SEMANTIC REGULARITY AND THE LIAR PARADOX

Nicholas J.J. Smith[1]

DRAFT: Inference and Meaning Conference, Melbourne, July 2004

§ 1. introduction. There are two main tasks in which writers on the Liar paradox might see themselves as engaged. The first task is saying what is going on in a natural language such as English when we say things such as 'This sentence is not true'. The second task is constructing consistent formal systems which have as much expressive power as possible, or which can express certain notions in which the author is particularly interested. Many authors attempt both tasks at once: they present a consistent formal system, and argue that it provides a model of natural languages such as English.

My task here is the first one. I do present a consistent formal system and claim that it provides a perfect model of natural languages such as English, but this system involves no surprises. It is none other than the standard framework of classical logic and model theory. The real weight of the argument lies in the claim that the classical framework—without alteration or addition—contains the resources to model what happens when we say in English 'This sentence is not true'.

Apart from the fact that it is one hundred percent classical, the solution to the Liar to be presented here has two other notable features. First, it does not generate a strengthened Liar paradox or *revenge problem*. Second, the entrenched belief which the solution asks us to relinquish—and of course it must ask us to give up some such belief, for the Liar would not be a *paradox* if it could be solved without giving up anything—(a) is a belief whose proper home is philosophy of language, not logic; (b) is a quite general belief about the operation of language, rather than a particular belief about a certain class of words—in particular it is not a belief specifically about 'is true'; and (c) is also the key to a host of other deep philosophical problems such as Quine's problem of the indeterminacy of meaning, Kripkenstein's sceptical puzzle, Putnam's paradox, the problem of empty names, and a recalcitrant problem about the semantics of vague language—the problem of false precision.

§ 2. the classical framework. I have said that my solution stays entirely within the bounds of classical logic and model theory. To begin, I

[1]Department of Philosophy, and Centre for Logic, Language and Computation, Victoria University of Wellington, PO Box 600, Wellington, New Zealand. <nicholas.smith@vuw.ac.nz> <http://www.vuw.ac.nz/phil/njjs/>

need to present the classical picture to which I shall remain faithful. This picture involves three parts: (i) A core part, which can be found in any standard logic textbook, such as Boolos et al. [2, chs 9, 10], Mendelson [15, ch. 2] or Shoenfield [23, ch. 2]. (ii) A gloss which tells us how properly to understand the core part. This gloss is not included in the standard textbooks, although it is included in works such as Halmos and Givant [8]. (iii) A minimal extension of the core part to enable us to deal with ordinary language, rather than the formal languages studied in the aforementioned logic textbooks.

§ 2.1. THE CORE PART. (This section contains entirely standard material. It is included (a) for the sake of having a particular presentation to which I can later refer, (b) for readers who are not entirely familiar with this material and (c) as a vivid reminder of just how little basic machinery the view of the Liar to be presented here presupposes. Feel free to skip the whole section, except for the last sentence.) The core part comes in two sections: syntax and semantics (or model theory). In the syntactical part, we specify a language: we choose some primitive symbols, and then say how they may be combined to form well formed formulas. The symbols are as follows:

- individual variables $x, y, z, \dots$
- individual constants $a, b, c, \dots$
- n-place predicate letters $P^n, Q^n, R^n, \dots$ for each $n \geq 1$
- the identity predicate $=$
- the propositional connectives $\neg$, $\vee$, $\wedge$, and $\rightarrow$
- the quantifiers $\exists$ and $\forall$
- the punctuation marks (,) and ,

We next define the notion of a *term* of the language: variables and individual constants are terms; nothing else is a term. Finally we specify the well formed formulas (wfs):

- if $t_1, \dots, t_n$ are terms and P^n is an n-place predicate, then $P^n(t_1, \dots, t_n)$ is a wf;[2]
- if $\mathcal{A}$ and $\mathcal{B}$ are wfs and y is a variable, then $(\neg\mathcal{A})$, $(\mathcal{A}\vee\mathcal{B})$, $(\mathcal{A}\wedge\mathcal{B})$, $(\mathcal{A} \rightarrow \mathcal{B})$, $((\exists y)\mathcal{A})$, and $((\forall y)\mathcal{A})$ are wfs;

[2]In the case of one-place predicates I shall in practice write Pa instead of $P^1(a)$, etc.

- nothing else is a wf.

In the semantical part, we define the notion of an *interpretation* of our language, and we say how to determine the truth of closed wfs on such an interpretation. An interpretation $\mathfrak{M} = (M, I)$ of the language consists in:

- A nonempty set M (the domain)
- An interpretation function I which assigns:
  - to each individual constant a, an object $a^{\mathfrak{M}}$ in the domain (its referent)
  - to each n-place predicate letter P^n, a set $P^n_{\mathfrak{M}}$ of n-tuples of members of the domain (its extension).

The truth value $[\mathcal{A}]_{\mathfrak{M}}$ of the closed wf $\mathcal{A}$ on the interpretation $\mathfrak{M}$ is specified recursively, as follows; the value is either 0, representing falsity, or 1, representing truth:

1. $[P^n(a_1, \ldots, a_n)]_{\mathfrak{M}} = 1$ iff $(a_1^{\mathfrak{M}}, \ldots, a_n^{\mathfrak{M}}) \in P^n_{\mathfrak{M}}$

   thus Pa is true iff the referent of a is in the extension of P

2. $[\neg\mathcal{A}]_{\mathfrak{M}} = 1 - [\mathcal{A}]_{\mathfrak{M}}$

   thus $\neg\mathcal{A}$ is true iff $\mathcal{A}$ is false

3. $[\mathcal{A} \vee \mathcal{B}]_{\mathfrak{M}} = \max([\mathcal{A}]_{\mathfrak{M}}, [\mathcal{B}]_{\mathfrak{M}})$

   thus $\mathcal{A} \vee \mathcal{B}$ is true iff $\mathcal{A}$ is true or $\mathcal{B}$ is true

4. $[\mathcal{A} \wedge \mathcal{B}]_{\mathfrak{M}} = \min([\mathcal{A}]_{\mathfrak{M}}, [\mathcal{B}]_{\mathfrak{M}})$

   thus $\mathcal{A} \wedge \mathcal{B}$ is true iff $\mathcal{A}$ is true and $\mathcal{B}$ is true

5. $[\mathcal{A} \rightarrow \mathcal{B}]_{\mathfrak{M}} = \max(1 - [\mathcal{A}]_{\mathfrak{M}}, [\mathcal{B}]_{\mathfrak{M}})$

   thus $\mathcal{A} \rightarrow \mathcal{B}$ is true iff $\mathcal{A}$ is false or $\mathcal{B}$ is true

6. $[\exists y\mathcal{A}]_{\mathfrak{M}} = \text{lub}\{[\mathcal{A}_y a]_{\mathfrak{M}^a_o} : o \in M\}$, where $\mathcal{A}_y a$ is the sentence obtained by writing a in place of all free occurrences of y in $\mathcal{A}$, a being some constant that does not occur in $\mathcal{A}$, and where $\mathfrak{M}^a_o$ is the interpretation which is just like $\mathfrak{M}$ except that in it the constant a is assigned the denotation o

   thus $\exists yPy$ is true iff something in the domain is in the extension of P

7. $[\forall y \mathcal{A}]_{\mathfrak{M}} = \text{glb}\{[\mathcal{A}_y a]_{\mathfrak{M}^a_o} : o \in M\}$

   thus $\forall y Py$ is true iff everything in the domain is in the extension of P.

The following piece of terminology is not part of the core view, but will be useful later: for any interpretation $\mathfrak{M}$, let $\mathbf{T}\mathfrak{M}$ be the set of wfs of our language which are assigned the value 1 on $\mathfrak{M}$ by the foregoing clauses.

§ 2.2. THE GLOSS. The semantics requires no comment, but two points need to be made about the syntax. First, the symbols of the language (i.e. ¬, ∃, the left bracket, the predicates and variables, etc.) are *individual objects*. Some philosophers prefer to think of these objects as types, whose tokens are the ink marks inscribed on the pages of logic books and elsewhere; others (myself included) prefer to think of them simply as particular objects, whose nature is unspecified (they *may* be identical to certain types of ink marks, or they may not be). It makes no difference which objects the symbols are: all that matters is that we have the right number of distinct objects. Second, the wfs are *finite sequences* of these objects, in the *mathematical* sense of 'sequence'—i.e. each formula is a *function* from some initial segment of the natural numbers to the set of symbols. A wf is *not* a bunch of symbols *lined up in a row*—if that was what a wf was, then we would need *two* negation objects (not to mention five left-bracket objects, etc.) to make the wf $((\neg P^2(a,b)) \wedge (\neg R^2(b,c)))$, whereas we have only *one* negation object, *one* left-bracket object, and so on. Wfs are not lines of ink marks. If they were, then 'Pa implies Pa' would not be a logical law: it would have the same status as 'Pa implies Rb', for (on the view under consideration) both would say that one atomic wf (one ink mark) implies another, *distinct* atomic wf (a distinct ink mark). Thus wfs cannot literally be touched or even seen (directly) on the pages of logic books. Rather, they are to be found with other abstract objects, such as numbers and sets and so on.

§ 2.3. NATURAL LANGUAGES. In order to see how the foregoing picture can be applied to natural languages such as English, we need to make a few additional comments, and introduce one new idea.

If wfs are mathematical sequences, which do not literally flow from our mouths or pens, then what is the relationship between ink marks, bursts of sound, chalk marks and so on on the one hand, and wfs on the other? I suggest the following picture. I turn on the light *by* flicking the switch. The event of my turning on the light and the event of my flicking the switch are one and the same event; but the switch is not the light, and turning on (what I do to the light) is not flicking (what I do to the switch). Similarly, I utter a wf *by* inscribing ink marks of a certain sort, or by producing sound waves of a certain sort. The event of my inscribing ink marks and the event of my uttering a particular wf are one and the same event; but the ink marks are not the wf, and inscribing (what I do to the ink marks) is not uttering (what I do to the wf). Likewise, the event of my producing sound waves and the event of my uttering a particular wf are one and the same event; but the sound waves are not the wf, and vocalising (what I do to the sound waves) is not uttering (what I do to the wf). (Those who think that wfs are types can agree with this, but also be more specific: they can say that the relationship between the ink mark which I *inscribe* and the wf which I *utter* is that of token to type.)

In ordinary language, we typically want to know whether a sentence is true simpliciter, not just that it is true on such and such interpretations and false on others. What is going on here is that when we utter a wf, we utter it relative to a particular interpretation: we *mean something*—some particular thing—by the wf we utter.[3] The particular interpretation that is relevant in a given case is often called the *intended interpretation*. This name is misleading however, and I prefer to follow Islam [9] and call it the *correct interpretation*.[4] I will suppose that every time you utter a wf, you invoke a particular interpretation of the language—the correct interpretation—and I will say that you utter the wf *relative to* that interpretation.[5] An utterance of a wf is true simpliciter if the wf uttered is true (i.e. is assigned the value 1 by the clauses in §2.1) on the *correct* interpretation (as invoked by that utterance of it), or in other words, if it is an utterance of a wf relative to an interpretation on which that wf is true (i.e. is assigned value 1). In a slogan: truth *simpliciter* is truth on the *correct* interpretation.[6]

Putting these ideas together, suppose you say 'Helen Clark is tall'. You thereby utter a wf Pa of the language. Now there is an interpretation on which a refers to Don Brash, and P has its extension the set of females, and on this interpretation the wf you uttered is false. However this interpretation is not the correct one: when you said 'Helen Clark is tall', you did not mean that Don Brash is a female. The correct interpretation of your utterance

[3] It is possible to make the right sounds (or shapes) without actually uttering wfs. Arguably there is a stage when they are learning language at which children do this.

[4] I prefer this term because it is quite possible for you to say something—i.e. mean something—which you did not intend to mean. In other words, a speaker's intentions are not the sole determinant of the meaning of what she says. An Australian might travel to the US and order "a pie with pepper", only to be surprised when she is presented with a capsicum pizza—but in the context, that is exactly what she asked for, even if she did not know what she meant, or mean to mean what she meant.

[5] Given the remark in note 4, this means that someone may utter a wf relative to a particular interpretation without intending to do so, or knowing that she has done so.

[6] Compare the supervaluationists' 'truth is supertruth', or in other words, 'truth simpliciter is truth on all admissible interpretations'.

assigns Helen Clark as the referent of a and the set of tall people as the extension of P—and on the correct interpretation, the wf you uttered is true (for Helen Clark is indeed a member of the set of tall people). Thus your utterance is true simpliciter.

Note that nothing in the foregoing requires that every time I inscribe ink marks in the shape 'John went to the bank' I utter the same wf of the language.[7] There might be quite complicated relationships governing which wf is uttered by which pattern of ink or sound in which context. Furthermore, nothing in the foregoing requires that every time I utter a particular wf of the language, the correct interpretation of my utterance is the same. Thus I might on one occasion say 'John' and utter a name a of the language, and on another occasion I might say 'John' and utter a different name b of the language—or I might utter the same name a but mean something different by it (e.g. John Howard, not John Woo).

In our investigations of philosophical issues to do with natural language, I believe we should treat the foregoing classical framework as we treat laws of physics when investigating natural phenomena: alterations to the framework should be countenanced only as an absolute last resort, if the phenomena *really* cannot be accounted for within the framework.[8] Now it may be that alterations to the classical framework *are* required. I myself think that accommodating the phenomena of vagueness really does require such alterations. But I shall argue in this paper that the Liar paradox demands no such alterations. As far as solving the Liar is concerned, there is no advantage to be gained by departing from the classical framework. And crucially, if departures from the classical framework are required to deal with other phenomena of natural language (e.g. vagueness, adverbial constructions, etc.), the solution to the Liar proposed here will still go through—mutatis mutandis—in the modified framework. In this sense, what I shall propose is not a specifically classical solution to the Liar, but a general template for producing a solution to the Liar within your favourite syntactic-semantic picture, illustrated with respect to the simplest and most widely-used picture available: the classical one.

§ 3. THE LIAR PARADOX AND ITS SOLUTION. Suppose I write:

A: Sentence A is not true.

The paradox here is as follows. If my sentence is true, then this means that what it says is the case *is* the case, and what it says is that it is not true, so

[7] On some versions of the view that wfs are types of ink marks—namely versions where the ink marks are typed by shape—this *is* required.

[8] Explaining my reasons for this view would take us too far afield from the topic of the present paper.

it is not true. If it is not true, then this means that what it says is the case is *not* the case, and what it says is that it is not true, so it is not the case that it is not true, that is, it is true. Yet of course the Liar must be either true or not true—for either it is true, or... it is *not*—so we have a contradiction.

Translating this intuitive reasoning into the terms of the classical framework gives us the following. When I write:

A: Sentence A is not true.

I thereby utter a certain wf $\neg Ta$ of the language, relative to a particular (correct) interpretation $\mathfrak{M}$. Now by 'sentence A' I mean the very sentence I uttered, so 'a' refers on $\mathfrak{M}$ to the wf $\neg Ta$ which I uttered; and by 'is true' I mean *is true*, so 'T' has as its extension on $\mathfrak{M}$ the set $\mathbf{T}\mathfrak{M}$. Now we are in the classical framework, so the wf I uttered must either have value 1 or value 0 on $\mathfrak{M}$. If it has value 1, then (by the clause for negation) the wf Ta has the value 0, so (by the clause for atomic wfs) the referent of a (which is the wf I uttered) is not in the extension of T, i.e. is not in $\mathbf{T}\mathfrak{M}$, so the wf I uttered has value 0 on $\mathfrak{M}$. But if it has value 0, then the wf Ta has the value 1, so the referent of a (which is the wf I uttered) is in the extension of T, i.e. is in $\mathbf{T}\mathfrak{M}$, so the wf I uttered has value 1. Contradiction.

This is not now a paradox: it is a proof. It is a proof that there *is no* interpretation $\mathfrak{M}$ on which the name a is assigned the wf $\neg Ta$ as its referent and the predicate T is assigned the set $\mathbf{T}\mathfrak{M}$ as its extension. This means that when I wrote what I wrote above and thereby uttered a wf $\neg Ta$, on the correct interpretation of my utterance either a did not refer to the wf I uttered, or T did not refer to the set $\mathbf{T}\mathfrak{M}$. In plain language, this means that when I write:

A: Sentence A is not true.

either 'sentence A' does not refer to the sentence I utter, or 'is true' does not pick out the set of sentences which are true on the correct interpretation of my utterance. Simpler still: either I do not refer to the sentence I utter, or I do not say of what I refer to that it is not *true*.

I believe this is the *solution* to the Liar paradox. There is no paradox, because there is no Liar sentence—that is, no sentence which says of itself only that it is not true. When you try to construct such a sentence, you fail: either you do not refer to what you wanted to refer to (the very wf you uttered), or you do not say of what you do refer to that it is not *true* (that is, the predicate you utter does not have $\mathbf{T}\mathfrak{M}$ as its extension, where $\mathfrak{M}$ is the correct interpretation of your utterance).

At this point I expect this all sounds absurd. My task in what follows will be to convince you that this is indeed the correct solution to the Liar paradox. I begin by responding to some objections, and eventually unearth the core belief which makes us resist the proposal just outlined. This is the belief, mentioned at the outset, which we need to give up in order to solve the Liar paradox—or rather, in order to feel that the solution just outlined really is the correct solution.

§ 4. FIRST OBJECTION. "Look, this is just nonsense! I *did* refer to my own sentence, and I *did* say of it that it was not true. After all, *what could have stopped me*? There were no guardians of classical logic present ensuring that I did not refer to forbidden things. Do you think that some spirit—Quine's ghost, perhaps—hovers around and ensures that no-one can make the name and predicate in the wf $\neg Ta$ refer in such a way as to generate paradox? That is complete nonsense. But in the absence of such mysterious constraints, there is nothing to stop us uttering Liars—and sometimes we do. Indeed, a little while ago, I uttered one myself!"

What stops us uttering Liars? Nothing! Of course there are no mysterious constraints of the sort just rightly dismissed as absurd. And yet we cannot utter Liars nevertheless. Why? Because there aren't any to utter. What we saw above was that there just *is no* interpretation $\mathfrak{M}$ of the language on which a refers to the wf $\neg Ta$ and T has $\mathbf{T}\mathfrak{M}$ as its extension. Thus, no constraints are needed to prevent us uttering wfs in such a way that such an interpretation is the correct interpretation of our utterances. Compare the barber who sets out to shave all and only those who do not shave themselves. What stops him succeeding in his quest? Nothing! No mysterious forces stay (or force) his razor hand; and yet he must fail. Or think of Juan Ponce de Leon searching Florida for the fountain of youth. What stopped him finding it? Nothing! The point is that there was no fountain of youth for him to find, and hence no constraints were required to stop him finding 'it'. Contrast the Liar with the case of the emperor's cat, which exists, but to which no-one is allowed to refer by name, on pain of death. The emperor has semantic guardians who ensure that no wf is ever uttered relative to an interpretation which assigns the emperor's cat as the referent of a name in the language. This is hard work, and the guardians are well rewarded. But this sort of case—where there exist interpretations which assign the emperor's cat as the referent of a name in the language, but these interpretations are forbidden—is quite different from the Liar case, where there just *are no* interpretations of the offending sort, and hence nothing special required to stop us uttering wfs relative to such interpretations.

§ 5. SECOND OBJECTION. "All right, let's suppose that on the correct interpretation $\mathfrak{M}$ of my utterance, either the name I uttered did not refer to the wf I uttered, or the predicate I uttered did not have the set of true-on-$\mathfrak{M}$

wfs of the language (as specified by the recursive definition in §2.1) as its extension. What then *did* my name refer to, or what *was* the extension of my predicate?"

I can best respond to this objection by drawing a parallel with the auto-infanticide paradox which arises in connection with backwards time travel. The paradox runs as follows. If backwards time travel were possible, then there would be nothing to stop a person travelling back in time and killing herself as a child. This would involve a contradiction: the time traveller both grows up to make a time trip, and does not grow up, because she dies as a child. So if backwards time travel were possible, there would be nothing to stop contradictions being true. Hence backwards time travel is impossible.

One reaction to this argument is, of course, to accept that backwards time travel is not possible. If, on the other hand, we wish to defend the possibility of time travel, then we need to show that it can occur *without* auto-infanticide occurring. To this end, some science fiction writers suppose that time travellers are accompanied by chaperones or Time Lords or chronology guardians who prevent the time travellers from changing the past: the chaperones act as bodyguards for the time travellers' younger selves, either preventing them being killed, or resurrecting them afterwards. Others posit mysterious contradiction-preventing forces which prevent time travellers from pulling triggers and getting pins out of grenades, or cause bullets to fly off course in mid air, and so on. But apart from being immensely unappealing in themselves, these responses are all over-reactions. As Lewis has shown, chaperones and mysterious forces, let alone outright bans on time travel, are unnecessary to avoid contradictions. No strange devices are required to stop the time traveller killing her younger self. Rather, she fails "for some commonplace reason" [13, p.150]: her gun might jam; a noise might distract her; she might slip on a banana peel; and so on. Nothing more than such ordinary occurrences is required to stop the time traveller killing her younger self. Hence backwards time travel does not imply the truth of contradictions, even in the absence of chaperones and special forces. Hence backwards time travel is *not* impossible.

This 'coincidences solution' of the paradox is structurally analogous to the resolution of the Liar outlined above. Auto-infanticide generates a contradiction; but auto-infanticide does not occur. Either the would-be committer of auto-infanticide fails to kill the person she is facing (she slips on a banana peel, etc.), or the person she kills is not in fact her younger self (some error has caused her to face another person who just looks like her younger self, etc.). A Liar sentence generates a contradiction; but Liar sentences do not exist. Either the would-be Liar sentence does not say of what it refers to that that thing is not true; or it does not refer to itself.

In each case the solution is negative: we are told that something goes wrong, and contradiction is thus avoided. Suppose we ask: "OK, but what exactly will go wrong? What *will* happen when the time traveller tries to kill her younger self?" I think it is quite clear that the coincidences solution of the time travel paradox is not deficient because it does not answer this question. Any one of innumerably many things could go wrong—your gun could jam, you could slip on a banana peel, a bird could intercept your bullet—and we just have no idea in advance exactly what will happen. The same goes in the case of the Liar. I say that when you try to utter a Liar sentence, something will go wrong, and either you will not refer to the wf you utter, or you will not say of what you refer to that it is not true. But I have no idea exactly what you *will* end up saying (i.e. meaning by what you say). Perhaps when you say 'This sentence is not true' you will refer to a wf other than the one you utter; perhaps you will refer to the Queen of England; perhaps you will refer to the wf you utter but say of it that it is not an *elephant*; and so on. The fact that my solution to the Liar does not include a specification of which of these possibilities will obtain *does not make it deficient*—just as the coincidences solution to the auto-infanticide paradox is not deficient because it tells us only that something will go wrong when the time traveller tries to kill her younger self, without telling us exactly what will go wrong.

§ 6. THIRD OBJECTION. "Point taken. But there is still a big difference between the auto-infanticide case and the Liar case. In the auto-infanticide case, we cannot say *in advance* what will happen when the time traveller tries to kill her younger self. But if we wait and watch, we can say, afterwards, what did happen. In the Liar case, we cannot even do that. Even after the fact of my utterance, all you can tell me is that either I did not refer to the wf I uttered, or I did not say of what I referred to that it was not true. This is mysterious in a way that the auto-infanticide case is not. Furthermore, this is what I had in mind when I made my previous objection! What I asked you last time was 'What then *did* my name refer to, or what *was* the extension of my predicate?' Note the tense here. You have not answered these questions; nor have you shown why you should not have to answer them."

As a thought experiment, imagine that there are *reference rays*, and suppose that you have a reference ray detector. When someone makes an utterance, your detector allows you to *see* the correct interpretation of her utterance. You see which wf she utters; you can see a ray coming off each name in this wf and hitting some object in the domain; and you can see a ray coming off each (n-place) predicate in this wf and striking some set of (n-tuples of) objects from the domain. Now suppose the situation is as follows. I write:

A. Australia is an island.

B. Sentence A is not true.

Through your detector, you see me utter two wfs, Pb and $\neg Ta$, relative to an interpretation $\mathfrak{M}$. You see a ray from b striking Australia, a ray from P striking the set of islands, a ray from a striking the wf Pb, and a ray from T striking the set of wfs $\mathbf{T}\mathfrak{M}$. Now I relabel my sentences as follows:

A. Sentence A is not true.

B. Australia is an island.

Through your detector, you see me utter two wfs, $\neg Ta$ and Pb, relative to an interpretation $\mathfrak{M}$.[9] You see a ray from b striking Australia, a ray from P striking the set of islands, and then you either see a ray from a striking the wf $\neg Ta$ and a ray from T striking some set other than $\mathbf{T}\mathfrak{M}$, or you see a ray from a striking something other than the wf $\neg Ta$ and a ray from T striking the set $\mathbf{T}\mathfrak{M}$. "Yes, but I wanted to know specifically what was struck—not just that it was something *other than* a given thing. So tell me, exactly which things get hit?" I don't know! You have the reference ray detector, not me—so you tell me! But seriously, the point here—and it is indeed a serious point—is that just because we cannot *find out* what I referred to when I uttered my would-be Liar sentence, this does not in any way undermine the claim that there are particular facts of the matter concerning what I picked out using my name and predicate. On the classical picture, the facts are there just as much as they would be if there were visible reference rays. If you had a reference ray detector, you would be satisfied, and would regard the auto-infanticide case and the Liar case as analogous; but just because there are not detectable reference rays, this does not make the cases disanalogous in any important way. The Liar case is like the case of the time traveller who attempts to commit auto-infanticide *at the bottom of a very deep hole, where none of us can see what happens, and where neither the older nor the younger version of the time traveller will talk about it afterwards.* In both cases, there are particular facts about what happens when we try to do the impossible (commit auto-infanticide, or utter a Liar wf); whether or not these facts are observable is just irrelevant.

§ 7. FOURTH OBJECTION. "It's not irrelevant at all! *We make meanings*! If no-one says what the name and predicate in the sentence 'Sentence A is not true' are to mean, then they do not mean anything. Likewise, if someone does say that they do mean some particular thing—for example, that

[9] As noted earlier, these may or may not be the same two wfs I uttered the first time.

the name uttered refers to the wf uttered, and 'is true' picks out the set of sentences which are true on the correct interpretation of that wf—then they do mean these things. Mean*ings* just do not run around independently of mean*ers*! There cannot just *be* these meaning facts out there, independently of what we want and of what we can even detect."

Now we get to the heart of the matter. The opinion just expressed—or rather, a more precise and less bold version of it which we shall see in a minute—is the core belief which we must relinquish in order to solve the Liar paradox—or more correctly, in order to recognise the solution proposed above as the correct solution. First we need to work out precisely what the core thesis is. Here's one view the objector might hold:

Semantic Omnipotence A name or predicate of a language L has a particular referent or extension if and only if the speakers of L decide that the name or predicate should have this referent or extension.

Four sorts of example will lead us to tone down, and clarify, this view.

Externalism. Consider natural kind terms such as 'water'. The ancients, who knew nothing of chemistry, referred to H_2O when they used the term 'water' (or their word for water, whatever it was). But they did not *decide* to use 'water' to refer to H_2O: after all, they had never heard of H_2O!

Reference borrowing. I never *decided* that the predicate 'is green' should mean anything in particular. Rather, I learned the proper use of this predicate, and presumably thereby came to mean by this predicate what those who taught me the language use it to mean.

Anaphora. I might decide that when I utter the following sentence, 'Bill' will refer to Bill, and 'he' will refer to Ben:

> Bill was really excited about going to the movie, but in the end
> he stayed at home.

I then utter the sentence (without any demonstration to accompany 'he'). Despite my decision, I referred to Bill when I said 'he'. We cannot just decide to refer to things willy-nilly: there are over-arching semantic laws—such as those governing anaphoric links—by which we live.

Indexicality. Suppose a poster saying 'Your country needs you!' is posted on a billboard, where, as it happens, only one person—Bob—ever reads it. On the occasion of Bob's reading the poster, the poster says that Bob's country needs *him*, i.e. Bob. But the poster maker did not decide to refer to Bob—she has never even heard of him!

In light of these four examples, let us refine our thesis:

Semantic Regularity There are reliable, principled relationships between our behaviour, mental states and physical environment on the one hand, and what we mean by our utterances on the other hand.

For example, if one points to various samples of stuff while intoning some new word, that word will come to refer to the natural kind underlying the samples, if there is such a natural kind. That's how 'water' got to refer to H_2O, and that's why 'water' on Twin Earth refers to XYZ. For those learning an existing language, the situation is even simpler. We just make the sound 'is green', and we utter a predicate which has the same extension as that uttered by our language-teachers when they made the sound 'is green'. Again, when you make an inscription of the form '*a* was blah, but in the end he did blah', then if '*a*' refers to a male person, 'he' does too. If you write 'you' on a poster, then on the occasion of someone reading the poster, 'you' refers to that person. And so on and on. This is a common-sense view. We cannot mean whatever we like whenever we like; however, there are regular and principled patterns relating the sounds we make in given circumstances to what those sounds mean in those circumstances.

My view is that the Liar forces us to reject this picture. Sometimes we go through all the right motions, but our words just don't come out meaning what we wanted them to mean. Mostly, when you say 'This sentence is ...', you refer to the wf you thereby utter; and mostly, when you say '... is true...' you pick out the set $\mathbf{T}\mathfrak{M}$, where $\mathfrak{M}$ is the correct interpretation of your utterance. But these relationships cannot be perfectly reliable, because when you say 'This sentence is not true', either you do not refer to the wf you thereby utter, or you do not say of it that it is not true.

Consider again the auto-infanticide paradox. There is no paradox. The attempt at auto-infanticide simply fails. But there is a strange consequence. If backwards time travel is possible, then there cannot be perfectly reliable killing machines and perfectly reliable methods for identifying persons. For suppose there were: suppose you have a super-gun which when pointed in a certain direction, invariably kills anything in front of it for a range of ten metres; and suppose you have a DNA test which allows you to identify people with perfect accuracy. Then you could find your younger self, and kill him or her. But you can't do that. However reliable your gun and identity test are, they cannot be *perfectly* reliable if there is backwards time travel: they cannot work in *every* circumstance, if one of the available circumstances involves you looking around in your own past for your own younger self, with the intention of killing him or her.

The same point can be made more forcefully by considering a simpler situation. Imagine a device—an *Earman rocket*—consisting of a firing mechanism and a sensor [5]. If no incoming probe is detected when the device is turned on, it waits a set number of seconds and then fires an outgoing probe in a specified direction at a specified velocity and then shuts down for 24 hours; if an incoming probe is detected when the device is turned on, it shuts down immediately. If there are closed timelike curves then *such a device cannot be perfectly reliable.* For if it is set to wait ten minutes, and aimed in such a way as to fire its probe along a closed timelike curve so that the probe strikes the sensor ten minutes before firing time, then *some part of the machine must malfunction.*[10]

The lesson of the Liar, on my view, is that our referential mechanisms—i.e. our devices of making certain noises or inscribing certain patterns in particular sorts of situation in order to utter wfs with particular meanings—are just the same. They cannot be perfectly reliable. There are circumstances in which they must malfunction. You go though moves—making certain sounds, inscribing certain shapes—which in any ordinary context would see you referring to the wf you utter, and saying of the thing you refer to that it is not true, but something goes wrong. The semantic mechanism fails, and you do not end up meaning what you wanted to mean.

At this point, I have unearthed the core belief—Semantic Regularity—underlying resistance to the classical solution to the Liar proposed above. If we accept Semantic Regularity, we will be opposed to my solution to the Liar; and as far as I can see, if we reject Semantic Regularity, then there is no *other* reason why we should resist my solution. Having thus fingered the culprit, my task now is to convince you that rejecting Semantic Regularity, and accepting my solution to the Liar, is the right way to go. I shall try to do this as follows. First (§8), I argue that there is no acceptable alternative. Second (§9), I argue that rejecting Semantic Regularity dissolves a number of other problems, apart from the Liar: most notably, semantic indeterminacy arguments such as those of Quine, Kripkenstein, Davidson and Putnam; the problem of empty names; and a recalcitrant problem about the semantics of vague predicates—the problem of false precision. Note that the dialectic is not 'Here's one advantage of rejecting Semantic Regularity (solving the Liar), and here are some more (solving the problems discussed in §9)', but 'When it comes to solving the Liar, there is no acceptable alternative to rejecting

[10]Assume the rocket has only one probe, and that it is deployed in an environment which is devoid of all probes other than those fired by the rocket itself, and is free of probe-deflecting obstacles and forces. These assumptions do not decrease the force of the example: just suppose the path from the firing part of the machine to the sensing part of the machine to be enclosed within the bounds of the machine itself, and suppose that part of what the machine is supposed to do is keep this path free of probe-deflecting obstacles and forces.

Semantic Regularity; but once you reject it, you get these other benefits for free'. Rejecting Semantic Regularity is *one* way of dealing with the problems discussed in §9; it is—or so I shall argue in §8—the *only* satisfactory way of dealing with the Liar once and for all.

§ 8. OTHER RESPONSES TO THE LIAR. Approaches to the Liar which start off taking a quite different tack from mine either (a) end up incomplete, or (b) and up resorting to hand-waving, or (c) plainly get the phenomena of English—which is what they are supposed to be modelling—wrong, or (d) end up asking us to reject a belief which is even more entrenched than Semantic Regularity, or (e) end up asking us to buy into something which amounts to a rejection of Semantic Regularity. I cannot survey *all* existing approaches to the Liar, and show that each falls into one of categories (a)–(e). Nor can I present a general argument to the effect that any *possible* approach to the Liar falls into one of these categories: the argument would have to be so general—in order to encompass all possible approaches to the Liar—that it is very difficult to see what it could take as a starting-point. What I will do is examine a number of the most important approaches to the Liar, and show that each falls into one of categories (a)–(e). This will lend inductive support to my claim; but it will also do more than this. Seeing *how* the views examined fail to offer a viable alternative to rejecting Semantic Regularity will make it hard to see how any other view could do better.

Recall that my project is to say what goes on in a natural language such as English when we say things such as 'This sentence is not true'. In this section I therefore assess other approaches to the Liar as contributions to this project, *not* as attempts to construct consistent formal systems which have such-and-such expressive capacities. Considered from the latter point of view, many of the approaches to the Liar which I shall discuss are entirely successful. Note also that in order not to make this paper excessively long, I shall assume some familiarity with the approaches that I shall discuss.

§ 8.1. KRIPKE. Consider Kripke's treatment of the Liar [11], which has replaced Tarski's as the orthodox treatment, at least amongst philosophers. Kripke's semantic framework is non-classical, in that instead of being assigned extensions, predicates are assigned a pair of an extension and an anti-extension, and these two sets need not exhaust the domain. Suppose we start with a classical interpretation of our language, that assigns an extension to every predicate *except* 'is true'. We now interpret sentences containing 'is true' by a recursive procedure. Sentences not containing 'is true' go into the extension of 'is true' if they are classically true. Non-sentences go into the anti-extension of 'is true'. As for sentences containing 'is true', 'S is true'—where S does not contain 'is true'—goes into the extension of 'is true' if S is true on the original interpretation and into the anti-extension of 'is true' if S is false on the original interpretation. Now we can treat ' 'S is true' is true', and so on. By following the recursive procedure we eventually get to a fixed point where no more sentences containing 'is true' are decided either way (i.e. are put into either the extension or the anti-extension of 'is true'). At this point, sentences such as the Liar have not been decided either way, i.e. they are neither true nor false. Such sentences are called *ungrounded*.

So far so good: we have been shown how to interpret the predicate 'is true' in a consistent and *regular* way. When we say 'is true', there is a reliable relationship between the noises we make and how those noises end up being interpreted. But in presenting his account, Kripke talks about some sentences being 'ungrounded'. Now there cannot be a predicate in Kripke's language which means 'is ungrounded'. For then we could construct a sentence 'This sentence is either false or ungrounded', leading to a new paradox (if it is true, then what it says is the case is the case, i.e. it is false or ungrounded; if it is false or ungrounded, then what it says is the case is not the case, and so it is neither false nor ungrounded, i.e. it is true). So Kripke just says: well, we cannot express this notion of ungroundedness in the language. There are three ways of taking this: Category (a): In English, we talk of ungroundedness; so Kripke's language is not a model of English; i.e. his solution is incomplete as an account of what is going on in English when we utter Liar-type sentences. Category (e): Kripke's language *is* a model of English, in which case we cannot express 'is ungrounded' in English. But hang on, we did express it in the course of presenting Kripke's theory. This can only mean that we cannot *reliably* express this notion: we cannot, whenever we want, utter a predicate which means 'is ungrounded'. Sometimes we will make the noises that would normally see us referring to the set of ungrounded sentences, but something goes wrong, and we fail to pick out this set (on pain of contradiction). But this is just to say that Semantic Regularity fails. Category (b): Someone might suggest that to deal with 'is ungrounded' we just apply Kripke's story again. This, however, is hand-waving. Until we see exactly how this is supposed to be done, this does *not* count as a complete account of what is going on in English when we utter Liar-type sentences.[11]

§ 8.2. BARWISE AND ETCHEMENDY. Consider the Austinian treatment of the Liar in Barwise and Etchemendy [1]. On this account, sentences (in general) express propositions. Propositions are determined by two constituents: a situation (i.e. a set of states of affairs) which the proposition

[11] Similar remarks to those about Kripke's view and the notion of ungroundedness apply to Gupta and Belnap's revision theory [7] and the notion of *stable truth*.

is *about*; and a type of situation. The proposition is *true* if the situation the proposition is about is of the type which figures in the proposition. For example, if you are playing cards with Claire and you say 'Claire has the three of clubs', you express a proposition which is about a situation which includes the facts of your card game, and which says that this situation is of the type in which Claire has the three of clubs; what you say is true if the situation really is of that type—in this case, if Claire really does have the three of clubs.

For any situation s, there is a proposition a_s—the *assertive Liar* for s—which says that s is of the type in which a_s is false, and another proposition d_s—the *denial Liar* for s—which *denies* that s is of the type in which d_s is true. On Barwise and Etchemendy's treatment, the assertive Liar is always false, and the denial Liar is always true. However there is a twist. The *fact* that the assertive Liar for s is false cannot, on pain of contradiction, be included in the situation s itself, and nor can the fact that the denial Liar for s is true be included in s. A consequence of this is that there is no Liar proposition (assertive or denial) about the whole world—i.e. about the set of *all* facts: for if there were, the fact that it was true (or false) *would* be in the situation the proposition was about (for this situation is, ex hypothesi, the set of *all* facts), and we would have a contradiction. Barwise and Etchemendy's response to this is to claim that no proposition at all can be about the whole world.

It might seem as though Barwise and Etchemendy face an obvious revenge problem here, in that they are talking about the world, while saying that no proposition can be about the world. But the situation is not so simple. They say that every proposition is about some situation—a situation being a set of states of affairs—and no proposition can be about the world as a whole—the world being the set of all actual states of affairs. But while a proposition cannot be *about* the world considered as a situation, a proposition can still *refer to* the world as an *object*. A statement about (in the *ordinary* sense) the world—for example, 'the world is the set of all actual situations'—could be construed as a proposition about (in Barwise and Etchemendy's sense) a situation which is a proper subset of the world, and which says of that situation that it is of the type in which the world (here referred to as an object) has the property of containing all actual situations.

I do not find this imagined reconstrual strategy entirely convincing, but for the sake of argument, let us grant that Barwise and Etchemendy's statements about (in the ordinary sense) the world can be handled within their framework. Nevertheless, when considering an account of the Liar as a model of English, we need to take into account not only what is said in presenting that account of the Liar, but also what the account makes sayable. The account cannot provide a correct model of English if (assuming the account is written in English) we cannot translate some statement in the presentation of the account into the terms of the account, in such a way that it can then be smoothly handled by the account. Furthermore, this applies not just to statements that can actually be found in the pages of the account, but to any statement whose meaning we can clearly grasp, once we have read and understood the account. If, after reading the account, we think through the implications of it, or reflect on its machinery, and express the results of our thinking in an English sentence S, then the fact that S is not literally to be found in the pages of the account does not save the account as a correct model of English, if it cannot handle S. Thus an account of the Liar must be able to handle both what is said in presenting the account, and what is made sayable by the presentation of the account.[12]

It is here that Barwise and Etchemendy run into trouble. It seems that just as I can utter a proposition about the situation that includes the facts in this room, or in this city, or in this country, or on earth, so too I can, if I want, utter a proposition (as long as it is not a Liar proposition) about the situation which includes the facts about *absolutely everything*. I can say, for example, abut this situation that it is of the type in which $2+2=4$, or that it is of the type which includes all facts. In saying that there are no propositions about the whole world, Barwise and Etchemendy would seem to be in the position of an IT consultant who configures your computer in various ways, and then says 'But you cannot change these configurations yourself'. Assuming you have an administrator account on the computer, this is literally false: you can make such changes. What the consultant really means is that any changes

[12]This test is arbitrary, in one important respect—but this does not affect the claim that passing the test is a *necessary* condition on any adequate account of the Liar. You might think: why do we require that account A be able to handle account A, and that account B be able to handle account B? If they are meant to be models of English, and both accounts are presented in English, then shouldn't A be able to handle B, and vice versa? Yes, they should. But if A cannot handle A, that in itself is enough of a problem for A. And typically, accounts of the Liar create problems for *themselves*, not for each other—a phenomenon which has come to be known as the *revenge problem*. This problem arises when, in presenting her theory, the theorist *uses* language which cannot be accounted for within the theory presented; or, more subtly, when, given an understanding of the theory, it is intuitively obvious that certain things are sayable (even if the theorist does not actually say these things in presenting her theory), but such things cannot be said in the language for which the theory provides an account. Thus the theory makes sayable things which cannot, according to the theory itself, be said. The *strengthened Liar paradox* is a specific instance—or family of specific instances—of the revenge problem which arises for accounts which posit extra truth values or truth value gaps. (The terms 'revenge problem' and 'strengthened Liar paradox' are used in various ways; this at least is how I use them.)

you make will be *unsupported*. Likewise, propositions abut the whole world are unsupported in Barwise and Etchemendy's framework. But there is no reason at all to think either that such propositions do not exist, or that we cannot utter such propositions. Indeed, there is plenty of reason to think otherwise: for it seems intuitively obvious (once we understand Barwise and Etchemendy's account) that we can say things about (in their sense) the whole world, and there is no principled reason why we should not be able to do so to counter this intuition (in contrast to the principled reasons why we cannot utter propositions about non-actual situations, or why we cannot utter *Liar* propositions about the whole world).

As the situation stands, Barwise and Etchemendy thus fall into category (a): their account is incomplete, because there are propositions of English—propositions about the whole world—which their account does not accommodate. Of course, this situation is easily remedied. For their only reason for banning propositions about the whole world is to avoid *Liar* propositions about the whole world, which generate contradictions. So they could easily admit that there are *some* propositions about the whole world, just not Liar propositions (on pain of contradiction). Now, however, they fall into category (e): rejecting Semantic Regularity. For now it cannot be that there are perfectly reliable mechanisms determining which situation the proposition one is uttering is about: for when you try to utter a Liar proposition about the whole world, you must fail, and the moves which in any other circumstance would see you uttering a proposition about the world as a whole will instead see you uttering a proposition about some situation which is only a proper subset of the world. No principled reason is given for this failure, other than that it is required to avoid contradiction: the failure is explained from the forbidden end result back down, not from basic principles on up. I suspect it was precisely to avoid such irregularity that Barwise and Etchemendy tried to ban *all* propositions about the world, rather than just Liar propositions. The problem with this ban is that it is ineffective: while Barwise and Etchemendy can refuse to support such propositions, they cannot make them not exist—and the ban then serves only to render their account incomplete.

§ 8.3. TARSKI. Tarski had two views of natural language: how it *is* (inconsistent), and how it should be (a hierarchy of metalanguages). There is considerable debate over what the former view amounts to exactly; in what follows I consider the latter view. Tarski's approach renders the Liar sentence non-well-formed. This proposal does not generate a revenge problem. However, as Kripke [11] and others have observed, it also renders non-well-formed various other sentences which seem perfectly meaningful. For example, consider the well-worn case where Dean says ‘All Nixon’s statements concerning Watergate are false’, and Nixon says ‘All Dean’s statements concerning Watergate are false’. If Nixon has made some straightforwardly true statements about Watergate, and Dean’s only statement concerning Watergate is the one quoted above, then it seems clear that Dean’s statements is false and Nixon’s is true. However Tarski cannot allow this. This puts his solution squarely in category (c): it simply gets the phenomena of English wrong.

§ 8.4. BURGE. According to the contextualist treatment of the Liar in Burge [4], ‘true’ is an indexical predicate whose extension varies systematically with context. Burge describes a hierarchy of interpretations, on each of which ‘true’ has a particular extension, represented using predicate constants ‘true_i’. Consider a typical Liar situation:

A: A is not true.

On Burge’s view, ‘true’ as uttered on this occasion has a particular extension, which we may represent using the predicate constant ‘true_i’. By the familiar Liar reasoning, A cannot be said to be true_i or not true_i. But now if we reflect on this fact, it seems natural to conclude that A is therefore not true. We might express our conclusion thus:

A is not true.

We have now uttered sentence A again. This time, however, ‘true’ has a different extension, which we may represent using ‘true_k’. We have not landed ourselves back in paradox, because A *is* true_k.

Paradox will however return if we say things such as ‘This sentence is not true at any level’, or ‘This sentence is not true_i for any i’.[13] Burge has a response to this problem. First, he says that the statement ‘This sentence is not true at any level’ “is not an English reading of any sentence in our formalization” [4, p.108]. But so what? Burge needs to show not only that *he* did not say ‘This sentence is not true at any level’, but that anyone who tries to say such a thing has somehow misunderstood his account: i.e. that his account does not make this *sayable*. Burge does try to show this. He says that we cannot “quantify out the indexical character of ‘true’ ” [4, p.108]; thus we just cannot say things such as ‘This sentence is not true_i for any i’.

[13]If ‘This sentence is not true_i for any i’ is true at some level, then what it says must be the case: but what it says is that it is *not* true at any level—including the one it was supposedly true at. So the sentence is not true at any level; but that is just what it says, which would seem to make it true after all. Similar remarks apply to ‘This sentence is pathological_i for all i’.

I do not find this claim convincing, and would say about it something similar to what I said about Barwise and Etchemendy's claim that we cannot talk about the whole world (Burge can refuse to support certain claims which his account makes sayable, but he cannot make them unsayable by stipulation: the only way to make them unsayable is to retract his whole account). But for Burge the problem is even worse. There is an *essential* part of his account in which he himself quantifies over the hierarchy of interpretations of 'true'. So far we have seen that Burge says that the extension of 'true' varies according to its context of utterance. Now we might wonder, how exactly is the extension of 'true' determined by context? If there is no principle to the variation, then we have a denial of Semantic Regularity. But Burge thinks there is a principle, which he calls Verity.[14] It says that the correct interpretation in a context is given by the *lowest* subscript possible, compatible with assigning truth conditions to the maximum number of sentences.[15] But to refer to an interpretation i as the lowest in the hierarchy which satisfies certain conditions is to say that *there is no* interpretation in the hierarchy which satisfies these conditions and is lower than i. So we *can* quantify over the hierarchy of interpretations: we do so when stating Verity. Why then can we not say 'There is no level i such that this sentence is true_i'? The only possible reason is: because that would generate a contradiction. So you can quantify over the hierarchy of interpretations sometimes (for example when stating Verity), but not at others (for example when trying to say 'There is no level i such that this sentence is true_i'). Burge's account yields no principle explaining why this should be so: the failure can only be explained from the forbidden end result back down, not from basic principles on up. Thus we are left with a denial of Semantic Regularity.

Note that simply abandoning Verity will not help. As noted, unless we have systematic principles saying which interpretation of 'true' is correct in which context, we have an implicit denial of Semantic Regularity in any case. Yet when it comes to stating such principles—which are supposed to tell us which *out of all the interpretations of 'true'* is correct in a particular context—it is very hard to see how one could do so without quantifying over all the interpretations of 'true'. If, finally, we say there *is* such a principle, but we cannot state it, then we are in category (b): hand-waving.[16]

[14]He thinks there is also a second principle, which he calls Justice [p.110].

[15]Assigning a sentence truth conditions is contrasted with deeming it pathological.

[16]Parsons [16] is in a similar spirit to Burge [4], but Parsons is less explicit than Burge on a number of key matters. Parsons repeatedly talks—rather elusively—of "systematic ambiguities". I am not sure whether to interpret this talk as an implicit denial of Semantic Regularity (category (e)), or as hand-waving (category (b)); but I *am* sure that these are the only plausible interpretations.

§ 8.5. PRIEST. We have seen various accounts which face revenge problems, and one (Tarski's) which does not—but which faces the equally devastating problem of being inadequate to the phenomena of English. Now consider dialetheism, the view that there are true contradictions. In particular, consider the treatment of the Liar paradox which says that the Liar is both true and not true—just as the Liar reasoning shows [17]. Bromand [3] has recently argued convincingly that this account of the Liar faces a strengthened Liar paradox. If so, then for reasons similar to those discussed above, the dialetheist solution is either incomplete, or implicitly rejects Semantic Regularity (or resorts to hand-waving to deny that it does either of these). It is not impossible, however, that dialetheism can solve its revenge problem. The fact that it can accommodate true contradictions sets this solution apart from those examined above, in a way that makes it seem possible that the dialetheist solution might genuinely be revenge free, where other solutions cannot hope to be. But suppose it is revenge free. It still falls into category (d): it asks us to reject core logical and semantical beliefs (such as the belief that contradictions cannot be true) which are even more entrenched than Semantic Regularity.

Someone might argue that Semantic Regularity is the most entrenched belief of all: if we cannot know that the meanings of our symbols are not changing unknowably as we speak, what can we know? But the denial of Semantic Regularity does *not* involve such wholesale semantic flux, as I shall explain in the next section. I deny that our semantic mechanisms can be *perfectly* reliable: there will always be situations in which they must fail—for example, situations in which one attempts to utter a Liar sentence. But this is a far cry from saying that our semantic mechanisms are generally unreliable. There is no reason to think they will fail in any contexts other than those in which they *have* to fail.

§ 8.6. THE PRESENT SOLUTION. Note that the view of the Liar which I have presented does *not* generate a 'reflection problem'. I say that the Liar is either true (if it says something such as 'this sentence is not *in German*') or false (if it says something such as '*Proposition *110.643 in Principia Mathematica* is not true'). Now we cannot go on from here and say: "ah then, but if it is true then it is false... and if it is false then it is true..."—because it is true (or false) precisely because it does *not* say *of itself* that it is not *true*—and it is only on the assumption that it *does* say this that the paradoxical reasoning can get under way. My whole point is that the Liar does not say what we wanted it to say, so the above reasoning is simply stopped dead in its tracks.

Nor does my account face any sort of revenge problem. I have not, in the course of my presentation, introduced any notions which cannot be expressed in the language modelled in my theory, or handled in the way that I claim we should handle the Liar paradox. It may seem as though I have. For example, I have talked of the set $\mathbf{T}\mathfrak{M}$, and yet the expression '$\mathbf{T}\mathfrak{M}$' cannot always pick out what I want it to pick out, on pain of contradiction: it is in no better position than the ordinary expression 'is true'.[17] But of course: this is *my* point, not an objection to it! The crucial thing is that I do not say anything which, on my view, cannot be said. On my view, almost anything can be said on its own. You can utter a predicate T relative to an interpretation $\mathfrak{M}$ on which T has $\mathbf{T}\mathfrak{M}$ as its extension. You can enumerate the expressions of the language and refer to each one by name. You can utter a predicate T relative to an interpretation $\mathfrak{M}$ on which T has as its extension the set of Gödel numbers of expressions of the language which are assigned the truth value 1 on $\mathfrak{M}$. And so on. It's just that you cannot always do everything at once. English does have the power to talk about the truth of its own sentences, and to refer to those sentences. It just does not have unrestricted power to do both at once. We have referential mechanisms which suffice to do each, but both mechanisms cannot work perfectly together. Sometimes, not all the things we say can mean what we want them to mean, at the same time—i.e. on the same interpretation. But there is nothing which I say in this paper—or which this paper makes sayable—which cannot, according to

[17] How exactly *does* the phrase 'is true' work, on my view? Well, my whole point is that we cannot give a *complete* account of this—that is to say, there just is no perfectly regular relationship between the use of this phrase and the semantic outcome of such use. We can only say: what saying 'is true' is *supposed* to do is utter a predicate T relative to an interpretation $\mathfrak{M}$ on which T has $\mathbf{T}\mathfrak{M}$ as its extension. Mostly this will indeed happen; but it cannot *always* happen (unless semantic malfunctions never occur with 'is true', only with other referential devices—which is possible of course), and when it doesn't happen, we just don't know what *does* happen instead. Compare the task of describing what a given gun does. We say that when you load it and pull the trigger it fires a bullet with velocity x in the direction in which the bullet is pointed. Mostly it does; but it cannot *always* do so. For example, if you try to use it to commit auto-infanticide, it may fail in unexpected ways. What about the T-schema? Well, it is not always satisfied by 'is true'. Here's the T-schema (where $\leftrightarrow$ is the material biconditional): $S \leftrightarrow Ps$. Let's say that a predicate P satisfies the T-schema on an interpretation $\mathfrak{M}$ if and only if every instance of the schema obtained by replacing 'S' with a wf, and 's' with a name which refers to S on $\mathfrak{M}$, is true on $\mathfrak{M}$. It's not hard to see that a predicate P satisfies the schema on $\mathfrak{M}$ if and only if its extension on $\mathfrak{M}$ is $\mathbf{T}\mathfrak{M}$ (assuming that there *is* a name for each wf on $\mathfrak{M}$). I have argued that the form of words 'is true' cannot invariably be used to utter a predicate T relative to an interpretation $\mathfrak{M}$ on which T has $\mathbf{T}\mathfrak{M}$ as its extension; thus, the form of words 'is true' cannot invariably be used to utter a predicate T relative to an interpretation on which T satisfies the T-schema. This should be no news and should require no separate discussion: for the T-schema just gives us a syntactic handle on the model-theoretic characterisation of truth (in terms of $\mathbf{T}\mathfrak{M}$) that I have been employing.

my view, be meaningfully said in the language modelled. In order to present my view, I only need to refer to various things—such as the symbols and wfs of our language, and the set $\mathbf{T}\mathfrak{M}$ for some interpretation $\mathfrak{M}$—individually. Thus I do not need to do anything which cannot, according to my own view, be done.

Nor does my view face Tarski's problem of being overly restrictive. There is no reason to think things will *generally* go wrong regarding the meanings of what we say. I assume that they will only go wrong when they *have* to. The auto-infanticide paradox does not show that Earman rockets will sometimes fail in contexts which do *not* involve closed timelike curves—only that they cannot always work perfectly in contexts which *do* involve closed timelike curves. Likewise, my view is that our referential mechanisms cannot work perfectly in *every* possible context—but there is no reason at all to think they will not work as they should in ordinary, non-paradoxical contexts. Things need only go wrong when there just *is no* interpretation of the language meeting all the constraints we seek to impose upon the correct interpretation; when there is such an interpretation, there is no reason at all to suppose that we do not speak relative to it. This is why I said at the end of the previous section that denying Semantic Regularity does not commit one to wholesale semantic flux. Generally, there is no reason at all to suppose that our words mean anything other than what we want them to mean. It is only when there just is no interpretation compatible with all our semantic desiderata that things have to go wrong.

Earlier, we saw that if we accept the classical framework, the view of the Liar which falls out is that either the Liar sentence does not refer to itself, or it does not say of what it refers to that that thing is not true. I advocated accepting this minimal claim as the solution to the Liar: there is nothing more to be said. The upshot is that we must reject Semantic Regularity. Sometimes our words do not mean what we want them to mean: not due to hidden complexities of our semantic mechanisms, operating behind the scenes to produce unforeseen results—i.e. not for some principled, bottom-up reason; but because our words *cannot* mean what we want them to mean, and so our semantic mechanisms simply break or malfunction, and some of our words get assigned meanings more or less randomly (but as noted, there is no reason to think that the malfunction extends any further than is necessary to avoid contradiction).

Many readers will have felt that this was (to put it mildly) the wrong approach. The correct account cannot be that our semantic mechanisms are classical, and so malfunction in face of the Liar. The true story must be that our semantic mechanisms are more complex than the classical picture allows. There are extra bells and whistles, which mean that what might look like a malfunction from the classical point of view is really just what was supposed to happen. Instead of simple mechanisms which sometimes fail, we have complex mechanisms which are perfectly reliable. To make this more concrete: if, for example, on some occasion when we utter 'is true', we do not speak relative to a classical interpretation $\mathfrak{M}$ on which 'is true' has $\mathbf{T}\mathfrak{M}$ as its extension, this is not because something went wrong: it is because 'is true' works in a more complex way than that!

The point of the present section has been to try and convince you that this line of thought is a red herring. We can introduce more complex semantic mechanisms surrounding the word 'is true'—mechanisms which can work perfectly, even in face of the Liar. But typically the problem is that in doing so, we use *other* words, and *these* words now face the problem that *they* cannot always mean what we want them to mean. Thus we just shift the lump in the carpet. We are never going to be able to describe a system of semantic mechanisms which assigns meanings to words in contexts and which operates perfectly reliably, even in face of Liar-type phenomena (just as there cannot be a perfectly reliable Earman rocket, in face of closed time-like curves). Once we accept this, we may as well stick with the tried and true, known and loved classical picture. Or at least, we may as well, as far as the Liar is concerned. If some other phenomena—such as vagueness—require an alteration to the classical framework, then so be it. We will have a new range of interpretations available. But once again, there will be things we want to say (Liar-type sentences) which cannot be interpreted in a way that makes them mean what we wanted. The Liar is thus in a quite different category from other semantic puzzles, such as the puzzle of modelling vague language. We will never find a semantic framework which makes available a range of interpretations such that anything we might ever utter in any context (including statements about the semantic framework itself) can always mean exactly what we want it to mean. So the Liar can never motivate moving from one semantic framework to another. Rather, its lesson is that *whatever* semantic framework we adopt, there can *never* be perfectly reliable mechanisms relating the noises we make in the contexts we make them to items made available in the semantic framework (assuming that the semantic framework is sufficiently rich to be a prima facie correct model of a natural language such as English). Assuming Semantic Regularity, the upshot is that no semantic framework whatsoever is correct—which would mean that there are no semantic facts at all, that nothing we ever say means anything. Better to let go of Semantic Regularity, than have it drag one to these depths as it sinks! The lesson of the Liar, then, is that Semantic Regularity fails.

§ 9. OTHER ADVANTAGES OF REJECTING SEMANTIC REGULARITY. I have argued the the Liar *forces* us to reject Semantic Regularity. If we do not reject it initially, the revenge problem forces us to reject it later (unless we just leave our 'solution' incomplete, or wave our hands, or reject some even more entrenched belief, or insist upon a 'solution' which plainly does not give an adequate account of the phenomena). I shall now describe some independent benefits which flow from the rejection of Semantic Regularity.

§ 9.1. SEMANTIC INDETERMINACY. A number of authors have argued that meaning is much less determinate than we ordinarily suppose. Consider for a start Quine's argument for the indeterminacy of translation [22, ch. 2]. Quine notes first that the evidence available to a radical translator underdetermines her choice of Jungle-English translation manuals: i.e. there are conflicting manuals which fit all the available evidence equally well. This in itself does not threaten the thought that there is a determinate fact of the matter concerning what the natives mean by 'Gavagai', and that we simply cannot discover what this fact is. But now Quine draws on what I call the *publicity premise*, which says that meanings are essentially public, so that the *facts* about meaning cannot transcend the publicly available *evidence* about meanings. Thus the radical translator has access to all the meaning facts; thus there *is no* determinate fact as to what the natives mean by 'Gavagai'.

Many people seem to think that the problem here is behaviourism—that it is the (non-public) mental states of speakers which fix determinate meanings for their words. Enter Kripkenstein [12], who challenges you to point to any fact at all about your past self (including facts about mental states) which shows that you meant addition and not quaddition by 'plus' in the past. Kripkenstein argues forcefully that no such facts can be produced. Of course, your present self is in no better position than your past self. The conclusion is that the meanings of the terms you use now are radically indeterminate. There is no fact about you which determines that you mean addition and not quaddition by 'plus', and hence there is no fact of the matter as to whether you mean addition or quaddition by 'plus'.

However, it only follows from there being no fact about you which determines that you mean addition and not quaddition by 'plus', that there *is no* fact of the matter as to whether you mean addition or quaddition by 'plus', if we assume Semantic Regularity. For suppose that there is a perfectly determinate fact concerning what I mean by 'plus': the correct interpretation of my utterances of 'plus' assign this term a particular binary function on the domain. Given that this fact is not determined by my behaviour and mental states in concert with my environment, Semantic Regularity fails. Thus, if we reject Semantic Regularity, we will not be troubled by Kripkenstein's sceptical puzzle. Although we cannot cite facts about us which determine what we mean—i.e. although we cannot describe principled relationships between what we say and do in what circumstances, and what we mean by what we say—still we do not need to abandon the idea that there *are* perfectly determinate facts concerning what we mean by what we say—and it was the threat that it posed to *this* idea that made Kripkenstein's puzzle worrying.[18,19]

§ 9.2. EMPTY NAMES. Consider empty names. On my view, there aren't any. Every time you make an utterance, you speak relative to a particular (correct) interpretation of the language. This interpretation is classical, which means, among other things, that it assigns an object in the domain to every name in the language. So consider what happens when you say 'Santa Claus is fat'. You utter a wf Fa of the language relative to an interpretation $\mathfrak{M}$ on which F has as its extension the set of fat things, and on which a refers to some particular object. Which object? I don't know. Why that object and not some other one? There is no reason: but once we have rejected Semantic Regularity, this fact will not trouble us.

This approach to empty names yields all of the advantages of Frege's fix—which was to choose a default referent, say the number 0 or the empty set, for any name whose sense does not determine a referent—while avoiding some of its worst disadvantages. The advantages are that we can retain classical logic and semantics, without addition or alteration. The main disadvantage of Frege's view is that certain statements such as '1 + Santa Claus = 1' (if the default referent is 0) or 'Santa Claus is a subset of every set' (if the default referent is the empty set) are guaranteed to be true, by virtue of our treatment of (would-be) empty names. This strikes many as absurd. On my view, on the other hand, on any particular occasion of utterance—which will invoke a particular correct interpretation—'Santa Claus' will refer to some particular object: but there is no reason to think that it will refer to the same object on every interpretation; and also, we will not in general know which object 'Santa Claus' refers to. Thus while, on any occasion of utterance,

[18]Note that this is different from the response to Kripkenstein given by Soames [27], according to which the facts about what we say and do, together with the environmental facts, metaphysically determine the semantic facts, but we cannot derive the semantic facts *a priori* from the facts about what we say and do together with the environmental facts. On this view, there *are* principled relationships between what we say and do in what circumstances, and what we mean by what we say, it's just that we cannot say what these relationships are. Thus Soames's response does not involve denying Semantic Regularity.

[19]Very similar remarks apply to other well known [19, 18, 20] and less well known [28] arguments for radical semantic indeterminacy, but I lack the space here to treat these cases in detail.

there will *be* odd truths of Frege's sort, there are no odd statements that are *guaranteed* to be true simply by virtue of our treatment of empty names, and which we can therefore *know* to be true.

I do not deny that if we were considering only the problem of empty names, and weighing up possible solutions to it, the present solution would not be overwhelmingly attractive. But remember the dialectic. We have to give up Semantic Regularity in face of the Liar paradox. With Semantic Regularity gone, so is the most obvious reason for objecting to the proposal that there are no empty names, because on every occasion of utterance, you speak relative to a particular correct interpretation, which—being classical—assigns a referent to every name in the language. Of course, someone might want to argue that the problem of empty names is like the problem of vagueness in being something that might well motivate a change of semantic framework: in particular, we might need to consider interpretations which do *not* assign referents to every name. Perhaps this is right—although it seems to me that the only real reason for rejecting the approach to empty names just presented would be an adherence to Semantic Regularity. But if so, the solution to the Liar proposed above still goes through, mutatis mutandis to allow for the new non-classical interpretations. Thus, the solution to the Liar presented in this paper gives us a free solution to the problem of empty names, but does not force us to accept this solution.

§ 9.3. FALSE PRECISION. Any theory of vagueness which uses *non*-vague language to present a semantics for vague language, faces the problem of false precision. This problem is widely known as it strikes the epistemicist and fuzzy accounts of vagueness; but it also confronts the supervaluationist, and indeed anyone who offers a non-vague theory of vagueness. In the case of the epistemicist—who thinks that classical semantics gives the correct account of vague language—the problem presents itself as follows. The epistemicist tells us that a vague predicate such as 'is tall' has a classical extension, which includes (say) persons greater than or equal to 6′2″ in height, and excludes all other persons. The problem here is that it seems quite clear that our practice does not fix the boundaries of 'tall' so precisely—there is nothing about what we have said and done in the past that fixes the boundary at 6′2″ rather than 6′1″ or 6′3″—and people therefore conclude that the boundary cannot *be* so precise. Of course, once we reject Semantic Regularity, we will reject this inference: sure, *we* do not fix the boundary precisely here or there, but that does *not* mean that it is *not* precisely here or there. Similar remarks apply to the fuzzy theorists' view that the degree of truth of the claim that balding Bob is bald is (say) 0.67. Why not 0.66, or 0.68? Surely our practice does not fix the degrees of truth of vague claims so precisely? Indeed, surely not: but once we reject Semantic Regularity, we will not conclude that therefore the degrees of truth of vague claims cannot in fact *be* so precise.

The problem is not confined to classical and fuzzy views of vagueness. Contrary to what some persons seem to think, classical and fuzzy semantic theory impose no more precision on vague discourse than does supervaluationist semantic theory, which posits a particular set of admissible interpretations—each of which is classical—and hence draws a sharp boundary around the persons x such that 'x is tall' is (super) true: a boundary which is far too precise to have been fixed by our practice with terms such as 'tall'. Of course there are ploys for dealing with higher-order vagueness within both the fuzzy and supervaluationist views,[20] but we should be under no illusions as to what these can achieve. They cannot possibly avoid the problem of false precision totally, unless they move to a semantic account presented in a vague metalanguage: for *any* non-vague theory of vagueness—actual or possible—must set out a range of statuses which sentences may have, and then say of each sentence (vague or otherwise) that it has a particular one of these statuses. There is simply no other way of presenting a non-vague semantics for vagueness.

The problem of false precision is thus very persistent. If we wish not to resort to a vague theory of vagueness—and I have argued in Smith [24, §§3.6.3, 3.8.2] that we should *not* resort to such a theory—then we will face a version if it. However, if we reject Semantic Regularity, then the problem of false precision is no problem at all.[21,22]

In the case of the Liar sentence 'This sentence is not true', we want to utter a sentence which has the logical form of a negation, and contains a singular term which refers to the very sentence we utter, and a predicate which picks out the set of true sentences of the language—true, that is, on the very assignment of semantic values to items of the language that gives

[20]On the former see Smith [25] and on the latter see Fine [6].

[21]This response to the problem of false precision is different from that of Williamson [29, p.275], [30, p.207], who—if I read him correctly—argues that although there is no known "recipe for extracting meaning from use"—and need be no such recipe at all—nevertheless meaning facts are determined by facts about use. This view does not involve denying Semantic Regularity; it is analogous to the view of Soames mentioned in n.18.

[22]Note that this still leaves the question of *which* non-vague theory of vagueness is correct wide open. Vagueness is in fact one case which I think the classical view cannot handle. I argue in [26] that in order to accommodate the phenomena of vagueness, we need to countenance degrees of truth, and in [25] I present a particular degree-theoretic account which, I argue, is superior to existing degree theories. But of course my argument against the classical view does not centre on the problem of false precision: it is just as much (little) of a problem for my own view as for the classical view.

the correct account of what we mean when we say ‘This sentence is not true’. There is no interpretation which meets all these desiderata. My view is that the correct interpretation is therefore chosen at random, from those interpretations which meet as many of the desiderata as possible. In the cases of semantic indeterminacy and vagueness, the problem is the opposite: *too many* interpretations meet our desiderata (nothing that we say or do fixes that we mean plus not quus, or fixes the lower bound of the tall persons at $6'2''$ not $6'3''$, etc.). But the same response applies: out of all the interpretations which are as good as any other, the correct interpretation is chosen at random. (In the case of empty names, the situation can be of either type: sometimes a name does not—apparently—refer because there are no eligible referents, and sometimes because there are too many.) We thus achieve semantic determin*acy* through semantic indetermin*ism*: on any occasion of utterance, there is one particular thing that we mean; but nothing about what we say or do, together with our context, determines that we mean this rather than something else. I do not deny that this is hard to accept. But the Liar is a very hard problem, and this is its lesson.

References.

[1] Jon Barwise and John Etchemendy. *The Liar: An Essay on Truth and Circularity.* Oxford University Press, New York, 1987.

[2] George S. Boolos, John P. Burgess, and Richard C. Jeffrey. *Computability and Logic.* Cambridge University Press, Cambridge, fourth edition, 2002.

[3] Joachim Bromand. Why paraconsistent logic can only tell half the truth. *Mind*, 111:741–49, 2002.

[4] Tyler Burge. Semantical paradox, 1979. In [14], pages 83–117.

[5] John Earman. Implications of causal propagation outside the null cone. *Australasian Journal of Philosophy*, 50:222–37, 1972.

[6] Kit Fine. Vagueness, truth and logic. *Synthese*, 30:265–300, 1975. Reprinted with corrections in [10], pages 119–50. Page references to latter.

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

[8] Paul Halmos and Steven Givant. *Logic as Algebra.* Number Twenty-One in The Dolciani Mathematical Expositions. The Mathematical Association of America, 1998.

[9] Amitavo Islam. The abstract and the concrete. Typescript, 1996.

[10] Rosanna Keefe and Peter Smith, editors. *Vagueness: A Reader*. MIT Press, Cambridge MA, 1997.

[11] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72: 690–716, 1975.

[12] Saul Kripke. *Wittgenstein on Rules and Private Language*. Harvard University Press, Cambridge, MA, 1982.

[13] David Lewis. The paradoxes of time travel. *American Philosophical Quarterly*, 13:145–52, 1976.

[14] Robert L. Martin, editor. *Recent Essays on Truth and the Liar Paradox*. Clarendon Press, Oxford, 1984.

[15] Elliott Mendelson. *Introduction to Mathematical Logic*. Wadsworth and Brooks/Cole, Pacific Grove CA, third edition, 1987.

[16] Charles Parsons. The liar paradox, 1974. In [14], pages 9–45.

[17] Graham Priest. *In Contradiction*. Martinus Nijhoff, Dordrecht, 1987.

[18] Hilary Putnam. Models and reality, 1977. In [21], pages 1–25.

[19] Hilary Putnam. Realism and reason. In *Meaning and the Moral Sciences*, pages 123–40. Routledge and Kegan Paul, London, 1978.

[20] Hilary Putnam. *Reason, Truth and History*. Cambridge University Press, Cambridge, 1981.

[21] Hilary Putnam. *Realism and Reason*, volume 3 of *Philosophical Papers*. Cambridge University Press, Cambridge, 1983.

[22] W.V. Quine. *Word and Object*. MIT Press, Cambridge, MA, 1960.

[23] Joseph R. Shoenfield. *Mathematical Logic*. Addison-Wesley, Reading MA, 1967.

[24] Nicholas J.J. Smith. *Vagueness*. PhD thesis, Department of Philosophy, Princeton University, 2001. UMI ProQuest Digital Dissertations no. AAT 3021982, `http://wwwlib.umi.com/dissertations/fullcit/3021982`.

[25] Nicholas J.J. Smith. Vagueness and blurry sets. *Journal of Philosophical Logic*, 33:165–235, 2004.

[26] Nicholas J.J. Smith. Vagueness as closeness. *Australasian Journal of Philosophy*, forthcoming. Draft version available at `http://www.vuw.ac.nz/phil/njjs/vac-draft.pdf`.

[27] Scott Soames. Skepticism about meaning: Indeterminacy, normativity, and the rule-following paradox. In A. Kazmi, editor, *Meaning, Reference and Truth*, Canadian Journal of Philosophy Supplementary Volume, 1997.

[28] J.R.G. Williams. A paradox in the theory of meaning. Typescript, `http://arche-wiki.st-and.ac.uk/~ahwiki/pub/Main/RobertWilliams/DefenceIndeterminacy.pdf`, 2004.

[29] Timothy Williamson. Vagueness and ignorance, 1992. In [10], pages 265–80.

[30] Timothy Williamson. *Vagueness*. Routledge, London, 1994.