have an argument for assessment-sensitivity over the mere local supervaluationism. The argument is simple – it is the best semantic theory consistent with local supervaluationism. Eklund doesn't endorse a semantic theory here at all, so it is hard to see why he would be opposed to mine unless he had some alternative in mind. Maybe he thinks that having no theory would be better than having the one I advocate. If so, then I didn't see an argument for that.

[25] Global supervaluationism, by contrast, advocates an alternative logical standard: an argument whose premises constitute a set G and whose conclusion is a sentence p is valid iff all the members of G are true in each interpretation, then p is true on each interpretation. Global supervaluationism requires a non-classical logic to handle certain vocabulary (e.g., 'determinately').

The inconsistency view definitely supports the assessment-sensitivity view, but that isn't my idea. I borrowed it from John MacFarlane's work on assessment-sensitivity and confusion.[26] Either way, Eklund's objection here seems to be based on a bad argument. He is right that inconsistency isn't necessary to support indeterminacy, but inconsistency does support assessment sensitivity, even though assessment-sensitivity entails indeterminacy. Saying that inconsistency isn't necessary to support indeterminacy is pretty weak. It just says there are other justifications for indeterminacy. And he is right, there are. But that has no bearing on whether inconsistency can justify indeterminacy and assessment-sensitivity. Eklund essentially argues that inconsistency isn't the only way to support indeterminacy so it cannot be a way at all. This is probably not what Eklund meant to do, but I don't see another reading of his text.

Overall, the claim that truth is an inconsistent concept plays a couple of related roles. It explains why extant solutions are so unsatisfying (they change the subject). It justifies introducing new concepts to deal with the problems caused by the liar – namely, that truth won't be able to fulfill its promise as an explanatory element in prominent semantic theories, once those theories are turned to the truth predicate itself. And finally the claim that truth is an inconsistent concept justifies the kind of semantic theory I advocate for 'true'.

Objection 6: Why do I favor replacement? I don't spend much time on explanatory role and I don't motivate replacements by appeal to their explanatory roles. I don't consider explanatory role of constitutive principles linking truth to other concepts. I do talk a bit about semantics, but I also have doubts about truth conditional semantics. Do I think that if truth has an expressive role, then it has an explanatory role?

[26] MacFarlane (2006).

I don't think that if truth has an expressive role, then it has an explanatory role. I didn't mean to give that impression. I figured (T-In) and (T-Out) were central in any explanatory role, but I focused on the role truth plays in explaining content in truth-conditional semantic theories. So I think I do motivate the replacements by their explanatory roles. And ultimately my reason for replacement is so that we can have something that plays this explanatory role a bit better in the same kind of semantic theory.

I did mention that I think dynamic semantic theories are going to upend a bunch of conventional wisdom in philosophy of language once philosophers of language pay attention them (which has begun in my opinion). But that doesn't in any way entail that truth-conditional semantics has no explanatory power. Newtonian mechanics still has tremendous explanatory power despite being dethroned over a century ago.

Ultimately, the argument for replacement is easy – a concept cannot do one of its legitimate jobs well. In the case of truth, the job is serving an explanatory role in truth-conditional semantics, and the reason it cannot do this job well is that it is an inconsistent concept whose inconsistency makes any standard truth-conditional semantic theory for 'true' inconsistent.

Greenough

Patrick Greenough focuses on the nature of concepts and my strategy for replacing the concept of truth for certain purposes. He formulates twelve distinct problems, which I reconstruct below interspersed with my solutions to each one.

***Problem One*: The replacement methodology on offer could be seen as a description of past (good) philosophical practice or it could itself be seen as a prescription for solving**

philosophical problems by replacing them. The first interpretation seems to be false. A great deal of extant philosophy has been conducted in a descriptive vein. The latter interpretation reveals some of the Marxist credentials of the project: philosophy needs to change in order to make progress. Which interpretation does Scharp have in mind? (I presume the latter.)

Solution: The descriptive/prescriptive distinction is especially important for my project. I advocate two distinct theories, one prescriptive and one descriptive. The prescriptive theory introduces two new concepts, ascending truth and descending truth, and offers an axiomatic theory of them, a way of interpreting this axiomatic theory in a broadly Davidsonian framework, and relationships between these new concepts and many of the concepts that are closely aligned with the concept of truth. The descriptive theory is a theory of truth, which consists of a semantics for the word 'true', the claim that the concept of truth is an inconsistent concept, and a theory of inconsistent concepts. The key is that neither of these theories relies on the defective concept of truth in any way. The semantics given for the word 'true', as part of the descriptive theory, use the concepts of ascending truth and descending truth. This semantic theory provides ascending truth conditions and descending truth conditions to all the sentences containing 'true', and the ascending truth conditions and descending truth conditions are the same when the sentence in question is not a paradoxical sentence like the liar.

Overall, I do not think that conceptual replacement projects of this sort are explicitly undertaken very often in western philosophy or in analytic philosophy in particular, but there are some obvious examples. Still, there are probably many cases where a philosopher has proposed some new way of thinking about something and that new way becomes so successful that it constitutes a new set of constitutive principles and so effectively introduces a new concept or new concepts. I don't think

Einstein thought of himself as replacing the concept of mass, but he did.[27]

***Problem Two*: Concepts can be defective in manifold ways. They can be intensionally defective: incomplete, confused, unsatisfiable, or even incoherent/inconsistent. They can be extensionally defective: too inclusive, too narrow, empty, or divided of reference. They can be too complex, too simple, too unspecific, or too vague. They can be too parochial or too elitist. They can be redundant or not fit to feature in any useful explanation. They can be superseded, hackneyed, or systematically misapplied. They can be loaded with ideological baggage or serve as ongoing devices for deceit, discrimination, or oppression. Given all this, conceptual incoherence is just one source of conceptual malfunction, and perhaps not the most prevalent or interesting source. So, why take philosophy to be mainly concerned with concepts which are defective in only one way—by being incoherent?**

Solution: There is a crucial distinction between inconsistent concepts like truth and unsatisfiable concepts like round square.[28] The concept of a round square is not defective – we can use it properly without contradicting ourselves or being committed to something false – just disapply it to everything. It is perfectly consistent to say that nothing is a round square. By contrast, the concept of truth is defective in that there is no consistent way to use it properly. No matter what we say about the liar sentence, we can reason, using the concept of truth, to a contradiction. In other words, the constitutive principles for the concept of truth, which we use to reason about truth, are inconsistent given the background assumption of certain basic inference rules and the existence of liar sentences. Not so for the concept of a round square. It is constitutive principles are something

[27] See Earman and Fine (1974) and Scharp (2013: ?) for discussion.

[28] See Stenius (1972), Chihara (1979), and Yablo (1993b) for discussions of the distinction.

like: round squares are round, round squares are square. These are not inconsistent with any facts about the world (even the fact that nothing is both round and square)—we would still need an additional constitutive principle like 'there are round squares' to get a contradiction. I am using 'defective' in a specific way that is roughly synonymous with 'inconsistent'. The key is that some of the constitutive principles for these concepts are false.

Why take philosophy to be mainly concerned with concepts that are defective in this sense? I am not arguing this on the basis of emptiness or complexity or deceit or oppression. None of those are conceptual defects in my sense. I didn't make this argument in *Replacing Truth*, but my reason is primarily that when one looks at the individual cases, one finds concepts that have constitutive principles that are inconsistent either internally or given the constitutive principles of other concepts, or given certain obvious facts about the world. My contention that philosophy is the study of what have turned out to be inconsistent concepts is a generalization based on evidence concerning all the particular cases. I included it in the book as a background for understanding the particular project in *Replacing Truth*. This material is the subject of a book in preparation.

***Problem Three*: Even if a replacement methodology for philosophy is called for, then philosophy will be concerned to offer replacement concepts across a wide variety of domains, both scientific and non-scientific. Take the concept of responsibility. If philosophy is in the business of replacing this concept with a better one, then we would surely pass the replacement on to legal theory (and related domains of study), and not to science. Likewise if the concepts of right, law, artwork, family, convention, freedom, for example, are to be replaced then we would not outsource the replacements to the branches of science. In other words, Scharp's replacement methodology, as stated, has some significant baggage: scientism. Does Scharp really conceive of philosophy in the Quinean**

tradition as the mere handmaiden to science or will he allow a more inclusive view?

Solution: I allow a more inclusive view. There are three related phenomena: (i) philosophy is shrinking, (ii) philosophy is outsourcing its subject matter, and (iii) philosophy is outsourcing its subject matter to the sciences. All three claims are true, but outsourcing is only one element of the shrinking, and outsourcing to the sciences is only one element of the outsourcing, Philosophy occasionally casts out something that had been part of its subject matter without that subject getting its own legitimate field of study. For example, early in western philosophy Socrates and Plato made a point of casting out rhetoric and sophistry. Later, early modern philosophers cast astrology out of philosophy. And the philosophers from the eighteenth and nineteenth centuries cast out theology. Today, neither rhetorical arguments, nor appeals to moon signs or God's omnipotence are legitimate moves in philosophical discussion, but they didn't get outsourced. Moreover, some of the topics that *are* legitimately outsourced do not become sciences. Law and international relations are good examples.

The establishment of science does not have to be the only exodus from philosophy for it to be the most significant and for it to be our role model. One can accept ths point without being committed to scientism.

Problem Three Plus: There is a serious worry whether the descending/ascending truth-predicates can function as devices for generalised endorsement/rejection respectively. For example, consider the claim: every claim in Scharp's book is descending true. Here we are using the descending truth predicate to record our assent to all the claims made in Replacing Truth. One of these claims is: the descending liar sentence is not descending true. (Indeed, this sentence is a theorem of ADT.) Since I endorse this sentence, and

descending truth is the device for endorsement, then I am committed to: "the descending liar sentence is not descending true" is descending true. But this claim quickly gives rise to paradox. (Cf. Scharp's discussion on pp. 286-7.) This is arguably just an instance of a more general worry that descending truth and ascending truth cannot, after all, function as consistent devices for endorsement andt rejection, respectively. Scharp (pp. 280-1) acknowledges this worry but replies that "[...] there is no such thing as a consistent device for endorsement. Descending truth is as close as one can get without having an inconsistent concept" (p. 281). In effect, the concept of endorsement needs replacing with a new concept endorsement*. Likewise, rejection needs replacing with rejection*. (It would have been useful to have seen this aspect of the theory spelt out in a bit more detail.)

Solution: I did not really have the space to pursue this replacement project in *Replacing Truth*, but the following ought to help flesh out the proposal.

It is common to say that truth predicates serve as devices of endorsement and rejection, and I have echoed these sentiments. One question is whether the replacement concepts serve these purposes. It seems like there is a straightforward argument that they do not. In particular, it seems difficult to see how to use ascending truth or descending truth to endorse a sentence like a descending liar (e.g., 'this sentence is not descending true'). If one calls it descending true, then one is thereby committed to it, but one's endorsement is not ascending true. If one calls it ascending true, then one is not thereby committed to it. So it seems impossible to say something acceptable that results in being committed to a descending liar. Parallel considerations hold for rejection.

My response to this objection is two fold. First, it isn't clear that ascending truth and descending truth need to serve as devices of endorsement and rejection in order to be acceptable replacements. Recall that the main aim of these replacement concepts is to do what we need for

natural language semantics. It might very well be that ascending truth and descending truth can be used to attiribute ascending and descending truth conditions without thereby serving as devices of endorsement and rejection. Still, it would be nice if they served these roles, so the second response is that no consistent replacement concepts whatsoever can serve these roles. Indeed, the conditions that have to be met in order for something to be a device of endorsement or a device of rejection are inconsistent (given relevant background information). Let E be a one place predicate that is a device of endorsement and let R be a one place predicate that is a device or rejection. In order for E to be a device of endorsement, it must be factive; that is, when one asserts that some sentence p is E, p itself follows from this utterance. Likewise, for R to be a device of rejection, it must be cofactive; that is, when one asserts that some sentence p is R, the negation of p follows from this utterance. So we have the following principles:

(Endorse 1) $E\langle p\rangle \rightarrow p$

(Reject 1) $R\langle p\rangle \rightarrow \neg p$

If these were the only relevant considerations, then it would be obvious that 'descending true' is a device of endorsement and 'not ascending true' is a device of rejection. However, there is an additional constraint. When one asserts that p is E, one's own utterance – that p is E – should be assertible if p is assertible. Likewise, when one asserts that p is R, one's own utterance – that p is R – should be assertibleif the negation of p is assertible. We can formulate these as two additional principles:

(Endorse 2) $\vdash p \rightarrow \vdash E\langle p\rangle$

(Reject 2) $\vdash \neg p \rightarrow \vdash R\langle p\rangle$

These additional principles invoke provability in the turnstile, which might seem to be a mischaracterization. My reasoning here is that provability from the theory in question is going to be

the basis for what one takes to be assertible on the basis of the theory in question and it is nicely behaved. Perhaps there is room for someone to say that a sentence that is not provable from the theory is still assertible, but I won't have anything to say about a view like this. Instead, if someone develops the same objection in this form, I will have to address it independently.

Let (Endorse 1) and (Endorse 2) be constitutive of a device of endorsement and let (Reject 1) and (Reject 2) be constitutive of a device of rejection. Given this background information, Montague's theorem demonstrates that devices of endorsement and devices of rejection are inconsistent concepts. That is, Montague showed that (Endorse 1) and (Endorse 2) are inconsistent.[29] A parallel argument shows that (Reject 1) and (Reject 2) are inconsistent. It follows that there is no coherent device of endorsement and there is no coherent device of rejection. The concept of a device of endorsement is defective, as is the concept of a device of rejection. Therefore, the demand that the replacements for the concept of truth ought to serve these roles is inappropriate. This completes my second response to the objection.

There is still a question of whether to replace the concept of a device of assertion and the concept of a device of rejection. I suggest that the second set of assumptions be dropped. It might seem like for any proposition that is correctly assertible, there is a way to endorse it indirectly by using a device of endorsement. For example, if I want to endorse the modularity theorem but I can't remember what it says, then I can assert 'the modularity theorem is true'. The modularity theorem itself follows from my assertion, and the sentence I uttered is assertible. It might seem like one can do this for any assertible proposition or collection of propositions, but this is a mistake. That is, it is not the case that one can simply endorse any assertible proposition by predicating something of it. There are certain assertible propositions that simply cannot be endorsed by uttering something

[29] Montague (1963).

assertible. Likewise, there are certain unassertible proposition that simply cannot be rejected by uttering something assertible.

If one makes this conceptual change, then the replacement concepts for the device of endorsement and device of rejection might be called endorsement* and rejection* (following Greenough). Then it is easy to see that 'descending true' is a device of endorsement* and 'not ascending true' is a device of rejection*. There is more to be said about how these considerations relate to the work on assertion, denial, endorsement, and rejection, but there is no space for elaboration here.

***Problem Four*: Suppose one thinks that truth has an expressive but no explanatory role: truth is Tschema for truth (and cognate schemas). Such is deflationism about truth. However, suppose the deflationist thinks that this device over-reaches: it doesn't do well when it comes to a range of paradoxical sentences such as the liar sentence. In the face of such paradox, one live option is to replace this device with a pair of devices, namely ascending and descending truth, together with the predicates "is ascending true", "is descending true". These replacement predicates, following Scharp, are fit to perform the expressive roles of the original truth-predicate. However, given deflationism, they are not fit for substantial philosophy theory—but that ought not to be a necessary condition for a replacement methodology. So, this seems to me to represent a perfectly coherent deflationary form of Conceptual Marxism. So, why exactly does Scharp think that a replacement strategy is inimical to deflationism?**

Solution: Simply being inconsistent is not a sufficient condition for replacement. One needs to show in addition that the inconsistency is an impediment to the concept's utility in some way. I argued this

point with the example of truth conditional semantics. Any "off the shelf" truth conditional semantic theory for the truth predicate is going to be inconsistent. It will imply that a liar sentence is in the extension of the truth predicate iff it is not. So not only is the concept of truth inconsistent, its inconsistency impedes its utility. Moreover, I showed that the replacement concepts, ascending truth and descending truth, can be used to fix this problem – they can do the job in truth conditional semantics that we thought truth could do. When we formulate semantic theories with ascending truth and descending truth, which attribute ascending truth conditions and descending truth conditions to sentences of the target language fragment, they are consistent even when applied to languages with truth predicates, ascending truth predicates, and descending truth predicates. So there are three elements to the replacement project: (i) showing the concept of truth is inconsistent, (ii) showing that its inconsistency poses a problem for one of its jobs, and (iii) showing that a team of replacement concepts can do this job instead. The job I on which I focused was serving an explanatory role in a semantic theory.

I didn't see a way to pursue this sort of strategy with an expressive job instead of an explanatory job. I also think that deflationism about truth is radically implausible, so I suppose I didn't try very hard. The deflationary refusal to admit one of the main tenets of one of the sciences – namely that truth plays an explanatory role in semantic theories in linguistics – strikes me as absurd. It certainly displays an arrogant attitude toward the sciences that is reminiscent of creationism or climate-change denial. Instead I prefer a modest philosophical attitude toward the sciences.[30]

Either way, I do not see the inconsistency in the concept of truth as much of an obstacle to its expressive jobs. Although there has been some discussion about whether deflationists are in a better or worse position to deal with the aletheic paradoxes, it is part of the folklore associated with the subject that the paradoxes rarely if ever pose a practical problem for communication. I think

[30] See Replacing Truth: 123-125.

Stephen Yablo once characterized the situation with a satirical question that is spot on: sure, truth works in practice, but does it work in theory? I just do not see the paradoxes as enough of a threat to the expressive role of truth to justify a replacement.

Moreover, one can think about replacement concepts – ascending truth and descending truth – and the properties they denote. One can ask whether these replacement concepts and these replacement properties are deflationary. Obviously the "no explanatory role" version of deflationism is no more plausible here than it was for the concept of truth, but there are other versions of deflationism: for example, that truth is transparent or that truth is logical. It turns out that ascending and descending truth might well be deflationary in some of these ways.[31] But they clearly fail to be deflationary in what is perhaps the most familiar way because they have explanatory roles in semantic theories that are a mainstay of the science of linguistics.

Problem Five: **Whether or not one accepts the Concept Identity Principle will be greatly influenced by the stand one takes on what kind of thing concepts are, where they live, and how they survive. Suppose concepts are three-dimensional enduring entities. Suppose further that a concept can persist through a significant revision to one or more of its core constitutive principles. Then the Concept Identity Principle is false. In effect, Scharp (implicitly) assumes that an endurantist view of concepts is wrong. Equally, one might take concepts to be four-dimensional perduring entities with temporal parts, where these temporal parts are composed of different sets of constitutive principles. On such a view, the Concept Identity Principle needs re-working such that a concept is individuated by its temporal parts, and temporal parts are individuated via the set of constitutive principles true**

[31] See Scharp (forthcoming) for discussion.

of the concept in question at a particular time. Scharp (implicitly) rejects such a perdurantist view since it permits conceptual persistence through a perdurantist conception of change—namely, whereby the concept is an aggregate of temporal parts, where these parts may be composed of different sets of (core) constitutive principles.

Solution: I do accept the concept identity principle, and I find this way of conceiving of concepts powerful and yet still firmly rooted in our practices. But it isn't required to make sense of replacing truth. For example, on the endurantist conception, we think of the concept of truth as something that can survive a change in its constitutive principles. On this conception, one might suggest giving up a constitutive principle for truth, but retaining the concept itself, is an option. So we ought to just reject all the constitutive principles that lead to paradox, right? So which one of (T-In) and (T-Out) ought we reject? Whichever one we decide on, the concept of truth after the change will not be able to perform its explanatory role in truth-conditional semantic theories. Let us say we reject (T-Out). So the concept of truth after the change is going to be a lot like the concept of ascending truth. Then we still need at least one replacement in order to do linguistics. We will need something like descending truth. Then we would have something like ascending truth and something like descending truth after this revision project, but we wouldn't have anything like what we think of as truth right now—nothing with (T-In) and (T-Out) as constitutive principles.

Moreover, I don't understand the rules of this game very well. Can we change the concept of truth to reject (T-In) and then switch it again to keep (T-In) and reject (T-Out)? Can we do this fast enough to just use truth in our semantic theories but with the understanding that, to use it, we have to keep switching its constitutive principles like this? Another question: on this endurantist view, how far can we go in rejecting constitutive principles? What if we reject all of them? How do we think of truth in that case? What if we adopt a new constitutive principle that only potatoes can be

true? I think there has to be some kind of limit, but it isn't clear what they are on the endurantist and perdurantist views outlined. On the other hand, the concept identity principle is very clear and leads to a useful and powerful account of concepts.

Problem Six: **Principle P1,**

(P1) If "S" is provable then "S" is true

is a constitutive principle for provability. To replace this principle with P2,

(P2) If "S" is provable with "S" is ascending true.

while keeping the concept of provability unchanged, is to be engaged with conceptual revision and not conceptual replacement, as we have just seen. The Concept Identity Principle enforces the result that to replace P1 with P2 means that we are now dealing with a different concept of provability. Crucially, we are not allowed to use the word "provable" to pick out this new concept, so we have to introduce some new vocabulary to refer to it. So, the replacement principle should in fact be:

(P3) If "S" is provable$_1$ then "S" is ascending true,

where "provable$_1$" picks out the replacement concept of provability. This begins to reveal how extreme Conceptual Marxism really is: replace the concept of truth and you must also replace the concept of provability, and indeed you must also replace the word "provable" too. Does Scharp acknowledge that this is an unavoidable feature of his view, a feature which shows that the view is more radical than the view officially advertised in Replacing Truth?

Solution: This is a great objection, and it would be really bad if it were true. Luckily, it isn't. So the short answer is that I do not acknowledge that this is an unavoidable feature of my view. In fact, this

objection overlooks a crucial feature of my views on constitutive principles. Namely, a constitutive principle is *constitutive for a particular concept*. In other words, constitutivity is a relation, not a property. There is no such thing as *being constitutive*. There is only *being constitutive for such and such concept*. This point was not emphasized in Chapter Two of *Replacing Truth*, although it is mentioned occasionally. However, it has a central place in the way the theory of constitutive principles is presented above. For example, I argue that (T-In) and (T-Out) are constitutive for truth. That is, if I am in a conversation and one of my companions denies an instance (T-In) or (T-Out), then that is a pro tanto reason to think that we do not mean the same thing by the word 'true'. If my companion, say, claims that snow is white but denies that 'snow is white' is true, then that is effectively a denial of an instance of (T-In). This should make me question whether we mean the same thing by 'true'. But it does not and should not make me think we do not mean the same thing by 'snow'. The reason is that (T-In) is constitutive for the concept of truth, but it is not constitutive for the concept of snow.

Recall that the theory of constitutive principles presented above is explicit about basing constitutivity on meaning reflection. Constitutive principles are those that, when denied, provide a reason to think that not all the participants in the conversation mean the same thing by one of the words used in that conversation. Thus, the fact that constitutivity is relative to a concept falls out of the theory of constitutive principles and is not some ad hoc move.

Consider another example. Imagine you are in a conversation with a person, Hilary, talking about cats and their quirky behaviors. At some point Hilary asserts 'cats are complex robots'. This assertion would probably be what I have called an interpretive red flag. That is, it probably would make you think that you and Hilary do not mean the same thing by one of the words involved. Which word? My guess is that you might think Hilary means something else by the word 'cat' or that Hilary means something else by the word 'robot'. However, you would not think Hilary means something else by the word 'complex'. If these considerations are right, then 'cats are not complex

robots' is constitutive for the concept of cat and for the concept of robot, but it is not constitutive for the concept of complexity. My claim is that this is a robust phenomenon that occurs across the range of constitutive principles. However, only empirical testing will be able to confirm this suspicion.

To return to Greenough's specific objection: I do not advocate replacing principle (P1) with principle (P2). I claim that (P1) is constitutive of truth. And I claim that (P2) is constitutive of ascending truth. Even if one takes (P1) to be constitutive of the concept of proof as well, no one thinks that (P2) is constitutive of the concept of proof because no one has even heard of ascending truth. Moreover, (P2) is stipulated by me to be constitutive of ascending truth – this is part of the project of linking the replacement concepts to other concepts in our conceptual scheme. But these linkages are constitutive only for ascending truth. They have no bearing on the identity of these other concepts. Overall, one can introduce the concepts of ascending truth and descending truth in the way I recommend without thereby changing any other concepts in our conceptual scheme. Replacing truth does not require changing truth and it does not require changing any other concepts or replacing any other concepts.

***Problem Seven*: However, by Scharp's lights, these cannot be the right replacement principles for (correct) assertion/belief. That is because, again, the Concept Identity Principle enforces the result that to replace a constitutive principle for a concept is to replace the concept. So, we can no longer use the same concept-word to pick out the new replacement concept. So, Scharp's replacement principles should be:**

(AB5) An assertion$_1$/belief$_1$ that p is ascending true if things are as they are asserted$_1$/believed$_1$ to be.

(AB6) An assertion$_1$/belief$_1$ that p is descending true only if things are as they are

asserted$_1$/ believed$_1$ to be.

(AB7) It is correct to assert$_1$/believe$_1$ that p only if p is ascending true.

(AB8) It is correct to assert$_1$/believe$_1$ that p if p is descending true,

where "assertion$_1$" and "belief$_1$" pick out the replacement concepts for assertion and belief, respectively. Surely Scharp is committed to the more radical revision resulting in AB5-8, rather than AB1-4?

Solution: This problem has the same form as the previous one and my solution is essentially the same as well. I am formulating constitutive principles *for ascending truth* and *for descending truth*. I am not suggesting any changes in the concept of belief or assertion. I can introduce two new concepts without changing the concept of belief or the concept of assertion. I am not suggesting that these concepts should get any new constitutive principles.

Of course, I do ultimately think that the concept of belief and the concept of assertion are inconsistent, and so they might need to be replaced (if their inconsistency impedes their utility). But that is an additional commitment that rests on additional evidence about belief and about assertion. It isn't required by the view defended in *Replacing Truth*.

Problem Eight: **The general lesson ought to be clear. If we replace the concept of truth with one or more surrogate concepts then any concept which is constitutively linked to truth via one of its core constitutive principles must be replaced too. Furthermore, we cannot use the old concept-word to pick out this new concept—we must introduce a new concept-word to refer to it. The problem is that not only is the concept of truth constitutively linked to the concepts of provability, assertion, and belief, it is also so linked to myriad other concepts such as the concepts of inquiry, objectivity, reality, knowledge, judgment, evidence,**

justification, confirmation, probability, fact, being, truthvalue, truth-bearer, reference, denotation, satisfaction, truth-condition, meaning, content, proposition, representation, necessity, possibility, contingency, and more. In turn these concepts are constitutively linked to a wider class of concepts which may well include, in the end, all concepts of central philosophical interest. All these concepts must be replaced too, together with their respective concept-words. The problem faced by Scharp thus proliferates very quickly. Again, this not only makes the view much more radical than the advertised view in Replacing Truth, it begins to make the view implausible. Would it not be better to ditch the Concept Identity Principle altogether and go for a less extreme replacement strategy whereby we can keep the old concept and the old concept-word, but merely replace the constitutive principles? Surely we should explore a more moderate left-wing option first before trying out the wholesale conceptual cleansing recommended by Scharp? (See below.)

Solution: Same reply. Constitutive principles are for particular concepts. There is no such thing as a constitutive principle in general – one that isn't tied to a particular concept. In introducing two new concepts, ascending truth and descending truth, I stipulated some constitutive principles for them in the form of an axiomatic theory and in the form of connections to other concepts to which truth has been linked. Neither the axioms of the theory (ADT) nor the connections to other concepts are constitutive for anything other than ascending truth and descending truth. Hence the replacement project does not depend on changing or replacing any constitutive principles for any other concepts. The key feature of constitutive principles that allows them this flexibility is that each one is constitutive *for a particular concept or word.*

***Problem Nine*: The replacement version of Eklund's view would resolve the sorites paradox**

via conceptual replacement. However, to effect such a replacement, the principle TOL would have to be replaced with some principle which was suitably weakened so as not to give rise to the paradox. Moreover, such a replacement principle would ensure that we are no longer dealing with the concept yellow, but a new concept—to be picked out by a new concept-word such as "yellow1". So much is enforced by the Concept Identity Principle. But then in order to address the sorites paradox in English, we need to replace every single vague concept with a new concept, and replace every single vague predicate in English with a new predicate which picks out this new concept. Unlike the liar paradox, which only arises when we formulate a liar sentence, vagueness is ubiquitous in natural language. So the required replacements are not simply limited to some special region of thought and talk. Upshot: once Conceptual Marxism is applied to vagueness the result is an even more extreme kind of conceptual cleansing. Scharp is then faced with a dilemma: either treat vagueness differently from how he treats the liar paradox and kindred paradoxes (and face a charge of ad hocness) or treat vagueness via his Conceptual Marxism (and face a charge of wholesale conceptual cleansing). Which horn of the dilemma will he take?

Solution: The first horn of the dilemma is only a problem if it involves treating similar cases differently, but the case of vagueness and the case of truth are very different. For example, the logical principles involved in deriving the sorites paradox are less entrenched than those involved in deriving the liar paradox – weaking classical logic to intuitinistic logic does not help. Moreover, even if I go with the "vague concepts are inconsistent" line, it doesn't require any conceptual change at all. The case for inconsistency and the consequences of inconsistency are distinct from the case for replacement and the consequences of replacement. Saying that concept X is inconsistent does not entail that concept X should be replaced in any way. The case for replacement of X is the

inconsistency of X *plus* the fact that X's inconsistency is preventing X from doing one of its jobs. So it might be that vague concepts are inconsistent but they don't need to be replaced. Or maybe some need to be replaced for certain purposes.

However, I am not committed to the claim that vague concepts are inconsistent. I am open to this possibility, but I am still thinking about the best way to make sense of vagueness in general and tolerance principles in particular (e.g., taking a penny away from a rich person does not make that person not rich). I am not convinced that tolerance principles are constitutive of vague concepts because people violate tolerance principles so often.[32] There is much more to be said here

Problem Ten: **Which of [the five] options [for replacing the concept of knowledge] should Scharp choose? [The options are: (i) revise the concept of knowledge to include ascending truth, (ii) revise the concept of knowledge to include descending truth, (iii) replace the concept of knowledge with a concept defined in terms of ascending truth, (iv) replace the concept of knowledge with a concept defined in terms of descending truth, and (v) replace the concept of knowledge with two knowledge-like concepts.] Given that he already has the concepts of ascending truth and descending truth in hand to address the liar paradox, then it looks as if one of the Mono-Replacement strategies [(iii) and (iv)] is the better option. The thought here is that a kind of maxim of minimal mutilation is in play: revise or replace as little as possible so as to save what can be saved of truth (and logic). Upshot: even though we need two new concepts to solve the liar, we don't need two new replacement concepts for knowledge to solve the knower (once we have two new replacement concepts for truth). Still, a residual problem remains: which of Option Three and Option Four are we to choose, and why?**

[32] See Raffman (2013, 2015) and Scharp (2015) for discussion

Solution: I do think that the concept of knowledge is inconsistent. And I think it is inconsistent in several different ways, as evidenced by the Fitch paradox, the Knower paradox, and the various skeptical paradoxes. Perhaps not all these inconsistencies get in the way of using the concept of knowledge, so they might not all necessitate a replacement. However, it does seem like the Knower causes serious problems for attempts to use say which claims about knowledge itself are known. So I do think that the concept of knowledge is not only inconsistent, but that it ought to be replaced for various purposes. However, I think that any discussion of replacing knowledge ought to take into consideration not only an approach to the knower paradox, but approaches to the other paradoxes affecting knowledge as well. Thus, I can only give a partial solution to this problem of Greenough's.

Greenough is right that options one and two are ruled out for me because they involved changing our existing concept of knowledge. Of options three, four, and five, I lean toward option five. My reason is that I am suspicious of attempts to parlay a solution of one paradox into a solution to another. These just usually don't work. For example, option three is compatible with knowing that p and knowing that not p (for certain propositions p like the ascending liar), and option four does not allow one to infer that we know certain things (like that the descending liar is not descending true). Another reason specific to this case is that the concept of safety and unsafety weren't designed to handle sentences other than those containing 'true'. It is for these reasons that, if forced, I would choose option five – replace the concept of knowledge with two concepts.

Problem Eleven: **Which of these options should Scharp choose between? [The two options are: (i) change the concept of truth to restrict the truth elimination principle, and (ii) change the concept of truth to restrict the truth introduction principle.] As it turns out, neither. Why is that? You've guessed it—it's because of our old friend The Concept Identity Principle.**

Once the elimination-rule for truth has been replaced, we are no longer dealing with the concept of truth. Likewise, for the introduction rule. In other words, there is no Mono-Replacement Strategy available not only for the concept of truth, but for the constitutive principles for truth. That's why Scharp's Conceptual Marxism is inevitably extreme—at least in so far as The Concept Identity Principle is taken to be inevitable. So, on pain of repetition, we may ask again: Is this Principle so inevitable?

Solution: The concept identity principle isn't inevitable, it is just better than any of its current alternatives. It might seem like anyone who endorses the concept identity principle cannot accommodate any kind change to our concepts, but I think this is a mistake. Consider the theory of constitutive principles and concepts I offered above. One important application for this theory of constitutivity is to conceptual engineering. There is a crucial distinction between conceptual revision and conceptual replacement. Conceptual *revision* occurs when a concept changes in some important way but still remains the same concept. Conceptual *replacement* occurs when new concepts are introduced. In Replacing Truth, I followed a conceptual replacement strategy. And given the way I am individuating concepts – according to their constitutive principles – it seems difficult to make any sense at all of conceptual revision. After all, any change to the constitutive principles of a concept will result in a new concept, not the same concept but altered in some way. However, the theory of constitutive principles and concepts is more versatile than it looks, and does allow one to explain at least certain kinds of conceptual revisions.

The constitutive principles themselves are modeled by a 4-tuple of person, sentence, word, and degree. We then use a global threshold – it applies to every constitutive principle – to figure out which sentences are genuinely constitutive because they have a degree that is above the pre-established threshold. However, we can make sense of conceptual revision as a change in the degree

of some constitutive principle that is not enough to cross the global threshold. For example, let us say we have three constitutive principles for 'bachelor': all bachelors are unmarried adult men, and all unmarried adult men are bachelors. Assume that each of these has a degree of 90%, and assume that the global threshold is 80%. So the concept of bachelor really does have these constitutive principles. Now imagine that something happens to weaken the person's confidence that one of these principles is constitutive of 'bachelor'. Perhaps the person thinks about whether the Pope is a bachelor even though the pope is unmarried. As a result, assume the constitutivity degree for 'all unmarried adult men are bachelors' drops from 90% to 85%. This is a genuine effect on the constitutivity of this principle for 'bachelor', but this change has no impact on which constitutive principles are genuinely constitutive for 'bachelor' because the change doesn't cross the global threshold. If it had dropped to 70%, for example, then the concept of bachelor would have changed. This would be a conceptual replacement, not a conceptual revision. Therefore, our model allows us to make sense of conceptual replacement and certain kinds of conceptual revision.

Final Problem: **In other words, if one is tempted by some kind of replacement strategy, and constrained by a maxim of minimal mutilation, then one should seriously consider ditching the Concept Identity Principle so as to allow the following constellation of claims: keep the concept of truth in troublesome contexts; keep the word "true" in troublesome contexts; keep the constitutive principle of truth-introduction (and the alethic principle of necessitation T-Nec); but revise the concept of truth so it no longer validates a rule of truth-elimination; finally, use a replacement concept—namely, descending truth—to function as a surrogate concept, equipped to function as a device for (indirect) assent. Surely this represents a less extreme, and more attractive proposal than the Bi-Replacement strategy offered in *Replacing Truth*?**

Solution: Greenough's strategy is not more attractive. First, it involves the dubious claim that we can somehow change our concept of truth. Even if we philosophers can make sense of this in theory – by ditching the concept identity principle – how exactly should it be implemented in practice? Presumably we would have to get funding to run a world-wide public service campaign – "Stop using truth that way! Ding, ding, ding, ding -- The More You Know". I don't see large-scale conceptual revision projects like this as remotely plausible. Everyday people simply won't change the way they use the word 'true' based on these kinds of considerations. Thus, from my perspective, this proposal is a non-starter.

By contrast, I am not trying to change *any* of the concepts we currently use. I am instead introducing two new concepts, and I am recommending that the handful of specialists who care about doing semantics for expressively rich languages use these new concepts instead of our old concept of truth. This sort of project does not require large-scale changes in the way humans use basic concepts like truth.

Second, Greenough's plan leaves us without a concept like the concept of truth as it is right now. Given the utility of the concept of truth (as it is now), his plan would result in a serious hole in our conceptual scheme. Despite being an inconsistent concept, truth is tremendously useful and it causes no problems at all in almost every case where it is used. It is only in a tiny number of applications – attributing truth conditions to certain sentences of expressively rich languages – that it causes troubles for us. Surely if Greenough somehow persuaded humanity to change its concept of truth in the way he lays out, we would immediately coin some new concept that would be exactly like the concept of truth as it is right now.

Consider an analogy that I find illuminating. We know that the concept of mass as it occurs in Newtonian mechanics is inconsistent. It was replaced by the concepts of relativistic mass and

proper mass about a century ago when Einstein and other physicists proposed and developed relativistic mechanics. However, the concept of mass is ridiculously useful, and very few applications require using the replacements for it. Imagine an analog of Greenough's plan for the concept of mass – it would leave us without our inconsistent concept of mass and we would be stuck with using relativistic mass and proper mass for all our projects that require some concept of mass or other. That means any time anyone wanted to design a house or a car or a bridge or calculate a stress or a force she would have to use relativity. Just think about how shockingly inefficient that would be. The same can be said for Greenough's plan for truth as well.

In conclusion, I want to reiterate my appreciation for the comments on *Replacing Truth* from Bacon, Eklund, and Greenough. It has been a tremendous pleasure to engage with these theorists, and I hope to continue benefitting from interactions with them for a long time to come.

Inquiry, 2018
https://doi.org/10.1080/0020174X.2017.1420012

Work Cited

Bacon, Andrew. (2015). "Can the Classical Logician Avoid the Revenge Paradoxes?" *Philosophical Review* 124: 299-352.

Bacon, Andrew. (2017). "Scharp on Replacing Truth," *Inquiry*

Beall, Jc. (2008). "Prolegomena to Any Future Revenge," in *The Revenge of the Liar*, Beall (ed.), Oxford.

Brandom, Robert. (1994) *Making It Explicit*. Cambridge: Harvard University Press.

Burgess, John. (2014). Review of Scharp (2013). *Studia Logica* 102: 1087-1089.

Chihara, Charles. (1979), "The Semantic Paradoxes: A Diagnostic Investigation," *The Philosophical Review* 88: 590–618.

Davidson, Donald. (1980). "Toward a Unified Theory of Meaning and Action." *Grazer Philosophische Studien* 11: 1-12.

Davidson, Donald. (1990). "The Structure and Content of Truth," in Davidson, *Truth and Predication*, Cambridge: Harvard University Press, 2005.

Earman, John, and Arthur Fine (1977). "Against Indeterminacy," *Journal of Philosophy* 74: 535–538.

Eklund, Matti. (2017). "Inconsistency and Replacement," *Inquiry*

Field, Hartry. (2006). "Truth and the Unprovability of Consistency," *Mind* 115: 567–605.

Field, Hartry. (2008). *Saving Truth from Paradox*. Oxford.

Fitch, Frederic. (1946). "Self-Reference in Philosophy," *Mind* 55: 64-73.

Friedman, Harvey, and Michael Sheard (1987). "An Axiomatic Approach to Self-Referential

Truth," *Annals of Pure and Applied Logic* 33: 1–21.

Greenough, Patrick. (2017). "Conceptual Marxism and Truth," *Inquiry*

Kripke, Saul. (1975). "Outline of a Theory of Truth," *The Journal of Philosophy* 72: 690–716.

Lewis, David. (1974). "Radical Interpretation," *Synthese* 27: 331–44.

Lewis, David. (1979) "Scorekeeping in a Language Game," in *Philosophical Papers* Vol. 1. Oxford: Oxford University Press. 1983.

MacFarlane, John. (2007). "The Logic of Confusion," *Philosophy and Phenomenological Research* 74: 700–708.

Mackie, J. L. (1964). "Self-Refutation--A Formal Analysis," *Philosophical Quarterly* 14: 193-203.

Moltmann, Frederike. (2015). "'Truth Predicates' in Natural Language," in Achourioti, Galinon & Martinez (eds.), *Unifying Theories of Truth*. Springer.

Montague, Richard. (1963). "Syntactical Treatments of Modality, with Corollaries on Reflection Principles and Finite Axiomatizability," *Acta Philosophica Fennica* 16: 153–67.

Raffman, Diana. (2013). *Unruly Words*. Oxford.

Raffman, Diana. (2015). "Responses to Discussants," *Philosophy and Phenomenological Research* 90: 483-501.

Roberts, Craige. (1996). "Information Structure: Towards an Integrated Formal Theory of Pragmatics," *Semantics and Pragmatics* 5, article 6: 1–69.

Scharp, Kevin. (2013). *Replacing Truth*. Oxford.

Scharp, Kevin. (2014). "Truth, Revenge, and Internalizability," *Erkenntnis* 79: 597-645.

Scharp, Kevin. (2015). "Tolerance and the Multi-Range View of Vagueness," *Philosophy and Phenomenological Research* 90: 467-474.

Scharp, Kevin. (forthcoming). Conceptual Engineering for Truth: Aletheic Properties and New Aletheic Concepts," *Synthese*.

Stalnaker, Robert. (1999). *Context and Content*. Oxford: Oxford University Press.

Stenius, Erik. (1972). *Critical Essays*. Amsterdam: North-Holland Publishing Company.

Williamson, Timothy. (2008). "'Conceptual truth'," *The Aristotelian Society, Supplement* 80: 1–41.

Yablo, Stephen. (1993). "Hop, Skip and Jump: The Agnostic Conception of Truth," *Philosophical Perspectives* 7: 371–96.