How to find an attractive solution to the liar paradox

Mark Pinder

mark.jonathan.pinder@gmail.com

This is the accepted version of a paper forthcoming in *Philosophical Studies*. The final publication is available at *Springer* via http://doi.org/10.1007/s11098-017-0928-z.

Abstract: The general thesis of this paper is that metasemantic theories can play a central role in determining the correct solution to the liar paradox. I argue for the thesis by providing a specific example. I show how Lewis's reference-magnetic metasemantic theory may decide between two of the most influential solutions to the liar paradox: Kripke's minimal fixed point theory of truth and Gupta and Belnap's revision theory of truth. In particular, I suggest that Lewis's metasemantic theory favours Kripke's solution to the paradox over Gupta and Belnap's. I then sketch how other standard criteria for assessing solutions to the liar paradox, such as whether a solution faces a so-called revenge paradox, fit into this picture. While the discussion of the specific example is itself important, the underlying lesson is that we have an unused strategy for resolving one of the hardest problems in philosophy.

Keywords: liar paradox; truth; metasemantics; magnetism; Lewis; complexity.

Metasemantic theories can play a central role in determining the correct solution to the liar paradox.

That is a bald statement of the general thesis that I defend in this paper. The consequences are important: there are vast numbers of conflicting solutions to the liar paradox in the literature, and new objective measures are needed to help us determine which (if any) is correct. Metasemantic theories—that is, theories that explain what fixes or determines the semantic values of words, complex expressions and sentences—provide such a measure.

The defence of the general thesis is indirect: I provide a specific example of how a metasemantic theory can help to determine the correct solution to the liar paradox. In particular, I show that David Lewis's (1983, 1984) reference-magnetic metasemantic theory (henceforth *Magnetism*) can

decide between two of the most influential solutions to the liar paradox: Saul Kripke's (1975) strong-Kleene minimal fixed point theory of truth (henceforth *MFP*) and Anil Gupta and Nuel Belnap's (1993) revision theory of truth (henceforth *RTT*). As we will see, on at least one plausible construal of Magnetism—that developed by J. Robert G. Williams (2007)—, it appears to favour MFP. The specific example is important in its own right: it demonstrates the surprising power of Lewis's Magnetism, provides new support for Kripke's MFP, and begins to explore an apparent alliance between the two theories. But the underlying lesson of what follows is that we have an unused strategy for resolving one of the hardest problems in philosophy.

Before proceeding, there is an important caveat. Many theorists have developed formal theories of truth that are not intended to describe the English term "true". For example, Kripke tells us that he

> do[es] not regard any proposal, including [his own], as definitive in the sense that it gives the interpretation of *the* ordinary use of "true", or *the* solution to the semantic paradoxes. […] The model […] is to be tested by its technical fertility. It need not capture every intuition, but it is hoped that it will capture many. (1975: 699)

In light of such comments, it is perhaps natural to understand Kripke as engaged in a prescriptive project, seeking to define a formal truth predicate adequate for the needs of, say, science and mathematics. Many extant formal theories of truth, although certainly not all, have likewise been offered in this spirit.[1] Now, whatever the merits of the prescriptive project, the discussion herein does not contribute directly to it. Rather, the present question is which (if any) of these formal theories of truth—whatever their authors' intentions—provides us with the correct solution to the liar paradox *as it arises in English*; and thus whether any such theory of truth is in fact a correct account of *truth*, understood in the ordinary, English sense. It is in this spirit that, I claim, metasemantic theories can play a central role

[1] Examples of theorists of truth who engage particularly explicitly in the more *descriptive* project include contextualists (Burge 1979; Glanzberg 2001, 2004) and inconsistency theorists (such as, in particular, Scharp 2013).

in determining the correct solution to the liar paradox. And it is in this sense that the following discussion should be understood.

1. The liar paradox and two solutions

The liar paradox arises when one reasons intuitively about the truth value of the so-called liar sentence, λ:

λ is not true.

If we suppose that λ is true, then, given that λ says that λ is not true, it seems to follow that λ is not true. On the other hand, if we suppose that λ is not true, then, as that is just what λ says, it seems to follow that λ is true. By case analysis, one can derive the conclusion:

λ is true and λ is not true.

This is a contradiction and, intuitively, should be rejected.

Two of the most influential (post-Tarski) approaches to the liar paradox are those developed by Kripke and Gupta and Belnap, and they will be my focus here.[2] I will sketch a version of Kripke's theory of truth and then a version of Gupta and Belnap's theory of truth.

(Strictly speaking, both Kripke and Gupta and Belnap define multiple theories of truth. My choice of which particular theories of truth to discuss is not essential to the underlying project; I aim to show, by providing a specific example, that metasemantic theories can play a central role in solving the liar paradox. The specific example involves Kripke's strong-Kleene minimal fixed point theory of truth, and the view that the set of Gupta and Belnap's *stable* truths is the set of truths. I make no claim that

[2] An approach similar to Kripke's was independently developed by Martin and Woodruff (1975); see also e.g. Leitgeb 2005; Field 2008. For alternative developments of the revision theory of truth, see e.g. Gupta 1982; Herzberger 1982; Yaqūb 1993.

these theories of truth are the most popular or plausible instances of the Kripkean or revision-theoretic approaches; nor do I claim that the details of the following discussion would carry over *mutatis mutandis* for other instances of the Kripkean and revision-theoretic approaches.)

Kripke's (1975) approach to the liar paradox involves restricting the validity of the law of the excluded middle: there is a class of *ungrounded* sentences, members of which lack a truth value and fall under neither "true" nor "not true". That is, ungrounded sentences are neither in the *extension* nor the *anti-extension* of "true".

If ungrounded sentences are deemed to lack truth values, then it is necessary to appeal to a three-valued logic. Kripke suggests the strong Kleene evaluation scheme, whose values are: true, false, and undefined. On this evaluation scheme: 'not-*p*' is true (false) if '*p*' is false (true), otherwise it is undefined; '*p* and *q*' is true if both conjuncts are true, false if either conjunct is false, and undefined otherwise; and the other connectives can be defined in the usual way. In particular, '*p* iff *q*' will be undefined if either or both of '*p*' and '*q*' are undefined.

Kripke proceeds by constructing the extension and anti-extension for a formal truth predicate, roughly as follows. We begin with an interpreted formal language, which does not contain any semantic vocabulary, but to which we add an initially uninterpreted predicate "Tr". Call the extended language *L*. Now, holding the interpretation of the non-semantic vocabulary fixed throughout, a hierarchy of extensions and anti-extensions (E_α,A_α) will be constructed for "Tr". To begin, $E_0 = A_0$ = the empty set. Then, for any α, $E_{\alpha+1}$ is the set of (codes of) sentences of *L* that come out true by interpreting "Tr" with (E_α,A_α); and $A_{\alpha+1}$ is the set of (codes of) sentences of *L* that come out false by interpreting "Tr" with (E_α,A_α).[3,4] When we obtain an infinite sequence (E_α,A_α), $(E_{\alpha+1},A_{\alpha+1})$ …, a *limit* stage, β, is defined by taking unions: E_β is the union of E_α for $\alpha<\beta$ and A_β is the union of A_α. For example, E_ω is the union of the sets E_n for finite *n*, and A_ω is the union of the sets A_n for finite *n*. It can be proved that there is a

[3] Here, the use of "true" and "false" can be understood classically: "grass is green" is true (false) just in case the referent of "grass" is in the extension (anti-extension) of "is green"; "Tr('grass is green')" is true (false) just in case "grass is green" is in the extension (anti-extension) of "Tr(*x*)"; etc. Truth values of complex sentences are fixed according to the strong Kleene evaluation scheme.

[4] Henceforth, I will largely leave the parenthesised reference to Gödel coding implicit.

minimal fixed point, i.e. a minimal value M such that $(E_M,A_M) = (E_{M+1},A_{M+1})$. E_M is then identified as the extension of "Tr", and A_M is identified as the anti-extension of "Tr". A sentence is *grounded* just in case it is a member of $E_M \cup A_M$; otherwise it is *ungrounded*. Let us call the theory that "Tr" is the *truth* predicate for L—that E_M contains precisely the sentences of L that are true and A_M contains precisely the sentences of L that are not true—*MFP*.[5]

Importantly, λ turns out to be ungrounded. Consequently, the relevant instance of the law of the excluded middle:

λ is true or λ is not true

is invalid. Thus, MFP blocks the use of case analysis in the derivation of the liar paradox.

Gupta and Belnap's (1993) revision theory of truth aims to model the reasoning engaged in by users of "true". Roughly, the idea is to reinterpret instances of the traditional T-schema:

X is true $\leftrightarrow$ *p*

(where "X" is to be replaced by a name of the sentence replacing "*p*") as providing partial definitions of "true". So, for example, where:

(1) (a) "γ" is a name of the sentence "grass is green"

(b) "T(γ)" is a name of the sentence "γ is true"

(c) "λ′" is a name of the sentence "λ′ is false"

we might have the following three definitional instances of the T-schema:

[5] Kripke probably does not endorse the strong thesis that I present here, as he ultimately suggests that, given his construction, λ comes out to be not true. ("The ghost of Tarski's hierarchy is still with us" (Kripke 1975: 714).) I leave this exegetical point to one side in the present discussion.

(2) (a) γ is true $\leftrightarrow_{def}$ grass is green

(b) T(γ) is true $\leftrightarrow_{def}$ γ is true

(c) λ' is true $\leftrightarrow_{def}$ λ' is false.

The connective "$\leftrightarrow_{def}$" is *not* to be understood as a material biconditional; rather, it is to be understood to indicate a rule for revising hypotheses about the truth values of sentences. One starts with a hypothesis about whether sentences are true or false, and then repeatedly revises one's hypothesis by using the definitional instances of the T-schema.

For example, suppose that one starts with the hypothesis, *h*, that γ is false, T(γ) is true, and λ' is true. One then uses (2) to revise *h*, by moving from the right-hand side of each of (2a)–(2c) to the left-hand side.

(3) (a) Using (2a): as (by empirical fact) grass is green, our new hypothesis is that γ is true.

(b) Using (2b): as (by *h*) γ is false, our new hypothesis is that T(γ) is false.

(c) Using (2c): as (by *h*) λ' is true, our new hypothesis is that λ' is false.

According to our new hypothesis, γ is true, T(γ) is false, and λ' is false. If one repeatedly revises hypotheses in this way, one obtains the following results.

(4) (a) After one revision, γ is true on all hypotheses.

(b) After two revisions, T(γ) is true on all hypotheses.

(c) For all consecutive revisions, λ' alternates between true and false.

These three results are not dependent on any initial hypothesis. Whatever the initial hypotheses, after some finite number of revisions, both "grass is green" and "'grass is green' is true" will be *stably true*, whereas the liar sentence will be *unstable*. We will understand the proposal here to be that the extension

of (the English word) "true" contains precisely the sentences that, according to the above framework, are *stable* truths.[6] Thus, γ and T(γ), but not λ′, are true. I will call this theory *RTT*.

Let us assume for argument's sake that either MFP is correct or RTT is correct: one of those theories correctly describes the extension (and anti-extension) of the term "true" in English. The question is: how do we decide which?

2. Magnetism

An answer may be provided by Lewis's (1983, 1984) reference-magnetic metasemantic theory.[7]

Lewis treats the objective *naturalness* of objects and properties as a *reference magnet*. The thought is that, somewhat metaphorically, reference is attracted to the *naturalness* of candidate referents. Two points serve to improve this famous metaphor. First, naturalness is not the only factor in fixing reference: an assignment of referents should also *fit* well with speakers' linguistic behaviour. We can think, then, of *both* naturalness *and* fit as reference magnets: as a result of the two forces, an assignment will end up in equilibrium. Second, it is not so much *reference* that is attracted by naturalness and fit, but rather whole *assignments of semantic values for languages*—henceforth *semantic theories*. So, while singular terms *ceteris paribus* refer to natural objects, predicates *ceteris paribus* express natural properties.

We get a better grasp on relevant conceptions of *fit* and *naturalness* as follows. First, in a Lewisian spirit, one might define the extent to which a semantic theory *fits* with a speaker's linguistic behaviour as the proportion of the sentences accepted by the speaker that, given the semantic theory, come out as *true*.[8] However, in the context of the liar paradox, especially given that we have distinguished between extensions and anti-extensions, it is natural to extend the relevant behaviour to include sentences rejected by the speaker (in the sense of denial). Then, the extent to which a semantic

[6] Thus, we simplify matters by understanding the proposed theory of truth to be Gupta and Belnap's *T**. (See their 1993: 218ff.)

[7] Schwarz (2014) argues that Lewis does *not* endorse Magnetism. I put this exegetical issue to one side throughout.

[8] See e.g. Lewis 1984: 222–224.

theory *fits* with a speaker's linguistic behaviour may be defined as the proportion of sentences that are either: *accepted* by the speaker and, given the semantic theory, come out as *true*; or *rejected* by the speaker and, given the semantic theory, come out as *not-true*. Not all sentences count equal: each is weighted by how clear-cut, how deeply believed or disbelieved, how intuitive or counterintuitive, etc., it is. The more sentences (weighted appropriately) assigned the appropriate truth value, the better.

Understanding the conception of *naturalness* in play will require more subtlety. I will follow J. Robert G. Williams's interpretation, according to which Lewis's conception of naturalness arises from independent considerations of the theoretical virtue of *simplicity*.[9] The idea, roughly, is as follows. First, theoretical terms should ultimately only refer to perfectly natural properties, where the perfectly natural properties are those targeted by fundamental physics (perhaps such as *being a quark*, or *having spin ½*). Second, *simpler* theories are *ceteris paribus* to be preferred over more complex theories, where one aspect of simplicity is *syntactic* simplicity. Here, a theory may be syntactically simpler than another if it, say, has fewer (tokens of) logical connectives and quantifiers, etc. Putting the two claims together, a theory T_1 is *ceteris paribus* to be preferred over a theory T_2 if, when both theories are spelt out in the language of fundamental physics, T_1 is simpler than T_2.

Now, consider a semantic theory, T_1, that has the axiom:

"green" applies to something iff it is green.

Note here that, on the right-hand side, "green" is *used*. It is thus, in the present context, a theoretical term. As such, to measure the simplicity of the semantic theory, "green" would have to be spelt out in primitive terms. Similarly so for another semantic theory, T_2, obtained by replacing the above axiom in T_1 with:

[9] At first sight, Lewis (1984: 228) appears to define naturalness of properties in terms of the *length of its definition in the language of fundamental physics*, but Williams claims that "[i]t is doubtful that this is Lewis's view" (2007: 377). See Williams 2007 for details.

"green" applies to something iff it is grue,

where something is *grue* just in case it is *observed before* t *and green, or not observed before* t *and blue*. Following Williams, one might conclude that T_1 is *simpler* than T_2: regardless of how "green" is defined in the language of fundamental physics, "grue" is defined *in terms of* greenness and will thus plausibly have a *more* complex definition in that language.

From this perspective, only the perfectly natural properties (i.e. those targeted by fundamental physics) are the metaphorical reference magnets; and they are 'magnetic' because (i) theories should ultimately be spelt out in perfectly natural terms, and (ii) simplicity, including syntactic simplicity, is a theoretical virtue. Let me note two advantages of understanding Lewis's view in this way. First, as Williams notes (2007: 377), it would be somewhat ad hoc to impose, without independent motivation, the constraint upon semantic theories that semantic values should generally be reasonably natural objects/properties. The appeal to simplicity provides such independent motivation. Second, as syntactic simplicity is only *one* aspect of simplicity, understanding Lewis in this way provides clear scope for extending his view. In particular, if we cannot determine the syntactic simplicity of competing theories—perhaps because we lack explicit reductions into the language of fundamental physics—, we might look to *other* aspects of simplicity to help us assess those theories. We will return to this idea later.

We can now characterise Lewis's view slightly more formally. I offer two principles, which I take jointly to introduce a (technical) relation of *attractiveness* amongst semantic theories; this relation shall then be used to characterise Magnetism.

> *The principle of fit*. For any two semantic theories for L, T_1 and T_2: if T_1 fits the linguistic behaviour of speakers of L better than T_2, then *ceteris paribus* T_1 is more *attractive* than T_2 (with diminishing returns).

The principle of simplicity. For any two semantic theories for L, T_1 and T_2: if, when spelt out in perfectly natural terms, T_1 is simpler than T_2, then *ceteris paribus* T_1 is more *attractive* than T_2 (with diminishing returns).

The bracketed qualifications "with diminishing returns" in the two principles are required because *perfect* simplicity or fit, at the expense of the other, is *not* attractive. Simplicity and fit are individually important, but neither is *all*-important: a balance between the two is preferable.

We can now state Lewis's metasemantic theory.

Magnetism. The semantic theory for L that is most *attractive* is thereby the *correct* semantic theory for L.

I here make the simplifying supposition that there is a *unique* maximally attractive semantic theory. The supposition need not hold, and there may be some residue indeterminacy. This is typically not taken to be a problem (Lewis states that "the sensible realist won't demand perfect determinacy" (1984: 228)) and I shall not discuss it here.

Although my aim is not to defend Magnetism, a quick note in its favour is in order.[10] The theory captures well the sense in which *natural language semantics*—the task of constructing semantic theories for natural languages—is an empirical enterprise. First, the principal data used in natural language semantics are truth-value judgments of competent speakers, where the competent speakers in question may be the theorists themselves and where these judgments may be weighted by how intuitive or clear-cut they are. And, reflecting this, the principle of fit ensures that the correct semantic theory will vindicate as many of those truth value judgments, weighted appropriately, as possible. Second, as with any other empirical theory, natural language semantics seeks to construct simpler theories where possible. This is reflected directly by the principle of simplicity. As such, Magnetism certainly complements, and can perhaps serve to underpin, the empirical enterprise of natural language semantics.

[10] See also Williams 2007: 371–378.

3. The most attractive theory of truth

Magnetism tells us how to decide between *semantic theories*. But Kripke and Gupta and Belnap provide us with *theories of truth*. So how can we apply the former to the latter?

The answer is to recognise that the two theories of truth identify particular properties, where the properties are candidate semantic values of the word "true". MFP identifies a property, $\text{TRUTH}_{\text{MFP}}$, possessed by just the sentences that, according to MFP, fall under "true".[11] And RTT identifies a property, $\text{TRUTH}_{\text{RTT}}$, possessed by just the sentences that, according to RTT, are stably true. The correct theory of truth, then, is the theory that identifies the property expressed by the English word "true". Or, for our purposes: the correct theory of truth is the theory that identifies the property that is assigned to "true" by the most *attractive* semantic theory for English.

Before we begin, let us assume that the principles of fit and simplicity will jointly result in the assignment of the intuitively correct semantic values to all *non*-semantic vocabulary. (This assumption is acceptable given that the present aim is not to defend Magnetism.) Given this assumption, we can look at what the principles of fit and simplicity tell us about the property expressed by "true". I argue that, while the principle of fit may *slightly* favour the claim that "true" expresses $\text{TRUTH}_{\text{RTT}}$, the principle of simplicity *significantly* favours the claim that "true" expresses $\text{TRUTH}_{\text{MFP}}$.

Let us look first at the principle of fit. Putting aside context sensitivity here and below, speakers by-and-large accept/reject a sentence just in case they accept/reject the truth of that sentence. So, by-and-large, a speaker accepts a sentence "…" just in case she also accepts the sentence "'…' is true" (or, perhaps, "it is true that …"); and, by-and-large, she rejects a sentence "…" just in case she also rejects the sentence "'…' is true" (or, perhaps, "it is true that …"). So, given the principle of fit, a semantic theory for English is likely to be more attractive if it by-and-large vindicates some version of the disquotation schema for English—where the disquotation schema is as follows:

[11] I simplify the issue here by putting aside anti-extensions. This simplification does not affect the results in this section.

"…" is true ↔ …

By the principle of fit, then, we expect the most attractive semantic theory to by-and-large vindicate some version of the disquotation schema. More precisely, we expect it to assign a property to "true" that by-and-large satisfies the disquotation schema. This is sufficient to ensure that the property assigned to "true" will by-and-large be possessed by the intuitively true sentences: roughly, a sentence falls under "true" just when that sentence says something that is the case. Importantly, this result is compatible with both the hypothesis that "true" expresses $\text{TRUTH}_{\text{MFP}}$ and the hypothesis that "true" expresses $\text{TRUTH}_{\text{RTT}}$: both theories of truth by-and-large satisfy the disquotation schema in this way.

So let us look at this issue at a finer grain by considering a few explicit examples. First, almost *any* sentence that ordinary speakers are likely to come across in ordinary circumstances will, according to both MFP and RTT, satisfy the disquotation schema and have its intuitive truth value. Thus, on both theories, we have:

(5) (a) "Grass is green" is true ↔ grass is green.

(b) "It's true that Sara is old, but she's still strong" is true ↔ it's true that Sara is old, but she's still strong.

and, unless we are exceedingly unlucky, we have:

(c) "Everything the Pope says is true" is true ↔ everything the Pope says is true.[12]

As such, for most sentences, our two hypothesised assignments of semantic values (that $\text{TRUTH}_{\text{MFP}}$ is assigned to "true" or that $\text{TRUTH}_{\text{RTT}}$ is assigned to "true") fit equally well.

[12] For (5c) not to hold, we might need, say, *n*-1 of the Pope's *n* assertions to be straightforwardly true and the remaining assertion to be "I sometimes say something false".

The two hypotheses come apart for somewhat less everyday examples, such as the following triad of sentences:

(6) (a) (6b) is true or (6c) is true.

(b) (6a) is true.

(c) (6a) is not true.

It is easy to show (although for brevity I will not do so here) that: on MFP, (6a–c) are ungrounded, (6a–c) have undefined truth values, and the corresponding instances of the disquotation schema have undefined truth values; while, on RTT, (6a) and (6b) are true and (6c) is false, and the corresponding instances of the disquotation schema are unproblematically true.

Let us suppose, then, that some instances of the disquotation schema are preserved by RTT but not by MFP. It might thus appear that the assignment of $\text{TRUTH}_{\text{RTT}}$ to "true" is, according to the principle of fit, the more attractive. However, in contrast to (5a–c), most ordinary speakers would rarely think about, and even more rarely form clear intuitions about, examples such as (6a–c). That is, (6a–c) are not the kind of sentences on which Magnetism will place great weight. The consequence is that, according to the principle of fit, an assignment of $\text{TRUTH}_{\text{RTT}}$ to "true" is at most only *slightly* more attractive than an assignment of $\text{TRUTH}_{\text{MFP}}$ to "true": they agree on the ordinary, everyday sentences that matter, but RTT does slightly better on some sentences that, so far as Magnetism is concerned, are largely irrelevant.[13]

If the preceding comments are right, then we can draw the following conclusion.

[13] This point is not undermined by the fact that *some* ordinary speakers *do* engage with logical puzzles of this kind, and *do* form intuitions about how to evaluate such sentences; nor is it undermined by the fact that there have been subtle philosophical discussions about the intuitive analysis of such examples (e.g. Cook 2002, 2003; Kremer 2002). Regardless of such facts, the vast majority of speakers do not have clear-cut, deeply held beliefs about such cases; for example, most ordinary speakers could be convinced that (6a–c) are untrue far more easily than they could be convinced that, say, (5a,b) are untrue.

(C1) The principle of fit at most *slightly* favours RTT over MFP. More precisely: the principle of fit deems a semantic theory that assigns $\text{TRUTH}_{\text{RTT}}$ to "true" to be at most (*ceteris paribus*) *slightly* more attractive than a semantic theory that assigns $\text{TRUTH}_{\text{MFP}}$ to "true".

Let us now look to the relative simplicity of semantic theories that differ only in whether they assign $\text{TRUTH}_{\text{MFP}}$, or $\text{TRUTH}_{\text{RTT}}$, to "true".

There are two difficulties that arise when we attempt to apply the principle of simplicity. Recall the underlying idea: (i) theories are ultimately to be spelt out in the language of fundamental physics; and (ii) simplicity, one aspect of which is syntactic simplicity, is a theoretical virtue. Both (i) and (ii) give rise to a difficulty. First, neither $\text{TRUTH}_{\text{MFP}}$ nor $\text{TRUTH}_{\text{RTT}}$ is obviously definable in the language of fundamental physics. Second, even if they were, we do not in fact have to hand *any* explicit definitions of $\text{TRUTH}_{\text{MFP}}$ and $\text{TRUTH}_{\text{RTT}}$; as such, we are not in the position to examine the relative syntactic simplicity of those definitions. Let us take the difficulties in turn.

First, then, it seems that we need definitions of $\text{TRUTH}_{\text{MFP}}$ and $\text{TRUTH}_{\text{RTT}}$ in the language of *fundamental physics*. But, on the face of it, *truth* simply does not appear to be the sort of property that can be defined in the language of fundamental physics—the property of truth appears to be *abstract*, whereas the language of fundamental physics appears to be tailored towards the *physical*.[14]

To resolve this difficulty, note the following: whereas it might not be appropriate to consider the property of truth as being definable in the language of *fundamental physics*, it does seem appropriate to consider the property of truth as being definable in the language of *arithmetic*. In particular, the latter allows us (via Gödel coding) to characterise and study the semantics of formally characterised sentences and, as such, is the obvious platform for constructing rigorous, candidate definitions of the property of

[14] Both Douglas Edwards (2013: 2, 7–9) and Brian Weatherson (2003: 11, 22) appear to implicitly accept that truth *can* be defined in the language of fundamental physics: they accept Lewis's conception of naturalness, and then allow that *truth* is a 'reasonably natural' property. However, as neither Edwards nor Weatherson attempt to explain how truth can be defined in the language of fundamental physics, I put this aside.

truth. On this line of thought, we should interpret Lewis's conception of naturalness in such a way so as to allow definitions in the language of fundamental physics *and* in the language of arithmetic.

There are (at least) three ways to spell out this line of thought. First, one might claim that, as the laws of physics make use of mathematics, the numbers (and addition, multiplication, etc.) are themselves perfectly natural by Lewis's lights. That is, the language of fundamental physics would *already contain* the language of arithmetic. If this is right, then the predicates "is a quark" and "is a number" might both express perfectly natural properties. Second, one might endorse the first suggestion in spirit but not in its detail; one might suggest that, rather than *numbers* being perfectly natural, it is the arithmetical primitives—i.e. the successor relation and 0—that are perfectly natural. If this is right, then the predicates "is a quark" and "is the number 0" might both express perfectly natural properties, but "is the number 1" would ultimately have to be spelt out as "is the successor of the number 0", and "is the number 2" would ultimately have to be spelt out as "is the successor of the successor of the number 0", and so on. Third, one might deny that either the numbers or the arithmetical primitives are perfectly natural, but instead claim that they can *all* be spelt out in perfectly natural terms. Such an approach might appeal to those who are inclined to *reduce* the language of arithmetic to the language of physics.[15]

I take it that those who endorse Magnetism (in the form under discussion here) are likely to accept one of the three suggestions in the previous paragraph: otherwise, such theorists would have little hope of accounting for mathematical (and, in particular, arithmetical) language. Let us suppose that this so. For present purposes, however, we may remain neutral about which of the three suggestions is to be preferred.

The second difficulty is that, without explicit definitions of $\text{TRUTH}_{\text{MFP}}$ and $\text{TRUTH}_{\text{RTT}}$, it is unclear how we can measure their simplicity. In particular, it is unclear how we can measure the *syntactic* simplicity of axioms that, in primitive terms, assign $\text{TRUTH}_{\text{MFP}}$ or $\text{TRUTH}_{\text{RTT}}$ to "true".

However, as discussed above, *syntactic* simplicity is only *one* aspect of simplicity. And there is another aspect of simplicity that is salient in the present context. We are concerned with two

[15] This may be Lewis's preferred strategy; see e.g. his 1993.

properties—$\text{TRUTH}_{\text{MFP}}$ and $\text{TRUTH}_{\text{RTT}}$—whose extensions can be formally constructed. As such, we can appeal to the branch of mathematics that studies the complexity of sets—where *complexity* is a measure of how hard it is to compute a given set. Let me briefly explain an important result, and then relate it back to Magnetism.

It has been proved that the extension of $\text{TRUTH}_{\text{RTT}}$ is significantly more complex than the extension of $\text{TRUTH}_{\text{MFP}}$.[16] Roughly, the results in question show that a Turing machine that knew all of the membership facts about the extension of $\text{TRUTH}_{\text{RTT}}$ (i.e. a Turing machine equipped with an oracle for that set) could output all of the membership facts about the extension of $\text{TRUTH}_{\text{MFP}}$, but not vice versa. (Here, I understand *membership facts* about a set, X, to consist of facts of the form: x_1 belongs to X; x_2 does not belong to X; etc.) That is: there is a Turing machine that knows all of the membership facts about the extension of $\text{TRUTH}_{\text{RTT}}$ that, input any (code of a) sentence, s, will output in a finite amount of time a "1" if s is in the extension of $\text{TRUTH}_{\text{MFP}}$ or a "0" if s is not the extension of $\text{TRUTH}_{\text{MFP}}$; but there is *no* Turing machine that knows all of the membership facts about the extension of $\text{TRUTH}_{\text{MFP}}$ that can similarly output the membership facts of $\text{TRUTH}_{\text{RTT}}$. We can gloss this result thus: the extension of $\text{TRUTH}_{\text{RTT}}$ is an order of magnitude more complex than the extension of $\text{TRUTH}_{\text{MFP}}$.

Now, the proposal is this: for arithmetical properties such as $\text{TRUTH}_{\text{MFP}}$ and $\text{TRUTH}_{\text{RTT}}$, we take the complexity of their extensions to provide an inverse measure of simplicity. That is, in light of the above result, we take a semantic theory that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" to be significantly *simpler* than a semantic theory that differs only in that it assigns $\text{TRUTH}_{\text{RTT}}$ to "true".

There are at least two ways that one might spell this out. First, one might suggest that the computational complexity of the extension of a property is an inverse measure of the simplicity of the property itself. The intuitive idea here might be this. Suppose that an agent has complete knowledge of which perfectly natural objects possess which perfectly natural properties. Then: the more steps that the agent may have to take to use this knowledge to determine whether some arbitrary x is F, the less simple the property of *F-ness*. For example, we might think that the property of greenness is simpler than the

[16] For precise statements of the results and technical details, see Burgess 1986 and Welch 2001.

property of grueness in the following sense: whatever it takes to determine whether some arbitrary *x* is green, it may take that *and more* to determine whether that *x* is grue. (I.e. in addition, it may also be required to determine whether *x* has been observed before *t*, whether it is blue, etc.) Computational complexity may be thought of as formalising and extending this intuitive idea. From this perspective, the above result would be understood as straightforwardly demonstrating that the property of $\text{TRUTH}_{\text{MFP}}$ is an order of magnitude simpler than $\text{TRUTH}_{\text{RTT}}$. One might then infer that a semantic theory that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" is *ceteris paribus* significantly simpler than a semantic theory that assigns $\text{TRUTH}_{\text{RTT}}$ to "true".

Second, consider an algorithm that, given any input sentence, would tell us whether that sentence is in the extension of $\text{TRUTH}_{\text{MFP}}$ or $\text{TRUTH}_{\text{RTT}}$. One might claim that such an algorithm is, in one good sense, to be understood as extensionally defining those properties. However, in the present case, there are no such algorithms that proceed solely by manipulating the input sentence according to a finite set of instructions. Rather, any algorithm that is to count as providing such an extensional definition of $\text{TRUTH}_{\text{MFP}}$ or $\text{TRUTH}_{\text{RTT}}$ must be endowed with an *additional* stock of information. As such, we might hope to make sense of *how simple the definitions of* $\text{TRUTH}_{\text{MFP}}$ *and* $\text{TRUTH}_{\text{RTT}}$ *are* in terms of *how little* additional information such extensional definitions would need. From this perspective, the result discussed above is that an order of magnitude less additional information is required to extensionally define $\text{TRUTH}_{\text{MFP}}$ than $\text{TRUTH}_{\text{RTT}}$—and thus that a semantic theory that assigns the former to "true" is *ceteris paribus* significantly simpler than a semantic theory that assigns the latter to "true".

There is obviously a lot more that could be said, and would ultimately have to be said, about these issues.[17] However, it is clear that there are strategies available to the supporter of Magnetism to

[17] For example, here are two issues that would have to be dealt with in a more complete discussion of this proposal. First, the technical result I have appealed to applies in the first instance to simple formal languages such as *L*, but I am concerned with significantly more complex *natural* languages. Ultimately, then, the technical result would need to be generalised if it is to be used in this way. Second, to the extent that some apparently gerrymandered properties are computationally simple, the present proposal may have some rather counter-intuitive consequences. While such issues ultimately deserve a full discussion, I put them to one side here.

incorporate the computational complexity results introduced above. And, whatever strategy is adopted, it will, if successful, lead to the following conclusion:

(C2) The principle of simplicity *significantly* favours MFP over RTT. More precisely: the principle of simplicity deems a semantic theory that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" to be (*ceteris paribus*) *significantly* more attractive than a semantic theory that assigns $\text{TRUTH}_{\text{RTT}}$ to "true".

Finally, let us now turn to consider the two conclusions (C1) and (C2) together.

According to (C1), RTT is at most *slightly* more attractive than MFP. According to (C2), MFP is *significantly* more attractive than RTT. As such, on balance, it seems that MFP is the more attractive. Of course, this judgment rests on certain assumptions about how the principles of fit and simplicity are to be weighted; but, given that Lewis intends natural properties to do some serious work in fixing the semantic values, it is fair to assume that the principle of fit is certainly not significantly more important than the principle of simplicity. As such, on any plausible construal of how the principles of fit and simplicity are to be weighted, MFP's failure to fit our linguistic behaviour with regard to (e.g.) unusual triads of sentences will be outweighed by the significantly superior simplicity of (the extension of) $\text{TRUTH}_{\text{MFP}}$.

We therefore draw the following conclusion.

(C3) On balance, a semantic theory that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" is (*ceteris paribus*) more attractive than a semantic theory that assigns $\text{TRUTH}_{\text{RTT}}$ to "true". According to Magnetism (and assuming that "true" expresses $\text{TRUTH}_{\text{MFP}}$ or $\text{TRUTH}_{\text{RTT}}$), "true" expresses $\text{TRUTH}_{\text{MFP}}$.

So Magnetism, as we have understood it, favours Kripke's MFP over Gupta and Belnap's RTT.

4. Other criteria

I now discuss an important style of objection that might be raised to the foregoing.

There are a range of criteria that are commonly used when assessing solutions to the liar paradox. For example: whether the solution faces a revenge paradox; whether the solution preserves classical logic; whether the solution permits a detachable conditional in the object language; and so on.[18] And any given theorist is likely to take some of these criteria to be important for determining the correct solution to the liar paradox.

This state of affairs may lead to following style of objection. *Criterion X is important for determining the correct solution to the liar paradox. Yet, as the above discussion shows, Magnetism appears not to leave room for criterion X to play such a role. Therefore, given the importance of criterion X, Magnetism cannot play a central role in determining the correct solution to the liar paradox (and thus it has not been shown that metasemantic theories can play a central role in determining the correct solution to the liar paradox).*

For each criterion, there are three basic strategies that one can take in response to the objection. First, one might explain why, contrary to the objection, the criterion need *not* play a role in determining the correct solution to the liar paradox. Or, second, one might argue that the criterion is, in an appropriate sense, subsumed by Magnetism. Or, third, one might argue that Magnetism *does* leave room for the criterion to play an important role. To see how the basic strategies might be developed, it is useful to focus on a concrete example. We will focus on the revenge paradox.

A solution to the liar paradox faces a *revenge paradox* if, using the resources introduced to formulate the original solution, one can construct a new liar sentence that gives rise to a close analogue of the original paradox. For example, making use of Kripke's technical concepts of truth and ungroundedness, we can construct $\lambda^{\dagger}$:

$\lambda^{\dagger}$ is ungrounded or not true.

Now, suppose $\lambda^{\dagger}$ is true. Then, given that $\lambda^{\dagger}$ says that $\lambda^{\dagger}$ is ungrounded or not true, it follows that $\lambda^{\dagger}$ is ungrounded or not true. Suppose instead that $\lambda^{\dagger}$ is ungrounded. Then, as that is what the first disjunct

[18] Respectively, see e.g. Scharp 2013, Williamson forthcoming, and Field 2008. See also Leitgeb 2007.

of $\lambda^{\dagger}$ says, it follows that $\lambda^{\dagger}$ is true. Suppose instead that $\lambda^{\dagger}$ is not true. Then, as that is what the second disjunct of $\lambda^{\dagger}$ says, it follows that $\lambda^{\dagger}$ is true. By case analysis, one derives:

($\bot^{\dagger}$) $\lambda^{\dagger}$ is true and $\lambda^{\dagger}$ is ungrounded or not true.

As truth and ungroundedness are exclusive, this is a contradiction.

A specific objection, then, may run as follows. *As* ungroundedness *can be expressed in English, MFP does not solve the liar paradox so much as shift it to another sentence. This threatens to seriously undermine Kripke's MFP. It is implausible that such a revenge paradox is irrelevant to determining the correct solution to the liar paradox. Yet Magnetism apparently favours MFP over RTT without leaving room for consideration of such revenge paradoxes. So it is implausible that Magnetism provides us with the appropriate strategy for determining the correct solution to the liar paradox (and thus it remains unclear whether metasemantic theories can play a central role in determining the correct solution to the liar paradox).*

As noted above, there are three basic strategies for responding to such an objection. In what follows, I briefly indicate in turn how each could be developed. First, then: one might explain why, contrary to the objection, revenge paradoxes need *not* play a role in determining the correct solution to the liar paradox.

There are at least two ways that this basic strategy could be developed. First, one might argue that revenge paradoxes are significantly less serious than sometimes suggested. A number of authors have already developed responses to the revenge paradoxes along these lines. For example: Beall (2007: 11–12) suggests that it is simply "too easy" to just take a revenge paradox to constitute an objection; Maudlin (2004) allows permissible assertion to outstrip truth, thereby blocking the inference from $\lambda^{\dagger}$ to the truth of $\lambda^{\dagger}$ in the derivation of ($\bot^{\dagger}$), and is happy to "learn to live with" (p. 177) the consequence that revenge paradoxes can nonetheless be constructed using his notion of permissible assertion; and Shapiro (2011) argues that, when a revenge paradox is aimed at a solution to the liar paradox, it effectively serves only to restate the problem that the solution initially solved, and thus can be

disregarded.[19] If any such response to revenge paradoxes is successful, then we would seemingly be justified in putting revenge paradoxes to one side when determining the correct solution to the liar paradox.

Alternatively, one might seek independent grounds for denying that the technical term "ungrounded" is part of English in any theoretically interesting sense. Independent grounds might be provided, for example, by the dominant generative tradition in linguistics and contemporary philosophy of language.[20] Very roughly, generative linguistics studies the innately acquired, generative procedures that underpin linguistic abilities in humans; for the generative linguist, a term is part of a natural language in a theoretically interesting sense principally if it can be learnt during the ordinary course of first language acquisition. On the presumption that the technical term "ungrounded" cannot be so learnt, the generative linguist may conclude that it is not part of English in a theoretically interesting sense. From such a perspective, consideration of such a sentence as $\lambda^{\dagger}$ may be simply irrelevant to the study of English—and *a fortiori* to the study of the liar paradox as it arises in English.[21]

Let us move on to the second basic strategy: one might argue that consideration of revenge paradoxes is, in an appropriate sense, subsumed by Magnetism. The idea would be that the principles of fit and similarity, together, determine how revenge paradoxes are to be resolved.

Consider, for example, ordinary speakers of English who lack technical knowledge of the literature on the liar paradox.[22] Call the specific variant of English spoken by such people "Ordinary

[19] Although, for extensive criticism of such attempts to resolve revenge paradoxes, see e.g. Scharp 2013 (esp. ch 4).

[20] See e.g. Chomsky 2000. For a philosophical discussion, see e.g. Ludlow 2011.

[21] Of course, much more would ultimately have to be said to spell out this particular line of thought. (E.g., why should the linguist have the final say about which aspects of natural language are theoretically interesting?) But the specifics should not detract from the underlying point, viz. that some of the criteria that are traditionally used to assess solutions to the liar paradox may be theoretically uninteresting (or even irrelevant) when we are focused on *natural* language.

[22] I put aside discussion of e.g. philosophical logicians. Philosophical logicians use, discuss and have deeply held beliefs about all sorts of sentences closely connected to liar and revenge paradoxes—including, e.g., sentences such as λ, $\lambda^{\dagger}$, $(\bot)$ and $(\bot^{\dagger})$, other sentences containing relevant technical vocabulary, sentences that purport to describe logical truths or rules of inference, etc. This complicates matters significantly, but does not give rise to serious theoretical concerns. Ultimately, to provide a plausible interpretation of such speakers' linguistic behaviour one may have to consider their languages individualistically:

English". Speakers of Ordinary English might accept or reject *some* sentences containing "ungrounded", as "ungrounded" is typically taken to have at least two non-technical meanings. For example, the speakers might accept "Mike's accusation was ungrounded" when the accusation lacked evidence, or reject "my lightening rod was ungrounded" when the rod was earthed, and so on. Such patterns of usage generally fit very poorly with the technical definition of "ungrounded". Given Magnetism it plausibly follows that, regardless of considerations of simplicity, "ungrounded" will thus not have its technical definition in Ordinary English. Plausibly, then, Magnetism implies that Ordinary English lacks the expressive resources to formulate $\lambda^{\dagger}$, blocking the derivation of $(\bot^{\dagger})$.

The details of the solution to the revenge paradox are not important here. What is important is the thought that, in one clear sense, Magnetism may already have the resources to solve revenge paradoxes. Perhaps, alternatively, speakers of Ordinary English *can* use "ungrounded" in its technical sense. But then, on various plausible assumptions—that ordinary speakers by-and-large reject explicit contradictions, or that the truth of $(\bot^{\dagger})$ would imply the truth of all sentences and ordinary speakers by-and-large do not accept all sentences, etc.—, it would nonetheless seem likely that the optimal balance of fit and simplicity would be a semantic theory on which, for one reason or another, $(\bot^{\dagger})$ is not true. And, even if $(\bot^{\dagger})$ did come out true, then the most attractive semantic theory would compensate for this elsewhere: perhaps "$\lambda^{\dagger}$" would have a different semantic value, or perhaps "and" and "not" would be nonclassical in such a way that $(\bot^{\dagger})$ would not be a contradiction, or something else.[23] One way or another, so goes the thought, revenge paradoxes will be solved. Given that ordinary speakers are largely coherent, a semantic theory that solves such paradoxes will generate sufficient gains in attractiveness

given the significant variation in relevant linguistic behaviour between philosophical logicians, Magnetism may yield different solutions to both liar and revenge paradoxes for different philosophical logicians. It would take us too far afield to explore this idea here.

[23] In such scenarios, it might turn out that Magnetism does not favour MFP. However, firstly, such scenarios strike me as unlikely, given that there will almost no loss in fit for a semantic theory that does permit the expression of ungroundedness; and, secondly, even if Magnetism turns out *not* to favour MFP, the underlying claim that Magnetism (and thus metasemantic theories more generally) can help us solve the liar paradox would still hold.

through increases in *fit* to outweigh any countervailing losses. In this way, Magnetism may subsume consideration of revenge paradoxes.

Finally, I turn to the third basic strategy: one might argue that, assuming Magnetism does not subsume consideration of revenge paradoxes in the above sense, Magnetism *does* leave room for revenge paradoxes to play an important role in determining the correct solution to the liar paradox. There are at least two ways that the basic strategy could be developed.

First, one might claim that revenge paradoxes provide a meta-level constraint on metasemantic theories. For example, such a constraint might be: a good metasemantic theory should *not* favour a solution to the liar paradox that gives rise to a revenge paradox. From this perspective, our overarching strategy for solving the liar paradox might be this: first, we might collect together a range of apparently viable metasemantic theories; second, for each of these metasemantic theories, we might establish which solution (or solutions) to the liar paradox it favours; third, by establishing whether the favoured solutions face serious revenge paradoxes, we might rule out the corresponding metasemantic theories.

If revenge paradoxes are to be taken to provide meta-level constraints on metasemantic theories, then the discussion in §3 will ultimately form part of a larger investigation into the correct combination of metasemantic theory and solution to the liar paradox. That is, it would not be viable simply to assume Magnetism and consequently endorse MFP. Rather, given that MFP faces a revenge paradox, the fact that Magnetism favours MFP would ultimately be a *criticism* of Magnetism—in virtue of which, perhaps, we would be justified in *rejecting* Magnetism.

Nonetheless, to be explicit, the two theses of this paper would stand. First, it would still be the case that Magnetism appears to favour MFP over RTT; this might simply be recast as an objection to Magnetism. And, second, it would still be the case that metasemantic theories could play a central role in determining the correct solution to the liar paradox; there would simply be additional, meta-level criteria (such as revenge) for deciding between metasemantic theories.

There is a second way in which the basic strategy—arguing that Magnetism does leave room for revenge paradoxes to play an important role in determining the correct solution to the liar paradox—might be developed. One might argue that Magnetism can be extended so as to take revenge paradoxes into account.

One way of extending Magnetism builds upon the idea of theoretical virtues. Recall that, following Williams, we have understood Lewis' conception of naturalness to be one aspect of the theoretical virtue of simplicity. And, although this was not explicit before, we might likewise have understood the principle of fit to arise from a theoretical virtue; the principle of fit may be construed as one aspect of what Kuhn (1977) called the theoretical virtue of *accuracy*. As Kuhn states, "within its domain, […] consequences deducible from a theory should be in demonstrated agreement with the results of existing experiments and observations" (1977: 321).

Now, cast in this way, Magnetism embodies the application of *general* criteria for theory choice to the *semantic* domain. There are, however, more than two theoretical virtues. For example, Kuhn listed three more: a theory should be *consistent* both internally and with other currently accepted views; a theory should have *wide explanatory scope*; and a theory should be *fruitful*, leading to new research findings. One way to extend Magnetism would be to build more theoretical virtues into it.

In order to take revenge paradoxes in particular into account, one might extend Magnetism by building into it the theoretical virtue of consistency. This can be done straightforwardly by supplementing the principles of fit and simplicity with a principle of consistency:

> *The principle of consistency*. For any two semantic theories for L, T_1 and T_2: if T_1 is consistent with a wider range of other currently accepted views than T_2, then *ceteris paribus* T_1 is more *attractive* than T_2 (with diminishing returns).

I assume that we can count the 'empty theory' as a currently accepted view, so that we can straightforwardly take the internal consistency of semantic theories into account.

Depending on the specifics of the case, there are two ways in which the principle of consistency allows Magnetism to take a given revenge paradox into account. First, a revenge paradox may show a corresponding semantic theory to be internally inconsistent. For example, the above derivation of ($\bot^{\dagger}$) may be taken to show that any sensible semantic theory for English that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" would be internally inconsistent. The principle of consistency would immediately deem such semantic theories to be highly unattractive. Second, a revenge paradox may only show a corresponding semantic

theory to be inconsistent with other, currently accepted views. For example, one might deny that the above derivation of ($\perp^{\dagger}$) shows that a sensible semantic theory for English that assigns $\text{TRUTH}_{\text{MFP}}$ to "true" would be internally inconsistent. Rather, one might take such a semantic theory to be inconsistent with: logic (which underpins the inferential steps required to derive ($\perp^{\dagger}$)); and/or the view that natural languages are extendable (which might underpin the claim that ungroundedness, in Kripke's technical sense, is expressible in English); and/or the view that truth is compositional (which might justify the interpretation of $\lambda^{\dagger}$, given the interpretation of its parts); and so on. Insofar as such views are currently accepted, the principle of consistency would deem such a semantic theory to be unattractive.

Thus, Magnetism could be extended to take revenge paradoxes into account. Of course, as it stands, (C3) would no longer hold. To draw a conclusion about the correct solution to the liar paradox, one would have to investigate the extent to which revenge paradoxes affect the consistency of semantic theories that assign $\text{TRUTH}_{\text{MFP}}$ to "true", and of semantic theories that assign $\text{TRUTH}_{\text{RTT}}$ to "true".[24] If the revenge paradox engendered by MFP is not significantly more serious than that engendered by RTT, Magnetism may nonetheless favour MFP over RTT.

5. Concluding remark

Lewis's Magnetism, it seems, can go a long way towards determining the correct solution to the liar paradox, apparently favouring Kripke's MFP over Gupta and Belnap's RTT. But there is a more general lesson: metasemantic theories can play a central role in determining the correct solution to the liar paradox.

The general lesson does not require one to adopt *Lewis's* metasemantic theory, or to consider the relevant instances of *Kripke's* or *Gupta and Belnap's* approaches to the liar paradox.

For example, consider the *cognitivist* view that semantic theories are grounded by appeal to the contents of a semantic module (e.g. Borg 2004, 2012). From this starting point, we might hold out hope that the semantic module encodes a semantic theory that could be taken to reveal the appropriate

[24] Scharp (2013: 86) uses a variant of "this sentence is either false or unstable" to raise a revenge paradox for RTT.

strategy for solving the liar paradox.[25] Or perhaps, when we look more closely at the semantic module, we will discover that the meaning of "true" is in some sense *inconsistent*, vindicating to some extent the so-called inconsistency theories of the liar paradox.[26] Or, alternatively, consider Davidson's (1973) view that meaning is fixed from the external standpoint of a radical interpreter: perhaps the radical interpreter would construct a semantic theory that treats "true" in accordance with some particular solution to the paradox. If we look from the standpoint of the radical interpreter, perhaps we will discover that "true" is context sensitive.[27]

The relationship between metasemantic theories and the liar paradox is underexplored. But there are clear grounds for optimism that the study of the relationship would prove fruitful—and, perhaps, we might ultimately be led to the highly attractive combination of a principled metasemantic theory along with the correct solution to the liar paradox.

References

Beall, J.C. 2007. Prolegomenon to future revenge. In Beall, J.C., (ed.) *Revenge of the Liar: New Essays on the Paradox*, Oxford: Oxford University Press, pp. 1–30.

Borg, Emma. 2004. *Minimal Semantics*. Oxford: Oxford University Press.

——. 2012. *Pursuing Meaning*. Oxford: Oxford University Press.

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

Burgess, John P. 1986. The truth is never simple. *Journal of Symbolic Logic* 51(3): 663–681.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Cook, Roy. 2002. Counterintuitive consequences of the revision theory of truth. *Analysis* 62(1): 16–22.

——. 2003. Still counterintuitive: a reply to Kremer. *Analysis* 63(3): 257–261.

[25] See Pinder 2014, 2015 for a discussion of some of the difficulties facing this approach.

[26] See e.g. Eklund 2002; Patterson 2009; Scharp 2013.

[27] See e.g. Burge 1979; Glanzberg 2001; 2004.

Davidson, Donald. 1973. Radical interpretation. *Dialectica* 27(3/4): 313–328.

Edwards, Douglas. 2013. Naturalness, representation and the metaphysics of truth. *European Journal of Philosophy* 21(3): 384–401.

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

Field, Hartry H. 2008. *Saving Truth from Paradox*. Oxford: Oxford University Press.

Glanzberg, Michael. 2001. The liar in context. *Philosophical Studies* 103(3): 217–251.

——. 2004. A contextual-hierarchical approach to truth and the liar paradox. *Journal of Philosophical Logic* 33(1): 27–88.

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

Gupta, Anil, and Nuel Belnap. 1993. *The Revision Theory of Truth*. Cambridge, MA: The MIT Press.

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

Kremer, Michael. 2002. Intuitive consequences of the revision theory of truth. *Analysis* 62(4): 330–336.

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

Kuhn, Thomas. 1977. Objectivity, value judgment, and theory choice. In *The Essential Tension: Selected Studies in Scientific Tradition and Change*, Chicago: University of Chicago Press, pp. 320–339.

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

——. 2007. What theories of truth should be like (but cannot be). *Philosophy Compass* 2(2): 276–290.

Lewis, David. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 62(3): 221–236.

——. 1984. Putnam's paradox. *Australasian Journal of Philosophy* 62(3): 221–236.

——. 1993. Mathematics is megethology. *Philosophica Mathematica* 1(1): 3–23.

Ludlow, Peter. 2011. *The Philosophy of Generative Linguistics*. Oxford: Oxford University Press.

Martin, Robert L., and Peter W. Woodruff. 1975. On representing 'True-in-L' in L. *Philosophia* 5(3): 217–221.

Maudlin, Tim. 2004. *Truth and Paradox*. Oxford: Oxford University Press.

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

Pinder, M. 2014. Borg's minimalism and the problem of paradox. In P. Stalmaszczyk (ed.) *Semantics and Beyond: Philosophical and Linguistic Inquiries*, Berlin and Boston: De Gruyter, pp. 207–230.

——. 2015. The cognitivist account of meaning and the liar paradox. *Philosophical Studies* 172(5): 1221–1242.

Scharp, Kevin. 2013. *Replacing Truth*. Oxford: Oxford University Press.

Schwarz, Wolfgang. 2014. Against magnetism. *Australasian Journal of Philosophy* 92(1): 17–36.

Shapiro, Lionel. 2011. Expressibility and the liar's revenge. *Australasian Journal of Philosophy* 89(2): 297–314.

Weatherson, Brian. 2003. What good are counterexamples? *Philosophical Studies* 115(1): 1–31.

Welch, Philip. 2001. On Gupta-Belnap revision theories of truth, Kripkean fixed points, and the next stable set. *Bulletin of Symbolic Logic* 7(3): 345–360.

Williams, J. Robert G. 2007. Eligibility and inscrutability. *Philosophical Review* 116(3): 361–399.

Williamson, Timothy. Forthcoming. Semantic paradoxes and abductive methodology. In Armour-Garb, B., (ed.) *The Relevance of the Liar*, Oxford: Oxford University Press.

Yaqūb, Aladdin M. 1993. *The Liar Speaks the Truth: A Defense of the Revision Theory of Truth*. Oxford: Oxford University Press.