Can The Classical Logician Avoid The Revenge Paradoxes?*

Andrew Bacon

December 22, 2014

Contents

1 Linguistic theories and revenge **4**
1.1 Revenge paradoxes 6
1.2 The impossibility of a classical revenge-free linguistic diagnosis of the liar . . 7

2 An outline of a non-linguistic theory **13**
2.1 Unclear truth conditions 15
2.2 Can we recover sentential clarity? 18
2.3 A theory of clarity 20
2.4 The theory DFS 23

3 Higher order unclarity **26**
3.1 Hierarchies of determinacy operators 27

4 Revenge and Assertability **30**
4.1 Assertion and higher order unclarity 32
4.2 Assertive uttering 33

5 Conclusion **35**

6 Appendix **35**
6.1 The consistency of DFS 35

*I owe a great deal of thanks to Cian Dorr, Peter Fritz, Jeremy Goodman, Harvey Lederman, Timothy Williamson, an anonymous reviewer and two anonymous editors for this journal, all of whom gave me substantial comments at various points in this paper's development. I'm also grateful to audiences at Bristol University, Brown University, LMU Munich, Yale University, Oxford University and the University of Southern California.

Abstract

Most work on the semantic paradoxes within classical logic has centred around what I call 'linguistic' accounts of the paradoxes: they attribute to sentences or utterances of sentences some property that is supposed to explain their paradoxical or non-paradoxical status. 'No proposition' views are paradigm examples of linguistic theories, although practically all accounts of the paradoxes subscribe to some kind of linguistic theory. This paper shows that linguistic accounts of the paradoxes endorsing classical logic are subject to a particularly acute form of the revenge paradox: that there is no exhaustive classification of sentences into 'good' and 'bad' such that the T-schema holds when restricted to the 'good' sentences unless it is also possible to prove some 'bad' sentences. The foundations for an alternative classical non-linguistic approach is outlined which is not subject to the same kinds of problems. Although revenge paradoxes of different strengths can be formulated, they are found to be indeterminate at higher orders and not inconsistent.

According to a naïve, but extremely attractive thought, the sentence 'p' and the sentence "p' is true in English' can be used interchangeably in English when appearing in extensional contexts.[1] Thus in particular, one ought to accept every instance of the following disquotational schema:

T. The sentence 'p' is true in English if and only if p.

Unfortunately this schema cannot be true in general, for if we let L be the sentence 'L is not true in English' an instance of this schema states that:

> The sentence 'L is not true in English' is true in English if and only L is not true in English.

But since L just is the sentence 'L is not true in English' we can derive this apparently absurd conclusion from Leibniz's law:

> L is true in English if and only if L is not true in English.

Although this conclusion was derived without assuming anything distinctively classical, if we do assume classical logic we can derive from this sentence any conclusion we like, and that really is absurd. We thus seem to have a paradox, for either we must relinquish this instance of T, or we must reject some of the subsequent reasoning, and both of these options seem radically counterintuitive.

What, exactly, would count as a solution to this paradox? For the classical logician it involves explaining why this instance of T, which holds good in so many cases, fails when applied to L. Such an explanation, while necessary, may well be insufficient for an adequate solution. Perhaps it is also impossible to know whether L is true or not, or futile to wonder or try to find out which it is, or perhaps it's always a bad idea to go about asserting L. If true, each of these things also deserve some kind of explanation too.

A standard strategy for answering such questions is to identify a class of sentences as problematic in the hope that restricting attention to sentences outside this class will guarantee that one will not run into trouble. In Tarski's words (Tarski (1969)) 'the appearance

[1] Here, and elsewhere in the paper, I take the 16th letter of the Latin alphabet to be a substitutional variable to be substituted for complete English sentences (thus the result of putting quotation names around that letter is to be thought of as a quotation name for a sentence, and not a quotation name for that letter.) Free occurrences of such variables should be read schematically: I am saying something about all of the substitution instances.

of an antinomy is [...] a symptom of disease', and the strategy in question is to identify that disease with some feature the sentence L and other problematic sentences share. Extant theories in this vein identify this disease with being ungrounded (e.g. Kripke (1975)), possessing a peculiar kind of context sensitivity (e.g. Parsons (1974)), having been introduced by impredicative (Russell (1903)) or inconsistent (Chihara (1979)) definitions, failing to follow analytically from non-linguistic facts (McGee (1990)) or being semantically unstable (Gupta and Belnap (1993)), to name but a few. I shall call all such approaches 'linguistic accounts' in virtue of their identifying the problematic phenomenon present in the semantic paradoxes with some property of the language or linguistic items used to express the paradoxes.[2] In each case sentences said to be in this class of diseased sentences fail to obey the T-schema, are unknowable, are unassertable, and so on, in virtue of having this status.

One thing you might expect from a linguistic demarcation of sentences is that it exhaustively characterize the paradoxical instances of T: that it correctly classify any sentence to which disquotational reasoning fails as diseased. The first section of this paper is devoted to showing that none of the theories listed above achieve this goal. Indeed, I argue quite generally that there can be no classification of sentences into healthy and diseased according to which ordinary disquotational reasoning can be carried out in cases where we are only concerned with healthy sentences. One can show, rigorously, that whatever one takes 'healthy' and 'diseased' to mean, if the theory proves that the T-schema holds for healthy sentences, it will have theorems which it will prove to be diseased. One can turn this around; if your theory of the disease is T, if T proves that each of its theorems are healthy then either T is inconsistent or it does not contain the restricted form of the T-schema for healthy sentences. This is of wide significance, I think, as a standard strategy for bracketing the semantic paradoxes in other areas of philosophy is to exclude the sentences regarded as problematic from consideration.

The impossibility of this project demonstrates that the aforementioned linguistic accounts at best give a *partial* classification of the sentences that cause problems for disquotational reasoning. There may be problematic instances of T involving a sentence which the relevant theory classifies as being disease-free. This strikes me as an important lacuna in theories that attempt to explain the failures of T in terms of features of the sentences that appear in the left hand side of that schema. Luckily there is a closely related project which I think is achievable: giving an exhaustive characterization of the failures of T via a classification of the things that appear on the right hand side of T – a classification of the truth conditions rather than the truth bearers. In the final three sections of this paper I lay the foundations for a non-linguistic theory of the problematic phenomenon responsible for the paradoxes.

Logically these two hypotheses are very different. Non-linguistic accounts of pathology seem to be very much underexplored, yet they appear to be in just as good a position as linguistic accounts to provide a solution to the paradoxes. One can still describe the conditions under which disquotational reasoning is licensed and when it isn't, one can outline which assertions are permissible and which aren't, when knowledge is possible, and so on; in

[2]One might think it incidental that these philosophers choose to theorize in terms of sentences rather than propositions, and that parallel theories could be developed for propositions instead, giving rise to non-linguistic versions of these theories. However, I think it is far from clear whether such parallel theories could be developed without assuming that propositions are effectively isomorphic in structure to sentences. Barwise and Etchemendy (1989), for example, provide an explicitly non-linguistic account of the paradoxes, but their propositions do not form a Boolean algebra and are highly non-wellfounded. At any rate, although my discussion might generalize to non-linguistic accounts that make these kinds of assumptions, the puzzles I raise here apply most clearly to linguistic theorists, and so these shall be my focus.

short none of the basic desiderata associated with an adequate solution rests on it providing a linguistic classification of the paradoxes.

Finally, let me briefly remark on the scope of this paper. My focus here shall be theories that attempt to explain the failures of T by placing some restriction on it. Of course there is the other response to the argument we mentioned at the beginning: to reject the classical reasoning that allowed us to infer arbitrary conclusions from this instance of T. For this kind of theorist the explanatory burdens seem less dire, for there are no failures of T to explain. Such theories, however, have their own costs, and thus the titular question, which is in my view still open, is of considerable interest to those who would rather exhaust the options before accepting those costs.[3]

1 Linguistic theories and revenge

Solutions to the liar paradox usually generate 'revenge paradoxes'; paradoxes structurally similar to the liar but involving the vocabulary the theorist employs to solve the liar.

Here is a standard recipe for revenge. In the course of providing a solution to the semantic paradoxes the theorist will introduce concepts which help elucidate the kind of phenomenon responsible for the antinomies. Extant examples of this strategy include the solutions listed earlier. In doing so the theorist thereby commits herself to the coherence of these notions and to structurally similar self-referential sentences involving them. If done correctly, the revenge paradox will show that the principles putatively governing the concepts employed in the solution commit the theorist to a contradiction. Theories that employ concepts like this are no better than theories that employ the original inconsistent conception of truth governed by the disquotational principles.[4]

This formula for generating revenge paradoxes is somewhat underspecified. Different theorists have different ideas about what a solution to the paradoxes is supposed to do, and therefore appear to have different commitments. It will be worth our while, therefore, to distinguish different kinds of concept which could lead a theorist into trouble. I will concentrate on two.

I take it that any solution worth its salt should identify which premise should be abandoned in a given derivation of a semantic antinomy. However it is a common goal among philosophers working on the paradoxes to want more than this – to want a *diagnosis* of what goes wrong in these derivations, and, more importantly, some sort of informative characterization of the cases where we can and where can't reason naïvely. Charles Chihara calls this the 'diagnostic problem of the paradox.'

> Alfred Tarski once remarked: "The appearance of an antinomy is for me a

[3] Non-classical approaches to the liar typically give up the law of excluded middle, $A \vee \neg A$, or the law of explosion which allows one to infer everything from $A \wedge \neg A$. Personally I am not moved by the loss of either of these principles. However I find it harder to make my peace with the other casualties of these theories: the principle $(A \wedge (A \to B)) \to B$ cannot in general be valid, nor can the inference from $A \to (A \to B)$ to $A \to B$, since both are susceptible to variants of Curry's paradox given some fairly modest background assumptions.

[4] One might want to distinguish this strong kind of revenge, which purports to show that the new concepts are inconsistent, from a weaker kind which tries to shows that the solution in question fails to correctly classify the paradoxes the new revenge sentences generate. For example one might introduce a consistent but highly restrictive notion of groundedness which is silent about what makes sentences of the form 's is grounded' grounded or not; while consistent it fails to say anything informative about the sentence which says of itself that it is not both grounded and true.

> symptom of disease." But what disease? That is the diagnostic problem. We have an argument that begins with premises that appear to be clearly true, that proceeds according to inference rules that appear to be valid, but that ends in a contradiction. Evidently, something appears to be the case that isn't. The problem of pinpointing that which is deceiving us and, if possible, explaining how and why the deception was produced is what I wish to call 'the diagnostic problem of the paradox'. (p. 590 of 'The Semantic Paradoxes: A Diagnostic Investigation,' Philosophical Review 88: 590-618, 1979) Chihara (1979)

Russell's theory of impredicative definitions, Kripke's theory of grounding, the Gupta-Belnap theory of circular definitions, McGee's theory of indefinite and definite truth, contextualist theories and Chihara's own inconsistency account of the paradoxes are some, among many, attempts to address the diagnostic problem. In Tarski's metaphor, these theories attempt to provide some informative characterization of the 'diseased' sentences, like the liar, a characteristic which is present in and responsible for the instances of the naïve theory which are inconsistent. Whatever form this takes, these theorists are committed to the coherence of a distinction between 'diseased' and 'healthy' sentences. We might call the paradoxes that are formulated using these notions the 'diagnostic form of revenge'.

I shall distinguish this from another form of the revenge paradox which instead focuses on paradoxes closely related to the notion of *assertability*. This is, for example, how Graham Priest presents the revenge paradox[5]:

> There is, in fact, a uniform method for constructing the revenge paradox [...]. All semantic accounts have a bunch of Good Guys (the true, the stably true, the ultimately true, or whatever). These are the ones that we target when we assert. Then there's the Rest. The extended liar is a sentence, produced by some diagonalizing construction, which says of itself just that it's in the Rest. (Priest (2007) 'Revenge, Field and ZF'.)

As we shall see later it will be important to distinguish these two forms. It is natural to think that if a sentence is diseased then it's not assertable – in which case it had better not be a consequence of your formal or informal theory. Given this assumption, which is widely held, the two forms of revenge amount to pretty much the same thing. However some theorists explicitly reject the connection between assertable and diseased sentences (see Feferman (1991), Maudlin (2004)); for these theorists this second form of revenge presents a separate and distinct problem.

Let me end by mentioning a third paradox that is sometimes discussed under the heading 'revenge': the strengthened liar. The strengthened liar says of itself that it's either false or gappy, where gappy means neither true nor false. Given some pretty uncontroversial logic, even by non-classical standards, this amounts to a sentence which says of itself that it's not true. The 'new' vocabulary needed to generate this paradox is therefore just the negation operator. Since everyone should be able to make sense of negation, I do not consider this to present a special problem over and above the problem of giving a consistent account of the semantic paradoxes in a language containing the usual connectives of propositional logic. Sometimes in discussions of three valued logic it is argued that there is some special kind of negation, 'exclusion negation' – having value 0 or $\frac{1}{2}$, which the theorist is committed to but cannot consistently accommodate. Perhaps the middle truth value represents an

[5]See also the presentation in Priest (2006), §1.7

indeterminate truth status caused by ungroundedness. In this case I think it would be better to assimilate the resulting paradox to the diagnostic form of revenge. Or perhaps one is a paracomplete theorist who explicitly rejects truth value gaps but who nonetheless posit assertability gaps (Soames (1999), or Field (2008).) Then one might assimilate the objection to the assertability form of revenge. Either way it is not the presence of negation in the language that poses a special problem, but the presence of these other notions.

1.1 Revenge paradoxes

The best way to frame this discussion is to assume a theory rich enough to talk about its own syntax and generate self-referential sentences (in this case arithmetic) which also contains vocabulary with which one can talk about truth and vocabulary with which one can diagnose the liar paradoxes. For example, in this simple language we can formulate a liar sentence (by the first two constraints) and say that something is wrong with it (by the last constraint.) Let $\mathcal{L}^-$ be the language of arithmetic, $\{+, ., 0, \prime\}$. We shall take $=$, $\neg$, $\vee$ and $\exists$ as logical constants and will adopt standard definitions of the remaining logical connectives. For each number n, we shall meta-linguistically represent the numeral, '0' succeeded by n '$\prime$'s by $\bar{n}$. $\mathcal{L}$ shall represent the language $\mathcal{L}^-$ augmented by the primitive predicates, Tr and H. $\ulcorner\phi\urcorner$ shall represent the Gödel numeral of the formula $\phi \in \mathcal{L}$, relative to some fixed Gödel numbering. In what follows I shall be considering theories in the language of $\mathcal{L}$ – sets of sentences that are closed under classical logic.

The predicate $H(x)$ is to be interpreted according to one's favored solution to the diagnostic problem, and should be substituted for whatever you take healthiness to be. It is assumed for the sake of argument that any satisfactory solution must at minimum classify sentences into those which are paradoxical, or 'diseased' to use Tarski's metaphor, and those which are not. The predicate H represents the sentences which are alleged not to be paradoxical, for example, 'snow is white' and 'London is the capital of France' – it may thus be substituted for 'grounded', 'semantically stable', 'definitely true or definitely false' and so on. We may also talk, when required, of the unproblematically true sentences, such as 'snow is white', but not the liar sentence or plainly false sentences like 'London is the capital of France'. Depending on the analysis, these sentences are 'grounded and true', 'stably true', 'definitely true' and so on. This notion can be represented with formula $H(x) \wedge Tr(x)$. When speaking informally I shall use the word 'diseased' to mean, by definition, 'not healthy'. We can simply represent the disease predicate in the language as $\neg H$.

The framework I have outlined allows us to evaluate a number of linguistic approaches to the diagnostic problem by appropriately reinterpreting H. For example.

1. Theories based on grounding, where $H(x)$ may be read as 'x is grounded.' See (Kripke (1975), Yablo (1982), Leitgeb (2005), Maudlin (2004)).[6]

2. The revision theory/the theory of circular definitions: $H(x)$ may be read as 'x is semantically stable.' (Herzberger (1982), Gupta and Belnap (1993), Yaqūb (1993))

3. McGee's theory of indefinite and definite truth McGee (1990). Here $H(x)$ would be read as 'either x or it's negation is definite'.

[6]Kripke is normally understood as endorsing a non-classical logic, and therefore the results below do not apply to him.

4. Contextualist or utterance based theories (Parsons (1974), Burge (1979), Williamson (1998), Gaifman (1992), Simmons (1993).) $H(x)$ can mean (depending on the theory) 'no utterance of 'ϕ' expresses a proposition' or 'no utterance of x expresses a proposition with a definite truth value'.[7]

These theories are all linguistic in the sense that it is sentences or utterances, not their truth conditions, which are the bearers of the disease. It is therefore possible to formalize these theories within the current framework. Contextualists who take utterances as the bearers of healthiness and disease may still acknowledge the distinction between sentences utterances of which are always healthy; for these theorists $H(\ulcorner\phi\urcorner)$ should be interpreted as 'every utterance of ϕ is healthy.'[8]

It should be noted that the language $\mathcal{L}$ is surprisingly modest. For most attempts to diagnose the liar paradox the predicate H, on the relevant interpretation, is only the end product. Such theories usually assume a much more substantial vocabulary such as a background language of set theory (see the grounding and revision theories) or philosophical English (such as the inconsistency theories) or a combination of the two. The problem of consistently accommodating all of this further vocabulary would be much harder, and could well leave the theorist open to new paradoxes.[9]

Other interpretations of H are possible, although we shall not consider them here. One would be a non-specific reading, in which H means 'the absence of whatever it is that is responsible for the paradoxes.' Another interesting interpretation to consider is 'would not give anyone in the set S cause to be concerned about the semantic paradoxes' for various choices of sets of philosophers S. The theorems to follow hold when H is substituted for these interpretations too.

1.2 The impossibility of a classical revenge-free linguistic diagnosis of the liar

Let me now describe to you the project of diagnosing the paradoxes in a little more detail. The hallmark of this kind of project is to identify some feature common to all problematic instances of T, and to accordingly diagnose the potential for T to fail as being *due to the presence of this feature*. Once this feature is identified we are in a position to start explaining why instances of T that have the feature are liable to lead to inconsistency, while instances that do not are not.

This is, in broad outline, the diagnostic project. A natural, albeit more specific, strategy for carrying out this project is to identify some feature of the sentences appearing in the left-hand side of T to play this role. Those who theorize with a linguistic predicate, such as the writers listed in the last section, seem to be engaging in something like this strategy. Insofar

[7] If utterances, and not sentences, are truth bearers then the discussion below will have to be modified slightly; I attend to this at the appropriate point.

[8] One might have hoped to include inconsistency accounts of the paradoxes in this list (see Chihara (1979), Yablo (1993), Eklund (2002), Azzouni (2007), Patterson (2009), Scharp (2013).) However it is not entirely clear what it means for a sentence to be healthy according to these views. (For reasons why they *should* have such a notion, see the discussion of adequate diagnoses in the next section.)

Scharp (2007), introduces a 'safety' predicate which is supposed to play the role outlined here. He introduces the concept by saying 'There are various technical ways of defining safety that I will not get into, but the rough idea is that if applying [disquotational reasoning] to a sentence leads to contradiction, that sentence is unsafe.' (Scharp's proposal is complicated by the fact that there are *two* truth predicates, and therefore two T-schemas.)

[9] We see this later with respect to the paradoxes of grounding (see Herzberger (1970).)

as you are engaging in the diagnostic project, and are attempting to do so by identifying some feature of the language involved, you should, I claim, endorse all instances of the restriction of T to sentences that don't have that feature. That is, you should endorse the following sententially restricted T-schema:

SRT $H(\ulcorner\phi\urcorner) \rightarrow (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$

If 'p' is healthy then 'p' is true if and only if p

SRT in effect guarantees that when S is disease free we are not susceptible to the paradoxes. SRT may well be derivable from other principles of one's theory of healthiness and truth, and an explicit definition of healthiness.

Why must a linguistic theorist engaged in the diagnostic project accept SRT? Take, for the sake of illustration, the notion of semantic instability. If there is a particular sentence, 'p', which is semantically stable, but is not true if and only if p, then we have misdiagnosed the problem. In Tarski's metaphor, 'p' has the symptoms of the disease – it features in a problematic instance of disquotational reasoning – but, at least according to the diagnosis in question, it's healthy. In the imagined scenario semantic instability is not the feature common to all problematic instances of T. Moreover, if this kind of scenario was possible, semantic stability would lose much of its explanatory power. The sentence 'p' is much like the liar in that the relevant instance of T fails, and so one would expect there to be a common explanation of this failure. However this common explanation cannot be semantic instability for, although the liar has it, this sentence doesn't.

To summarize, then, a linguistic diagnosis does two things: (a) provides a diagnosis of the paradoxes by identifying a feature common to all problematic instances of T, (b) more specifically, does so by identifying some feature of the language involved in those instances. To achieve both of these they must endorse SRT. Unfortunately, there is a serious problem for the project of achieving the second of these things, (b), within classical logic. If T is an arithmetical theory which satisfies two minimal constraints (is closed under classical logic and contains SRT) then T proves that its own theorems are diseased. It is natural to put this as a trilemma:

Theorem 1.1. *Let T be any set of sentences in the language $\mathcal{L}$. Then one of the following must be true:*

(i) T either does not contain some axiom of classical logic, does not contain some axiom of Peano arithmetic, or is not closed under the classical rules of inference.

(ii) T does not contain every instance of the schema $H(\ulcorner\phi\urcorner) \rightarrow (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$.

(iii) There is a sentence, γ, belonging to T such that $\neg H(\ulcorner\gamma\urcorner)$ also belongs to T. In other words, T proves that one of its theorems is diseased.

Proof. In order to show this we assume that neither (i) nor (ii) hold, and construct a sentence, γ, such that (1) γ is a theorem of T and (2) $\neg H(\ulcorner\gamma\urcorner)$ is also a theorem of T.

The sentence γ is a familiar revenge sentence, 'I'm either unhealthy or untrue', constructed via diagonalisation.

1. $\gamma \leftrightarrow (H(\ulcorner\gamma\urcorner) \rightarrow \neg Tr(\ulcorner\gamma\urcorner))$ (Diagonal lemma).
2. $H(\ulcorner\gamma\urcorner) \rightarrow (Tr(\ulcorner\gamma\urcorner) \leftrightarrow \gamma)$ by (SRT)

3. $H(\ulcorner\gamma\urcorner) \rightarrow (Tr(\ulcorner\gamma\urcorner) \rightarrow (H(\ulcorner\gamma\urcorner) \rightarrow \neg Tr(\ulcorner\gamma\urcorner)))$ by 1, 2 and weakening biconditional.
4. $H(\ulcorner\gamma\urcorner) \rightarrow \neg Tr(\ulcorner\gamma\urcorner)$ from 3 (by the classical tautology: $(p \rightarrow (q \rightarrow (p \rightarrow \neg q))) \rightarrow (p \rightarrow \neg q)$.)
5. $H(\ulcorner\gamma\urcorner) \rightarrow (H(\ulcorner\gamma\urcorner) \rightarrow \neg Tr(\ulcorner\gamma\urcorner))$ from 4.
6. $H(\ulcorner\gamma\urcorner) \rightarrow \gamma$ by 5. and transitivity of $\rightarrow$
7. $H(\ulcorner\gamma\urcorner) \rightarrow Tr(\ulcorner\gamma\urcorner)$ by lines 6. and 2. and classical logic
8. $\neg H(\ulcorner\gamma\urcorner)$ by lines 4. and 7.
9. γ by 1., 8. and classical logic.

□

We can turn this into an inconsistency argument against theories which endorse a rule of necessitation for 'healthiness.' Bear in mind, however, that the necessitation rule is much stronger than what we need to prove theorem 1.1.

Corollary 1.2. *No consistent classical theory containing Peano arithmetic that proves all of its theorems to be healthy can have SRT among its axioms or theorems. In other words, if T is classical, contains SRT, Peano arithmetic, and is closed under the following rule of necessitation, then T is inconsistent.*

***H-Nec**: If $\vdash \phi$ then $\vdash H(\ulcorner\phi\urcorner)$*

Let me stress that I do not take H-Nec to be a required of a good theory of truth and healthiness. However, this corollary is useful when considering theories that do endorse this principle.

I will shortly argue that theorem 1.1 poses a problem for the project of giving a linguistic diagnosis of the paradoxes, as described by (a) and (b), in classical logic. A moderate response to this result would be to look for an alternative diagnosis; to find a feature of the problematic instances of T that doesn't reference the sentences appearing on the left-hand side. I shall spend the rest of the paper defending one such alternative, but before I do that let me briefly consider another slightly more radical response.

The radical response is to take the above results to show that we should give up on the project of giving an exhaustive diagnosis of the paradoxes altogether.[10] Although this seems like a fairly negative conclusion, let me mention two less ambitious projects you might be engaged in instead:

(1) The project of giving a *partial* diagnosis. A feature that explains why some instances of T are liable to fail, even though there are some failures it does not explain.

(2) The project of giving a *liberal* diagnosis. A feature that is present in all instances of T that are liable to fail, but also includes some instances that are completely unproblematic.

[10]This kind of response can come in several flavors. One might think that the project is outright misguided, given the revenge paradoxes, or one might think that there is a property of healthiness out there, although it is impossible to communicate the diagnosis to anyone because it is not possible to express this property in a language.

The first kind of project leaves open the possibility of the existence *false negatives*: things that are diagnosed healthy but are really diseased. Theories falling under the second horn – condition (ii) of theorem 1.1 – could, if they wished, conceive of themselves as carrying out this type of project.[11] The second type of project leaves open the possibility of *false positives*: healthy sentences that get diagnosed as diseased. Theories falling under the third horn, condition (iii), could be understood as carrying out this type of project: some apparently healthy sentences, sentences which the theorists assert themselves, are classified as diseased.

Needless to say, I find neither of these less ambitious projects to be satisfactory, for neither really meet the explanatory burdens we began with. For example, a classic example of a liberal diagnosis would be one in which only sentences free of the truth predicate are counted as healthy. This account falls under the third horn of our theorem: theorems of classical logic that contain the truth predicate, such as 'if 'snow is white' is true then 'snow is white' is true', are classified as diseased. However this diagnosis, like any liberal diagnosis, is a deeply unsatisfying account of the problem for it does not have the resources to explain why we get into trouble with the liar in particular and not, say, the sentence "snow is white' is true' which also contains the truth predicate. It should be clear that the mere presence of the truth predicate does not satisfactorily explain why the liar is problematic. Similar things can be said about partial diagnoses.[12] Although it's certainly true that everybody is committed to something deeply unsatisfying when discussing the liar, this seems like a particularly central theoretical shortcoming.

We shall now apply theorem 1.1 to various different readings of H, with an eye to working out which horn the relevant interpretation falls under. We shall begin with the fairly generic use of the word 'pathological' within philosophy. Many philosophers use the word 'pathological' in a way that would commit them to a version of SRT. It is, for instance, commonplace for a philosopher to assert a disquotational principle and then follow it with a footnote saying 'of course, I mean to exclude pathological sentences from being substituted into this schema.' For example in Horwich (1994) Horwich writes 'the axioms of the minimal theory are all propositions of the form 'p' if and only if p – at least, those which do not fall foul of the 'liar' paradoxes'. I think one of the upshots of theorem 1.1 is that this strategy for bracketing the paradoxes is not in general safe – Horwich's principle, for example, commits him to accepting sentences which fall foul of the liar paradox. Whatever you take 'doesn't fall foul of the paradoxes' to mean, it can be substituted as H in theorem 1.1.

My real targets, however, are those who want ultimately to diagnose the paradoxes whilst retaining classical logic. In each case, I shall show that the relevant theory does not succeed in diagnosing the paradoxes by falling either into horn (ii) or (iii) of our theorem. For example in McGee (1990) McGee proposes a theory according to which the liar sentence is indefinite. A central feature of indefinite sentences is that they are unassertable and their truth is unknowable. The thought that all your commitments should be definite is captured in McGee's framework by a principle of definiteness introduction; a principle that allows you to infer that 'ϕ' is definite from ϕ.[13] If we substitute $H(\ulcorner\phi\urcorner)$ for '$\ulcorner\phi\urcorner$ is definitely true or definitely false' it is easy to see that H-Nec follows from this rule, so plugging this interpretation into corollary 1.2 it follows, given that McGee's theory contains both H-Nec

[11] Let me mention one way of justifying diagnoses with false negatives: you might think that the notion of 'healthiness' is indefinitely extensible. There is a no single comprehensive diagnosis, but rather a sequence of partial diagnoses, each one more comprehensive than the last, but each with false negatives. This is just one way of insist that there is no such thing as a general diagnosis under which all the paradoxes fall.

[12] See our earlier discussion of a sentence, 'p', that was semantically stable, but not true if and only if p.

[13] See his strong adequacy condition in chapter 8.

and is consistent, that it does not have SRT. Thus McGee's theory of definiteness is not diagnostically adequate – being definitely true or false does not guarantee freedom from paradox – and so is at best a partial diagnosis of the paradoxes.

A similar point applies to the theory of semantic stability, endorsed in various forms by Herzberger (1982), and Gupta and Belnap (1993). Here we instead substitute $H(\ulcorner\phi\urcorner)$ for '$\ulcorner\phi\urcorner$ is semantically stable'. In Gupta and Belnap (1993) Gupta and Belnap show how to consistently add this predicate into the object language. According to their construction the predicate '$\ulcorner\phi\urcorner$ is semantically stable' obeys H-Nec, so as before we may conclude that this theory does not have SRT. It follows that the theory cannot rule out the existence of a semantically stable sentence for which the corresponding instance of the T-schema is inconsistent in their theory.

Since theorists of this kind believe that their commitments should be healthy (on the relevant readings of 'healthy') they cannot assert SRT. What are we to make of these theories? On the one hand there is something quite strange about them for they are not in a position to deny assertions of the following form:

> Either the sentence 'p' is true but it's not the case that p or the sentence 'p' isn't true even though p; but despite all this strangeness 'p' is perfectly healthy according to this diagnosis.

One can mitigate this strangeness by insisting that one's theory is only a partial account of healthiness – one that is supposed to cater for the liar-like sentences but not revenge-like sentences like γ. In other words, one can simply disavow the diagnostic project as I have defined it. However this response is deeply unsatisfying, for not only is it not general enough to account for two almost completely parallel paradoxes, it also leaves one wondering what on earth the more general notion is if not definiteness or semantic stability. Lack of generality means a lack of explanatory power: one has to ask why we should be interested in the more limited notions in the first place if they cannot adequately explain the failures of the T-schema.

Note that theorem 1.1 generates problems for theories in which *utterances*, instead of sentences, are the bearers of disease; in these variants one must read $H(\ulcorner\phi\urcorner)$ as 'every utterance of ϕ is unhealthy' and $Tr(\ulcorner\phi\urcorner)$ as 'every utterance of ϕ is true'. This is pertinent for contextualist and other utterance based theories such as Parsons (1974), Simmons (1993), Williamson (1998), Gaifman (1992) and others.

A common and natural type of utterance based theory identifies the relevant notion of unhealthiness with a failure of that utterance to express a proposition. According to these theories sentences involving semantic vocabulary have a special kind of context sensitivity. It is not that they express different propositions in different contexts, it is rather that in some contexts they do not express propositions at all (although in 'good' contexts they behave disquotationally.) Theories like this are susceptible to a particularly simple revenge paradox, an instance of our general theorem, since they are committed to certain instances of the principle that if some utterance of 'ϕ' is true, then ϕ.[14] A particular instance of this can be reasoned out as follows:

> If an utterance is true then it says something.
>
> If an utterance is true and says that p then p.

[14]Of course not every instance of this schema is true. Ordinary context sensitive sentences which express different propositions in different contexts do not conform to this schema.

> If an utterance of 'no utterance of L is true' says anything at all it says that no utterance of L is true
>
> Therefore: if some utterance of 'no utterance of L is true' is true, then no utterance of L is true.

Here L = 'no utterance of L is true'. By the above fact it follows that if some utterance of L is true, no utterance of L is true. Thus, by reductio, it follows that no utterance of L is true.

Anyone who endorses these premises is committed to the claim that no utterance of L is true, and therefore should surely be able to assert the fact that no utterance of L is true. Yet when the theorist tries to express this commitment, by uttering the sentence 'no utterance of L is true' she fails. Since according to her own view there are no true utterances of the sentence 'no utterance of L is true' – even her very own utterances of this sentence fail to be true; presumably by failing to express a proposition.

Note, of course, that obvious analogies can be drawn between all of the paradoxes we have discussed in this section, and paradoxes discussed in the literature. Existing treatments of revenge, however, tend to be quite piecemeal, with individual paradoxes resting on particular features of the account of healthiness. The important moral of this discussion is that each of these arguments are instances of a general theorem: the problem is endemic to theories that attempt to exhaustively classify the problematic instances of T by the sentences appearing in the left-hand-side, and is therefore not a problem that can be solved by searching for a more sophisticated interpretation of 'healthy'.

Let me end my discussion by considering those who explicitly disavow the connection between assertable and diseased sentences (see for example Feferman (1991), Maudlin (2004).[15]) The interest of theorem 1.1 is rather that adopting a linguistic diagnosis of the paradoxes *forces* us to prize these two notions apart. For those who already distinguish them, it is hardly a troublesome consequence.

As it happens, those who identify the disease with ungroundedness tend not to worry about asserting ungrounded sentences. For example, they will typically assert 'the liar is neither true nor false', a sentence which is, by their own account, both untrue and ungrounded (Maudlin (2004).)

Since ungroundedness is typically taken to preclude truth, to endorse this kind of theory one must assert untrue sentences. In general, then, there is a challenge for this kind of theorist to explain what they're doing when they assert. Only Maudlin has attempted to do this (Maudlin (2004) chapter 5), but as Priest observes Priest (2005), the theory of assertability seems quite arbitrary; you're not allowed to assert an ungrounded sentence if it is atomic, but you may assert an ungrounded sentence if it is the negation of an atomic sentence. In Priest's words 'truth is the aim of assertion. Once this connection is broken, the notion of assertion comes free from its mooring, and it is not clear why we should assert anything.'

Secondly, and more importantly, there are other revenge paradoxes which these theorists do not escape. In distinguishing between assertable and diseased sentences they distinguish the diagnostic form of revenge from the assertability form of revenge outlined at the beginning of this section. In making this distinction they are forced to treat both paradoxes

[15]Maudlin notes that some classical inferences do not preserve truth on this view, and so chooses to call these inferences 'invalid'. However theorem 1.1 still applies to Maudlin because he still asserts and reasons as though he accepted classical logic – one can simply understand T to represent the set of sentences Maudlin asserts.

separately. Maudlin attempts to address these in Maudlin (2007), but does so at the cost of giving up his theory of assertability.

As well as the assertability form of revenge, there are paradoxes of grounding which do not rely on the principle of necessitation. Herzberger shows that there are sentences, s, which are grounded by all and only the grounded sentences; for example he suggests that the sentence 'every grounded sentence is true or false' is grounded by all and only the grounded sentences. But (i) if everything that grounds s is grounded then s is grounded and therefore, (ii) since s is grounded by *all* grounded sentences s grounds itself. It follows that s grounds itself and is thus ungrounded after all, a contradiction.[16]

Theorem 1.1 is thus certainly not the last word on the matter. One can get around it by allowing oneself to assert diseased sentences. However such theories are susceptible to the assertability paradoxes and, at least in the case of extant theories of this type, the paradoxes of grounding.

2 An outline of a non-linguistic theory

In the last section I outlined some very general problems. Now I want to sketch what I think is the most promising avenue for avoiding these paradoxes. The key idea involves rejecting the assumption that the distinction between the healthy and unhealthy instances of the T-schema must be grounded in a distinction between the kinds of sentences we can substitute into the left hand side of the T-schema. The T-schema associates sentences with truth conditions; on the opposing view it is rather the truth conditions, not the sentences, that must be vetted for disquotational reasoning. In order to properly understand this it is crucial to have two assumptions out in the open. These are:

CL. Classical logic.

CS. If s says that p then s is true if and only if p.

The first assumption I list only to emphasize the scope of this paper; one can certainly avoid theorem 1.1 weakening the logic, but it is of interest to know whether we can diagnose the paradoxes without weakening classical logic. CS, on the other hand, encodes a principle (practically a definition) of classical semantics: that what a sentence says (or means or expresses) is its truth conditions. Whatever these are, they must *at least* be materially necessary and sufficient for the sentences truth.[17]

If one can make sense of quantification into sentence position then the expression 'true' can be defined in terms of 'says that': s is true just in case for some p, s says that p and p. If one made a standard assumption in classical semantics, namely that every sentence

[16]This is known as Herzberger's paradox (Herzberger (1970)), which is closely related to Mirimanoff's paradox concerning the set of all well-founded sets Mirimanoff (1917). See also, for example, Leitgeb (2005) Lemma 14.9: 'There is no sentence $\phi \in \mathcal{L}_{Tr}$ s.t. ϕ depends on Φ_{lf} essentially (that is, on the set of sentences that depend on non-semantic states of affairs).' Leitgeb uses this to show that one cannot define groundedness (Φ_{lf}) arithmetically. However it seems to show that one couldn't even define the set of grounded sentences, for a set theoretic language, set theoretically at all. Leitgeb, in personal communication, has further told me that he does not think that one could accommodate a binary grounding relation in one's object theory, even though it may be possible to have a groundedness predicate in the object language. See also McGee, theorem 5.11, which also seems to suggest that one cannot formulate an informative definition of grounding in the object language.

[17]For simplicity I have ignored context sensitivity, however a principle similar to CS presumably must hold once both saying and truth are relativized to a single context.

means at most one thing, then one can prove CS from this definition.[18] In summary, then, CS is partly substantive – to ensure its truth one needs to assume that every sentence says at most one thing – and partly verbal – it is a matter of choosing to use the words 'true' and 'says that' in a related way.

While it is well known that classical logic proves that we cannot have all instances of the T-schema,

T. 'ϕ' is true if and only if ϕ

it is less often observed that the disquotational saying schema

S. 'ϕ' says that ϕ

is also problematic. To my knowledge this was first shown in Prior and Prior (1971). Prior's argument required one to be able to quantify into sentence position and required the assumption, mentioned above, that every sentence says at most one thing.[19]

In fact you can refute S directly from CS without invoking anything as controversial as quantification into sentence position. First note that we can prove that no sentence can say of itself that it is not true.[20] For if s said that s is not true then by CS s would be true if and only if it is not true. Secondly note that if we considered the sentence L='L is not true', the saying schema, S, would entail that L says that L is not true, and this is the kind of thing we have just shown to be impossible.[21]

Much of what I say about sentences and truth conditions or operators and predicates will make little sense if the reader assumes they are related by schemata such as S or T. It is therefore absolutely crucial in what follows to bear in mind that neither T nor S hold unrestrictedly.

A straightforward way to make sense of this result would be if some sentences didn't say anything at all: if 'p' doesn't say anything then, in particular, it doesn't say that p. Some of what I say later is compatible with this interpretation of the result that S and T have false instances. However it is not a sensible interpretation given everything I have said so far – it is, in effect, a version of the 'no-proposition' response to the semantic paradoxes, which is a linguistic theory (and thus appears to be subject to theorem 1.1), and is often noted to be subject to the problem that statements of the view itself don't express propositions.

The alternative view, which also rejects S, maintains the assumption that each sentence says something, although in some cases denies that the proposition the sentence 'p' says is the proposition that p. In fact, the view that there are certain sentences that express a proposition but not the 'expected' one, is practically forced on us by theorem 1.1.

One might at this juncture wonder what the liar does say, if not what it seems to say. Questions like this will need answers eventually, however for now let me just make the point that we shouldn't require an adequate theory to tell us what every sentence means. Even

[18]Indeed, the principle that every sentence means at most one thing is unnecessarily strong. One needs only the substantially weaker schema: if s says that p and s says that q then p if and only if q.

[19]The premises of the argument do not preclude the possibility that some sentences don't say anything. Although Prior doesn't mention this explicitly, it actually only requires the assumption that each sentence only says materially equivalent things.

[20]This contradicts a common way of reporting what self-reference achieves. Self-reference cannot produce a sentence that says of itself that it is not true, for I am claiming that is logically impossible, although it can produce a name for the sentence that is identical to the result of concatenating that name to the string ' is not true'.

[21]Note that, although we know L doesn't say of itself that it's not true, we don't have any reason to think that there aren't other sentences that say that L isn't true.

the cases in which sentences say what they seem to say record empirical facts. For example, while the sentence 'snow is white' says that snow is white, this fact depends on how we use the word 'snow' and 'white'. The liar sentence is slightly different in that we can rule some answers out *a priori*, but even so, to determine what it does say we'd need to pay some attention to how people use certain semantic vocabulary, and for our purposes in the following, this is a question we can be relatively neutral about.[22]

2.1 Unclear truth conditions

Despite the paradoxes we have just discussed, it is hard to deny that there is a distinction between two kinds of sentence. Some sentences are clearly true, 'snow is white' for example, and others are neither clearly true nor clearly false, such as the liar sentence or the truth teller. None of the paradoxes we consider cast doubt on the existence of such a distinction. However, whatever clear truth is we know it cannot satisfy both SRT and a necessitation principle.

Taking these remarks at face value, 'clearly' does not function as a predicate, and it is not sentences that are clear or unclear. In 'S is clearly true' 'clearly' functions as an adverb, modifying the truth predicate just as 'possibly' does in 'S is possibly true', and 'not' in 'S is not true'. In philosophical discourse we often regiment adverbial modification by way of a sentential operator, so that for example 'Hector is a possible skater' can be regimented as 'it's possible that Hector is a skater'. This way we may apply clarity, possibility, negation, and so on even to sentences that are not in subject predicate form.

'True' is not the only predicate that 'clearly' modifies, thus while clear truth certainly is linguistic, in some sense, clear tallness, say, need not be. Consider

1. Snow is clearly white.

2. Clearly, snow is white

Neither 1, nor its regimentation using an operator expression, 2., seems to be about language in any way. Prior made this point very succinctly in the case of tense operators, noting that 'When a sentence is formed out of another sentence or other sentences by means of an adverb or conjunction, it is not *about* those other sentences, but about whatever they are themselves about.' (Prior (1993).)

Unless the proposition that p itself happens to be about language, when we say that it's clear that p we are not talking about language any more than we would be if we said it's necessary that p. In the contexts where we will apply this distinction, however, the proposition that p will often concern language so it pays to be especially careful about the distinction. Consider

3. 'Snow is white' is clearly true

4. Clearly, 'snow is white' is true.

[22] A natural extension of the theory DFS I develop in later sections, however, does tell us something about what L says. Although L does not say that L is not true, there will be some F which is 'truth-like' in the sense that a disjunction is F if and only if one of the disjuncts is F, a negated sentence if F if and only if the sentence isn't F and so on, and moreover L says that L isn't F. The relation between the proposition that L expresses and the proposition that L isn't true can also be modelled quite naturally within the theory of symmetries I develop in Bacon (MS).

While it seems that 1 is equivalent to 3 and 2 is equivalent to 4, 3 and 4, unlike 1 and 2, concern English sentences and so these equivalences are at best contingent, since we could have used 'snow is white' to mean something unclear. Even if 'snow is white' had meant something unclear, snow would still be clearly white.

I hope to have introduced the notion of clarity in such a way that we may have a neutral understanding of the notion without having a theory of it. Anyone who has thought about the paradoxes enough to recognize the difference between the claim that snow is white and the claim that L is true, where $L =$ 'L is not true', should be able to understand the distinction. And the regimentation of this distinction using an operator expression seems reasonable given that other non-theoretical ways of introducing the distinction involve some kind of adverbial modification of 'true'. For example, we might said that 'snow is white' is definitely true, straightforwardly true, determinately true and so on.

Unclarity, on this picture, is not something sentences possess but something truth conditions possess. A sentence may possess unclear truth, or may express an unclear proposition but may not be unclear itself; thus if I say that Harry is bald then it is what I have *said* that is unclear, not the sentence I used to say it. With this distinction in hand a subtle assumption being made in diagnoses of the naïve theory of truth is revealed. Consider

'p' says that p in English

'p' is true in English if and only if p

The left hand part of both of these equations concerns language, whereas the right hand parts concern the world in some sense; together they match sentences with their truth conditions. The assumption being made is that we shouldn't accept instances of these schemata which involve sentences with certain disquotation preventing features; for example we should not accept instances in which the left side involves an ungrounded sentence, or a semantically unstable sentence. That is, we are placing a 'language side' restriction as opposed to a 'world side' restriction.

But what of the other possibility? Might there be truth conditions which are not suitable for standing in an 'enquotational relation' to a sentence?[23] Could it be that unclarity in a proposition sometimes prevents it from being expressed by the sentences you would expect it to be expressed by, and unclear propositions are therefore unsuitable for disquotational reasoning?

I do not want to adjudicate between possible explanations of the unsuitability of certain instances of the disquotational schemata. However, it is clear that there is an unoccupied position in logical space that identifies the truth conditions, rather than the truth bearers, of the naïve theory as culprits: it is, for example, the proposition that the liar isn't true, not the liar itself, that lacks the feature that is distinctive to safe applications of the T-schema.[24] So it is of great relevance that on this kind of view standard revenge arguments do not motivate any theory with the form of SRT which tells us when a *sentence* can have disquotational truth conditions:

SRT. If 'p' is clearly true or clearly untrue then 'p' is true if and only if p.

[23]This is a figure of speech – I do not think there can be such a relation in general, but it is clear what is meant by this in particular instances: that the sentence 'p' is true if and only if p.

[24]Note that for all I've said, the liar sentence might express a proposition (indeed, this is the view I prefer). However, we know, from our considerations of schema S and T, that whatever proposition that is, it is not the proposition that the liar isn't true. This is why it is important to distinguish between whether the 'culprit' is the sentence 'L is not true' or the proposition that L is not true.

They motivate instead principles which tell us when the truth condition for a sentence is 'enquotational':

RT. When it is clear that p or it's clear that not p, then 'p' is true iff p.

RT simplifies a little: assuming the factivity of 'it's clear that', RT is equivalent to the conjunction 'if it's clear that p then 'p' is true and if it's clear that not-p then 'p' isn't true' with the latter conjunct being redundant if we assume that nothing with a true negation is true.

When theorizing about the liar one cannot be too careful about use and mention. Here the pedantry pays off: RT is not equivalent to SRT, for it is enquotational in nature rather than disquotational. The most one can prove from RT, for example, is that if a sentence 'p' is clearly true or clearly untrue then "p' is true' is true if and only if 'p' is true.

What other principles should we expect to govern the notion of clarity? One constraint, which we motivated in our discussion of assertability in the last section, is that it is in general bad to believe and assert unclear propositions. The necessitation rule, Nec, ensures that everything the theory commits you to is clear.[25] What is clear is closed under classical logic, so we should endorse the closure principle for clarity, K, below. Finally note that the informal notion of clarity is factive, so we add the schema T. The resulting modal system is called KT which contains formal principles corresponding to the following (stated in English)

K If it's clear that if p then q and it's clear that p then it's clear that q.

T If it's clear that p then p.

Nec. If you can prove that p, then you can prove that it's clear that p.

Call the logical system you get by adding suitably formalized analogues of RT to KT: BCT – the basic logic of clarity and truth. BCT looks like it should be susceptible to versions of theorem 1.1. However it is not: BCT is a consistent theory (BCT is a fragment of the theory whose consistency we prove in theorem 2.1; see Appendix.)

Given this fact it is of course natural to wonder what happens with the revenge sentences. Let's introduce the name R for the sentence 'R is not clearly true'. The first thing to observe is that the standard way of introducing the revenge problem, described below, rests on a confusion of use and mention in the present framework:

> Suppose that R is clearly true. Since we can disquote clear truths, and since R='R is not clearly true', it follows that R is not clearly true. This contradicts our assumption.
>
> So R is not clearly true. But we've just proved the sentence 'R is not clearly true', which is just R, so R is clearly true. Contradiction.

The mistake in this argument is found in the line 'we can disquote clear truths'. If it is clear that 'p' is true then RT licenses the enquotational inference to "p' is true' is true, but *not*

[25] Note that the appropriateness of a necessitation rule like this depends crucially on which principles we take to be part of the background theory. We will later expand the background theory to include some clear truths that are not clear at all orders. If we extended the necessitation rule to this background theory, we would be able to prove that each theorem of the theory is clear at all orders, and so the unrestricted necessitation rule will not be appropriate in this expanded context.

the disquotational inference to p.[26]

The most natural way to modify the argument so that we can apply RT is to instead assume that it's clear that R is not clearly true (that is, instead of assuming that 'R is not clearly true' is clearly true.) The argument would then proceed as follows

> Suppose that it is clear that R is not clearly true. Then, of course, R is not clearly true. But since we can enquote in clear cases, we also have 'R is not clearly true' is true. 'R is not clearly true' just is R so R is true. So we have that R is true but not clearly true.

Is this a contradiction? No, we know that there are true propositions which are unclear: either the liar is true or it isn't, but either way it is unclear. Of course, no-one is forced to endorse the claim that R is true but not clearly true on the basis of this argument – the point is that nothing contradictory follows from the assumption that it is clear that R is not clearly true. It seems, therefore, that once we are careful about use and mention and are careful to distinguish enquotational principles from disquotational principles, the revenge paradox does not get off the ground.

2.2 Can we recover sentential clarity?

It is natural to be puzzled by the fact that such a slight change in vocabulary – the use of an operator rather than a predicate – can allow us to avoid the revenge paradoxes. A good place to begin easing this feeling of puzzlement would be to work through what happens to the sentential predicate that states that a sentence is clearly true in the present theory. This can be defined in this theory by simply composing the clarity operator with the truth predicate.

Note that if you can prove a sentence in BCT, you can also prove that it is clearly true. Since BCT is consistent, the conditions needed to apply corollary 1.2 hold: BCT does not prove a version of SRT restricted by clear truth. Thus we cannot rule out the possibility that there is a sentence, 'p', such that (i) 'p' is clearly true but (ii) it's not the case that ['p' is true if and only if p]. Clear truth, therefore, cannot play the diagnostic role of exhaustively identifying the problematic instances of the T-schema that I have set out to provide.

This is perfectly fine by my lights: I never intended clear truth to provide such a classification – this is achieved in my theory by a restriction on the right-hand side of the T-schema stated in terms of the clarity operator, rather than a restriction on the left-hand side by the notion of clear truth. However this observation does highlight the fact that there is an important difference between saying that 'p' is clearly true and saying that it's clear that p.

The general observation that this can happen should not be too surprising. We already have reason to suspect that 'it's clear that p' and 'it's clear that 'p' is true' can come apart. To move from one to the other is, after all, just to employ an instance of naïve reasoning which we already know to be problematic: the inference that allows the free substitution of 'p' for "p' is true' and vice versa in arbitrary formulae. (This substitution rule would allow us to infer the T-schema, "p' is true if and only if p' from the tautology 'p if and only if p'.)

[26] How might disquotation for clear truths fail? 'p' would be a clear truth if, say, it clearly said that q, and it was clear that q. Thus 'p' would be true if and only if q. But without assuming the saying schema, S, we cannot assume that the proposition that p and the proposition that q are identical, or even materially equivalent. In particular, we cannot conclude that 'p' is true if and only if p from the fact that 'p' is a clear truth.

One might still be worried on the grounds that there ought to be at least *some* predicate which has the same logical role as each operator, even if it is not clear truth in the case of the clarity operator. Surely, one might think, there ought to be a predicate – some notion of being a clear sentence – that satisfies the schema: 'p' is a clear sentence just in case it's clear that p.

This style of reasoning is also easily dispensed with. It is in general incorrect to think that one can find a predicate corresponding to every operator. Take negation for example: there is no predicate corresponding to the negation operator. If $F(\ulcorner\phi\urcorner)$ were such a predicate then there would be a sentence, γ, such that $F(\ulcorner\gamma\urcorner) \leftrightarrow \gamma$ is a theorem of arithmetic by the diagonal lemma. If this predicate really did correspond to negation we could conclude that $\neg\gamma \leftrightarrow \gamma$, which is impossible.

A similar worry one might have in the vicinity is that it at least *appears* as though there is a general way to introduce properties of sentences which commit you to SRT; a way of introducing the property that does not depend on your preferred way of diagnosing the paradoxes. Here is a naïve, albeit ultimately flawed way to introduce such a property (we shall attempt to refine it later): the property a sentence, 'p', has when it is true if and only if p.[27] This at least appears to be a property of sentences, but yet this property appears to commit us to SRT. Even someone who disavows the project of diagnosing the paradoxes by classifying sentences is committed to this notion, or so it seems, and therefore everyone is committed to something like SRT, and the problems that come with it.

I think that the above thought ultimately rests on an abuse of our convention regarding the use of schematic letters; we are clearly not using it in the above as merely short hand for writing a number of structurally related sentences. However, although the suggestion shouldn't really be expressed in this way, one might ask whether there is a coherent thought in the vicinity that is being gestured at. One way to make it precise would be to introduce quantification into sentence position, and define a sentence, S, as healthy iff, for some p, S = 'p' and S is true if and only if p. Note, however, that this proposal involves quantifying into quotation marks – the usual way of treating quantification into sentence position would only bind the last occurrence of 'p', since the occurrence in quotation marks is not free.[28] Alternatively, one might try an infinite disjunction: 'either S = 'snow is white' and S is true if and only if snow is white, or S = 'L is not true' and S is true if and only if L is not true and ...' where there is a disjunct for every sentence of the language. However, in standard infinitary languages sentences are well-founded: there cannot be a disjunction, ϕ, such that, for every formula of the language, ψ, ψ appears as a subformula of a disjunct of ϕ, for otherwise ϕ would be a proper subformula of itself ϕ. Lastly, one might consider the property a sentence, 'p', has when the disquotational sentence "'p' is true if and only if p' is true. Unlike the above proposals, this is a perfectly well-defined property of sentences, however we cannot derive any instance of SRT under this interpretation unless we can move between "'p' is true if and only if p' is true' and "p' is true if and only if p'; these inferences are instances of the T-schema and are not in general valid.

These arguments are all fairly abstract. There are other, completely independent, reasons for thinking that no predicate of sentences can play the right diagnostic role. To demonstrate this, let T be any sentence that means in German that T is true in German. For my purposes

[27] Another variant would be, the property a sentence, 'p', has when it says that p. This suggestion is due to an anonymous editor.

[28] The alternative is to try and make sense of quantifiers that can bind both inside and outside of quotation marks. On the face of it this seems incoherent, however this idea is explored by Wray in Wray (1987). However, no such theory will be consistent with classical logic.

it does not matter what T is (setting T equal to the German sentence 'T ist auf Deutsch wahr' would do.) T is a truth-teller for German, much like the sentence $T' =$ 'T' is true in English' is a truth-teller for English.

Is T true in German or not? The answer, one would have thought, is that it's unclear in the same way that it's unclear whether T' is true in English. Note that in classifying these two cases as alike we relied on the fact that what is unclear in each case is not a sentence of English or German, but the proposition that that sentence is true in English, or German, respectively. The adverb 'clearly' can modify any predicate, for example 'white' in the earlier example, but also 'true in English' and 'true in German'. Thus we may say:

T is not clearly true in German.

T' is not clearly true in English.

The thing that is unclear is whether T is true in German – it is not the German sentence T itself that is unclear. The former is the kind of thing that can be said, rather than the kind of thing that is used to say things. It is not easy to see how this important parallel between the German and English truth-teller could be made using the predicate 's is definite in English.' The claim that T is indefinite in English, the obvious parallel to the claim that T' is indefinite in English, will not do: T doesn't contain a reference to a sentence of English so it is not indefinite in English.[29]

2.3 A theory of clarity

In this section I introduced the word 'clearly' in a non-partisan way, intending the discussion to be neutral between the many different interpretations we could substitute for 'clearly.' While everything I said there would be compatible with a purely epistemic reading of 'clearly', for example, I shall now turn to developing my own positive theory of clarity.

The kind of phenomena that we are interested in includes not just paradoxicality, characterized by failures of the various disquotational schemata, but the more general kind of phenomenon associated with non-paradoxical, but somehow factually defective discourse involving truth. Consider for example:

L_1 L_2 is not true in English

L_2 L_1 is not true in English

While there is no inconsistency involved here, even if we assume the two relevant instances of the disquotational schema, something is not quite right. There seems to be no plausible semantic constraints, including the relevant instances of T, which would decide whether or not L_1 or L_2 is true or not. Furthermore the non-semantic facts do not help decide whether they're true or not either. I suggest that there is no fact of the matter whether L_1 or L_2 are true or not.

Once we have seen that some discourse involving truth is factually defective it is natural to ask whether talk involving paradoxical liar-like sentences can be non-factual too. It is

[29] We could break the parallel between the English and German truth-teller by instead saying 'T is true in German' is indefinite in English, or by saying that T is indefinite in German, but these are subject to other problems. For example the former would not encapsulate the idea that T is not clearly true in German: T would remain not clearly true in German, even if English speakers used 'T is true in German' in such a way as to make it definite – perhaps by using it to mean that snow is white.

perhaps easier to evaluate whether cases like the truth-teller and L_1 and L_2 involve non-factuality since there are multiple truth values we could assign that are consistent with any plausible semantic constraint, including the relevant instances of the T-schema. Once we have given up the T-schema, however, isn't any truth-value assignment game? I think this would be too quick: the paradoxes provide no reason to give up on the constraint that possible truth-value assignments obey the normal compositionality clauses for the extensional connectives, and indeed, these constraints follow from the supposition that everything expresses a proposition, and CS. These constraints would guarantee that the liar was either true or false and not both, but they would not tell us which, inviting the same thoughts we had about the truth-teller. I think it is natural to think that the truth value of the liar sentence is also unconstrained. Just like the truth-teller it seems unclear what kind of evidence would settle the question of whether or not the liar is true, it seems like we shouldn't assert that it's true or that it's false, and there generally doesn't seem to be any fact of the matter concerning its truth or falsity.

The kind of predicament we find ourselves in when we have all the evidence we could have about a subject matter and we still seem unable to decide whether p is often associated with cases in which it is borderline whether p. A natural conjecture is that there is a general phenomenon, indeterminacy concerning that subject matter, which encompasses both the kinds of situations we find ourselves when confronted with questions about whether T true in German or whether L_1 is true in English, and so on, and with questions about whether a borderline bald man is bald or not. A characteristic feature of cases of indeterminacy is that they involve an inability to know whether p even when all the facts are available to you. Furthermore there is a certain kind of irrationality involved when someone is strongly opinionated about something they think is indeterminate, or cares intrinsically about the indeterminate. Facts such as these situate indeterminacy within a theory of rational propositional attitudes. The source of my ignorance concerning whether T is true in German, my inability to rationally care, believe, and so on, that T is true in German are not facts about the German or the English language – I need not be acquainted with either language. Indeterminacy therefore a more specific way of understanding the notion of clarity that I introduced informally. Moreover satisfies one of the desiderata for providing an interpretation of clarity, as introduced informally in section 2.1: it is a theory of *propositions* and provides us with a way of assessing which propositions are suitable truth conditions to put in the disquotational schema.

This way of thinking about things is perhaps more familiar in the philosophy of vagueness. While the majority of theories of vagueness appear to be linguistic theories – theories that identify vagueness with some linguistic property (examples falling under this this camp include semantic indecision theories (e.g. McGee and McLaughlin (1995)), metalinguistic safety accounts (Williamson (1994)) and inconsistency theories (Eklund (2005))) – there is precedent for the operator view in Field (2000) and Fine (1975).[30]

So what further things do we need to say about clarity to guarantee that it is a species of indeterminacy as described above? Perhaps a fully adequate account would provide an explicit definition of clarity. We do not actually need to be this specific: we can introduce this operator implicitly to someone who doesn't understand it by outlining its inferential properties and the role it plays in a theory of rational propositional attitudes. A full defense

[30]Note that, although Fine is routinely cited in connection with linguistic accounts of vagueness, his formal theory and informal explication of the technical terms indicate that an indeterminacy operator is part of Fine's basic ideology, as a opposed to a linguistic definiteness predicate (even if the latter might be explained or defined in terms of the former by combining it with a linguistic truth predicate).

of this theory is developed elsewhere (see Bacon (MS)), so I shall simply present it below; little of what I will say depends on the specific details.

> Logical The operator 'it's clear that p' is governed by the modal logic KT, or some extension of it:
>
> > Closure If it's clear that if p then q, and it's clear that p, then it's clear that q.
> >
> > Factivity If it's clear that p, then p
> >
> > Necessitation If one can prove p from classical logic and these three principles, one can prove that it's clear that p.
>
> Alethic Clarity licenses disquotational reasoning. If it's clear whether p, then 'p' is true in English just in case p.[31]
>
> Epistemic If it's unclear whether p, then it's not rationally known whether p.
>
> Doxastic No rational person who has a credence of 1 that it's unclear whether p assigns a credence of 1 or 0 to p.
>
> Pragmatic
>
> > Assertion If it's unclear whether p then you are not in a position to assert that p and you are not in a position to assert that $\neg p$.
> >
> > Questions It is not permissible to ask whether p if you have been told that it's unclear whether p.
>
> Bouletic It is not rational to care intrinsically about the indeterminate: if p is a maximally strong clear proposition, you should be indifferent between any two propositions entailing p.
>
> Attitudes No rational person who believes that it is unclear whether p wonders whether/hopes that/fears that, [...] p.

Since the above aspects are distinctive to the questions we take there to be no fact of the matter about, call the above principles the determinacy role. The fact that unclarity satisfies it suggests that unclarity is a species of indeterminacy.

The role described above is not the only reasonable way to go about developing a theory of indeterminacy. In Field (2000) Hartry Field argues for alternative constraints on our doxastic attitudes. He argues among other things that:[32]

> Reject If you know that it's unclear whether p you should reject p and reject $\neg p$.
>
> Credences Your credence in p should be the same as your credence in Δp. In particular if $Cr(\nabla p) = 1$ then $Cr(p) = Cr(\neg p) = 0$.

[31] By 'it's clear whether p' I just mean 'either it's clear that p or it's clear that not p'.

[32] Unfortunately, Field's theory is not suitable for my purposes because his theory of attitudes seems to commit him to the $\mathsf{4}$ principle for determinacy, which is inconsistent in our theory (see section 6.) Field's theory requires that one's credence in Δp match ones credence in $\Delta\Delta p$.

Here Δ is to be understood as 'it's clear that' and ∇ as 'it's unclear whether'. The role I have described states that when you are certain that it's unclear whether p you should have an intermediate credence, whereas according to Field we should have a credence of 0 in both p and its negation. Abstracting from the details: both proposals provide us with a non-linguistic account of unclarity that would provide a very natural interpretation for the clarity operator. One advantage Field claims in favor of his constraints is that they apparently rule out a purely epistemic interpretation of the clarity operator – a theory in which unclarity is just an incurable kind of ignorance – for mere ignorance does not require one to have the pattern of credences predicted by Field's theory. In my theory this is achieved by the principle Bouletic – one may clearly care intrinsically about things you are ignorant about.[33]

It is worth pointing out that the above theory rules out other possible interpretations of the word 'clearly' that would seem to trivialize the theory. For example, the theory BCT, and the richer theories we consider in the next section, are all consistent with the uninformative interpretation of 'clearly p' as 'p and 'p' is true in English'. There is, of course, an apparent problem about quantification into the clearly operator on this interpretation, although this might well be surmountable. However, given the determinacy role we can decisively rule this interpretation out. For example, a monolingual Chinese speaker who knows perfectly well that snow is white, might be under the misapprehension that the sentence 'snow is white' is not true in English. If the suggested interpretation were correct, such a person would be in a position to deduce that it was unclear whether snow is white. But given the determinacy role, it would follow that she should be uncertain whether snow is white, should refrain from asserting that snow is white, shouldn't wonder whether snow is white, and so on. I take it that someone in the situation described is not committed to any of these things, and that we cannot interpret 'clearly' in this simplistic way if it is to accord with the determinacy role.

2.4 The theory DFS

The determinacy role stipulates that clarity interacts with truth in a certain way: it requires that it obey the basic theory of clarity and truth, BCT, we mentioned in section 2.1. However our remarks in section 2.3 motivated a much stronger compositional theory of truth and clarity which we shall now formulate.

Let $\mathcal{L}^-$ be the language of arithmetic with a function symbol for each primitive recursive function f. Let $\mathcal{L}$ be $\mathcal{L}^-$ augmented with a truth predicate, Tr, and primitive clarity operator, Δ. We shall retain the conventions we set in place in section 1 regarding Gödel numbering. The standard 'dot' notation shall be adopted for representing the syntactic operations; for example, I shall write the function taking the Gödel number of a sentence to Gödel number of its negation $\dot{\neg}$, and so on. If x is the Gödel number for ϕ then I shall write $x[\dot{y}/z]$ for the Gödel number of $\phi[y/z]$. The numeral for the Gödel number of a formula ϕ is written $\ulcorner\phi\urcorner$, and there are primitive recursive functions $\dot{\neg}$ ($\dot{\wedge}, \dot{\vee}$ and so on) taking the Gödel number of a formula to the Gödel number of the result of prefixing negation to that formula (and similarly for $\dot{\wedge}, \dot{\vee}$ etc.) In arithmetic we can define predicates $Sent$, At and Ver saying that a number is the Gödel number of a sentence, atomic arithmetical sentence and an atomic arithmetical truth respectively. With this in place we can formulate a class of theories that that will be the focus of our discussion: I will present a sequence of theories, which I'll call DFS_n for each n, which get successively stronger as n increases. I will begin by

[33] I defend this feature of the theory elsewhere. See Bacon (MS).

describing the 'limit' of these theories which I've named DFS, which contains each of these theories. Although DFS has some nice properties I will later discuss some reasons to go for one of these weaker theories. The weaker theories can be obtained from DFS by restricting the necessitation principles mentioned below.

PA. Peano arithmetic including full induction (i.e. the induction schema may take formulae containing the truth predicate and the clarity operator.)

ΔNec If $\vdash \phi$ then $\vdash \Delta\phi$

K $\Delta(\phi \to \psi) \to (\Delta\phi \to \Delta\psi)$

T $\Delta\phi \to \phi$

BF $\forall x \Delta\phi \to \Delta\forall x\phi$

RT $(\Delta\phi \vee \Delta\neg\phi) \to (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$ when ϕ is closed.

At. $\forall x(At(x) \to (Tr(x) \leftrightarrow Ver(x)))$

C$\to$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x \dot{\to} y) \leftrightarrow (Tr(x) \to Tr(y)))$

C$\vee$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x \dot{\vee} y) \leftrightarrow Tr(x) \vee Tr(y)))$

C$\wedge$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x \dot{\wedge} y) \leftrightarrow Tr(x) \wedge Tr(y)))$

C$\forall$. $\forall x(Sent(x(\bar{0}/v)) \to (Tr(\dot{\forall} vx) \leftrightarrow \forall y Tr(x[\dot{y}/v])))$

C$\neg$. $\forall x(Sent(x) \to (Tr(\dot{\neg} x) \leftrightarrow \neg Tr(x)))$

CΔ. $\forall x(Sent(x) \to (Tr(\dot{\Delta} x) \leftrightarrow \Delta Tr(x)))$

Nec If $\vdash \phi$ then $\vdash Tr(\ulcorner\phi\urcorner)$[34]

Conec If $\vdash Tr(\ulcorner\phi\urcorner)$ then $\vdash \phi$

Let me begin by citing an important result (for the proof see the appendix.)

Theorem 2.1. *The system* DFS *is consistent.*

In fact one can go further and show that it doesn't prove any false arithmetical sentences. If you restrict induction to its arithmetical instances it proves no new arithmetical theorems.[35] It's also consistently augmentable with propositional quantification (and full second order logic.)

DFS has as its truth theoretic basis the compositional theory of truth known as FS (see Halbach (1994)) which consists of the principles of DFS that do not contain Δ – in other words, the compositionality axioms (axioms beginning with a 'C') apart from CΔ , At, and Nec and Conec. Compositionality is a fundamental principle of modern semantics and it is a significant drawback of many rival theories that they do not have it.[36] A noteworthy

[34]The rule Nec is actually redundant; it is a derived rule of BCT.

[35]These facts follow from the results of Halbach (1994). One can interpret DFS into the system FS Halbach considers by defining $\Delta\phi := \phi \wedge Tr(\ulcorner\phi\urcorner)$ (of course, this interpretation is ruled out in the broader theory by the determinacy role outlined above.)

[36]McGee, for example, drops C$\forall$, the revision theory drops both C$\forall$ and C$\neg$, and Feferman's KF does not have C$\neg$.

consequence of compositionality, i.e. of the C-axioms, is the principle of bivalence: every sentence is either true or false (here we follow the convention of calling a sentence with a true negation 'false'.)[37] Note also that compositionality extends also to the intensional connective Δ. (These principles prove especially useful if we wanted to augment the language in such a way that it could state its own semantic theory.)

We have already motivated the principles K, T and RT in our discussion of BCT (although I have more to say about $\Delta\mathsf{Nec}$.) RT tells us when we can and can't enquote and disquote. In any modal logic based on classical quantification theory the converse of BF, CBF, will already be a theorem, as will be the claim that there is no unclear existence: $\Delta\forall x\Delta\exists y x = y$. The principle of conecessitation, Conec, states that all sentences whose truth is provable in DFS, are in fact provable in DFS. Although I have no strong positive arguments for BF or Conec in the theory, they are natural principles and it is of interest that they are consistent with the remaining theory. At any rate, the result of removing BF and/or Conec from DFS is obviously also consistent given theorem 2.1 so we can treat these principles as optional for the time being.

The unrestricted principle of necessitation for clarity, $\Delta\mathsf{Nec}$, and the principle of necessitation Nec for truth, are much more tricky. In what follows I shall just focus on $\Delta\mathsf{Nec}$ (this focus is reasonable given the observation that Nec is in fact redundant given $\Delta\mathsf{Nec}$ and RT.) The basic role of $\Delta\mathsf{Nec}$ is to ensure that the *axioms* (that is, everything apart from the rules Conec, Nec and $\Delta\mathsf{Nec}$) listed above, and their logical consequences, are themselves clear. This point is crucial since I want to assert these axioms, and their consequences, and it would not be appropriate to do so if some of them were unclear. In fact, I *also* want to go further and assert that these principles and their consequences are clear (I in effect committed myself to this when I asserted the preceding sentence) and I can only do this appropriately if they are clearly clear. Thus presumably we also want a theory that proves that these axioms and their consequences are clearly clear. And, of course, in order for *this* assertion (the one in the previous sentence) to be appropriate we need the axioms to be clearly clearly clear. I think I could proceed in this way for a little while, justifying further iterations of clarity by appealing to the appropriateness of earlier assertions, but as this continues the justifications get weaker and weaker and it becomes less and less clear that my assertions are appropriate. Indeed, I become a little bit less confident about my assertion that the principles are clear than I am about my assertion of the principles, and this lessening of confidence will build up as I continue to assert higher orders of clarity.

This is all motivated by the thought that being clear at all orders is a status that is hard to come by. Perhaps logical and conceptual truths are clear at all orders, but one might have thought that principles about the truth of sentences, which are contingent on the way we use language, are not like this. If so, then principle $\Delta\mathsf{Nec}$ is not acceptable, for it guarantees that if ϕ is a consequence of the above axioms then so is $\Delta\phi$, $\Delta\Delta\phi$, and so on forever. However, the theory contains contingent semantic principles which depend on the way we use the symbols $\neg$, $\vee$ and so on[38], and given the preceeding remarks it is unlikely that such principles will be clear at all orders.

The suspicion that the necessitation principles $\Delta\mathsf{Nec}$ (and the derived Nec) are too strong is, in a sense, verified by the fact that these principles lead to an ω-inconsistency. Although one cannot prove contradictions from DFS in a finitary proof system, it becomes inconsistent given natural infinitary rules. This can be seen in two ways. Firstly the theory contains

[37] This follows from C$\vee$, C$\neg$ and the truth of the principle of excluded middle, which we get by applying Nec to the principle of excluded middle.

[38] If we had used '$\vee$' to mean conjunction then the axiom C$\vee$ would not have been true.

the theory I called FS which is known to be ω-inconsistent by a theorem due to McGee (see McGee (1990).) Secondly the full theory DFS contains a distinct ω-inconsistency that doesn't involve any of the principles it shares with FS:

Theorem 2.2. $\mathsf{BF} + \mathsf{T} + \Delta\mathsf{Nec} + RT + C\Delta$ *are* ω*-inconsistent in PA.*

The proof of this is in the appendix.

On the other hand, theories that contain a limited amount of necessitation can be formulated. Consider the rule

> RNec If ϕ is provable in $\mathsf{KT} + \mathsf{BF} + \mathsf{RNec}$ (i.e. from the pure logic of clarity) then infer $\Delta\phi$

RNec allows us only to necessitate the theorems of classical logic and $\mathsf{KT} + \mathsf{BF}$; conceptual truths that are plausibly clear at all orders. Let X_n be the set of sentences provable in DFS with at most n applications of the necessitation principles in a proof, plus RNec which allows arbitrarily many applications if the sentence is provable in the pure logic of clarity. Then the theory DFS_n is just the closure of X_n under classical logic. The presence of the restricted necessitation principles effectively allows us to recover the reasoning above, in which a few iterations of the clarity of the truth theoretic axioms can be justified (enough to justify all the assertions I need to make at any rate.) However without the unrestricted necessitation principles we cannot infer that the truth theoretic axioms are clear at all orders; the best we can do is infer that they are clear at all orders less than some upper bound.

Theorem 2.3. *For each n, the theory* DFS_n *has a standard model.*

3 Higher order unclarity

In the previous section we introduced a distinction between clear and unclear cases of truth and offered a positive theory of it. One of the benefits of this way of drawing the distinction is that 'clearly' is not a predicate in its own right and can just as well be used to state that someone is clearly bald as it can be to state that a sentence is clearly true. This freedom allows us to formulate questions about iterations of the clarity operator. We argued that the distinction between being true and being untrue is not always a clear one, as the liar paradox demonstrates. A natural follow up question concerns the distinction between being a clear truth and an unclear truth: is this distinction a clear one? The proposition that the liar isn't true is a clear example of an unclear proposition; we can in fact prove that it is unclear from the minimal theory we designated BCT, so, by necessitation, it is clearly unclear. The liar paradox, therefore, does not witness the unclarity of the distinction between clear and unclear propositions. Here we shall argue that the revenge sentence R='R is not clearly true' does.

The claim at stake is whether it is always a clear matter whether p is clear or not. This can be broken down into two claims; that when it is clear that p, this fact is itself clear, i.e.

> 4 If it's clear that p then it's clear that it's clear that p.

and that when it's unclear that p this fact is itself clear, i.e.

> 5 If it's not clear that p then it's clear that it's not clear that p.

In general the distinction between clear Fs and unclear Fs is not a clear one. This phenomenon manifests itself quite evidently with the related phenomenon of vagueness. When presented with a Sorites sequence for the property of being a heap, such as a sequence of piles of sand beginning with a few grains and ending with a large mound, we may similarly introduce the distinction between the piles which are clearly heaps and those which are not. The distinction is evidently there, as the enormous pile at the end of the sequence is not only a heap, but a clear heap, whereas the first pile is clearly not a heap, and therefore not a clear heap. But when do the clear heaps stop being clear heaps? This distinction, between the clear heaps and the unclear heaps, is no clearer than the distinction between heaps and non-heaps. At some point there must be a clear heap which is not clearly a clear heap, or an unclear heap which is not clearly an unclear heap; either we have a failure of $\mathsf{4}$ or $\mathsf{5}$ (and most probably both).

The situation is somewhat similar in the context of the revenge sentence. The assumption that R is clearly clearly true allows us to apply disquotational reasoning to R in BCT.[39] Thus if R is clearly clearly true then R is true if and only if R is not clearly true. But also, if R is clearly clearly true then R is both true and clearly true – contradiction. So R is not clearly clearly true. Thus one can prove in BCT:

> If R is clearly true, then this fact is itself unclear.

What happens if R is not clearly true? I will spare you the details, but here the most we can prove is:

> If R is not clearly true, then this fact is itself either unclear, or it is unclear whether it is unclear.

Either way the naive iteration principles for clarity fail. The revenge paradoxes, in this setting, do not result in inconsistency, but second or perhaps third order unclarity.

Stronger liars can be obtained by iterating the number of clearly operators. Let us write 'clearly^n' to indicate n 'clearly's in a row, and consider the sentence $R_n =$ 'R_n is not clearly^n true'. Similar patterns of higher order unclarity arise here; for example we can prove that if R_n is clearly^n true, it's unclear whether R is clearly^n true, and similar things.[40] In other words, the nth revenge sentence gives rise to certain combinations of $n+1$th order unclarity in BCT.[41]

(Although we cannot consistently add both $\mathsf{4}$ and $\mathsf{5}$ to BCT together, one might wonder whether one could have just one or the other. Although this might be possible for the minimal theory BCT, neither $\mathsf{4}$ nor $\mathsf{5}$ can be consistently added to the richer theories DFS, or DFS_n for $n \geq 2$; for those interested proofs can be found in the appendix.[42])

3.1 Hierarchies of determinacy operators

This diagnosis of the revenge paradoxes bears a striking resemblance to Hartry Field's recent account of revenge, in which the revenge paradoxes also exhibit indeterminacy at

[39] To apply RT we need to show that either it's clear that R is not clearly true or it's clearly not the case that R is not clearly true. Our assumption that R is clearly clearly true entails the second disjunct.

[40] The structure of higher order unclarity in BCT and the theories we consider later are actually quite complicated; a full description of the properties of each R_n sentence is not possible here.

[41] If we were to augment the language with infinite conjunctions we could continue iterating into the transfinite. These can be treated in different ways, although I think there are independent reasons, discussed in 2.4, for thinking that no proposition about the truth of any sentence is clear at all orders.

[42] The latter goes via the fact that the related principle B – $\phi \rightarrow \Delta\neg\Delta\neg\phi$ – cannot be consistently added to these theories.

higher orders. Although there are many points of departure between the present theory and Field's – most notably Field's endorsement of all instances of the T-schema and his rejection of classical logic – many of the issues are parallel.

One of these issues, raised by Graham Priest in his critical notice of Field's 'Saving Truth from Paradox', turns crucially on the way that Field understands the hierarchy of determinacy operators in his theory. Priest argues that Field cannot express the general notion of defectiveness that the liar and its relatives have. Since this argument turns on an important difference between my view and Field's, and marks a significant advantage of mine, I shall quote Priest at length:

> "There are two jobs for the notion of defectiveness to do: [a] we must be able to say of certain sentences, e.g., the liar, that they are of this kind; and [b] we must be able to talk about such sentences in general and say things about them. [...].
>
> Now, Field's D operator does the job of [a] in many cases. As we have seen, for a number of defective sentences, A, like the Liar, we can say truly $\neg DA \wedge \neg D\neg A$, or at least $\neg D^{\alpha}A \wedge \neg D^{\alpha}\neg A$, for some suitable iteration of 'D's – maybe into the transfinite. But the D operator cannot do the job of [b]: as α increases, the extension of $\neg D^{\alpha}$ gets larger and larger (p. 238), so for no α does the extension of $\neg D^{\alpha}$ comprise all the non-(determinately true) sentences. Where $Q = \neg D^{\alpha}Tr(\ulcorner Q \urcorner)$, Q is not in the extension of $\neg D^{\alpha}$. Nor is it possible to define a predicate whose intuitive meaning is something like $\bigvee\{\neg D^{\alpha}Tr(x) \mid \alpha$ is an ordinal$\}$, since, as Field shows, the precise definition of this depends on some ordinal notation, and will therefore take us only so far up the ordinals. [...].
>
> Indeed, without the notion, one cannot even formulate the driving thought behind Field's own solution: that the LEM fails because of the existence of indeterminacies. To declare all general claims about indeterminacy unintelligible is an act of ladder-kicking-away desperation of Tractarian proportions." Priest, 'Hopes Fade for Saving Truth' Priest (2010)

Field's picture is that there is a (large) hierarchy of predicates, and each level only does a partial job of characterizing the true notion of defectiveness or being diseased which includes the liar, the revenge liar, and the further iterations.

This is problematic for Field since it suggests that there is a notion – which both Field and Priest informally call 'defectiveness' – which Field cannot express. Informally, a sentence is defective when it falls under one of Field's partial predicates; the union of all Field's partial defectiveness predicates therefore seems to be the notion that is being informally employed, but which is not expressible within Field's framework. More to the point, this reveals that the project that Field is engaging in is what I earlier labeled the project of giving a partial diagnosis (more precisely, he is giving multiple partial diagnoses). I suggested there that this project is not entirely satisfactory.

Let me note at this juncture that there isn't a straightforward reason to think that the revenge liar is defective.[43] A poor reason would be to say that it's also self-referential: we know that self-reference is neither necessary nor sufficient for defectiveness.

[43] Remembering that 'defective' or 'unclear' is the notion I've so far been characterizing by the determinacy role – its inferential and normative properties – and is not necessarily a notion governed by the intuitions Priest is appealing to.

The view that I am defending is not the view that first, second, and the higher orders of indeterminacy all express partial, albeit increasingly more complete, diagnoses of the paradoxes. The notion that Priest and Field informally call 'defectiveness' can be completely expressed using the indeterminacy operator. I therefore do not accept the assumption that Priest and Field make: that the revenge liar is defective, or more specifically, that it's indeterminate whether R is true, where R='R is not determinately true'. I reject this because this claim is *itself* indeterminate; asserting this would be akin to asserting that someone is bald when they have a borderline number hairs. Recall that the determinacy role states that if it's unclear whether p one should not assert either p or its negation. Furthermore, since I have suggested that it's an indeterminate matter whether the revenge liar is determinately true or not, it would simply be wrong to assert that the revenge liar is indeterminate. So, *pace* Priest and Field I would object to the claim that the revenge liar is indeterminate in exactly the same way that I would object if someone asserted that the liar is true.

The relevant notion of indeterminacy, the one I've axiomatized above and which governs assertability, knowability and so on, is perfectly well captured by the operator ∇. Like many useful but vague concepts it admits indeterminate instances. It would be a mistake, however, to assert that these indeterminate cases of indeterminacy are straightforwardly indeterminate. That would be to assert the indeterminate - to assert of anything that is indeterminately F that it is F is to assert the indeterminate, and it would be an equivocation to use the word 'indeterminacy' when one really means 'indeterminacy at some order'. The equivocation is plain for all to see when the analogous objection is levelled at the view defended here: 'as α increases, the extension of $\neg D^{\alpha}$ gets larger and larger, so for no α does the extension of $\neg D^{\alpha}$ comprise all the non-(determinately true) sentences.' This would not make sense if we were using the operator 'D' (or 'Δ' in my theory) and the word 'defective' interchangeably, for when $\alpha = 1$, $\neg D^{\alpha}(x) = \neg D(x)$ comprises all the defective sentences simply by the fact that D was introduced to mean 'defective' (and $\neg\Delta$ applies to all and only the defective propositions in my framework).

Note that if we did interpret Priest's use of the word 'defective' to just *mean* indeterminate at some order or other, his final remarks no longer hold any bite. It may well be the case that not much is determinate at all orders.[44]

Given that I have been denying that the higher iterations of the indeterminacy operator play an important role in the workings of the liar paradox, and in particular, denying that they serve the purposes required by a solution to the semantic paradoxes, it would be fair to ask in which respects the second order, third order, ..., nth order indeterminacy differ.

It is easy to verify that if an operator O – a potential interpretation of 'Δ' – satisfies the axioms of the theory DFS, then so will the iterations of this operator: OO, OOO, and so on. This naturally raises the question of what the intended interpretation of 'Δ' is, and if there is anything we can say to narrow down this infinite list of candidate interpretations. Note that this problem is distinctive to DFS. My preferred response is to reject the consequences of DFS_n when n gets sufficiently large – on that view large iterations of the determinacy operator behave differently from smaller iterations. However, another more generally applicable way to distinguish the iterations is to appeal to the role that Δ plays in a theory of propositional attitudes; in this respect I have suggested that it satisfies the determinacy role. As it is currently stated, the determinacy role consists of several conditionals stating that indeterminacy poses a distinctive obstacle to knowledge; that one cannot rationally

[44] At any rate, this is certainly true of the models of DFS_n described in the appendix.

believe or care about a proposition you believe to be indeterminate and so on.

However one could also maintain, as I will in the next section, that the iterations of determinacy do not play this distinctive role in regulating belief and knowledge. It is consistent with the determinacy role, for example, that one be in a position to know that p even when it is not determinately determinate that p (so long as it is in fact determinate that p.) In which case it is clear that second order indeterminacy does not present a distinctive obstacle to knowledge in the same way that indeterminacy does. This is consistent with the idea that we have simply introduced the word 'indeterminacy' as a name for *whatever* that distinctive obstacle to knowledge is, or for whatever it is that regulates your beliefs in such and such a way (and so on). If, for example, the distinctive barrier to knowing whether p characteristic of cases like the liar is present in exactly the second order indeterminate propositions, then on this view we weren't applying the word 'indeterminate' properly in the first place: the proper use should latch onto the notion we've been incorrectly calling 'second order indeterminacy'.

Thus it is possible to use the role of determinacy in a theory of rational propositional attitudes to distinguish the determinacy operator from its iterations. We can contrast this view with an alternative position. According to the alternative there are infinitely many operators that satisfy the determinacy role: if O is an operator that satisfies the determinacy role so does all of its iterations. This appears to be Field's position in a number of writings (see, for example Field (2003).) My view maintains that there is only one true notion of defectiveness at play here, and the others are just iterations of this notion. According to this alternative position, endorsed by Field and others, higher order indeterminacy is also a distinctive source of ignorance. Indeterminacy and second order indeterminacy play the same role in regulating propositional attitudes. However the idea that indeterminacy is not the only distinctive obstacle to knowledge invites the thought that whatever the obstacle is in all these cases we could surely introduce an expression for it (this is, effectively, Priest's intuition.) And if so, why not just use a more expansive use of the word 'indeterminate' or 'defective' from the beginning? Field's position that there are all these distinct kinds of obstacles to knowledge, assertion, rational belief and so on, but there's no umbrella phrase for them (on pain of paradox) is reminiscent of the Tarskian strategy of denying the expressibility of seemingly intelligible notions. The alternative suggestion, which seems to be extremely natural, is that there is an umbrella expression for this, although it's one that does not iterate trivially.

4 Revenge and Assertability

This brings us to another benefit this approach has over its rivals. As we noted earlier one of the forms a revenge paradox can take involves the propositional attitudes of acceptance and rejection, or, sometimes, permissible assertion and denial. The problems are often discussed in the context of non-classical solutions (see Soames (1999), Field (2008), Beall (2009), Priest (2006).), but they seem to be just as bad for classical solutions (see especially Feferman (1991), Maudlin (2004).)[45] For example, paracomplete logicians – logicians who relinquish the law of excluded middle – are at pains not to accept (or assert) that the liar is neither true nor untrue since, by the deMorgan laws, this would commit them to accept the

[45]The problem of assertion and acceptance does not apply exclusively to philosophers in this list. I think these issues are just as problematic for McGee or the revision theory, although they are not discussed as frequently.

contradictory claim that the liar is both not true and not not true. In order to express their disavowal of LEM they *reject* (deny) the claim that the liar is true or untrue. Rejecting (denying) p is not to be reduced to acceptance (assertion) of the negation of p. These theorists need then to say something about the sentence: A='A is not assertable.'

How might paradoxes involving assertion and acceptance arise in this setting? A similarly central distinction between assertable propositions exists in the present framework: like the paracomplete logician we agree that one should not assert that the liar is true, or that it's untrue since it is unclear. In general, we might endorse the following schema from the determinacy role

> ASSERTION: If it's unclear whether p then you are not in a position to assert that p and you are not in a position to assert that $\neg p$.

Perhaps this is a basic fact about unclarity. However it is natural to think that it could be explained by the fact that unclarity precludes knowledge. We should not assert that the liar is true or assert that the liar is untrue because we do not, and cannot, know which it is.

Note that ASSERTION is an externalist norm in the sense that one is not always in a position to know whether one is complying with it. This can happen for quite mundane reasons – for example if you do not know whether Harry is clearly bald because you simply do not know how much hair he has. In these cases someone else in a better epistemic position than you would be able to evaluate your assertion for compliance.[46] However we must also make room for the possibility that sometimes no-one else is, or even could be, in a position to evaluate whether you have complied with ASSERTION or not; for example if it is unclear whether p is unclear or not.

How might assertability paradoxes arise in this framework? Note first that the problematic notion is a normative one, not a descriptive one. The sentence A = 'it's not the case that so-and-so has at some point asserted that A is true' is clearly true or clearly false depending on what the person in question has asserted.[47] The problem supposedly arises when we want to talk about which propositions I should and shouldn't assert. Now evidently there are cases where I shouldn't assert p even though it's clear that p; I shouldn't assert that there are an even number of stars within a hundred light years of the sun, nor should I assert that there are an odd number, since I do not know either way, even though it is a clear matter whether or not there are an even number of stars within a hundred light years of the sun.[48] However the proposition that the liar isn't true seems different; it seems to be *inherently* unassertable. The reason that it is unassertable is to do with unclarity about whether the liar is true or not. (Note: being inherently unassertable does not mean being necessarily unassertable – contingent liars are inherently unassertable because one couldn't be in a position to assert with the non-semantic facts as they actually are.)

Is there a problem of expressing the distinction between inherently unassertable propositions in a semantically closed language without falling afoul of assertability paradoxes? A proposition is inherently unassertable when the source of its unassertability is the kind of unclarity we find generated by semantic paradoxes; the problematic notion of unassertability is directly controlled by the clarity operator. Thus we get a partial converse of ASSERTION

(*) p is inherently unassertable if and only if it's unclear whether p.

[46] Usually an assertion such as this will fail to comply with the norm of asserting only what you know.

[47] Paradoxes formulated using the operator 'So-and-so is disposed to assert that p at t' look like they are equally unproblematic.

[48] Assuming, for simplicity, that there no borderline stars.

It seems that, as long as we can consistently introduce the clarity operator into the language, we can also introduce the notion of inherent unassertability into the language. And furthermore, it seems that the assertability paradox A='it is inherently impermissible to assert that A is true' should pattern with the revenge paradox R='it is not clear that R is true' in that it should be unclear whether it is inherently impermissible to assert that A is true.

Intuitively there are a bundle of tightly related phenomena which arise in the cases of unclarity. When p is such a case there is a peculiar source of ignorance and uncertainty, it is impermissible to assert or question whether p, and so on; we introduce the word 'unclear' to distinguish propositions that associate with that bundle of properties. What happens when it's unclear whether p is clear or not? Since 'unclear' was introduced as a way of picking out propositions with that bundle of properties, to say it's unclear whether p is clear or not is to say that it's unclear whether p has that bundle of properties.

This leads us to the following natural more general conjecture: in epistemically ideal situations, second order unclarity usually comes along with unclarity about what attitudes you should adopt, and about whether you should assert. If, for example, it's unclear whether p is clear or not, and you are in these idealized circumstances, then it's also unclear whether you are in a position to know whether p, unclear whether it's permissible to assert p or not, unclear whether it's reasonable to wonder whether p, and so on and so forth. This marks another important difference between the present interpretation of the determinacy operator and Field's. In his informal writing it is clear that Field only takes an assertion to be appropriate if it is determinate at all transfinite orders. Yet the latter notion is not expressible by Field on pain of paradox. The fact that it is a theoretically central concept – it is the property against which we evaluate assertions and attitudes for appropriateness – makes its inexpressibility all the more uncomfortable. By contrast, on the present proposal the central concept is just that of determinacy which is something that is quite clearly expressible in the theory.

4.1 Assertion and higher order unclarity

With this in mind, let us consider an example. Assume that it's unclear whether it's clear that p. May we assert p, or should we abstain from assertion? A naïve but persuasive thought is that second order unclarity in p indicates a *kind* of badness in p and one should accordingly refrain from asserting p or its negation in all contexts just as one should refrain in cases of first order unclarity, such as the liar and the truth-teller. This thought, however, is not sustainable. Second order unclarity indicates unclarity over whether p is "bad" in the relevant sense, and thus, unclarity over whether one should assert or refrain from asserting p.

Suppose that it is clear that I am in a context in which I have a reason to assert whether or not p, and that I am knowledgeable of the relevant non-semantic facts. If unclarity is the only barrier to knowledge (that is, if second (and higher) order unclarity is not a barrier as well) then this amounts to saying that I am only ignorant in the unclear cases.

According to the present view what I am in a position to know and assert, given I am in an epistemically ideal situation, is exactly what is clear. In other words, in this scenario we have that, clearly:

> I'm in a position to assert that p if and only if it is clear that p

By fairly uncontroversial reasoning – namely, that Δ obeys the K axiom from §2.4 – it follows

that if it is unclear whether it's clear that p, it is unclear whether you are in a position to assert that p. (Similarly, it is unclear whether you are in a position to know that p (provided you are not ignorant of any of the relevant non-semantic facts.)[49])

What should I do in this situation? Should I assert or shouldn't I? I am suggesting there simply is no answer to these questions, there is no fact of the matter, and it would be misguided to continue asking. One could quite rightly imagine getting exasperated at this advice; after all, you have to do *something*, assert or do nothing, in that situation and knowing that there's no matter of fact which option I should pursue is not the least bit helpful.

Although this is clearly a frustrating situation to find oneself in I think that it is no more esoteric than more humdrum cases of vagueness in normative claims. Many normative properties are vague. One might for example, think that it is wrong to kill a person but not an embryo that is not a person. No doubt it is vague at which point one begins to be a person, and thus, according to our toy ethical theory, it will be vague at which point during pregnancy it ceases to be permissible to abort. Similar points apply to the norms of assertion. Consider a Sorites sequence for the property of being bald. I take it that it is permissible to assert that people with no hairs at all are bald, and that it is not permissible to assert that people with 1000000 hairs are bald, but surely there will be cases, people with a certain distribution and number of hairs, where it is borderline whether it's permissible to assert that they're bald. If Harry is such a case then there's simply no fact of the matter whether it's permissible or impermissible to assert that Harry is bald.

Another potential source of discomfort about the current theory might stem from the fact that the conditions for proper assertion are not always known to the subject. Surely, the objection might go, the conditions for proper assertion should be transparent, in the sense that both conditions for proper assertion, and the conditions for improper assertion, are luminous:

> If it is permissible for A to assert that p, then A is in a position to know that it is permissible for A to assert that p
>
> If it is not permissible for A to assert that p, then A is in a position to know that it is not permissible for A to assert that p

Surely, the objection might continue, it must be possible to operationalize assertion: one should be able to provide rules which tell you when it is and isn't OK to assert such that one is in principle always in a position to know whether one is complying with these rules.

I do not find this objection convincing. What the Sorites sequence above shows is that, quite independently, there will cases where there it is unclear whether you may assert p or not. Since unclarity bars knowledge these will be cases where you are not in a position to know whether p is assertable or not. In Williamson (2008) Williamson criticizes the view that normative rules should be operationalizable in the context of epistemology. I think

[49] In Caie (2012) Michael Caie argues that whenever one ought to believe that it is indeterminate whether p, one ought to be such that it is indeterminate whether you believe that p. The view I am endorsing here is inconsistent with this principle since I think one should withold judgment when you believe that it's indeterminate whether p. The primary difference, therefore, is that for me indeterminacy concerning whether you are in a position to know or assert p arises due to second order indeterminacy in p whereas for Caie it is due to first order indeterminacy. The other important difference is that I am not recommending that anyone seek out the state of indeterminately believing p, I am only claiming that, as a matter of fact, it is sometimes indeterminate what you are in a *position* to know; the latter consequence is quite modest, and is plausibly already predicted by the vagueness of 'in a position to know'.

much of what Williamson says about epistemology when your knowledge is 'inexact', cases where you plausibly fail to know what your knowledge is because it is vague what your knowledge is, must apply to the epistemology of the semantic paradoxes too.

4.2 Assertive uttering

We have focused our attention on the verb 'assert' which takes a that-clause in its right argument and relates a person to a proposition. There is also an important relation that holds between a person and a sentence when that person makes a certain kind of noise, perhaps with certain kinds of intentions. I shall call this 'assertive uttering.' These two relations are clearly very different. One asserts *by* assertively uttering a sentence, but the relationship is complicated by the fact that different sentences can be used to assert the same thing, and the same sentence in different languages and different contexts can be used to assert different things. Since one might worry that there are paradoxes that can be formulated using the notion of assertive utterance which cannot be formulated using assertion it is worth addressing this issue.

The most important rule governing assertion that we have discussed is the norm Assertion. Like the rule 'do not assert p if you do not know that p' it provides an objective standard for evaluating assertions, one that is not always available to the asserter. In general there is no simple rule which would guarantee that you conformed to Assertion, and would be such that you could always know that you're following the rule. We do not always know whether or not we know p, and we do always not know whether or not p is unclear. The fact that Assertion (or the knowledge norm) provides an external standard of assessment means that sometimes it is impossible for *anyone* to know whether someone is complying with Assertion (or the knowledge norm.) When it's unclear whether you know p or unclear whether it's clear that p (due to vagueness, perhaps, or the kind of second order unclarity the paradoxes generate) then it is impossible for anyone to know if someone has asserted p in conformity with Assertion, or with the knowledge norm.

An interesting observation about the rules of assertion is that they are often language independent. For example the knowledge norm, Assertion, the maxim 'be relevant', and other Gricean maxims appear to be generally applicable whatever your mode of communication is: they apply equally, whether you are speaking German, French or sign language.

On the other hand, when deciding which sentence to assertively utter in a given context, (and a given language, L) you must somehow incorporate your knowledge about what it would be appropriate to *assert* in that context, as determined by Gricean norms of conversation, Assertion, and so on, with your particular linguistic knowledge telling you which sentence of L would result in you achieving this assertion. In order to have a theory of proper and improper assertive uttering, then, we must appeal to the speaker's *linguistic* knowledge in a way that we did not need to for our theory of proper assertion.

Here it is natural to think that we need specific linguistic knowledge about which sentences mean which propositions; if you have determined, via the very general norms of assertion that p would be the best proposition to assert then we should strive to find a sentence of L which can be assertively uttered to achieve an assertion of p. Even then there will be multiple sentences within that language that express p, some sentences express different things in different contexts, so there is plenty more to be said about this. The rules for appropriate assertive utterances are therefore much more complicated and less susceptible to systematic theorizing than the rules for assertion. However, it is natural to think that something like this story is correct, in which case we can partly reduce the theory of assertive

uttering to our theory of assertion, as governed by the principle ASSERTION amongst other things. Thus:

> Assertively utter s in a context c only if you are in a position to assert the proposition that would be asserted by an utterance of s in c.

It is natural, therefore, to think that there would be no special problem in constructing a theory which contains its own theory of assertive uttering, as well as assertion. The expressive requirements needed seem to be (i) that your theory can state when it is and isn't permissible to assert p (see section 4) and (ii) that the theory can say when a sentence s uttered in c would result in p being asserted. The latter would be achieved if the theory contained a 'saying' connecticate, which states when a sentence means that p in a given context. Both of these are easy to accommodate.[50]

The important thing to note about this reduction is that in assertively uttering 'p' in English you have not always asserted that p. This connection relies on the contingent fact that 'p' means that p in English. But the principle can fail when 'p' contains indexical expressions, and more to the present point, when 'p' contains semantic vocabulary. You would quickly find yourself confronted with paradoxes if you simply translated principles involving the locution 'it's permissible to assert that p' with the locution 'it's permissible to assertively utter 'p"; these only approximate one another when 'p' says that p in English.

5 Conclusion

Unlike the majority of diagnoses of the paradoxes, the present view takes a sentential operator rather than a linguistic predicate as primitive, and characterizes it by its role in a theory of rational propositional attitudes. The virtues of this approach are two-fold. Firstly, unlike its linguistic counterparts, the theory is in a position to address the diagnostic form of revenge. We have shown that it is possible to consistently characterize the problematic instances of disquotational reasoning, allowing us to give a diagnosis of these failures without falling afoul of the revenge paradoxes. Indeed the revenge paradoxes, rather than leading to inconsistency, lead to indeterminacy at higher orders. Secondly the specific theory of determinacy outlined here, partially given by the role the operator plays in regulating propositional attitudes, puts us in a position to solve the assertability form of revenge. The theory straightforwardly identifies assertability in epistemically ideal circumstances with determinacy, providing us with an extremely natural model of assertion. Again, instead of an outright inconsistency we merely found that it is sometimes indeterminate whether one ought to assert; a conclusion that seems plausible on independent grounds.

6 Appendix

6.1 The consistency of DFS

The consistency of DFS is shown by a model theoretic construction akin to the revision theoretic construction of Gupta and Belnap (1993).[51]

[50] The theories DFS and DFS_n, and their consistency proofs, can be modified fairly straightforwardly to accommodate a 'saying' connecticate.

[51] See also Herzberger (1982) and Friedman and Sheard (1987).

A Kripke model for the language $\mathcal{L}$ is a quadruple $\langle W, R, D, [\![\cdot]\!].\rangle$ consisting of a set of worlds, W, an accessibility relation, R, for interpreting Δ, a domain, D, and interpretation function $[\![\cdot]\!]_w$ which assigns each term, function and predicate symbol an element, function, or relation over D of the respective type, relative to each world w.

In our model both W and D will be the set of natural numbers, Rxy the relation that holds if $x = y$ or $x = y + 1$, and each primitive piece of arithmetical vocabulary is assigned its standard interpretation relative to every world, for example $[\![+]\!]_w = \{\langle\langle x, y\rangle, x + y\rangle \mid x, y \in \mathbb{N}\}$, $[\![0]\!]_w = 0$, etc, for every $w \in W$. I shall denote the element $n \in W$ as w_n to distinguish it from the same element n in the domain of natural numbers D. An assignment function, v, is a function from variables of $\mathcal{L}$ to D.

All that is left is to specify the extension of Tr at each world w_i, $[\![Tr]\!]_{w_i}$. We shall do this in a moment.

We can define what it means for a formula of $\mathcal{L}$ relative to an assignment to hold at a world in a Kripke model of this kind, which we can write $w_n, v \Vdash \phi$, or just $w_n \Vdash \phi$ if ϕ is closed. I shall omit the details of the definition since it is standard. What is important is that whether a formula holds at a world, w, depends only on what holds at worlds which are R accessible to w, or R accessible to worlds which are R accessible to w, and so on. In our model this just amounts to this: whether ϕ holds at w_n depends only on what happens at worlds w_i for $i \leq n$. Moreover, whether ϕ holds at w_n never depends on what's going on at w_m for $m > n$. Thus the following definition of the extension of the truth predicate at a world w_n is not viciously circular:

$$[\![Tr]\!]_{w_0} = \emptyset$$

$$[\![Tr]\!]_{w_{n+1}} := \{\ulcorner\phi\urcorner \mid \phi \text{ is closed and } w_n \Vdash \phi\}$$

To get a picture of the construction see the diagram below

$$\emptyset \qquad \{\ulcorner\phi\urcorner \mid w_0 \Vdash \phi\} \qquad \{\ulcorner\phi\urcorner \mid w_1 \Vdash \phi\} \qquad \{\ulcorner\phi\urcorner \mid w_2 \Vdash \phi\} \ldots$$

All the axioms of DFS hold at w_1 (although the rules do not in general preserve truth at w_1). In fact every theorem of DFS_n holds at w_{n+1}. It follows that DFS_n has a standard model, w_{n+1}, for each n.

Say that a sentence eventually holds in this model if there exists an i such that the sentence holds at every world w_j for $j > i$. The consistency of DFS is established by noting that every axioms holds at every world to the right of w_1, and therefore eventually holds, and that if ϕ eventually holds, so does $Tr(\ulcorner\phi\urcorner)$ and $\Delta\phi$ and if $Tr(\ulcorner\phi\urcorner)$ eventually eventually holds so does ϕ. Furthermore the classical rules of inference preserve eventual holding; so every theorem of DFS eventually holds. The consistency of DFS is obtained by noting that $1 = 0$ does not eventually hold.

It is also straightforward to see that the world w_{n+1} is a standard model for the theory DFS_n, thus the restricted theories DFS_n are not only consistent but consistently closable under the ω-rule.

Note, as we mentioned earlier, that there are restrictions on the kind of iterations of Δ are allowed in this theory. In particular none of B, $\mathsf{4}$ or $\mathsf{5}$ can be added to this theory.

Theorem 6.1. DFS + 4 *is inconsistent.*

Proof. 1. $\gamma \leftrightarrow \neg Tr(\ulcorner \Delta\gamma \urcorner)$ instance of the diagonal lemma.

2. $\Delta\gamma \rightarrow \Delta\Delta\gamma$ by 4.
3. $\Delta\Delta\gamma \rightarrow Tr(\ulcorner \Delta\gamma \urcorner)$ by RT.
4. $\Delta\gamma \rightarrow Tr(\ulcorner \Delta\gamma \urcorner)$ 2 and 3.
5. $\Delta\gamma \rightarrow \neg Tr(\ulcorner \Delta\gamma \urcorner)$ by T and 1
6. $\neg\Delta\gamma$ by 4 and 5.
7. $Tr(\ulcorner \neg\Delta\gamma \urcorner)$ by necessitation of line 6.
8. $\neg Tr(\ulcorner \Delta\gamma \urcorner)$ by truth functionality of $\neg$.
9. γ by 1
10. $\Delta\gamma$ by necessitation for Δ, which contradicts 6.

□

Theorem 6.2. DFS + B *is inconsistent.*

Proof. 1. $\gamma \leftrightarrow Tr(\ulcorner \Delta\neg\gamma \urcorner)$ instance of the diagonal lemma.

2. $\gamma \rightarrow \Delta\neg\Delta\neg\gamma$ by B
3. $\gamma \rightarrow Tr(\ulcorner \neg\Delta\neg\gamma \urcorner)$ by 2 and RT.
4. $\gamma \rightarrow \neg Tr(\ulcorner \Delta\neg\gamma \urcorner)$ by truth functionality of $\neg$.
5. $\gamma \rightarrow Tr(\ulcorner \Delta\neg\gamma \urcorner)$ by 1.
6. $\neg\gamma$ by 4 and 5.
7. $\Delta\neg\gamma$ by necessitation for Δ
8. $Tr(\ulcorner \Delta\neg\gamma \urcorner)$ by necessitation of line 7.
9. γ by 1, which contradicts 6.

Since DFS+5 entails B it follows that we cannot add 5 to DFS either. □

Note that both of these arguments can be carried out in DFS_n for $n \geq 2$.

Theorem 6.3. BF + T + ΔNec + *RT* + *C*Δ*., are* ω*-inconsistent in PA.*

Proof. It is possible to arithmetically define a function g such that you can prove $g(0, x) = x$ and $g(n+1, x) = \dot{\Delta} g(n, x)$. For presentational reasons it is more transparent, if not strictly rigorous, to write $\ulcorner \Delta^k \gamma \urcorner$ for $g(k, \ulcorner \gamma \urcorner)$.

Also note that Nec is a derived rule once you have T + ΔNec + RT.

Using the diagonal lemma choose γ so that $\gamma \leftrightarrow \neg\forall n \Delta Tr(\ulcorner \Delta^n \gamma \urcorner)$.

1. $\neg\gamma \rightarrow \forall n \Delta Tr(\ulcorner \Delta^n \gamma \urcorner)$ by the diagonal lemma
2. $\forall n \Delta Tr(\ulcorner \Delta^n \gamma \urcorner) \rightarrow \Delta\forall n Tr(\ulcorner \Delta^n \gamma \urcorner)$ by BF

3. $\Delta\forall nTr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow Tr(\ulcorner\forall nTr(\ulcorner\Delta^n\gamma\urcorner)\urcorner)$ by RT and T.

   (a) $\forall nTr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow Tr(\ulcorner\Delta\Delta^k\gamma\urcorner)$ for arbitrary k, by UI.

   (b) $Tr(\ulcorner\Delta\Delta^k\gamma\urcorner) \rightarrow \Delta Tr(\ulcorner\Delta^k\gamma\urcorner)$ by CΔ.

   (c) $\forall nTr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow \forall n\Delta Tr(\ulcorner\Delta^n\gamma\urcorner)$ by a), b) and classical quantificational theory.

   (d) $\forall n\Delta Tr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow \neg\gamma$ by Diagonal lemma.

   (e) $\forall nTr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow \neg\gamma$ by c and d.

   (f) $Tr(\ulcorner\forall nTr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow \neg\gamma\urcorner)$ by Nec.

   (g) $Tr(\ulcorner\forall nTr(\ulcorner\Delta^n\gamma\urcorner)\urcorner) \rightarrow Tr(\ulcorner\neg\gamma\urcorner)$ by C$\rightarrow$.

4. $\neg\gamma \rightarrow Tr(\ulcorner\neg\gamma\urcorner)$ by 1-3 and g.

5. $Tr(\ulcorner\neg\gamma\urcorner) \rightarrow \neg Tr(\ulcorner\gamma\urcorner)$ by C$\neg$

6. $\neg Tr(\ulcorner\gamma\urcorner) \rightarrow \neg\Delta Tr(\ulcorner\gamma\urcorner)$, T axiom for Δ.

7. $\neg\Delta Tr(\ulcorner\gamma\urcorner) \rightarrow \neg\forall n\Delta Tr(\ulcorner\Delta^n\gamma\urcorner)$, UI, $n = 0$.

8. $\neg\forall n\Delta Tr(\ulcorner\Delta^n\gamma\urcorner) \rightarrow \gamma$ by diagonal lemma.

9. $\neg\gamma \rightarrow \gamma$ by 4-8.

10. γ

But now by k applications of ΔNec to γ, one application Nec and then ΔNec again we can prove $\Delta Tr(\ulcorner\Delta^k\gamma\urcorner)$ for arbitrary k. But since we can prove γ we can prove $\neg\forall n\Delta Tr(\ulcorner\Delta^n\gamma\urcorner)$. So the principles listed in this proof are ω-inconsistent. □

References

Azzouni, J. (2007). The inconsistency of natural languages: how we live with it. *Inquiry 50*(6), 590–605.

Bacon, A. (MS). *Vagueness and Thought*.

Barwise, J. and J. Etchemendy (1989). *The liar: An essay on truth and circularity*. Oxford University Press, USA.

Beall, J. (2009). *Spandrels of truth*. Oxford University Press, USA.

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

Caie, M. (2012). Belief and indeterminacy. *Philosophical Review 121*(1), 1–54.

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

Eklund, M. (2002). Inconsistent languages. *Philosophy and Phenomenological Research 64*(2), 251–275.

Eklund, M. (2005). What vagueness consists in. *Philosophical Studies 125*(1), 27–60.

Feferman, S. (1991). Reflecting on incompleteness. *The Journal of Symbolic Logic 56*(1), 1–49.

Field, H. (2000). Indeterminacy, degree of belief, and excluded middle. *Nous 34*(1), 1–30.

Field, H. (2003). The semantic paradoxes and the paradoxes of vagueness. In *Liars and Heaps*, pp. 262–311. Oxford University Press.

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

Fine, K. (1975). Vagueness, truth and logic. *Synthese 30*(3), 265–300.

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

Gaifman, H. (1992). Pointers to truth. *The Journal of Philosophy 89*(5), 223–261.

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

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic 35*(3), 311–327.

Herzberger, H. (1970). Paradoxes of grounding in semantics. *The Journal of philosophy 67*(6), 145–167.

Herzberger, H. (1982). Naive semantics and the liar paradox. *The Journal of Philosophy 79*(9), 479–497.

Horwich, P. (1994). *Truth*, Volume 8. Dartmouth Publishing company.

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

Leitgeb, H. (2005). What truth depends on. *Journal of Philosophical Logic 34*(2), 155–192.

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

Maudlin, T. (2007). Reducing revenge to discomfort. In *Revenge of the Liar*, pp. 184–196. Oxford University Press.

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

McGee, V. and B. McLaughlin (1995). Distinctions without a difference. *The Southern Journal of Philosophy 33*(S1), 203–251.

Mirimanoff, D. (1917). Les antinomies de russell et de burali-forti: et le problème fondamental de la théorie des ensembles.

Parsons, C. (1974). The liar paradox. *Journal of Philosophical Logic 3*(4), 381–412.

Patterson, D. (2009). Inconsistency theories of semantic paradox. *Philosophy and Phenomenological Research 79*(2), 387–422.

Priest, G. (2005). Review of truth and paradox: Solving the riddles. *Journal of Philosophy 102*(9), 483–486.

Priest, G. (2006). *In contradiction: a study of the transconsistent*. Oxford University Press, USA.

Priest, G. (2007). Revenge, field, and zf. In *Revenge of The Liar*, pp. 225. Oxford University Press.

Priest, G. (2010). Hopes fade for saving truth. *Philosophy 85*(1), 109.

Prior, A. (1993). Changes in events and changes in things. In R. Le Poidevin and M. MacBeath (Eds.), *The philosophy of time*, pp. 35–46. Oxford University Press.

Prior, A. and A. Prior (1971). *Objects of thought*. Oxford.

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

Scharp, K. (2007). Replacing truth. *Inquiry 50*(6), 606–621.

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

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

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

Tarski, A. (1969). Truth and proof. *Scientific American 220*, 66.

Williamson, T. (1994). *Vagueness*. Routledge.

Williamson, T. (1998). Indefinite extensibility. *Grazer Philosophische Studien 55*, 1–24.

Williamson, T. (2008). Why epistemology cant be operationalized. In *Epistemology: New Philosophical Essays*, pp. 277–300. Oxford University Press.

Wray, D. O. (1987). Logic in quotes. *Journal of philosophical logic 16*(1), 77–110.

Yablo, S. (1982). Grounding, dependence, and paradox. *Journal of Philosophical Logic 11*(1), 117–137.

Yablo, S. (1993). Definitions, consistent and inconsistent. *Philosophical Studies 72*(2), 147–175.

Yaqūb, A. (1993). *The liar speaks the truth: A defense of the revision theory of truth*. Oxford University Press, USA.