Paradoxes and the Foundations of Semantics and Metaphysics

by

Matti Eklund

Submitted to the Department of Linguistics and Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2000

Author ....
Department of Linguistics and Philosophy
14 August, 2000

Certified by ....
Stephen Yablo
Associate Professor
Thesis Supervisor

Accepted by ....
Vann McGee
Chair, Department Committee on Graduate Students

Paradoxes and the Foundations of Semantics and Metaphysics

by

Matti Eklund

Abstract.
Numerous philosophical problems, otherwise quite different in character, are of the following form. Certain claims which seem not only obviously true, but even constitutive of the meanings of the expressions employed, can be shown to lead to absurdity when taken together (perhaps in conjunction with contingent facts about the world). All such problems can justly be called paradoxes. The paradoxes I examine are the liar paradox, the sorites paradox, and the personal identity paradox posed by the fission problem.

I argue that, in all of the cases examined, the claims that jointly lead to absurdity really are constitutive of the meanings of the expressions employed, in the following ways. First, semantic competence with the expressions involves being disposed to accept these claims. Second, the claims are reference-determining, in that the semantic values of the expressions employed are constrained by the condition that these claims should come out true, or as nearly true as possible. If a claim or principle is constitutive of meaning in both of these ways, I call it *meaning-constitutive.* When the meaning-constitutive principles for some expressions of a language are inconsistent, I call the language inconsistent. This is a stipulative definition; but it accords well with what other theorists who have talked about languages being inconsistent, for example Alfred Tarski, have had in mind.

In chapters one and two, I argue that our language is inconsistent. In chapter three, I relate my theses to Frank Jackson's and David Lewis's views on how reference is determined.

Another problem posed by the liar paradox concerns important theses in the philosophy of language. The liar reasoning shows that under certain conditions, a natural language cannot contain a predicate satisfying the T-schema. But many theses in the philosophy of language presuppose that truth satisfies the T-schema. I resolve this conflict in a Tarskian way: by saying that truth then is expressed only in an essentially richer metalanguage. However, I argue that taking this route means having to embrace the existence of absolutely inexpressible properties – and even to embrace the conclusion that some properties of which we appear to have concepts are absolutely inexpressible. All this is dealt with in chapter four.

In the fifth chapter I show that my arguments of the previous chapters have (dis)solved the liar paradox. And finally, in the sixth chapter, I discuss the philosophical significance of truth and logic, and argue that these questions are significant only if understood in a new way. In this last chapter I also discuss the implications of the liar paradox for metaphysics; more specifically, its implications for the issue of how metaphysical claims are justified.

Thesis supervisor: Stephen Yablo
Title: Associate Professor of Philosophy

Contents

Acknowledgments

Many thanks to my main advisor, Stephen Yablo, and the other members of my committee, Michael Glanzberg and Robert Stalnaker, for very helpful comments and criticism.

Other people who deserve thanks for commenting, criticizing and proofreading the chapters in my thesis, at one stage or other of their development, are Alyssa Bernstein, Andrew Botterell, Linus Broström, Alex Byrne, Adam Elga, Michael Fara, Ned Hall, Patrick Hawley, Sarah McGrath, Anthony Newman, Graham Priest, Agustín Rayo and Judith Thomson.

Very special thanks, of a different kind, are due to my wife and my daughter, Helga Sverrisdóttir and Silja Mattadóttir, and to my parents, Lennart Berglund and Camilla Eklund.

Chapter 1

Inconsistent Languages

1.1 Introduction

In this chapter, I shall argue for the hypothesis that we are sometimes led to accept untrue and even jointly inconsistent propositions by virtue of our semantic competence.

This hypothesis, if true, serves to explain a host of puzzling features of the sorites and liar paradoxes. It shows that standard work on the paradoxes fails to address the main theoretical question posed by them; and in addition, it (unexpectedly) shows how classical logic and bivalence can be retained at least in the face of the sorites.

If my hypothesis about semantic competence is true, this is likely to have significant consequences also regarding other philosophical problems. In chapter 2, I shall show how, given this hypothesis about semantic competence, a certain problem regarding personal identity, the fission problem, can be nicely dissolved. But the project of the present chapter is to argue for the hypothesis about semantic competence by considering the sorites and liar paradoxes.

1.2 The paradoxes

The sorites paradox goes as follows. We have a series of color patches, ranging from red to orange. The first patch looks red (to normal observers under normal conditions), the last orange. Adjacent patches are observationally indistinguishable with respect to color for normal observers in normal circumstances. These stipulations are clearly coherent. But now consider the following argument: First, the first patch looks red, by the stipulations of the thought experiment. Second, if patch n looks red, then patch $n+1$ looks red as well, since adjacent patches are observationally indistinguishable with respect to color. But then, by elementary reasoning, we can conclude that the last patch looks red. But the last patch looks orange, not red.

As for the liar paradox, consider a liar sentence, a sentence λ that says of itself that it is not true:

λ: λ is not true.

Suppose first that λ is not true. Since this is precisely what λ says, λ is then true. Hence, by reductio, λ is true. But if λ is true, then λ is not true: for λ says that it is not true, and if λ is true then what it says is the case. Contradiction.

We can think of these paradoxes as *arguments*; presumably, unsound arguments. But the sorites and the liar are perplexing in a way that other unsound arguments are not. We are puzzled by their unsoundness. It is not clear to us *how* these arguments can be unsound. In this chapter, I shall present a particular hypothesis about why these paradoxes are so perplexing.

A number of theorists have argued that the sorites and the liar arise because the expressions employed are "incoherent" or "inconsistent".[1] In the case of the liar, this has even been called the "received view".[2] But talk of expressions being inconsistent is hardly self-explanatory. And on the most obvious ways of understanding the claim, it is scarcely plausible. If we understand the claim that the truth predicate is inconsistent to mean that the truth predicate is false of everything, or perhaps is not meaningful at all, then the claim that the truth predicate is inconsistent is pretty obviously false. Similarly if we take the inconsistency of the truth predicate to consist in there being sentences the predicate both does and does not apply to. For then if 'true' is inconsistent, the law of non-contradiction is not valid. But still I believe that the theorists who have suggested that the paradoxes arise because of incoherence or inconsistency in the expressions employed are getting at something both correct and important.

Let us call the untrue premises and invalid steps in an unsound argument the *culprits* of the argument. Then what I shall argue is that being competent with the expressions

[1] In the case of the sorites, see e.g. Dummett (1975), Quine (1981), Rolf (1981) and Horgan (1995). Regarding the liar, see Tarski (1983) and Chihara (1979).

[2] McGee (1997), p. 403.

employed in the paradoxes involves being disposed to accept the culprit as true or valid: so long anyway as you do not have what you take to be evidence against its truth or validity. If you refuse to accept the culprit as true or valid, when not having what you take to be defeating evidence against its truth or validity, you manifest lack of semantic competence.[3] Competence involves being disposed to accept the culprit. Let us say that an unsound argument *exerts pull* just in case speakers' semantic competence disposes them to accept the culprit or culprits in the argument. I shall frequently talk about the belief-forming dispositions a speaker has by virtue of her semantic competence. Let us call these the speaker's *competence dispositions.* My main theses and arguments will concern semantic competence. It will therefore be useful to tie talk of meaning closely to what semantic competence involves. (There are, of course, other uses of 'meaning'.)[4]

Here is a way of illustrating the basic idea. It is occasionally noted that there are two distinct ideas underlying the notion of *analyticity*: on the one hand the (metaphysical) idea of truth by virtue of meaning; on the other hand the (epistemological) idea of statements we are justified in believing solely by virtue of our semantic competence. The theorists who have noted this (Paul Boghossian and Jamie Tappenden) have both gone on to argue that the metaphysical notion of analyticity and the epistemological notion of analyticity need not go together. Among other things they have argued that the metaphysical notion does not make any sense, while the epistemological notion does. What I shall argue here is, in effect, that there is a different reason notions of analyticity come apart: epistemologically analytic statements may be jointly inconsistent.[5]

Roughly, the picture is this. The sorites and the liar arise because our semantic intuitions

[3] Competence, of course, comes in *degrees*. For many expressions we use and are generally competent with, there are some aspects of the meanings of these expressions that we do not know.

[4] I talk throughout about *linguistic expressions* and what *semantic* competence involves. I could as well carry out the discussion in terms of *concepts* and what *conceptual competence* involves. In some places, this would make the discussion read more naturally.

[5] See Boghossian (1997) and Tappenden (1993 and 1993b). Tappenden, like me, holds that epistemologically analytic statements may fail to be true. But he still holds that epistemologically analytic statements are always guaranteed to be *non-false*. (For criticism of Tappenden on analyticity, see chapter 5.)

I shall by and large set aside Quinean worries about analyticity here, but a couple of remarks are in order. First, if Quine is right then there can be no serious lexical semantics at all: lexical units do not have intrinsic semantic features. But everyone concern ,d with solving the paradoxes is concerned with what to say about the semantic features of vague predicates and of the truth predicate. So if Quinean worries should be taken seriously, they present problems for all parties concerned.

Second, it deserves emphasis that I am only committed to epistemological analyticity. It is certainly not obvious (I would say: it is false) that commitment to epistemological analyticity carries commitment to metaphysical analyticity. And arguably, the most convincing Quinean criticisms concern analyticity in the metaphysical sense.

are inconsistent, where our semantic intuitions are the intuitions of truth and validity that we have by virtue of our semantic competence. Our semantic intuitions are not indefeasible. Upon noting that your semantic intuitions lead you to accept an inconsistent corpus of statements, you can refuse to take these intuitions at face value, rejecting one or more of them as non-veridical, without thereby manifesting lack of semantic competence. (Rather the opposite: you positively manifest logical skills.)

The reason the sorites and the liar are perplexing is then that we are disposed to accept the culprits in these arguments by our competence with the expressions employed. As I shall go on to explain in the next section, this suggestion can serve to cash out the claim that expressions are incoherent or inconsistent.

1.3 Why believe that the paradoxes exert pull?

Why should we believe the account I have just outlined? In this section and the next, I shall go through some considerations in its favor. In this section I shall show that my account serves nicely to explicate the rough idea that the paradoxes arise because the expressions employed are 'incoherent' or 'inconsistent'. Someone already attracted to this idea ought on this basis to find the account plausible. In the next section, I shall present some more general considerations, aimed to convince also those who do not find incoherence suggestions attractive. The positive case for my account will not be completed in this section. The arguments here will primarily serve a motivating function. Later (primarily in section eight) I shall present further positive arguments.

I shall center the discussion in this section around a recent discussion of the perhaps most well-known claim of this general kind: Tarski's claim that the liar paradox shows natural language to be inconsistent.

In his recent (1999), spelling out the intuition that the liar paradox arises because the notion of truth is incoherent, Scott Soames says,

> Nothing in the statement of [the liar paradox] by itself constitutes an attack on the legitimacy or coherence of our ordinary notion of truth. Nevertheless, the intractability of the paradox could be used as the basis for such an attack. One line of reasoning to this effect might be put roughly as follows: The notion of truth, as we ordinarily understand it, requires all instances of [the disquota-

> tion schema] to be true. However, [the liar paradox] shows that this leads to contradiction. Therefore, the notion of truth, as we ordinarily understand it, is incoherent...[6]

Soames goes on to explain the suggestion that the notion of truth, or, as I shall put it, the meaning of 'true', requires that the disquotation schema[7] be valid, with special reference to Tarski. Tarski appears to have had in mind something like the suggestion Soames mentions when saying that the liar paradox shows natural language to be "inconsistent".[8] Soames easily refutes the position he attributes to Tarski and apparently takes this refutation to show that the suggestion that the notion of truth is "incoherent" ought to be rejected.

We need not go into the specific details of Soames' interpretation of Tarski. The basic outline is the following. The laws of classical logic are valid in natural language, in the sense that some expressions of natural language express the classical logical operations, and natural language contains the means for naming all its expressions. Moreover, Tarski held that the meaning of 'true' requires that the disquotation schema be valid, and that accordingly the disquotation schema is valid.[9] So a contradiction can be validly derived from principles true in English, and natural language accordingly contains true contradictions. This is what it means to say that natural language is inconsistent. As Soames points out, given the assumption that the laws of classical logic hold for natural language, every sentence of natural language is then true (and false). Soames concludes that Tarski's view must be rejected, since its consequences are so unpalatable.[10]

There are at least two reasons for thinking that Soames' Tarski – whether he is the real Tarski or not – is badly confused. Not only is his position very implausible; it is internally incoherent. *First*, if every sentence of natural language is true, there seems to be no interesting sense in which the laws of classical logic hold for natural language or in which the disquotation schema is a valid schema: *all* sentences of the object language are

[6] Soames (1999), p. 51.

[7] The disquotation schema is the schema,

$\ulcorner p \urcorner$ is true if and only if p.

[8] See Tarski (1983), p. 164f.

[9] Following Soames I shall, when representing the train of thought represented in the quote, talk about what meanings 'require'. Such talk is only meant to capture this loose train of thought: the real philosophical work consists in cashing out such talk, and it is this that will concern me.

[10] See Soames (1999), pp. 52-55. In (1967), Hans Herzberger presented a broadly similar interpretation of Tarski.

true. Perhaps this problem can be overcome. It may be possible to delineate a special status for classical logic and the disquotation schema in natural language after all. But *second*, as Soames points out, the conclusion that there are true contradictions is unhappy, as the meanings of 'true' and 'not' would appear to require that there be no such things. Soames does not argue for this claim but takes it to be obvious, which seems reasonable. Moreover, Tarski, defending classical logic, ought himself to have found it reasonable. After all, the reason *ex falso quodlibet* – the principle that from a contradiction every proposition follows – is a valid form of inference in classical logic is, intuitively, that it is impossible for contradictions to be true. But if the meanings of 'true' and 'not' require that there not be true contradictions, then there are no true contradictions – at least if we assume, as Soames' Tarski does, that if the meanings of some expressions "require" that S is true, then S is true. So *by his own lights*, Soames' Tarski ought not to conclude that there are true contradictions.[11]

What I think this shows is that we must not assume that if the meaning of 'true' requires that the disquotation schema is valid, the extension of 'true' is such that this schema is valid. A different, more plausible picture is available: the meaning of 'true' requiring that the disquotation schema be valid only implies that a disposition to accept the disquotation schema is part of semantic competence with the predicate: the liar paradox exerts pull.

Nothing I have said is meant to imply that Tarski's real view was that the liar exerts pull. If Tarski had something specific in mind when he said that natural language is inconsistent (which I doubt), we can only make guesses as to what it was. But interpreting Tarski as holding that the liar exerts pull is charitable and apparently respects the intuition that underlies his reasoning.

I shall make use of Tarski's locution of languages being "inconsistent", interpreting it in accordance with my suggestion on Tarski's behalf. A language L is inconsistent just in case the principles – by which I will mean sentences and inferences – partially constitutive of the meanings of expressions of L are jointly inconsistent, where a principle is constitutive of the meaning of an expression just in case competence with the expression involves being disposed to accept the principle as true or valid. Talk of inconsistency expressions and sets

[11] Aficionados of reasoning with contradictions will notice that there is a way out for Soames' Tarski: he can say that there both are and are not true contradictions. But clearly, what from a classical point of view the meanings of 'true' and 'not' require is not only that there are no true contradictions but also that it is not the case that at the same time there are true contradictions.

thereof is to be understood in the same way.[12,13]

1.4 The culprits

The suggestion that the paradoxes arise because the expressions employed are inconsistent can be vindicated if the paradoxes exert pull. For those already attracted to the thesis that the paradoxes arise because of inconsistency of the expressions employed, this should be ample motivation for the thesis that the paradoxes do exert pull. But not everyone is so attracted.

I shall now present rather generally accepted diagnoses of which the culprits in the sorites and the liar are, and then argue that our competence with the expressions employed in the paradoxes disposes us to accept the culprits. If the diagnoses of the paradoxes are correct, the paradoxes then exert pull.

It is generally (if not universally) accepted that the culprit in our version of the sorites paradox is the premise which states that for every patch k in the series, if k looks red

[12] This definition of "inconsistent language" is similar in spirit to that in Tappenden (1993b). Tappenden characterizes a language as inconsistent just in case the *pre-analytic* sentences of that language are jointly inconsistent, where a sentence is pre-analytic "if one need only know the sentence to know that it is not to be counted false" (p. 238).

To illustrate my characterization of inconsistent languages, consider the following putative objection: "Why does not the fact that 'is an even odd number' is a predicate of English already show that English is 'inconsistent' in the sense outlined? After all, *being an even odd number* is an inconsistent concept."

Reply: Let x range over natural numbers. We can suppose the meaning-constitutive principle for 'even' to be: x is even iff x is divisible by 2; and we can suppose the meaning-constitutive principle for 'odd' to be: x is odd iff not so divisible. The meaning-constitutive principle for 'even and odd' is then: x is even and odd iff x both is and is not divisible by 2. And to believe what is after the colon is *not* to have inconsistent beliefs; what is after the colon even seems *true*.

The expression 'even and odd' *would* be inconsistent if there was some *other* meaning-constitutive principle for it that entailed that there exist numbers that are both even and odd.

[13] In (1979), Tyler Burge says, criticizing the idea that languages can be inconsistent:

> The best ground for dissatisfaction [with this idea] is that the notion of a natural language's harboring contradictions is based on an illegitimate assimilation of naturـl language to a semantical system. According to that assimilation, part of the nature of a "language" is a set of postulates that purport to be true by virtue of their meaning or are at least partially constitutive of that "language". Tarski thought he had identified just such postulates in natural language as spawning inconsistency. But postulates are contained in theories that are promoted by people. Natural languages *per se* do not postulate or assert anything. (Burge 1979, p. 169)

This is fine conceived as a criticism of the idea that the claim that ordinarily would be expressed by "natural language is inconsistent" is true, or even makes sense. The way we normally talk about languages, languages are not the kind of thing that can be inconsistent. But Burge seems also to want to argue against the idea that we can endow talk of the inconsistency of language with meaning in an intuitively natural way. But it seems one can, as in the main text, in a natural way find principles associated with expressions such that the consistency of a language can be identified as the consistency of the associated principles.

then $k + 1$ does too. This premise we shall call the sorites premise, following established usage. The reason for taking the sorites premise to be the culprit is simply that there seems to be no other likely candidate. The logic employed in the argument is fairly weak, and denying the premise that the first patch looks red seems desperate. But as I now shall argue, our semantic competence disposes us to accept the sorites premise as true in the imagined situation.

In his discussion of vagueness (1975), Crispin Wright introduced the notion of *tolerance*. A predicate of color (size, age, etc.) is tolerant just in case there is some positive degree of change in respect of color (size, age, etc.) that never suffices to affect the justice with which the predicate is applied to an object, and yet the applicability of this predicate is affected by some sufficient change in respect of color (size, age, etc.).[14] Wright also presented a case for taking many vague predicates to be tolerant and for taking their vagueness to be closely linked to, perhaps to consist in, their tolerance. The idea was that the use and usefulness of predicates like 'is a heap', 'is a child', 'is red', etc. is linked to their tolerance. Whether something is a heap or not is something to be settled by *casual* observation, and this is crucial to the usefulness of 'heap': a precise counterpart of 'heap' would not serve the purpose served by 'heap' equally well.

The reasoning here could well be doubted: one should not too readily accept that just because we commonly or always choose to call or not call something a heap on the basis of casual observation, small enough differences in size cannot matter to the applicability of the predicate. It could simply be that in all cases we are interested in, the matter can be settled through casual observation.

But in the case of a predicate like 'looks red', more convincing considerations can be presented. The meaning of 'looks red' appears to require that its applicability be determinable by mere observation: it is in this sense observational. But then it seems that 'looks red' must be tolerant. For then differences in color that cannot be detected by mere observation

[14] The applicability of many vague predicates does not vary only with respect to a single property, but with respect to a set of properties. The characterization of tolerance is easily modified to handle that. We might say, e.g.:

> Where Φ is the set of determining properties for a predicate F, there is some difference with respect to some or all of the properties in F that sometimes matters to the applicability of F, but some sufficiently small difference with respect to some or all of the properties in Φ never matters to the applicability of F.

See for example Varzi (1995), p. 50.

should not make a difference to the applicability of the predicate.[15]

But the sorites premise, the culprit in the sorites paradox, is true provided (1) adjacent patches are observationally indistinguishable with respect to color, and (2) if two objects are thus indistinguishable, then if one looks red the other does too.[16] That (1) is true is simply a feature of the set-up of the thought experiment. Accordingly, (2) cannot be true; for if it is, the sorites premise is true and the sorites argument comes out sound, which is absurd.[17] But we just saw that reflection on "looks red" yields the conclusion that (2) should come out true. So on the one hand, it appears that 'looks red' must be tolerant; on the other hand it cannot be.

But suppose that a speaker's semantic competence can dispose her to accept principles that are in fact untrue or invalid. Then we can still respect the intuition behind the claim that 'looks red' and other observational predicates are tolerant. We can say that our semantic competence involves being disposed to accept that these predicates are tolerant. This is compatible with saying that, on pain of inconsistency, no predicate can actually be tolerant.

So we are disposed by our semantic competence to accept the culprit in the sorites paradox: and hence the sorites paradox exerts pull (provided the suggested diagnosis of this paradox is correct). I shall now argue that the same goes for the liar paradox.

According to a widely accepted diagnosis of the liar argument, the culprits of the argument are steps which rely for their validity on the assumption that all instances of the disquotation schema are true. (We can go from the supposition that λ is not true to the claim that λ is true and vice versa in the liar argument only if we assume that the relevant instances of this schema are true.)

But even someone who accepts this diagnosis of the liar reasoning would have to agree it is extremely natural to take the disquotation schema to be valid: one even has to remind oneself that the liar argument does rely on that assumption. Predicating 'true' of a sentence

[15] Hence, strictly, when I say that the sorites paradox arises because the expressions employed are inconsistent, I really should say that the particular version of the sorites that I consider arises for that reason. I believe the diagnosis generalizes to other versions of the sorites, but I will not argue that point here.

[16] I set aside difficulties arising from the fact that 'looks red' is context-sensitive. Some theorists (e.g. Williamson (1994), p. 174) have argued that the context-sensitivity of 'looks red' blocks the sorites reasoning in this case. That puts very much strain on the notion of 'context' involved.

[17] As Wright of course notes in his (1975). Wright concludes that we must explain away the appearance of tolerance. He holds that faulty assumptions about meaning are what lies behind the mistaken impression that it is part of the meaning of vague predicates that they are tolerant.

is, or is intended to be, to say that what the sentence says is the case actually is the case. But that is exactly what is effected by assertorically uttering the sentence.

Hence a sentence S and a sentence which says that S is true *should* come out materially equivalent. But if we accept the diagnosis presented, then we must abandon the presumption that they in fact are equivalent. Again we have a conflict between what it appears that the meanings of expressions require and what can actually be the case. And again this conflict can be resolved if we accept that the meanings of some expressions can require some sentences to be true without these sentences being true. Thus, the meaning of 'true' can require that S and "S is true" come out materially equivalent without S and "S is true" actually being materially equivalent. All that follows from what the meaning of 'true' requires is that a speaker's competence with the predicate disposes her to accept that this equivalence generally holds. Hence, given the suggested diagnosis of where the liar reasoning goes wrong, the liar paradox exerts pull.

I expect one response to my arguments that our competence dispositions compel us to accept the culprits in the sorites and the liar as true and valid, respectively, to be that if we really are disposed by our competence to accept the diagnosed culprits as true and valid, that only shows that the culprits have not been correctly identified.

Nothing can strictly disprove this claim, save going through every possible diagnosis of the paradoxes. But the culprits I have considered seem to be the obvious ones; and if our semantic competence disposes us to accept *them* it is more than likely that it disposes us to accept also the less likely culprits. Besides, more importantly, I think one is prepared to react as envisaged only if one believes that there are a priori reasons to think that our semantic competence cannot dispose us to accept principles which are in fact untrue or invalid. I shall go on to show that there is no principled reason to think that our semantic competence does not dispose us to accept untrue or invalid principles. The envisaged response is then decidedly less attractive.

There is also the phenomenon that we would regard someone who fails to be puzzled by the paradoxes, someone who fails to regard each premise and step as initially attractive, as simply having failed to grasp something. This phenomenon admits of different explanations. But one straightforward explanation is provided by the thesis that the paradoxes exert pull. For this thesis entails that someone who fails to be puzzled by the paradoxes thereby manifests lack of competence with the expressions employed. This gives further reason to believe that the paradoxes exert pull; a reason not conditional upon any particular diagnoses of which the culprits are.

1.5 How can meaning be such that the paradoxes exert pull?

I have been motivating the thesis that the paradoxes exert pull: that we are attracted to the culprits by virtue of our semantic competence. The motivating considerations have been centered around features of the paradoxes. One may think, however, that the suggestion is a non-starter – that it is a priori that it cannot be by virtue of semantic competence that one accepts a principle that is untrue or invalid. In this section and the next, I shall consider this worry. First I shall reinforce it, showing (briefly) that exertion of pull is indeed ruled out at the outset by all popular accounts of semantic competence, as standardly presented. But then I shall show how at least some of these accounts can be slightly modified so as to be compatible with exertion of pull; and that nothing in the general rationale underlying these accounts demands that they not be so modified.

First, I shall briefly show that truth-conditional semantics, conceptual-role semantics, and Fregean accounts of meaning as standardly conceived all rule out pull exertion from the outset.

(1) Consider first truth-conditional semantics, under which to know the meaning of a sentence is to know the truth condition of the sentence. The truth conditions of the sentences employed in the paradoxes cannot be such that all premises in these arguments are true and all inferences are valid. For then the paradoxes would be sound arguments. But then what a speaker knows by virtue of knowing the truth conditions of the sentences employed in the paradoxes cannot dispose her to accept the culprits in them.[18]

(2) Conceptual-role semantics, according to which to know the meaning of an expression or set of expressions is to have the right belief-forming dispositions, typically assumes that all inferences a speaker is disposed to accept by virtue of her semantic competence are truth-preserving.[19]

[18] This is not to suggest that if knowledge of meaning is knowledge of truth-conditions, then we ought not to be taken in by the paradoxes at all. My claim is only the weaker one, that if knowledge of meaning is knowledge of truth-conditions, then our knowledge of meaning cannot essentially involve being disposed to be taken in by the paradoxes, and hence this thesis about semantic competence is inconsistent with the paradoxes exerting pull.

[19] See e.g. Peacocke (1992), p. 19.

(3) Fregean accounts of meaning, according to which the semantic values – contribution to truth conditions[20] – of expressions are whatever satisfies the conditions specified by the sense of the expression, require that senses of meaningful, non-empty expressions be consistent in that the conditions specified by the sense be jointly satisfiable.[21] If these conditions are not jointly satisfiable, then nothing satisfies them, and the expressions fail to have semantic values altogether.

The argument that truth-conditional semantics is not compatible with exertion of pull can also be used to show that if the sorites and the liar do exert pull, this is something standard accounts of these paradoxes are unable to explain. Theorists writing about the paradoxes are typically concerned only with the semantic values of the expressions employed. The solutions to the sorites paradox most commonly presented are, for example, supervaluationism, intuitionism, many-valued logic, and degree theories of truth.[22] These accounts are all *accounts of truth conditions*, in that these accounts are about the truth conditions of the sentences concerned and the semantic values of the subsentential expressions concerned. And the work on the liar paradox, most of it formal in nature, is about giving acceptable truth conditions to the sentences employed, and other sentences in which the expressions employed occur.[23] But no account of truth conditions can explain why these unsound arguments exert pull. Exertion of pull is a matter of what semantic competence requires. In order for an account of truth conditions to explain exertion of pull, a speaker must appreciate the pull of the paradoxes by virtue of her knowledge of truth conditions. But the truth conditions of the sentences employed in the paradoxes cannot be such as to make all premises and steps true and valid, respectively. But then a speaker's knowledge of truth conditions cannot require her to (lacking what she takes to be defeating evidence) accept each premise and step in the paradoxes. Thus, if indeed the paradoxes do exert pull, standard work on them fails to address this phenomenon: and it seems that if they do exert pull, that is an extremely central and puzzling fact about them.

There is thus a conflict between on the one hand pull exertion, and on the other hand, standard views on semantic competence and standard accounts of the paradoxes. If we

[20] Which we may think of as, for example, intensions: functions from possible worlds to extensions.

[21] See e.g. Dummett (1981), p. 227.

[22] Regarding supervaluationism, see Fine (1975) and Dummett (1991). Regarding intuitionism, see Putnam (1983). For many-valued logic, see Tye (1994). For degree theories of truth, see Edgington (1992).

[23] See, for example, Barwise and Etchemendy (1987), Gupta and Belnap (1993), McGee (1991), Simmons (1993), and the papers in Martin (1984).

conclude that the paradoxes exert pull, we shall have to rethink our positions on the other matters. This conflict might make us skeptical toward the idea of pull exertion. However, I shall now show how rather minor modifications of, at least, conceptual-role semantics and Fregean accounts of meaning allow us to accommodate pull exertion. The assumption commonly made by conceptual-role theorists that the inferences a speaker's competence disposes her to accept are all truth-preserving is, we shall see, unfounded. Moreover, there is a way of revising Fregean semantics so that the semantic values of the expressions of a language are not what satisfies *all* conditions specified by the senses of the expressions, but rather what comes *closest* to satisfying these conditions. And Fregean semantics thus modified is as faithful to the reasons for embracing it as is the original version.[24]

I shall return later (at the end of the next section) to the suggestion that Fregean semantics can be revised as suggested. First, let us consider a modified conceptual-role theory according to which (i) competence dispositions can fail to be truth-preserving, and (ii) the competence dispositions in which semantic competence consists are defeasible. Such a theory is obviously compatible with exertion of pull. The one worry is: is such a modified conceptual-role theory *workable*?

It is because I want to accommodate pull exertion that I allow the belief-forming dispositions in which semantic competence consists to be defeasible. But it is important on any account according to which semantic competence consists in having some kind of belief-forming dispositions that these belief-forming dispositions be defeasible. There are long-standing philosophical disputes over the status of some purported logical laws. We would not want to say that in every such dispute, either some disputant lacks semantic competence or the disputants speak past each other, speaking different languages. It is possible to reject some law of logic without thereby manifesting semantic competence. Accordingly, any conceptual-role theorist in her right mind ought to leave open the possibility that two speakers of the same language may disagree about whether some sentence is a law of logic, without either of them manifesting lack of semantic competence. One natural and attractive way of leaving open this possibility is to say that two speakers speak the same

[24] In connection with conceptual-role semantics, I should remark that according to certain 'holistic' versions of conceptual-role semantics, *all* inferences in which an expression is employed are constitutive of the meaning of the expression. But then, unless all our belief-forming practices are jointly consistent, some inferences that are not truth-preserving will have to be constitutive of meaning.

language just in case they have the same competence dispositions.[25,26]

1.6 How semantic values are determined

As noted, on common versions of conceptual-role theories, the inferences speakers are disposed by their semantic competence to accept are held always to be truth-preserving. The semantic values of expressions are then taken to be constrained by the fact that these inferences should all come out truth-preserving. But if, as on the modified conceptual-role theory, semantic competence can consist partly in dispositions to accept and make inferences that are not truth-preserving, the semantic values of expressions cannot be so constrained. For the modified conceptual-role theory to be viable, we must find some other way in which semantic values of expressions are constrained by what principles are constitutive of meaning.

I shall give a limited response to the problem of how semantic values are determined. I shall identify two assumptions commonly tacitly made in the literature and show that if we take these assumptions on board, we can say how, on the modified conceptual-role theory, the semantic values of expressions are determined by the meanings of the expressions.[27]

[25] It may be suggested that a theorist who denies a logical law that is in fact valid only has a mistaken *theory* about her use of the expressions employed in stating the logical law, and she actually does not use the expressions employed in stating the logical law in any non-standard way. If so, the problem discussed in the text does not even arise: the problem arises only on the assumption that the dissenting logicians' use of the relevant expressions differs from that of the rest of us.

However, although there may be people such that it is *only* their theoretical views on the meanings of the logical constants that are false, there clearly *can* be dissenting logicians who let their theoretical views influence their use. Those dissenters are the topic of the discussion in the text.

[26] There are other reasons for being attracted to the view that a principle can be, in some sense, part of the meaning of an expression even if speakers' adherence to it is only defeasible. One sort of skepticism about the notion of analyticity, common among philosophers, is that even should there be a principled analytic/synthetic distinction, the analytic truths are very few and far between. It is said that few purported analytic truths are in reality epistemically necessary – they could turn out to be false – and epistemic necessity is held to be a necessary condition for analyticity. To take a time-worn example, "cats are animals" is not epistemically necessary, and hence, the reasoning goes, not analytic. Since few principles are analytic, few principles are part of the meaning of an expression, for if a principle is part of the meaning of some expression it is analytic. Linguists working in lexical semantics, on the other hand, typically assume that there is a substantial network of intrinsic meaning connections. (See e.g. the exchange between Fodor & LePore (1998) and Pustejovsky (1998).)

Given the present account of meaning, there is hope for reconciliation: for a principle can then be constitutive of meaning without being epistemically necessary, analytic, or even true or valid.

[27] The semantic value of an expression is not exclusively a matter of what goes on within speakers. It depends as well on features of the world. It is generally accepted that the semantic values of natural kind terms and proper names depend on what objects the use of these expressions is causally linked to. More controversially, David Lewis has suggested that the semantic values of expressions depend, to some extent, on what candidate values are *natural* in some general sense. (See Lewis (1983), pp. 370-77.) I shall however disregard the input from the environment, at least in the discussion in the main text.

The assumptions are: (a) *Respect.* Other things equal, the semantic values of expressions are such as to make speaker judgments maximally correct.[28] Hence, when we, as theorists, conclude from a speaker's judgments what the semantic values of expressions of her language are, we should seek to respect the speaker's judgments in so far as possible.[29,30] (b) *Approximation.* When rendering speaker judgments "maximally correct", different judgments are differently weighted; making speaker judgments maximally correct is not to make maximally *many* speaker judgments correct. (The assumption of approximation is often left implicit.)

A response of this kind suffices, as these assumptions are nearly universally made. In particular, commentators on the paradoxes seek to present accounts of truth conditions that respect speaker judgments as far as possible (respect), and recognizing that their accounts of truth conditions cannot make true every intuition, the discussion of which account of truth conditions to opt for proceeds in terms of which account comes closest to making true all speaker judgments (approximation).

Making the assumptions of respect and approximation, we can give the following account of how semantic values are determined by the principles constitutive of meaning. The semantic values of expressions are such as to make the meaning-constitutive principles come out as nearly correct as possible, where for a principle to come out correct is for it to be true or valid, respectively. In an ideal case, the meaning-constitutive principles should all come out correct, but if I am right about the sorites and the liar, these paradoxes show that not all meaning-constitutive principles can be correct.

Even if, in principle, it is possible to compare the extent to which assignments of semantic values come close to rendering correct the meaning-constitutive principles, it should

[28] Or to minimize unexplained untruth of speaker judgments, or something like that. For present purposes, the differences between these claims are unimportant.

[29] "Speaker judgments" in the statement of the assumption of respect can be understood in two different ways. First, "speaker judgments" can mean, generally, judgments that people, who ipso facto are speakers, make. Second, "speaker judgments" can be understood to mean: judgments people make on the basis of their linguistic competence. It will not matter for present purposes how we understand it, except in the next paragraph, where the former understanding is presupposed. It will not hurt to understand "speaker judgments" in the former way throughout.

[30] One must not confuse the assumption of respect, understood in the former way, with the thesis that a principle of charity is true, at least not under common ways of understanding principles of charity. The assumption of respect says only that *for one reason or other*, semantic values are such as to make speaker judgments true. Charity principles say that it is a kind of *fundamental methodological constraint* on interpretation that the interpreter should interpret the interpretees as having mostly true beliefs. Such principles would thus provide one particular *explanation* of why speaker judgments should come out true. (See Warmbrod (1991).)

not be expected that there always will be a unique best way to maximize the correctness of these principles.[31] This state of affairs will occupy center stage later. Let an *acceptable assignment* of semantic values to expressions of a language L be an assignment that succeeds maximally well in rendering correct the meaning-constitutive principles. L being an inconsistent language is one reason for there being many acceptable assignments of semantic values to expressions of L; there is no reason, however, to believe it to be the only reason.[32]

To recap, one problem regarding modifying a conceptual-role theory so as to allow that inferences that are not truth-preserving can be constitutive of meaning, was that it was not clear how semantic values are determined from meaning-constitutive principles. I have now shown how this problem can be overcome.

Armed with our account of how semantic values are determined by meaning-constitutive principles, we can now also justify the claim, made earlier, that a broadly Fregean account of meaning can accommodate the possibility of inconsistent senses – in a Fregean sense of the word 'sense'. Let the sense of an expression be the principles that competence with the expression disposes a speaker to accept. Then the semantic values of the expressions of a language are determined by the senses of the expressions of the language in that the semantic values are whatever comes *closest* to satisfying the conditions laid down by the senses.

Just as in the case of a conceptual-role account, nothing in the rationale for the Fregean account of semantic competence rules out the modification suggested. For the Fregean, the sense of an expression is a set of conditions that determines reference. A competent speaker trying to find out what the referent of a particular expression is, tries to find out what satisfies the conditions specified by the sense. But this still holds true, if the Fregean

[31] Above I said I would set aside issues of how semantic values depend on features of the speakers' environment. It should be noted, however, that attention to the role of the environment helps resolve a problem that can be presumed to arise. It is evident that the semantic values of expressions of inconsistent language-fragments will be highly indeterminate, if semantic values are determined as outlined. This indeterminacy can be somewhat limited by the input of the environment.

[32] Talk of "acceptable assignments" is familiar from supervaluationist analyses. But note that the acceptable assignments considered here are quite different from the acceptable assignments - SV-assignments, let us call them - the supervaluationist talks about. For the supervaluationist, the reason there are many SV-assignments for a natural language is that the meanings of some expressions are *incomplete*: they can be *extended* without being *changed*, as Kit Fine puts it (1975, p. 267). The SV-assignments correspond to all the possible completions of meanings of expressions of the language. They are not, as the acceptable assignments discussed here, meant to be as faithful as possible to the meanings expressions actually are endowed with. (As illustrated by, for example, the fact that although supervaluationists normally do not accept bivalence, all the particular SV-assignments are bivalent.) In chapter 3, there is a (slightly) more extensive comparison of the theses I argue for here and supervaluationism.

account is modified as outlined.[33]

I have considered the problem that it may seem a priori that it cannot be by virtue of semantic competence that you believe some things that are false. In light of the fact how easy it is to modify accounts of semantic competence, consistently with their respective rationales, to render them compatible with inconsistency induced by competence, this worry can be set aside.

1.7 Discipline and unsoundness

The discussion of how semantic values are determined also serves to dispose of two other problems for the thesis that the paradoxes exert pull. These two problems I shall call *discipline* and *unsoundness.*

What motivated the assumption that semantic competence can consist partly in dispositions to accept inferences that are not truth-preserving was that this would explain certain features of the sorites and the liar, and hence certain behavior of the expressions employed in this reasoning. But the expressions employed in these arguments normally occur outside of paradoxical settings and their use is there unproblematic: coherent and well-disciplined. It may seem that the account of the paradoxes that I have given manages to explain why the paradoxes exert pull only at the cost of rendering this fact inexplicable. If our semantic competence involves being disposed to accept jointly inconsistent claims, why do we not, as competent speakers, feel compelled to accept inconsistent verdicts also on sentences whose truth-values seem unproblematic? I call this the problem of *discipline.* Moreover, how do we explain the fact that the paradoxes are not after all sound arguments? If competence with 'true' and 'looks red' involves being disposed to accept principles that lead to contradiction, why are their semantic values not such as to make some contradictions true? This I call the problem of *unsoundness.*

The worries are not precise. It is not as if there is some apparently watertight argument to the effect that if competence with some language involves being disposed to accept jointly inconsistent principles, then this fragment of the language is not disciplined, or contradictions come out true. But still these worries arise quite naturally. Crispin Wright

[33] Cf. the so-called *cluster theory* of proper names, according to which sets of descriptions are associated with proper names, but the referent of a proper name need not satisfy all descriptions associated with it; it need only satisfy (e.g.) sufficiently many of them and more of them of them than any other object satisfies.

(1975) argues that no practice can be governed by inconsistent rules. It appears to be roughly the problem of discipline he has in mind.[34] Gupta and Belnap (1993) press a similar objection to Charles Chihara's view that "there are generally accepted conventions that give the meaning of 'true' and which are expressed by [the instances of the disquotation schema]",[35] with these conventions jointly with other conventions leading to contradiction. Gupta and Belnap's complaint is that "it does not have the resources to explain the meaning and use of ordinary everyday sentences containing the word 'true'".[36] In particular, they use Curry's paradox[37] to show that from the statements Chihara says are conventions giving the meaning of 'true' any statement whatsoever follows – even without the use of *ex falso quodlibet*.[38] But if that is the case, Gupta and Belnap claim, Chihara is unable to account for the fact that most of our use of the truth predicate is unproblematic.

The problems of discipline and unsoundness are both resolved given our account of how semantic values are determined. The problem of unsoundness is solved provided the paradoxes come out as unsound arguments under every acceptable assignment of semantic values. That is, it is solved provided some assignment that renders incorrect some premise or step in the argument leading up to the contradiction succeeds better in rendering correct speaker judgments than does any assignment that renders some contradictions true. But that hardly needs arguing for. The principle of non-contradiction is so well-entrenched that whenever there is a perceived clash between it and some other principle, the other

[34] Wright (1975), e.g. p. 362f. See also Wright (1987), p. 235f.

[35] Chihara (1979), p. 611.

[36] Gupta and Belnap (1993), p. 13f.

[37] Here is Curry's paradox. Consider the sentence

(1) If sentence (1) is true then God exists.

The instance of the disquotation schema for this sentence implies

(2) Sentence (1) is true if and only if (If sentence (1) is true then God exists).

Now suppose

(3) Sentence (1) is true.

By (2), we have

(4) If sentence (1) is true then God exists.

By modus ponens, we can conclude that God exists. Since from the supposition (3), we can conclude that God exists, we have the conditional

(5) If sentence (1) is true then God exists.

By (2) again, we can conclude that sentence (1) is true, and by modus ponens on this conclusion and sentence (1) itself, we can conclude that God exists.

[38] Gupta and Belnap (1993), p. 14f.

principle will have to go.[39] As for discipline, note that a great many sentences, presumably all sentences we ordinarily use, will receive the same truth-value under every acceptable assignment of semantic values. But then, supposing that speakers aim to speak the truth and have an implicit grasp of how the correctness of the principles constitutive of meaning is maximized, discipline does not present a problem.

1.8 Further positive arguments

Having outlined how semantic values can be determined by meaning-constitutive principles if some unsound arguments can exert pull, we can now adduce further positive considerations in favor of the thesis that the sorites and the liar exert pull.

First, recall the Tarskian idea that the meaning of 'true' somehow "requires" that the disquotation schema be valid. I proposed that we can respect this idea if we assume that competence with 'true' involves being disposed to accept the instances of the disquotation schema as true. We can now say more. The disquotation schema also constrains the semantic value of 'true' in that although the disquotation schema is not valid, this schema should come out as close to valid as possible, given other constraints. The same goes for the idea that the meanings of vague predicates require that they be tolerant.

Second, it has proved notoriously difficult to give an account of the semantic values of the expressions centrally employed in the sorites and the liar. It has proved so difficult that it even appears that there are principled reasons why no fully satisfactory account of these semantic values can be given. Thus, it is not uncommon for theorists writing on the liar and the sorites to declare that any acceptable account of the truth conditions of the sentences employed in the statement of these paradoxes must be to some extent stipulative. This is decidedly *odd*. For these expressions do have semantic values antecedent to the stipulations, and the semantic values are consistent, in the sense that no contradictions are made true by them.

But this state of affairs is readily explained given the account of the determination of semantic values I have outlined. The judgment that no satisfactory account of semantic

[39] More would need to be said if the immediate opponent was someone who took seriously the possibility of true contradictions. But the objection from unsoundness presupposes that no contradiction can be true, and asserts that the present account fails to respect this fact. (See chapter 5 for criticism of the view that there are true contradictions.)

values is to be had is explained by appeal to the fact that no account can make all meaning-constitutive principles correct.

Third, and last, consider jointly inconsistent definitions and stipulations. Imagine a hypothetical language that is like English except that the logical particles have been introduced by the stipulation that they are to satisfy the proof-theoretic rules that characterize the classical logical operations and the truth predicate is introduced by the stipulation that the disquotation schema is to come out valid. It is easy to see that the stipulations *fail*, in the sense that there can be no expressions satisfying the conditions laid down. But the expressions introduced by means of these stipulations would not or need not be meaningless. In fact, the hypothetical language could work very much like English. Moreover, the stipulations, though in a sense failing, determine the meanings of the expressions introduced.

When a set of stipulations is inconsistent and it becomes generally known that they are, the stipulations will perhaps no longer be accorded the status of being meaning-giving. But imagine a case where it either is not discovered that the stipulations are inconsistent, or where the stipulations, though generally known to be inconsistent, still are treated as meaning-giving. The expressions introduced can still be used meaningfully, and what meanings they have is determined, at least in part, by the stipulations.

In so far as we are inclined to doubt that the stipulations fail while the expressions they introduce are still meaningful and the meanings of the expressions are determined by the stipulations, that is because we do not see how these claims could be jointly true. But suppose we accept the suggestion that a disposition to accept untrue and invalid sentences and inferences can be constitutive of semantic competence, and that the semantic values of the expressions of a language are such as to make the competence dispositions as nearly correct as possible. Then we can say that the inconsistent stipulations determine the meanings of the expressions introduced in that (a) dispositions to accept these stipulations are constitutive of semantic competence, and (b) the semantic values of the expressions introduced are constrained by the stipulations in that the stipulations should come out as nearly true as possible.[40]

[40] My proposal regarding the semantic values of expressions of inconsistent language-fragments is rather similar to David Lewis' (1970) proposal regarding the denotations of theoretical terms used in false theories. (For a closer comparison with ideas of Lewis's, see chapter 3.)

Lewis' basic proposal regarding the denotations of theoretical terms is, to put it very roughly, that they denote whatever satisfies the formulae that result when variables are substituted for the theoretical terms of the theory. This immediately leads to the problem: what about theoretical terms in false theories? Lewis'

1.9 A (partial) defense of classical logic and bivalence

In this last section I shall, focusing on the sorites paradox, show that the view I have argued for by no means entails that classical logic and semantics, i.e., classical logic plus the principle of bivalence, should be abandoned. My view that the sorites paradox arises because natural language is inconsistent is not only compatible with adherence to classical logic and semantics in face of the sorites, but perhaps even serves to justify such adherence.

Of course, there are many other problems for a proponent of classical logic and bivalence – for example the liar paradox. The reason the suggestion concerning the sorites paradox still is of interest is that it shows that accepting the radical view on the sorites presented here does not provide any reason to abandon classical logic and bivalence; but rather the contrary.

So consider the thesis that classical logic and bivalence hold valid in spite of the sorites. I shall focus on the claim that bivalence holds, as there seems to be no good reason to hold that bivalence but not the law of excluded middle is valid in the face of the sorites. Throughout the discussion I shall assume that there is a presumption in favor of classical logic and bivalence, and show that the sorites paradox should not make us abandon that presumption.

Upholding the principle of bivalence in face of the sorites faces two main problems. First, the thesis implies, for example, that out of two color patches that cannot be distinguished by mere observation it is true that one looks red and false that the other looks red. But this,

answer is *if* there is a true theory ***sufficiently similar*** to the false theory and in which these theoretical terms occur, the theoretical terms denote whatever makes that true theory true. If there is no such true theory, the theoretical terms do not denote anything.

Just as, according to Lewis, theoretical terms in false theories very 'far' from the truth fail to denote anything, we can say that if the principles constitutive of the meaning of a given expression are very far from being true and valid, respectively, then the expression has no semantic value. An interesting example is Prior's (1960) connective 'tonk', introduced by means of the stipulation that it satisfy the following rules of inference:

> From p, infer p *tonk* q;
> From p *tonk* q, infer q.

It is evident that if the language into which 'tonk' is introduced is non-trivial, in that neither all nor no sentences are assertible, it cannot be that both these rules of inference are valid. There is accordingly a clear sense in which the stipulations fail, if 'tonk' is conceived as introduced into a natural language. It has been much discussed in the literature just what to say about 'tonk' and what the example of 'tonk' tells us about meaning. On the present view, the answer is clear. The two rules of inference are constitutive of the meaning of 'tonk' even though they cannot both be valid. But 'tonk' has no semantic value. No assignment of a semantic value to 'tonk' comes even remotely close to satisfying the principles that are constitutive of the meaning of the connective.

many people think, must be wrong: our semantics must respect the fact that the transition between the red-looking patches and the non-red-looking patches is, somehow, *smoother*. Accordingly, many theorists have proposed introducing non-classical truth-values, intermediate between truth and falsity, and truth-value gaps. Second, if it is true that patch k looks red and false that patch $k + 1$ does, then, many have argued, something about the use of the expression 'looks red' (or *whatever* determines the meaning of 'looks red') must determine that the line between the patches that look red and those that do not is exactly there. But it has seemed utterly implausible that our use of this expression does determine that.

If the present account of meaning and semantic values is correct, then these two objections can be satisfactorily met.

The *typical* response to the first problem for someone who wishes to defend bivalence is to emphasize that the same problem that arises for her arises in different forms also for everyone who resolves it by introducing truth-value gaps and/or intermediate truth-values. Even if we introduce truth-value gaps or extra truth-values, it remains that for some color patches k and $k + 1$, "patch k looks red" will be true and "patch $k + 1$ looks red" will be either 'gappy' or have some truth-value other than truth. But then we are faced with the same problem that faces the proponent of bivalence: it appears that non-perceptible differences in color ought not make a difference to whether it is true or (in one way or other) *untrue* that a particular patch looks red, any more than such differences in color ought to make a difference as to whether it is true or *false* that a particular patch looks red. Similarly if we introduce further intermediate truth-values.

As it stands, this response is only moderately convincing. Showing that the same problem that arises for you arises also for someone who introduced truth-value gaps and intermediate truth-values is not sufficient to show that you can disregard the problem. It could be that some more radical revision of the traditional view is necessary. Perhaps vague predicates do not apply to anything, or lack semantic values altogether, or perhaps the idea that a predicate partitions the domain into classes must be abandoned for vague predicates.[41] But given the present account of semantic competence the response can be bolstered. We know that *every* assignment of semantic values to vague expressions will fail to respect at

[41] Peter Unger (1979) has defended the first view, Mark Sainsbury (1996) the third. Frege appears to be a proponent of the second view.

least *some* principle constitutive of meaning. The response on behalf of bivalence can then be recast as saying that it is not only the defender of bivalence who fails to render true some meaning-constitutive principle; and since her account fares better than alternative accounts in other respects (since there is a presumption in favor of classical logic and bivalence), her account, the response goes, is the best.

As regards the second objection, the proponent of bivalence can say that it simply misses the target. This objection is relevant only if truth is identified as truth under all acceptable assignments: if the predicate 'true' (under each assignment) applies to exactly the sentences that are true under all assignments. But if instead one holds that meaning-constitutive principles are best respected if truth, under each acceptable assignment, satisfies the disquotation schema (liar-type exceptions aside), then the principle of bivalence does come out valid under every acceptable assignment after all. Given the present apparatus of acceptable assignments, one should distinguish between two levels of indeterminacy. Let us say that a sentence is *first-level indeterminate* if it comes out lacking a classical truth-value under one or more acceptable assignments, and *second-level indeterminate* if it comes out having different truth-values under different assignments. The present point is then that second-level indeterminacy is compatible with the principle of bivalence coming out valid under each acceptable assignment.[42]

The general point here is familiar from discussions of supervaluationism. Some theorists have combined acceptance of supervaluationism with taking the truth predicate to satisfy the disquotation schema under each acceptable assignment: and thus they have upheld bivalence (since bivalence follows from classical logic plus the disquotation schema).[43]

The reason the point still is worth stressing, is that taking this line has been associated with embracing a "deflationary" conception of truth, according to which there is nothing more to this notion than what is stated by (the instances of) the disquotation schema (or some variant thereof).[44] Though nothing I say rules out such a conception of truth, there is nothing that forces it upon us.[45] All that is required for truth to be bivalent is that when the

[42] The distinction between first-level and second-level indeterminacy is discussed at greater length in chapter 3.

[43] I should again warn that the supervaluationists' assignments are different creatures from mine; see footnote 32.

[44] See e.g. Horwich (1990) and McGee & McLaughlin (1995).

[45] Actually, some of what I have argued in this chapter can be of help for the deflationist. The liar obviously poses a problem for the deflationist, as the liar apparently shows that the disquotation schema is not valid. But if my view on semantic competence is accepted, the deflationist can still say that our competence with

semantic value of 'true' cannot both be such as for 'true' to satisfy the disquotation schema and for 'true' to be coextensive with 'true under all acceptable assignments', the meaning-constitutive principles for 'true' come out more nearly correct if the semantic value of this predicate is such that it satisfies the disquotation schema and the principle of bivalence.

It is by no means obvious that truth is not to be identified with truth under all acceptable assignments. Perhaps it should so be identified, in which case bivalence is not valid after all. The reason for emphasizing the possibility that truth satisfies the disquotation schema under all acceptable assignments is to illustrate that from the view defended here it by no means follows that classical logical and semantic principles must be abandoned, but rather the contrary. (And whether the suggested defense of bivalence works or not, it still seems to be the most plausible defense of bivalence.)

Theorists who have argued that the sorites paradox arises because the expressions involved are – in some way or other – inconsistent have typically taken their accounts to rule out classical logic and bivalence. For example, Dummett says that "there can be no coherent...logic" of vague expressions and Horgan opts for a particular version of three-valued logic.[46] An apparent exception is Quine (1981), who both holds that the sorites paradox shows that the expressions employed are incoherent and opts for classical logic and bivalence. But Quine's position is that vagueness, being incoherent, is a deficiency of natural language, and that natural language hence must be *replaced* by a precise language in which bivalence and the laws of classical logic are valid.

On the present view, there is in the following sense no coherent logic of vague expressions: every consistent account of truth-conditions of sentences of vague discourse is bound to falsify at least one principle speakers are disposed to accept by virtue of their competence with vague expressions. *Both* classical logic and all varieties of non-classical logic are thus ruled out. But it is still very misleading to say, with Dummett, that there can be no coherent logic of vague expressions. For sentences with vague terms have consistent truth conditions, i.e. truth conditions that are never both satisfied and not satisfied. And there can be some logical and semantic principles that are valid under every acceptable assignment of semantic values. I have argued that for all that the problem of vagueness shows, it may well be that

the notion of truth is exhausted by our being disposed to accept the disquotation schema. This is different from her original view, but in the same spirit.

[46] See Dummett (1975), p. 320 and Horgan (1995), pp. 112ff.

the laws of classical logic and the principle of bivalence are valid.

1.10 Concluding remarks

I have argued that the sorites and liar exert pull: that it is by virtue of our semantic competence that we are disposed to accept the culprits in these arguments. The general thesis is that we can accept untrue and invalid sentences and principles by virtue of our semantic competence.

This general thesis can provide the key also to other, entirely different philosophical problems. Many philosophical problems can be presented in the form of paradoxes: we have a set of individually plausible claims that jointly have unacceptable consequences. Mark Johnston has shown how to view the problem of personal identity in this way, and Michael Rea has shown how a whole class of problems about essence and identity over time can be viewed in this way.[47]

The present account of semantic competence suggests a possible resolution of such problems: like the paradoxes we have considered, these problems might arise because meaning-constitutive principles are jointly inconsistent. That would explain the difficulties in finding intuitively reasonable solutions to them.

In the next chapter, I shall consider the so-called fission problem of personal identity, and apply the present account of semantic competence to that problem.

[47] Johnston (1989) and Rea (1995).

Chapter 2

Personal Identity and Conceptual Incoherence

2.1 Introduction

The fission problem shows that natural and intuitive versions of the psychological and the physical criterion of personal identity are incoherent. It is standardly assumed that we therefore must abandon (these versions of) these criteria. I shall argue for a different conclusion. I shall argue that it is our concept of personal identity that itself breaks down in fission cases. Then, insofar as a proposed criterion of personal identity gives incoherent verdicts about fission cases, it merely reflects a feature of the concept, and this is of course a point in favor of the criterion.

My investigation into these problems will serve also another purpose. Discussion of the problem of personal identity often revolves around intuitions about highly counterfactual scenarios. The fission cases are cases in point. This methodology is sometimes questioned. It is argued that we do not have intuitions about the application of our concepts in scenarios that are so wildly counterfactual. And sometimes the incoherence of our intuitions about these cases is used by those skeptical of the methodology to show how unreliable the method of considering highly counterfactual is.[1] Of course, this criticism of the debate over personal identity is echoed in many other areas of analytic philosophy. Debates over causation,

[1] Some theorists about personal identity who have voiced skepticism about the method of considering wildly counterfactual cases are Carol Rovane (1993) and (1998) and Kathleen Wilkes (1988).

reference and knowledge come to mind.

My conclusions about the concept of personal identity can be used in support of the method of considering counterfactual scenarios. The critics are right, I argue, that our intuitions about some of these counterfactual scenarios are incoherent. But that does not entail that our intuitions about these scenarios fail to reflect genuine features of the concept.

2.2 Problems concerning personal identity

In this section, I will go through the fission problem for personal identity. I shall show that the fission cases demonstrate that intuitively reasonable versions of the psychological and the physical criterion lead to absurdity in certain cases. I will go through all the argumentative moves fairly quickly, as the novelty of my approach lies not in the presentation of the fission problem, but in the consequences I will go on to draw from this problem.

In a recent paper (1997), Judith Thomson characterizes the psychological and the physical criterion roughly as follows:

> (1) person x at t = person y at t' if and only if y at t' is psychologically continuous with x at t,

and

> (2) person x = person y if and only if x's body = y's body.[2]

The psychological criterion as stated appears to be incoherent as a criterion of identity. For there appear to be possible cases where two distinct people at a later time are both psychologically continuous with one and the same person at an earlier time. If the psychological criterion is right, there are then two distinct persons identical to one and the same person.

[2] Thomson (1997), pp. 204 and 208. I have revised Thomson's formulation of the psychological criterion slightly. The present formulation of the psychological criterion differs from Thomson's in its apparent commitment to person-stages: a commitment Thomson seeks to avoid. I see problems, irrelevant to present purposes, with Thomson's attempt to avoid person-stages, and though I think such commitment can be avoided, this brings unnecessary complications.

This is Thomson's criterion of psychological identity:

> person x = person y if and only if there are times t and t' such that y at t' is psychologically connected with x at t (p. 208).

(In the main text I use 'continuous' rather than Thomson's 'connected'.) The problem is that this criterion – as opposed to the psychological criterion as more commonly stated – does *not* violate the transitivity of identity: rather it unequivocally yields the different absurd result that in (what intuitively are) fission cases, the resulting persons are really identical. (Just let t and t' be two times *before* the fission.)

But that violates the transitivity of identity. Let us refer to this kind of problem case for the psychological criterion as *psychological splitting.*[3]

Thomson rejects the psychological criterion on the basis of considerations of this kind, and opts for the physical criterion.[4] But, one may ask, can one not construct equally problematic cases for the physical criterion, cases of *bodily splitting*? It may seem not. For the physical criterion of personal identity says that personal identity is bodily *identity*, and trivially no identity relation fails to be transitive.

But suppose we were to seek a criterion of *bodily* identity. Would not similar problems arise? Suppose we presented a criterion of bodily identity of the following form:

(3) body x at t = body y at t' if and only if y at t$'$ bears relation Q to x at t.[5]

Then suppose a living human body is struck by lightning and splits down the middle. Half the body is destroyed, the other body half – somehow – lives on. This scenario is fantastic, but, importantly, neither conceptually nor metaphysically impossible. It seems to me clear that the body struck by lightning lives on: the half that lives on after lightning has struck is the same body as the body that existed before the lightning. (I will present considerations in support of this contention shortly.) Hence, it is a constraint on relation Q that Q obtains between the body after lightning strikes and the body before lightning strikes: human bodies can survive losing half of what constitutes them.

But then of course we run into problems analogous to those that arose for the psychological criterion of personal identity. For suppose lightning strikes a living human body, the body splits vertically in two, and both halves live on, separated. Then after lightning strikes there are two distinct bodies. Call them NewBody1 and NewBody2. Call the body existing before lightning struck OldBody. Then, given the noted constraint on Q, both NewBody1 and NewBody2 are by (3) identical to OldBody. But NewBody1 and NewBody2 are distinct.

I passed over rather quickly the crucial claim that a living human body indeed can survive the destruction of half of what constitutes it. Is it really obvious that a body survives being struck by lightning, when half the body is destroyed? The matter is in fact

[3] Cases of psychological splitting, and cases of bodily splitting like the ones I shall go on to discuss, have been widely discussed in the literature. See e.g. Wiggins (1967), p. 50, and Parfit (1984), p. 254ff.

[4] See Thomson (1997), pp. 221ff. Also Bernard Williams (1970, p. 179f) has used the fission problem specifically against the psychological criterion.

[5] I use 'Q' because 'R' in the context of discussions of personal identity is appropriated by Parfit.

rather complicated, due to the fact that whether a material object survives the loss of some of what constitutes it depends not only on *how much* it loses, but also on how what it loses is distributed. A shoe perhaps does not survive losing, say, the left half of what constitutes it; but that does not mean that it could not survive losing half of what constitutes it, if what it lost were differently distributed.[6] To simplify matters, let us consider only material objects losing the right or the left half of what constitutes them; that is, consider only what I shall call *vertical semi-destruction.*

A reason for considering vertical semi-destruction is that we are interested in the fission cases, and the possibility of persons' surviving vertical semi-destruction is more relevant to the problem of fission than their surviving other forms of semi-destruction would be. For survival of vertical semi-destruction more clearly suggests the possibility of fission.

It can be argued that not all material objects survive vertical semi-destruction. I have already mentioned shoes. One may think that since a vertical half of a shoe cannot perform the function typical of shoes (it cannot be used as footwear), it is not a shoe. Cans and cups are additional examples of the same kind. And does a *lump of clay* survive vertical semi-destruction? I am really not sure. It may depend partly on the *manner* in which it is vertically semi-destroyed. If it happens gradually, then I think I would be inclined to say that the lump of clay does survive it. If it happens instantly (e.g. lightning strikes) I am at least somewhat more uncertain.[7]

Now suppose a living thing, for example a tree, is vertically semi-destroyed. Does the tree continue to exist? This seems clear: we would say that *the tree survived* being struck by lightning. Similarly, I believe we would say that the human body survived being struck by lightning, or that the human organism survived being struck by lightning.

We rather seldom talk about human bodies or human organisms as opposed to *people.* But we would certainly say of a *person* that *she* survived being struck by lightning in a case such as the one envisaged. And then the proponent of a physical criterion of personal identity had better *hope* that the same goes for bodies.[8]

[6] Suppose the shoe has a very thick sole, and most or all of the matter it loses is from the sole.

[7] The claims here are tentative. Very probably, matters are still more complicated. (Is it that obvious that a vertical half of a shoe cannot perform the function typical of shoes? It would do it very poorly, but that is another matter. And even if a vertical half of a shoe cannot perform the function typical of shoes, is it so clear that it is not a shoe? Is being able to perform this function at t a necessary condition for being a shoe at t?)

My aim in this paragraph in the main text is only to give some indication of the various difficulties.

[8] Also in the case of living organisms it may matter *how* the vertical semi-destruction occurs. I do not

In both the case of trees and the case of human bodies, there is a tacit assumption which should be made explicit. It is that trees and human bodies survive vertical semi-destruction when *the organism* lives on; the life continues. It is because a biological life can continue through vertical semi-destruction and because we take the continuation of the biological life to be sufficient for the continued existence of a tree or a human body that we take it to be possible for trees and human bodies to survive vertical semi-destruction.[9]

I have avoided stating the criterion of bodily identity in terms of physical continuity, for I want to avoid giving the impression that anything in my argument depends on taking physical continuity to be the criterion of bodily identity. However, it is natural to take physical continuity to be the criterion of bodily identity, and I see no harm in making this natural assumption. Accordingly I shall henceforth take physical continuity to be the criterion of bodily identity (or be the criterion of bodily identity, *present problems aside*). This is in order to have a simple way to talk simultaneously about the physical and the psychological criterion. Skeptics can take "physical continuity" as a place-holder for whatever should go in its stead.

The proponent of the physical criterion is likely to point out that just as the manner of the semi-destruction and the distribution of the parts that are destroyed can be relevant to whether some physical object survives vertical semi-destruction, what happens to the parts lost in the semi-destruction can be relevant. For example, whether some physical object of a particular kind (a tree, a body, etc.) survives vertical semi-destruction can depend on whether the half lost is simply destroyed or forms another object of this particular kind. Thus, whether NewBody1 is or is not the same body as OldBody can depend on whether the part or parts that OldBody lost when lightning struck still form a body or not.

The proponent of the physical criterion can thus present something like the following criterion of bodily (and hence, by extension, personal) identity:

(4) body x at t = body y at t' if and only if y at t' is physically continuous with

know how amoebas split; but let us suppose for the sake of the argument that the splitting amoebas can be regarded as cases of vertical semi-destruction. There is virtually no temptation to say that amoebas survive the split - even should one half immediately be destroyed and only one half survive. It seems to me that this is because amoebas split as a matter of reproduction. Consider the unfortunate amoeba Alfred, who is vertically semi-destroyed not in the course of normal reproduction but because of some accident: for example, lightning strikes. In this case, I am inclined to say that Alfred survives.

[9] It is not necessary, of course, for the continued existence of a human body that the biological life continues. We do after all speak of "dead bodies".

x at t and no other body at t' is physically continuous with x at t.[10]

We should note that if (4), or a criterion like it, is acceptable as a criterion of bodily identity, it is not clear why a similarly reformulated psychological criterion of personal identity cannot be made to work. That is, if (4) is acceptable as a criterion of bodily identity, we have yet to see what is wrong, if anything, with the following as a criterion of personal identity:

> (5) person x at t = person y at t' if and only if y at t' is psychologically continuous with x at t and no other person at t' is psychologically continuous with x at t.

(A potential problem with both (4) and (5), for someone who wishes to give a genuine, complete analysis of bodily or personal identity, is their mention of "other bodies" and "other persons". One response to this problem is to say that we are only concerned with the problem of diachronic identity and can presuppose a criterion of synchronic identity. In the present context, I can also say that I am concerned not with any kind of reductive analysis of personal identity but only with purported truths about personal identity. The criteria (4) and (5) clearly are not uninformative despite their failures qua reductive analyses of personal identity. In fact, at least one of them is false.)

Moreover, the criteria (4) and (5) both have an immediate consequence that is rather counterintuitive. It seems intuitively clear that whether person (or, for that matter, body) x existing at t is identical to person (body) y existing at t' ought not to be a matter of what *other* persons (bodies) exist at t'. Let us call this the presumption of *locality.* Criteria (4) and (5) both violate locality, for obvious reasons.[11]

The presumption of locality has been challenged prominently by Robert Nozick. Nozick asks us to compare the following two hypothetical scenarios.

> (6) When the Vienna Circle had to leave Austria, the only members who remained active were two or three members who settled in Istanbul and continued to meet there and called themselves the Vienna Circle.

[10] A version of the same general idea has it that when the second condition of (4) is not satisfied, it is *indeterminate*, and not false, whether body x at t = body y at t'. (See Rieber (1998).) I shall not discuss this possibility separately: the same remarks that apply to (4) would apply to it thus reformulated.

[11] Given the way I have characterized (4) and (5), it may seem trivial that these criteria are ruled out. They both appear to fly in the face of the necessity of identity. The criteria could be formulated more carefully. I shall, however, stick with the present, more intuitive formulations of them. For I shall not bring up the problem of necessity of identity. For discussion, see Shoemaker (1984), pp. 115ff, Sosa (1990), p. 301f, and Johnston (1989), pp. 379ff.

and

> (7) When the Vienna Circle had to leave Austria, two or three members settled in Istanbul and continued to meet there and continued to call themselves the Vienna Circle. However, unbeknownst to them, the majority of the Vienna Circle moved to the USA, remained active there and went on to call themselves the Vienna Circle.

Nozick suggests that, intuitively, in the first case, (6), the group in Istanbul is the Vienna Circle and in the second case, (7), the group in the USA is the Vienna Circle. But then the presumption of locality is false as regards groups (or whatever kind of object the Vienna Circle is). By analogy, Nozick reasons, it may well be abandoned also as it applies to persons.[12]

But criteria formulated to respect the intuition Nozick appeals to – criteria such as (4) and (5)[13] – have rather counterintuitive consequences. Let me just mention two.

First, suppose that at time t I am struck by lightning and split in half. The doctors try to save the lives of the now separated halves, Lefty and Righty as we may call them. Lefty and Righty are equally continuous with me, in whatever dimension or dimensions are relevant. The doctors manage to save Lefty. They manage to keep Righty alive for a few trembling moments, time t' being one of them, but then Righty dies. At time t'' only Lefty has survived. I believe we would say that Lefty is me. But by (4) and (5), Lefty is alive at a time at which I am not – that is, when Righty is alive – so Lefty cannot be me.[14]

Second, consider the following case, from Johnston (1987). Suppose we have a *brain-state transfer machine* that can "continue the mentality of a given patient A in each of the

[12] See Nozick (1981), pp. 29-70. On the basis of his arguments, Nozick claims that person x at t is identical to person y at t' just in case y at t' is the closest of the sufficiently close continuers of x at t, where both physical and psychological continuity matters. (This is an instance of Nozick's "closest-continuer schema".) Nozick's preferred criterion of personal identity is thus different from both (4) and (5). However, for present purposes Nozick's criterion and criteria (4) and (5) can be grouped together. For Nozick's argument extends to (4) and (5); and the problems I shall discuss likewise concern all three criteria.

[13] Or variants such as those mentioned in footnote 10.

[14] I discuss cases of people split by lightning because I am concerned with physical criteria where it is the continued existence of the body rather than the brain that matters. If instead we were to consider physical criteria of personal identity where it is the continued existence of the brain that matters we could consider the more familiar fission case where one of A's hemispheres goes into the body of one of his brothers and the other goes into another brother of his. Such a case is sometimes contrasted with the *failure case*, where one of the hemispheres is dropped on the floor. The failure case is analogous to that of when only one of my body-halves survives being struck by lightning. Such failure cases actually constitute counterexamples to Nozick's account: for we may assume that there is a time at which the hemispheres are separated while both are functioning.

bodies of two patients B and C", and that

> ...the original A-body dies as a result of the machine's operation, and, after the machine operates, the person associated with the B-body and the person associated with the C-body are each sufficiently close psychological continuers of A. Suppose also, however, that typically the person associated with the C-body is a considerably better psychological continuer of A than the person associated with the B-body. Finally, suppose that this extra continuity is more than enough to compensate for the extra ten-minute delay involved in "reading" A's psychology into the C-body.[15]

So much for set-up of the experiment. Now for the problem for Nozick's theory.

> Now imagine that the machine has been operating and A's psychology has been..."read into" the B-body. The person associated with the B-body gets up and walks around and thinks to himself, "I am A. I did not just come into existence." He sees the machine beginning to "read" A's psychology into the C-body. He knows that if the process is allowed to continue the result will be that the person associated with the C-body will be a better continuer of A than he is. So he turns off the machine.[16]

If Nozick's account is right, then the action the person associated with the B-body performs will decide whether he or the person which might come to be associated with the C-body is A. But, as Johnston remarks, "...surely our intuition is that the B-body person's thought "I am A. I did not just come into existence," is made true or false by what has happened *up to and including* the time at which that thought occurs. Surely no *subsequent* act by the B-body person can make this thought true or false".[17]

[15] Johnston (1987), p. 67f.

[16] Johnston (1987), p. 68. As Johnston's thought experiment is set up (and as corresponding thought experiments in the literature are set up), it speaks to (5) but not to (4). But it is easy to see how to revise it so as to apply to (4). Thus, suppose we have a fission machine that works as follows. First it completely halts your life processes; say, it freezes you down. Then, while you are frozen, it splits you into two halves, and it infuses life first into the one and then into the other. As it happens, the person associated with the half that comes to life last is always much more physically continuous with the original person (it is for example less damaged by the process). Again, something like Johnston's thought experiment can be carried out.

[17] Johnston (1987), p. 68; my emphases. It again deserves emphasis that, as mentioned in footnote 11, there are no *logical* problems with taking Nozick's line. Taking this line does not, for example, violate the necessity of identity. The only problem is that, as made out in our two counterexamples to Nozick's view, it has *counterintuitive* consequences.

Nozick might respond that the cases just discussed are problematic for *everyone*, not only for him: it is hard to see what could possibly be a satisfactory account of these cases. For example, I take it that what we *want* to say, in the first case, is that Lefty is me after Righty has died, and consequently that Lefty is me also when Righty is alive. But suppose we give this account, and consider the state of affairs *before* Righty has died. What we are committed to saying is that it is because Righty *later* dies that Lefty at this earlier time is me: but that is bizarre. Our intuitions in this case are problematic on *all* accounts. And Nozick might accordingly say that this kind of case therefore cannot be used in an argument against *his* theory in particular.

I must put off a fuller discussion of this response on behalf of Nozick until later (I believe there is something to it). Here I will say only the following: to note that a problem that arises for you arises also for certain other views serves only to show that those holding these other views have no case against you. More specifically, it remains that criteria (4) and (5) violate some of our intuitions about personal identity.

Nozick is not the only theorist who has abandoned the presumption of locality. Certain other theorists, for example Derek Parfit, have also abandoned it.[18] My reason for considering only Nozick at this point will become apparent later, when we have the theoretical tools at hand for adequately stating a certain important difference between, on the one hand, Nozick's attitude toward locality and, on the other, Parfit's attitude toward it.

2.3 The problem generalized

If we accept my arguments up to this point, we find ourselves in a rather problematic situation. Not only must we reject all proposed versions of the psychological and the physical criterion as unsatisfactory, but it appears that *any* account of personal identity must be rejected for similar reasons. For suppose I am struck by lightning, and one of my vertical halves lives on and the surviving person is physically as well as psychologically continuous with me. In order to capture our concept of personal identity, an account of personal identity must respect the fact that the survivor is me. If, furthermore, an account of our concept of personal identity must respect locality, any account of personal identity is bound to run into absurdity.

[18] Parfit (1984), pp. 255-60, 267-70.

The psychological and the physical criterion face other problems, which I will not discuss here. (For example, one may think that for principled reasons, independent of the fission problem, no reductive criterion can hope to succeed.)

But the fission problem does not in fact concern only these criteria of personal identity. What the fission cases show is that intuitions we have about personal identity are inconsistent. We have the intuition that it is not the case that in fission, the original person survives as both of the resulting persons. But we also have the intuition that persons survive vertical semi-destruction, and we have the intuition of locality. In fact, the psychological criterion (1) and a physical criterion such as (3) conceived as an identity criterion for persons can be seen as *succeeding* in capturing our notion of personal identity just to the extent that these criteria run into trouble in just the way our intuitions about personal identity run into trouble.

It would be wrong to think that we have ruled out all possible ways of avoiding the threatening contradiction. For example, some theorists have suggested with regard to fission cases that, appearances to the contrary, two persons existed also before the fission, and it has been suggested that, appearances to the contrary, there is only one person around also after the fission.[19] Neither suggestion is ruled out by anything I have argued so far. I shall not argue against these claims head on.[20] Rather, I shall present an entirely different picture.

In the next section, I shall repeat the basic theses from chapter 1. In later sections I shall apply these theses to the fission problem.

[19] Lewis (1976) defends the former suggestion, and Parfit discusses the latter (without endorsing it). In Johnston (1989), there is a more careful discussion of fission cases than the one I have defended here. Johnston presents four (exceedingly plausible) principles that jointly entail that fission cases cannot be dealt with. The four principles are:

> (B*) If in one possible world w a process p secures the survival of a person x then in any world w' in which p occurs and is intrinsically exactly as it is in w, in that world w' p secures the survival of x.
>
> (C) A person never survives spatially separated from himself in the fashion of a universal.
>
> (D) Since persons are solid and not interpenetrable, no two persons can be in exactly the same place at exactly the same time.
>
> (E) At no time is a person constituted by two independently functioning human bodies.

[20] But see footnote 34.

2.4 Conceptual incoherence

Since the sorites paradox perhaps serves best to explain and motivate my semantic theses, we might well center the discussion around a particular version of this paradox.

Consider a series of color patches. The first looks red and the last looks orange (to normal observers in normal circumstances). Adjacent patches are ***observationally indistinguishable*** with respect to color, in the sense that someone with the normal visual capacities of humans cannot detect a difference in color between them by observation of these two patches alone.[21]

The stipulations made so far are clearly coherent. But now consider the meaning of 'looks red' and other broadly observational color predicates. It appears to be a feature of the meaning of 'looks red' that if two objects are observationally indistinguishable with respect to color, in the sense characterized, then if 'looks red' applies to the one it also applies to the other. In this sense, 'looks red' is observational.[22]

But now we can derive an absurd conclusion: (1) Patch 1 looks red. (2) If patch k looks red, then, since adjacent patches are observationally indistinguishable, patch $k + 1$ looks red. (3) Hence all the patches in the series look red.

There is (virtually) no doubt that the *culprit* – untrue premise or invalid step – in this argument is the second premise, commonly called the *sorites premise.* But because of the observationality of 'looks red', it appears that this premise would simply have to be true.

What is more, consider someone who fails to find this premise even initially attractive. Shc would be regarded as having failed to understand something about the meaning of 'looks red'.

To account for this, I proposed the following theses about semantic competence. Semantic competence essentially involves being disposed to accept certain sentences: or at any rate to be disposed to accept them so long as one does not, for example, possess what one takes to be evidence against their truth. These sentences we can call *constitutive of meaning* of the expressions involved. Now, what the sorites reasoning shows is that some sentences constitutive of meaning can fail to be true and even be jointly inconsistent. When the language-fragment we employ contains expressions for which the sentences constitutive

[21] Two patches x and y observationally indistinguishable in this sense may thus be distinguishable to the naked eye by other means: for example, there may be a third color patch z such that x but not y is observationally indistinguishable from z.

[22] 'Looks red' presumably has a wide range of different uses. What I say here may not be true about 'looks red' on *all* uses: it is sufficient, however, that it is correct about some uses of 'looks red'.

of meaning are jointly inconsistent we can be led into absurdity by our semantic competence, as exemplified by the sorites arguments. (Needless to say, my claim about the sorites paradox is not uncontroversial. But already from these brief remarks it should at least be clear why an account of this kind should be *attractive.*)

Let us say that an *expression* (language, language-fragment) is *inconsistent* just in case the sentences constitutive of the meaning of the expression (or the expressions of the language or language-fragment) are jointly inconsistent. *Conceptual incoherence* is to be understood analogously to the inconsistency of expressions.

When a language-fragment is inconsistent, what should we say about the semantic values – contributions to truth conditions – of the expressions of this language-fragment? My suggestion is that the semantic values are such as to make the sentences constitutive of meaning as close to correct as possible. Since it is unlikely that there will be a unique way of making the sentences constitutive of meaning as nearly correct as possible, there will likely be many equally acceptable assignments of semantic values. (Note that I do *not* say that when a language-fragment is inconsistent, the sentences of this language-fragment will fail to have truth-values, or will all be false, or that there will be some true contradictions in this language-fragment. On the contrary, those theses should seem quite unattractive given my view – as given all other reasonable views.)

It is of course difficult to say what closeness comes to. And it will presumably be a vague matter what comes closest to making meaning-constitutive sentences true. But all that is needed for my purposes is that somehow we do have a grasp of proximity of theories. Naive set theory is inconsistent, and nothing satisfies the claims this theory makes about 'sets', but the (class of) entities that satisfy the claims of Zermelo-Fraenkel set theory come closer to satisfying naive set theory than do, say, the (set of) natural numbers. And in physical science, there are of course cases of false theories satisfied by nothing, but such that the entities currently presumed to exist come sufficiently 'close' to satisfying the claims of these false theories that they reasonably can be said to be what these theories are about, and are the referents of the theoretical terms of these false theories. (Consider early versions of the atomic theory.)[23]

Having set out and motivated my theses about competence and reference-determination, I shall now apply them to the fission problem. I have not argued for these semantic theses

[23] I discuss this in more detail in chapter 3.

here.[24] But that they have nice consequences with regard to the fission problem in itself supports them.

2.5 The incoherence of the concept of personal identity

My suggestion regarding personal identity is that a criterion such as (1), or (3) reformulated as a criterion of personal rather than bodily identity, could be constitutive of the meaning of talk about personal identity, even though it states something that is not actually true of personal identity. The relation between these criteria and the notion of personal identity is then like the relation between the sorites premise and the expression 'looks red'.

Let P be some proposed candidate for being the relation of personal identity: psychological or physical continuity, for example. If the suggestion that some concepts are incoherent is right, then it may be that our semantic competence disposes us to accept that x at t is the same person as y at t' just in case y at t' bears P to x at t, even if, because of the problems considered, it cannot be true that x at t is the same person as y at t' just in case y at t' bears P to x at t. The theorists who have identified the relation of personal identity as P may accordingly have correctly identified the *meaning* of talk of personal identity, in the sense of having identified what semantic competence with discourse about personal identity involves, even if their analyses fail as accounts of the necessary and sufficient conditions for personal identity to obtain.

Again I should emphasize that there may be *other* reasons for being extremely doubtful about one or both of the psychological and the physical criterion. The fundamental point is that it is precisely by virtue of our competence with the concept of personal identity that we have the intuitions whose joint inconsistency is demonstrated by reflection on fission cases. This point is independent of the possibility of a reductive analysis of the concept.

It seems to be a kind of presupposition of our discourse and thought about personal identity that fission cases do not occur. Speaking loosely, when this presupposition does not hold, our discourse and thought about personal identity *breaks down*. I have already mentioned the broadly Wittgensteinian criticism of philosophy. The spirit of this criticism is echoed in Quine's famous remark, concerning the use of fantastic examples in the literature on personal identity, that "The method of science fiction has its uses in philosophy, but

[24] See chapter 1 for extensive arguments.

at points in the Shoemaker-Wiggins exchange and elsewhere I wonder whether the limits of the method are properly heeded. To seek what is "logically required" for sameness of person under unprecedented circumstances is to suggest that words have some logical force beyond what our past needs have invested them with".[25]

What my view about semantic competence can deliver is the means to make these rather intuitive considerations more precise. We can say that a proposition ϕ is a *meaning presupposition* underlying the use of a certain language-fragment just in case the sentences constitutive of the meanings of expressions of this language-fragment are jointly satisfiable only when ϕ is true. In cases where the meaning presupposition does not hold true, the belief-forming dispositions bound up with our semantic competence with discourse about 'persons' can lead us into contradiction: and discourse about 'persons' breaks down, as it were.

If the concept of personal identity is incoherent in the way outlined, then it should be clear how the problems we ran into earlier can be explained. Let P be some proposed candidate for being the relation of personal identity, either psychological or physical continuity. Suppose then that competence with the concept of personal identity, presumably in conjunction with contingent facts about the world[26], involves a disposition to accept that

(8) person x at t = person y at t' if and only if y at t' bears P to x at t.

If, as in our actual environment, there is always only one person y at t' such that y at t' bears P to x at t, everything runs smoothly, in the sense that (a) our judgments (about actual cases and relevant possibilities) are consistent, and (b) there is no problem in assigning truth conditions to statements about personal identity (when these statements are about actual cases and relevant possibilities): we can treat as true the statement that person x at t is the same person as person y at t' just in case P obtains between y at t' and x at t. But with respect to cases where it does happen that two distinct persons at t' both bear P to x at t, all judgments will conflict with some sentence constitutive of meaning and so to be counterintuitive in some way. A judgment not consistent with (8) trivially violates (8).

[25] Quine (1972), p. 490. Quine may legitimately be interpreted as occupying a position conflicting with mine. He may be interpreted as holding that our concept of personal identity does not have any view, so to speak, on the outlandish cases discussed in the literature. The same remarks apply to this position as apply to the position possibly defended by Johnston, discussed in the next section.

[26] See footnote 30.

A judgment made in accordance with (8) will violate the transitivity of identity.[27]

2.6 Johnston on the lessons of the fission problem

In (1989), Mark Johnston argues that our intuitions about the fission cases lead to inconsistency. Johnston identifies four principles (presented in footnote 19) that are all intuitively true but lead to contradiction. He says that their "restrictions to ordinary cases are very plausibly taken to be true", but that "[o]n the strength of our ordinary practice of reidentifying people and the fact that relative to this ordinary practice [these four principles] hold up, we do not have sufficient reason to subscribe to unrestricted versions of these principles". Johnston's conclusion is that the fission case is "a case of indeterminacy".[28]

It is unclear to me to what extent Johnston would agree with the view presented here. His claims seem to admit of an interpretation under which his view is compatible with and rather similar to the one presented here. But they encourage a different interpretation, according to which our concept of personal identity has views only on what is going on in relatively ordinary cases, and does not say anything about extraordinary cases, e.g. fission cases. Insofar as we have any intuitions about fission cases, we are merely mistakenly extrapolating from ordinary practice.

The reason it is hard to see which interpretation of Johnston is correct is that it is not clear what he means by "sufficient reason to subscribe to unrestricted versions of these principles" in the passage quoted above. If he means simply that we do not have sufficient reason to take unrestricted versions of these principles to be true, then what he says is

[27] The picture that emerges seems to be roughly what John Perry (1993) has in mind.

In response to fission cases (for psychological criteria of personal identity) Perry says, criticizing Bernard Williams, that Williams illicitly requires that the relation with which personal identity is identified should be an equivalence relation not merely as a matter of fact but also "as a matter of logical or metaphysical necessity" (p. 430). As opposed to Williams, he holds that the psychological criterion may get matters right so long as this relation is, in the actual world, an equivalence relation.

Perry's ideas are suggestive, but rather problematic. If personal identity is *not* an equivalence relation as a matter of *metaphysical* necessity, it is not an identity relation at all. And moreover, some statements about personal identity made with respect to possible worlds where the relation of psychological continuity is not an equivalence relation have truth values; and it is hard to see what, on Perry's analysis, these truth-values can be. Moreover, a consequence of Perry's analysis would seem to be that the correct analysis of a certain notion depends on facts about the actual world. But surely, we *could* have the same notion of personal identity as we now have, even if we were in a possible world where psychological continuity is not an equivalence relation. (As Perry notes, our actual notion of personal identity would be less useful if we employed it in such a world. But that does not mean that it is metaphysically impossible for us to have this notion in such a world.)

[28] Johnston (1989), p. 392f.

compatible with what I say. If he instead means that we do not even have sufficient reason to take unrestricted versions of these principles to be, in some way, part of the concept of personal identity, then he is in disagreement with me. I believe Johnston's view is the latter view. But however that may be, let me consider this latter view in its own right, and briefly explain why I think the view I have argued for is superior.

Our concepts are generally applicable to cases outside their normal or past range of application. It is hard to say why this is so in a non-theory-laden way, but when our concepts evolve they somehow do so in such a way as to have views on matters outside their normal or past range of application. The question then arises of exactly which matters a concept has views on. It is perhaps natural to assume, as does Johnston – or the possible Johnston I consider – that if the principles associated with a given concept yield incoherent conclusions when applied to certain cases outside the concept's normal range of application, then the concept does not have views on these cases, and the assumption that the principles can be applied there is illegitimate. But given the account I have sketched above, one *need* not make this (natural) assumption. And in light of the fact that, as Johnston seems to agree, we *do* have shared intuitions about these extraordinary cases, the view I have sketched seems preferable.

2.7 Intuitions

The word 'intuition', which could serve a useful function in philosophical theorizing, has unfortunately been used in many different ways. Without much effort, one can come up with four different senses in which it is not infrequently used: (1) naive, pretheoretical convictions; (2) beliefs quite generally; (3) what Reason (with capital 'R') delivers; (4) semantic intuitions: beliefs speakers have by virtue of their semantic competence (and for those of us who do not mind talking about concepts, there are conceptual intuitions, understood on the model of semantic intuitions).[29]

I shall adopt the convention of using 'intuition' in accordance with (1). 'Belief' seems a fine word for 'intuition' in the sense of (2). Intuitions in the sense of (3) can be called "rational intuitions" and intuitions in the sense of (4), both semantic and conceptual, I shall

[29] The word "intuition" is also ambiguous in the way "belief" is, in that sometimes it refers to a specific mental episode and sometimes it refers to the content of that episode. I shall not, however, worry about this extra ambiguity.

call "competence intuitions". All competence intuitions are intuitions; but not vice versa. The same goes for the relation between intuitions and rational intuitions.

The argument up to this point is easily and usefully restated in terms of intuitions and competence intuitions.

What the arguments presented early on showed was that our *intuitions* about personal identity are inconsistent. I have diagnosed this inconsistency as an inconsistency in our *competence intuitions.* This conclusion is not *forced* upon us by anything I have said. But it provides a good account of the resilience of the intuitions that are jointly inconsistent.

Ordinary intuitions of a non-linguistic nature, no matter how widely shared, would seem to be over-ridable by similarly non-linguistic evidence. But it is very hard to see what evidence, other than conflicting intuitions having similar status (whatever that status is), plus logic, could possibly defeat the intuitions about the persistence conditions of persons.[30]

The kind of account I have given of the problems about personal identity can be given also of certain problems about artifact identity, such as those illustrated by the ship of Theseus. We can say that it is a meaning presupposition underlying our talk of sameness of ships that it not happen that for some two ships S and S′ at t' and one ship O at t, S is causally continuous with O, having arisen from O by gradual replacement of planks, and S′ is at t' constituted by the same material as O was constituted by at t. And of course there is nothing special about ships: the same would go for other artifacts as well as various other kinds of objects.

[30] In the main text I tend to talk as though the persistence conditions of persons were determined exclusively by our concept of person, and as though our intuitions about the conditions under which persons persist are exclusively competence intuitions. However, at least given some views about personal identity, the persistence conditions of persons are determined at least in part by facts about the actual world. Consider the sentence "cats are robots". We know that cats are not robots, and can conclude that as a matter of metaphysical necessity, cats are not robots. But it could have turned out that cats are robots. Now, one may reasonably hold that given the actual facts,

cat x = cat y just in case x is the same organism as y;

but had cats turned out to be robots the persistence conditions of cats would obviously have been different. Similarly, humans could arguably have turned out to be robots, and this could have had analogous relevance to the persistence conditions of persons. If a physical criterion of personal identity is correct, then if persons are organisms, the persistence conditions are one way, whereas if persons were machines the persistence conditions would have been different. (Note that if a psychological criterion of personal identity is correct, the persistence conditions of persons do not seem to be similarly dependent on the actual facts.)

Claims about the persistence conditions of persons should be taken to be conditional upon (generally accepted) hypotheses about the actual facts.

In this connection I should also say a few words about the "soul view" on personal identity: the view that sameness of person is sameness of soul. I think it is reasonable to say about this view that if there were, in the actual world, souls associated with human bodies, the soul view might be true, but since there are no souls in the actual world, some other criterion of personal identity is correct.

Philosophers of a certain bent – presumably most philosophers concerned with identity over time – would respond to the problem of the conflicting intuitions about personal identity, and the corresponding problems about artifacts, by trying to find some *general metaphysical principle* given which at least one of the conflicting intuitions is non-veridical: a *reason* why *the world* is such as to make one rather than another of the conflicting intuitions non-veridical, independently of how firmly entrenched various intuitions of ours are.

Here is an example of metaphysical principles of the kind indicated, loosely adapted from the discussion in Lowe (1983):

> (1) An object can be a proper part only of one object (of a particular kind) at a time;[31]
>
> (2) Interrupted existence is impossible: once an object has ceased to exist it cannot resume its existence.

From the general principles (1) and (2), we can, Lowe says, derive a solution to the problem of the ship of Theseus. O at t is identical with S, not S′, at t'. Here is the argument, in brief. Principle (2) requires that for O to exist at all at t', it must exist at all times between t and t'. But consider a time soon after t, when only some parts of the original ship have been replaced. At that time, the only ship that exists is the one at sea. The removed parts are too few to constitute a scattered ship. If a scattered ship exists, some of its parts are also parts of the ship at sea at this time soon after t. But by principle (1), that is impossible.

It is not obvious that this account works, even if we were to accept principles (1) and (2). At any rate, more would need to be said in defense of it.[32] My aim here is just to use (1) and (2) and their employment to illustrate the strategy of appealing to general metaphysical principles to solve the puzzles of identity over time.

Nothing I have argued or claimed *rules out* the existence of metaphysical principles playing this role. But if the picture I have presented is correct, then: (a) Even if such principles exist, an account of personal identity or artifact identity which leaves out the fact of *conflicting competence intuitions* is clearly *incomplete*, leaving out important features of the

[31] Or perhaps a suitably restricted version of (1) is more plausible; something of the form

> For objects *of certain specified kinds*, an object can be a proper part only of one object of that kind at one and the same time.

[32] See Lowe (1983), for such further defense.

concepts concerned. (b) Much more importantly, the assumption that such general metaphysical principles must exist is *unfounded.* We can in principle give an acceptable account of the truth values and truth conditions of statements about identity over time without appeal to such principles. I believe there is a tacit assumption, among theorists concerned with the problems we have been discussing, as well as other metaphysical problems, that there *must* be some reason, of the kind described, for which intuition should be rejected as non-veridical. But on the account I have outlined, what determines which intuition should be rejected are rather considerations about what maximizes the correctness of sentences constitutive of meaning.

There are also more general problems concerning appeal to metaphysical principles of the kind we have considered. Suppose I am right that meaning-constitutive sentences can be untrue. Then consider the question: what status do the metaphysical principles appealed to in order to defeat the intuitions have? If they too are only meaning-constitutive, appeal to them does not fundamentally affect the situation. If they are supposed to have some other status, for example, that of rational intuitions, the question is what the source of our trust in them is. One need not be a hard-core empiricist to prefer an account of the justification of our judgments about metaphysical matters in terms of the epistemologically relatively unproblematic source of semantic or conceptual competence to an (unspecified) account seemingly destined to be fundamentally mysterious.[33]

2.8 Do persons survive fission?

Suppose my hypothesis about our competence with the concept of personal identity is correct. What consequences should we draw about the results of fissions: do persons survive fissions or not? Recall that on my view, the semantic values of expressions are such as to make the sentences constitutive of meaning as nearly correct as possible. There will presumably be many equally acceptable assignments of semantic values, as there presumably

[33] In a discussion of the liar paradox (1970), Robert L. Martin discusses what a solution to this paradox should involve. He says that it should tell us which assumption or assumptions in the reasoning are untrue; and moreover a reason must be given for taking this assumption to be untrue: a reason beyond the mere fact that this assumption, in the presence of other plausible assumptions, leads to absurdity. What I am arguing here is that such reasons cannot be given with respect to all philosophical problems. Sometimes, when individually plausible claims jointly lead to absurdity, and one or more of them must be abandoned, nothing more can be said about why they must be abandoned than that they conflict with other initially plausible claims.

will not be a unique way of maximizing the correctness of the sentences constitutive of meaning.

Suppose that B and C at t' both bear P to A at t, where P is the relation such that our competence with the concept of personal identity disposes us to (lacking evidence to the contrary) take x at t and y at t' to be the same person just in case y at t' bears P to x at t. For example, P might be physical or psychological continuity. There are then several reasonable hypotheses about the class of admissible assignments of semantic values to statements about whether B or C is the same person as A. Here are some examples. (i) Under some acceptable assignments of semantic values, the sentence

(9) B is the same person as A

is true, under the other acceptable assignments, the sentence

(10) C is the same person as A

is true. (ii) Under each acceptable assignment of semantic values, neither (9) nor (10) is true, and A does not survive the split. (iii) Under some acceptable assignment of semantic values (9) is true, under some (10) is true, and under some A does not survive the split.[34]

As regards (i) and (iii), I should remark that it certainly is arbitrary to suppose that one of B and C but not the other is identical with A. It seems wrong to say that one but not the other is A, for they both have equal claim to being A.[35] However, this only rules out the possibility that one of B or C is identical with A under all acceptable assignments. It is compatible with the possibility that under some assignments B is identical with A and under others C is identical with A.

[34] My arguments in the early sections did not speak to views on personal identity according to which what intuitively are the two different persons resulting from the fission really are one and the same person (the view discussed by Parfit); nor to views according to which there exist two persons also before the fission, it is only that before the fission they spatially coincide (Lewis). For all I argued, such views – let us call them *the extreme views* – could be fine.

Having set out my positive view, let me briefly state my position with respect to the extreme views.

The extreme views are perfectly coherent. We could have used a notion of person such that one of the extreme views was true. But it seems that we do not employ such a notion of person, and that the only reason to embrace an extreme view would be that no other view is workable. But my view on the truth-conditions of statements about personal identity is a non-extreme view, in the sense of being a view which does not differ as radically from common sense as the extreme views with regard to which judgments on personal identity it deems true, which seems workable. There is therefore no reason to believe an extreme view to be correct.

[35] See e.g. Parfit (1984), p. 256.

Regardless of whether (i), (ii) or (iii) is accepted, the presumption of locality is not respected. But the reason locality is not true is that the most reasonable ways of rendering correct the sentences constitutive of meaning involves abandoning locality.

This brings us back to my somewhat cryptic remark above that Nozick's denial of locality differs in some fundamental respect from Parfit's rejection of locality. What Nozick holds is that locality misrepresents the way we think of the survival and identity of persons. By contrast, consider Mark Johnston (1987):

> To the extent that we have [intuitions that Nozick's suggestion is wrong], we are not to be construed as responding to cases as if we were adherents of the closest-continuer schema, at least as applied to people. To that extent Nozick's reaction...is inadequate, *even if it were the case that philosophical theorists should ultimately adopt the closest-continuer schema in some form in the light of all the evidence about personal identity.*[36]

Johnston's claim is analogous to my suggestion that although the meaning of talk about personal identity somehow demands that locality be true, the truth-conditions of claims about personal identity may make locality false. Denying locality means doing violence to intuitions constitutive of the notion of personal identity. Nevertheless locality may in the end have to be abandoned.

The way I read Parfit, his view is closer to that presented by Johnston in the above quote than to that of Nozick. Parfit seems to hold that in a fission case, *no* answer to the question of which resulting person, if any, is identical to the original person is *completely faithful* to our concept of personal identity.[37]

After presenting my arguments against Nozick above, I suggested that Nozick might respond by saying that *all* accounts of personal identity will have some counterintuitive consequences. I hinted that there might be something to this response; and I am now in a position to say what. Rejecting locality may, and I believe does, serve to maximize the correctness of the sentences constitutive of meaning, even though rejecting locality means rendering some sentences constitutive of meaning untrue.

[36] Johnston (1987), p. 68; his emphasis.

[37] See Parfit (1984), pp. 255-60. Parfit's view on the fission case is that we would probably find that the best description of the case is that neither resulting person is identical to the original person, but no description is entirely happy.

Given the present account of meaning, and the separation of meaning from truth conditions, questions such as that of the nature of personal identity potentially split into two importantly different problems. First, there is the problem of identifying the conditions under which our semantic competence disposes us to judge that x at t is the same person as y at t'. Second, there is the problem of giving the truth conditions of sentences of the form "x at t is the same person as y at t'". The lesson is general. Philosophical debates over a particular concept F can in principle split into two fundamentally different questions. *First,* there is the question of which principles are constitutive of the concept: what competence with F can require one to be disposed to accept. *Second,* there is the question of truth conditions and truth values of claims about Fs and F-hood.[38] What I claim is that the fission cases show that this possibility is actualized. The problem of personal identity really does split into two different problems.

2.9 Psychological vs. physical criteria

The debate *between* proponents of psychological criteria of personal identity and physical criteria of personal identity is often centered around hypothetical cases that ought to decide the matter, if anything decides it: cases where the two criteria yield different verdicts. For example, cases of the following kind: at t' there are two persons y and z such that y at t' is psychologically continuous with x at t and z at t' has the same brain/body as x at t. Proponents of psychological criteria argue that y at t' is identical to x at t; proponents of physical criteria argue that z at t' is identical to x at t. (I shall refer to such cases as *mind/body switching cases.*)

Our intuitions appear not to yield clear verdicts in cases of the kind envisaged. As is made out very clearly in Bernard Williams' important paper "The Self and the Future" (1970), our intuitions can go either way in such cases, depending on how they are presented – and neither presentation seems clearly superior.[39]

[38] This distinction will have to be made on any broadly Fregean view, where the sense of an expression is something more fine-grained than the intension (in the possible-worlds sense) of the expression; but when the principles constitutive of the concept need not be true, the distinction takes on a greater importance.

[39] See also chapter 2 of Rovane (1998). Williams ends up favoring the physical criterion over the psychological criterion, but that is for an independent reason: he thinks that the fission cases present problems only for the psychological criterion. Rovane defends the psychological criterion. But that is because she holds that the most reasonable way of revising our understanding of personal identity is to make it conform to a version of the psychological criterion.

If my theses about semantic competence are on the right track, it could be that there is a sense in which proponents of *both* criteria get it right. For it may be that our semantic competence disposes us to take x at t to be identical to y at t' just in case x at t *either* is appropriately psychologically related to y at t' *or* is appropriately physically related to y at t'. This presupposes, in the sense of meaning presupposition, that there will not be more than one person at t' to whom y is psychologically or physically related. This would explain our perplexity when confronted with hypothetical cases in which physical and psychological criteria yield different verdicts: we are perplexed because our use of the notion of personal identity presupposes that these criteria always go together.[40]

In his paper, Williams takes great care to present cases of the kind described above in such a way as to beg no questions against any of the participants to the dispute. Thus he takes care not to describe the case either in terms of "persons changing bodies", which would beg the question in favor of a psychological criterion, or in terms of "persons changing brains", which would beg the question in favor of the physical criterion. But note how natural *both* of the question-begging descriptions of the case are. Unless we are in a philosophical context and taking care not to beg any questions, both descriptions could (at least when taken separately) be accepted as natural and correct.

It is standardly assumed that since psychological and physical continuity do not always go together, we must make a choice between the psychological and the physical criterion, since a criterion like

(11) person x at t = person y at t' if and only if y at t' is either psychologically or physically continuous with x at t

or

[40] In chapter 2 of her recent book (1998), Rovane argues that our *intuitions* about personal identity are *incoherent* and that we therefore must be *revisionists* about personal identity. Let me briefly spell out what I think is wrong about Rovane's position, in order to bring out how it contrasts with mine.

Rovane does not say that the intuitions we have by virtue of our competence with the notion of personal identity are incoherent. For all she says, it is only that our pre-theoretical beliefs about personal identity are incoherent. And her "revisionism", if demanded by the incoherence of our intuitions about personal identity, accordingly would say only that some of our pre-theoretical beliefs about personal identity would have to be abandoned.

But even if our pre-theoretical beliefs about personal identity are incoherent, the fact remains that our assertions and thoughts about personal identity (before any revision has taken place) have truth-values. Rovane rightly takes the incoherence of our pre-theoretical beliefs about personal identity to show that some of these beliefs must be abandoned. But she goes wrong when she takes this by itself to license revising our concept of personal identity. For it remains that for all she has shown, some of our beliefs about personal identity are made true by the concept plus the world and some are made false: and it is a factual, descriptive question which beliefs are true and which false.

(12) person x at t = person y at t' if and only if y at t' is both psychologically and physically continuous with x at t

hardly strikes anyone as plausible, in view of for example the fact that it is hardly the case that (intuitively) both original persons die in the mind/body switching cases or that both original persons survive as both the resulting persons.

This standard assumption is justified if our concern is with the question of the truth conditions and truth values of statements about personal identity (and this is indeed most people's concern), but not if we are concerned with the meaning of talk about personal identity. For then we can still allow that psychological and physical continuity both count as sufficient and/or necessary for personal identity. The mind/body switching cases simply reveal that it is a meaning presupposition underlying talk of personal identity that psychological and physical continuity always go together.

I motivated the suggestion that the notion of personal identity is incoherent by arguing that the physical and the psychological criterion of personal identity both lead to absurd or contradictory results in certain cases. I should now admit that those arguments were not essential to my case. Hopefully they served their motivating function, but it ought to be clear that we can, as it were, kick away the ladder. Suppose that contrary to what I argued, only the psychological criterion is incoherent. It could still be that our competence with the concept of personal identity requires us to accept (1) – the psychological criterion as originally stated – and that the semantic value of "is the same person as" is constrained accordingly. That is, even were the physical criterion coherent and the psychological criterion not, it could be that the psychological criterion best captures the notion of personal identity.

2.10 Concluding remarks

The most natural versions of the psychological and the physical criterion of personal identity are *supported* by the fact that they run into absurdity when applied to fission cases. At least if they are meant to capture the concept of personal identity, as opposed to state principles which actually hold true of persons in all possible scenarios.

This conclusion serves also to defend the method of considering wildly counterfactual cases from one kind of criticism often directed against it: the criticism that since our intuitions about these cases are incoherent, the verdicts delivered by this method cannot be trusted. We must here ask whether the method primarily is supposed to tell us which principles actually are true of persons in all possible cases, or to tell us about the nature of the concept of personal identity. At least as used for the latter purpose, the method can well be valuable, in spite of the fact that our intuitions are incoherent.

The arguments in this chapter have also served to justify further the account of semantic competence presented in chapter 1. In this chapter, I have shown how, given this account of semantic competence, we can dissolve a host of philosophical problems. Philosophical problems often arise because claims which, taken separately, seem not only obvious but also fundamental to the concepts employed jointly lead to absurdity (that is, they lead to the denial of a claim that seems equally fundamental to the concepts involved). Such problems, or at any rate many of them, can be seen to arise because the language, or conceptual scheme, that we employ is inconsistent.

Chapter 3

When Folk Theories are False

3.1 Introduction

In the first two chapters I have argued for a particular thesis about semantic competence – that we are sometimes disposed to accept untrue claims by our semantic competence – and used this thesis to diagnose and resolve some rather different philosophical problems. As discussed in chapter 1, the thesis about semantic competence requires for its acceptability a related claim about reference-determination: briefly, that a principle can be reference-determining even if untrue. In this chapter, reference-determination will be the main focus. I will show that in order that the philosophical problems discussed be resolved as earlier outlined, only the claims about reference-determination are really necessary. Considerations of semantic competence need not enter into it. Toward the end of the chapter, I shall return to the issue of semantic competence.

3.2 Folk theories

Some influential philosophers – e.g. Frank Jackson[1] and David Lewis[2] – hold that many notions studied in philosophy get their content through *folk theories* implicitly defining them. I shall here spell out their idea and some of its immediate consequences. In particular, I shall consider the case of false folk theories. I shall argue that it is likely that the folk theories defining some important philosophical notions are false, and that this gives

[1] Jackson (1998).

[2] See especially Lewis (1972), (1997).

rise to an unrecognized or at least underappreciated form of indeterminacy, which I shall call FTF-indeterminacy (for Folk-Theory False indeterminacy). This all but dissolves the philosophical problems concerned.

I shall here show how this kind of indeterminacy arises, distinguish it from other kinds of indeterminacy, and show how recognition of FTF-indeterminacy provides keys, or at any rate clues, to important philosophical problems. I shall show that it is likely that the liar paradox and certain problems related to personal identity arise because of FTF-indeterminacy, and I shall present an analysis of the phenomenon of vagueness centered around the notion of FTF-indeterminacy. These are the only examples of philosophical problems that I shall consider in any detail, but I shall indicate why the kind of diagnosis suitable to these problems is likely to be correct also for many other problems.

Provocatively stated, my thesis is that many philosophical problems arise and persist because fundamental theories or conceptions we have of the world are mistaken; and because these theories or conceptions are reference-determining, these problems can be expected not to have determinate solutions, for the expressions employed in stating them will be referentially indeterminate in a crucial way.

Most of the discussion will proceed on the largely undefended assumption that the semantic thesis of the folk-theoretic philosophers (as we may call the theorists concerned), that there exist reference-determining folk theories for many ordinary expressions, is correct. In the next to last section, however, I will generalize the conclusions reached, and show how other theoretical assumptions lead to the same conclusions.

The basic idea of the folk-theoretical philosophers is usefully divided into three theses: (i) theories implicitly define their theoretical terms, (ii) certain widely shared beliefs constitute 'folk theories', (iii) in particular, many expressions that express concepts or denote phenomena studied in philosophy are implicitly defined by folk theories.

First, as regards theories implicitly defining their theoretical terms, see David Lewis (1997):

> If, without benefit of any prior definition of 'entropy', thermodynamics says that entropy does this, that, and the other, we may factor that into two parts. There is an existential claim – a 'Ramsey sentence' – to the effect that there exists some quantity which does this, that, and the other (or near enough). And there is a semantic stipulation: let that which does this, that, and the other (or near

> enough), if such there be, bear the name 'entropy'. Here is another way to say it: the theory associates with the term 'entropy' a theoretical role. It claims that this role is occupied. And it implicitly defines 'entropy' as the name of the occupant of the role.[3]

'Entropy' is defined by thermodynamics as that which satisfies the relevant theoretical claims of thermodynamics. The parenthetical "or near enough" will concern us later.

Second, as regards folk theories. These are, as a first approximation, widely shared and firmly entrenched beliefs about the subject matter of which they are theories: beliefs so firmly entrenched that they are regarded as platitudes. They are, moreover (as I shall explain shortly), the beliefs we actually draw upon when making judgments about the application of a particular expression or concept. According to the folk-theoretic philosophers, these beliefs can be regarded as theories of the subjects at hand, implicitly defining the ordinary terms.

Folk theories are largely only tacitly believed, and it may be a very difficult task to spell out their actual content. Even for very simple expressions, it is hard to make explicit associated beliefs which are such that the referents of these expressions are plausibly the only things that make these beliefs true. Why, then, believe that we have the tacit beliefs posited? The reason is that if we are fully competent with a predicate 'F', we are (*ceteris paribus*) good at making judgments about whether something is F or not: we must have some beliefs which we draw on when making such judgments. As Lewis remarks (with reference to Smart (1961)) in a discussion of color terms, "we can sort dyed bits of wool [according to color] and then shuffle them together, and then sort them again into just the same heaps as before".[4] This ability is explained, Lewis says, by our tacit knowledge of a folk theory about color.

Our folk theories are the beliefs we draw upon when making judgments about whether a certain concept applies in a given situation. And not only are these beliefs often implicit: very firmly entrenched explicit beliefs can fail to be among them, and can in principle even *conflict* with them. Consider, for example, the notion of personal identity. Arguably, many people do explicitly believe, and have explicitly believed, that sameness of person is the same as sameness of *immaterial soul.* But Mark Johnston argues that when we – even

[3] Lewis (1997), p. 326..

[4] Lewis (1997), p. 325f.

those who believe in souls – make judgments about personal identity, we do not draw upon any beliefs about souls. He says for example that our certainty about persons being the same over time is hard to explain if we suppose that personal identity is identity of the soul. For we have no direct means of establishing soul identity.[5] Whether Johnston is right or not (I believe he is), he draws attention to a genuine possibility: that our explicit metaphysical beliefs about personal identity (no matter how widely shared and firmly entrenched) need not be what we draw upon when making judgments about personal identity. If so, then these metaphysical beliefs are not part of our folk theory of persons.

Third, it is claimed that among the ordinary terms defined by folk theories are those expressing concepts discussed by philosophers. Thus, color terms, value terms, mental terms, causation, etc. have been studied via the folk theories in which these terms are said to occur.

This third thesis, that notions studied by philosophers are defined by folk theories, is not a consequence of the second thesis, that many ordinary notions get their content from the folk theories defining them. To get from the second thesis to the third, we also need the premise that the notions studied by philosophers are indeed the ordinary notions. Other understandings of what philosophers are concerned with are possible: for example that the notions employed in philosophical theorizing are most often more or less *technical*.

But a reason for taking philosophy, as actually practiced, to be typically concerned with our ordinary notions, stems from consideration of accepted philosophical methodology. We are often supposed to consult our 'intuitions' about possible cases, and it may reasonably be argued this method makes sense only if we are concerned with ordinary notions. For then the intuitions can be taken to reflect our competence with these notions. (To say that philosophy, as actually practiced, is concerned with our ordinary notions is not to embrace the normative claim that what we should be concerned with are the ordinary notions.)

If there are folk theories of philosophical notions, then for obvious reasons, the structure and content of these theories form a natural topic for philosophers. For since folk theories implicitly define terms occurring in them, it follows that provided the terms do refer at all, the theories are true (or near enough to being true). We are fairly certain that color terms, value terms, mental terms, etc. refer, so we can be fairly certain that the folk theories are,

[5] See Johnston (1987), p. 69ff, and Johnston (1989), p. 372ff.

nearly enough, true.[6] And folk theory seems to be the kind of thing we can study by – relatively – a priori means.[7]

However, if philosophy, or a large part of it, is conceived as suggested, the philosophical enterprise is open to the following kind of criticism.[8] Consider, as an example, an ordinary semantic term, such as 'meaning'. Suppose that this term is implicitly defined by a folk theory, folk semantics, as the folk-theoretic philosophers would have it. And suppose further that folk semantics is sufficiently close to being true for 'meaning' to have a semantic value. Meaning (in the ordinary sense) can then be studied as outlined. But suppose also that in the actual empirical science of language, i.e. linguistics, the ordinary notion of 'meaning' has no place: it does not play enough of an explanatory role to get into a science of language. What, then, is the *interest* of the envisaged study of meaning? It may perhaps be of, say, *anthropological* interest, since it may tell us how the folk think about language: but why study how the folk think about language if what we are interested in is the phenomenon of *language* itself? (Compare: studying how the folk think about physical phenomena is not how physicists go about their job.)

This kind of criticism can presumably be leveled against the philosophical enterprise also under other understandings of it, but on the Jackson-Lewis construal of it, and its emphasis on what is commonly believed among the folk, it becomes especially pressing.

An accommodating response to this criticism is that even though specific features of folk theories may not be of great interest (for the reasons given), study of the general structure of folk theories can be of interest: common features these theories may have, independent of specific details about how the folk think, that are of interest simply because of the reference-determining role of folk theories. Much of this chapter will be devoted to making good this claim.

There may also be less accommodating responses to the criticism presented. But my aim here will simply be to argue that even if this criticism should be essentially sound,

[6] In all these cases, there are skeptics. Most notoriously, eliminativists argue that mental terms do not refer. Eliminativism will be discussed toward the end of section three.

[7] Some philosophers who emphasize the study of folk theories (in the present sense), prominently Frank Jackson, have based rather grand philosophical programs on this idea. But adherence to the general idea that folk theories implicitly define ordinary terms does not by itself carry commitment to this philosophical program.

[8] This criticism is prominent in Noam Chomsky's later philosophical writings. See e.g. Chomsky (1995), passim, especially the characterizations of the enterprise of philosophy as actually practiced, in this case philosophy of language, as "ethnoscience" and as the study of our "folk theories".

interesting consequences (not anthropological) may be drawn from consideration of folk theories.

3.3 When folk theories are false

The idea that ordinary terms are defined by their folk theories is clearly related to the idea from old-fashioned conceptual analysis that the meanings of (many) expressions are given, by and large, analytic statements in which these terms figure. But there are important differences. Analytic statements are all *true*. But the statements of folk theories need not all be true. They may be false for empirical reasons, or they may conflict with true metaphysical claims, or they may be jointly or internally inconsistent. This is of some consequence.

What, if anything, does an expression 'τ' refer to when the theoretical statements implicitly defining the term are not all true?[9] A simple-minded idea would be that 'τ' then simply does not refer. But it is easy to see that this idea is mistaken. We surely want to say, regarding some terms occurring in scientific theories now perceived to be false, that these terms refer, although some claims made about what they refer to are false. Similarly for folk theories. It could be that some firmly entrenched beliefs about color (beliefs which are part of our folk theory) are not true, but color terms still refer. The proper thing to say about the reference of terms like 'τ' is that they refer to whatever comes closest to making true the theoretical statements in which 'τ' occurs, provided there is something that comes sufficiently close. In Lewis' phrase, 'τ' refers to the *best deserver* of the label. If nothing comes sufficiently close, 'τ' does not refer.[10]

Talk about *closeness* here is of course loose, vague, etc. But we have enough of a handle on it to make at least some observations.

It is possible that if 'τ' is implicitly defined by a false theory but still refers, it is indeterminate what 'τ' refers to, in that there are different candidate semantic values that all come equally close to satisfying the theoretical statements in which 'τ' occurs. Let us say that if a term is for this reason indeterminate in reference (or is *in this way* indeterminate

[9] Two notes on terminology. First, 'τ' need not here be a singular term: it can well be an expression of a different grammatical category. Second, in spite of this I shall speak of 'τ' as *referring* to its semantic value. If you are offended by talk of, say, predicates referring, just substitute a more neutral verb as you see fit.

[10] See Lewis (1970).

in reference) it is ***FTF-indeterminate*** **(*folk-theory false indeterminate*).**

Here is an abstract – and unrealistic – scenario: 'τ' occurs in theoretical statements $\phi_1, ..., \phi_n$. These statements cannot be jointly true, but for each i between 1 and n, there is some candidate semantic value that makes all of these statements except ϕ_i true. These candidate semantic values are all equally good candidates for being the semantic value of 'τ'. It is then indeterminate which of these semantic values is the semantic value of 'τ'. A clear example of a scientific term that exhibits this kind of indeterminacy is 'mass' as used in Newtonian physics. Newton's theoretical claims defining 'mass' are not true; and there are two different kinds of mass (rest mass and relativistic mass) posited by modern physics that both come equally close to satisfying the theoretical claims associated with 'mass' by Newtonian physics.[11]

A term implicitly defined by a folk theory can be indeterminate in reference for other reasons beside FTF-indeterminacy. For example, it could be that all statements implicitly defining the term are true, but there are several different entities all satisfying these statements. Or it could be that it is not determinate which statements are part of a particular folk theory, or that the contents of the statements of the folk theory are indeterminate. Such types of indeterminacy, being more familiar, will not be a major topic here.

In asserting that the falsity of folk theories can give rise to FTF-indeterminacy as outlined, I go along with Bedard (1993) and Lewis (1997). In (1970), Lewis' position was that when a theory defining a term is false, the term's semantic value is what comes closest to making the theory true, provided there is a *unique* such semantic value. If several candidate semantic values come closest to making the theory true, the term does not refer. Bedard (1993) suggested instead that if several candidate semantic values all come maximally close to making the theory true, then the defined term is indeterminate, in that under different acceptable assignments, it has different semantic values. In particular, if x is among the candidate semantic values that come maximally close to making the theory true, then under some acceptable assignment the semantic value of the defined term is x. Lewis' latest published view on the matter, in (1997), is that unless the resulting indeterminacy in reference is too wild, Bedard's account is right; whereas in cases where the indeterminacy would be too wild, the term simply fails to refer.[12]

[11] For details, see Field (1973), p. 463ff and (1994), p. 422ff.

[12] Bedard and Lewis consider only the case of *singular terms* being defined by corresponding folk theories,

Jackson's view on the possibility of folk theories being false differs, at least superficially, from Lewis's. Jackson emphasizes the study of folk theories, but curiously, when he considers how folk theories define terms, he talks not about our *actual* theories, but about our *best* theories of the matter. Discussing the example 'fish', Jackson says that "x is a fish iff x has the important properties out of or descended from or explanatory of F_1, F_2, F_3,..., according to the best true [biological] theory", where F_1, F_2, F_3,... are the properties ascribed to fish by the current folk theory.[13]

There is reason to be skeptical about Jackson's claim, as literally stated. First, the best true biological theory need not employ any concept like 'fish' at all: its taxonomy might be quite different. And neither need it contain any claims either descended from or explanatory of the beliefs that fish have the properties F_1, F_2, F_3, etc. Second, Jackson presents his semantical claims in the course of an argument for *cosmic hermeneutics*: the view that (provided physicalism is true) knowledge of everything described in purely physical terms and everything that is a priori suffices for knowledge of everything whatsoever. But if the reference of words like 'fish' is determined as Jackson says it is, it is hard to see how his semantic claims can be used in an argument for cosmic hermeneutics. For even if the best true biological theory does contain claims descended from or explanatory of the beliefs that fish have the properties F_1, F_2, F_3,..., our knowledge of what this biological theory will say does not seem a priori.[14] But an important assumption of the argument for cosmic hermeneutics is that linguistic knowledge is a priori.[15]

These two points indicate that for reasons of charity, we should interpret Jackson's talk of "best true biological theory" in a perhaps slightly non-natural way. For knowledge of the meaning of 'fish' to be a priori, the "best true biological theory" must be recoverable, in some suitably a priori way, from the actual folk theory, the contents of which by assumption are known a priori by the folk in question.

Suppose that an actual folk theory – stated in physicalistically acceptable terms – is false. Knowledge of everything physical and of the folk theory is, by Jackson's lights, sufficient

and outline ways of reducing other cases to this one.

[13] Jackson (1998), p. 35.

[14] I borrow the term "cosmic hermeneutics" from Byrne's (1999) discussion of Jackson-style ideas. Byrne in turn borrows the term from Horgan (1983).

[15] On some characterizations of the a priori, it is true by definition that linguistic knowledge is a priori, no matter how linguistic knowledge is further characterized. But if Jackson were to rely on such a characterization of the a priori to defend the possibility of cosmic hermeneutics, the view would be rather more trivial than at first it seems.

to know that the folk theory is false. Now, Jackson may reasonably (given his theoretical lights) further suppose that knowledge of everything physical and of the folk theory is also sufficient for knowing how to assign semantic values to the terms characterized by the actual folk theory in such a way as to make this theory come out as nearly true as possible. This knowledge would then also suffice to arrive at the true theory most similar to the actual folk theory: the "best true biological theory" in a rather non-standard sense.

Everything here is of course very 'iffy'. But if there are significant problems here, that is Jackson's problem, not mine. All I have said these last few paragraphs has been on behalf of Jackson's philosophical use of the existence of folk theories. And the main point is that the views of the reconstructed Jackson are very like those of Lewis (1997). If the term 'fish' is FTF-indeterminate in reference on the view of Lewis (1997), there are, correspondingly, many "best true biological theories" defining the term on the view of the reconstructed Jackson. And, although Jackson could in principle adopt the view of the Lewis of 1970 and take 'fish' to have no reference, it seems more reasonable to take 'fish' to be indeterminate in reference: in effect, FTF-indeterminate.

The possible falsity of folk theories has received much attention in one area: the discussion of 'eliminativism' or 'eliminative materialism'. Eliminativists argue that intentional terms are defined by folk psychology, our folk theory of the mental, and that because this theory is false, or perhaps because this theory is *very* false (see below), intentional terms do not refer. Putting their conclusion in the material mode: there are no such things as beliefs and desires (and, presumably, neither is there such a thing as the property of being a belief or the property of being a desire).

Sometimes in the writings of eliminativists it seems as if they assume that the falsity of folk psychology is sufficient for intentional terms not to refer. And sometimes it seems instead that they hold that it is because our folk theory is so widely off the mark that intentional terms do not refer.

In the former case, their theory relies on an (unargued) very radical semantic thesis, a semantic thesis far more radical than that of Lewis (1970).[16] In the latter case, their theory is compatible with all the reviewed positions on the falsity of folk theories, but relies on a

[16] I don't know that any eliminativist has explicitly subscribed to the thesis that the mere falsity of folk psychology as a theory is sufficient for intentional terms not referring. But the later Stich (1992, p. 254f) when criticizing his earlier self ascribes this thesis to the latter.

stronger assumption about folk psychology.[17] In any case, consideration of eliminativism is unlikely to shed light on the condition for a folk-theoretically defined term to refer. Either the eliminativist's semantic thesis is too radical to be plausible, or it is so weak as to fail to constitute an alternative to the positions of Jackson and the various versions of Lewis.[18]

3.4 Application to some philosophical disputes

FTF-indeterminacy is an interesting phenomenon, because provided the Jackson-Lewis account of how reference is determined is correct, it is quite plausible that many resilient philosophical problems arise and persist because of such indeterminacy. And furthermore, these problems can be dissolved once it is recognized that FTF-indeterminacy is what lies at their root. (I say *dissolved*, not *solved*, for reasons that will become apparent.)

The idea is that (a) the folk theories implicitly defining the expressions central to the statements of the philosophical problems concerned are false; (b) these expressions are, as a result, FTF-indeterminate; (c) this provides the key to dissolving these problems.

Here are two examples. First, *the liar paradox*. If the idea that many ordinary terms are implicitly defined by folk theories is correct, it seems extremely plausible that the disquotation schema (or some variant) is at least part of the folk theory implicitly defining 'true'. What the liar reasoning shows is that the disquotation schema is not valid. The semantic value of 'true' is then whatever comes *closest* to satisfying the folk-theoretical claims defining the term: nothing satisfies all of these claims. The predicate 'true' is then likely to be FTF-indeterminate, since it is unlikely that there will be a unique 'best deserver' of the label 'true'. (Or, more precisely, the expressions crucially employed in the various versions of the liar paradox, 'true' and logical expressions, will be FTF-indeterminate, for the folk theories associated with these expressions cannot all be true. The expressions are hence FTF-indeterminate, and their semantic values will be whatever comes closest to satisfying these folk theories, taken jointly.)

This is of some consequence for proposed accounts of truth and the liar. To adequately

[17] Thus, Churchland (1981) argues that folk psychology is *radically* mistaken.

[18] Concerning eliminativism about the mental I should mention another reason for thinking that the eliminativist is committed to thinking not only that our folk theory of the mental is false but that it is *wildly* false. The eliminativist does not wish to say only that beliefs and desires do not exist, but also that there exists nothing with the intentionality of beliefs and desires. So neither must there be schmeliefs and schmesires, slightly different from beliefs and desires but sharing their intentionality.

describe the import of the claim that 'true' is FTF-indeterminate, let me first introduce a distinction between *theories* of truth and the liar, and *philosophies* of truth and the liar. A theory of truth and the liar is an account of the semantic values of 'true' and other expressions centrally employed in the liar reasoning; a philosophy of truth and the liar, in my stipulated sense, is a philosophical justification for taking the semantic values of 'true' and other expressions employed to be what they are. A proposed account of truth and the liar can typically usefully be separated into two parts: a theory and a philosophy.

Consider now some popular proposed accounts of the liar paradox: for example, Kripke's theory and the revision theory of Gupta and Belnap.[19] The specific details of these two theories do not matter for the claims to follow. All that the reader unfamiliar with the liar literature would need to know is that these are two theories of truth primarily justified by their according with our pre-theoretic intuitions as well as they can while remaining consistent; and among these pre-theoretic intuitions of ours are, it is generally agreed, that the instances of the disquotation schema should come out true.

The account of Kripke and the account of Gupta and Belnap are different: they make conflicting claims about the extension of the truth predicate. The proponents of these accounts argue against each other. But if 'true' is FTF-indeterminate, it may be that *both* Kripke's theory and the revision theory come maximally close to satisfying the theoretical claims defining 'true': and here I of course talk about the *theories* put forward by Kripke and by Gupta and Belnap. Both theories make our false folk theory of truth come out as nearly true as possible. So both theories constitute acceptable assignments of semantic values to the relevant expressions. It is FTF-indeterminate whether Kripke's theory or the revision theory is correct.

The reason neither theory seems fully satisfactory is that neither theory makes our folk theory of truth wholly correct. But then, on the other hand, no theory can respect all of the claims of our folk theory of truth. The claim that there is no definite account of truth is of course often made, but the appeal to a reference-determining folk theory, and the fact that no property satisfies all of its claims, serves to make more precise this rather loose claim.[20]

But although the theories of Kripke and of Gupta and Belnap may each in this way

[19] Kripke (1975) and Gupta and Belnap (1993).

[20] Kripke himself asserts in his (1975) that no "definite account of truth" is to be had. It is from him I borrow the phrase.

be maximally satisfactory, the philosophies of Kripke and of Gupta and Belnap are by no means vindicated. The justification for taking the theories proposed by these theorists to be maximally correct is that the expressions employed in the liar are FTF-indeterminate, and the proposed theories are among the assignments of semantic values that come closest to satisfying the false folk theories associated with these expressions. The considerations put forward by Kripke and Gupta and Belnap are rather different. Kripke appears to see his theory as vindicated by consideration of the step-wise way in which we arrive at truth evaluations; Gupta and Belnap take the concept of truth to be "circular" and its meaning to be given by a "revision rule" rather than an ordinary rule for application.[21]

The problem of *personal identity* can be dissolved in a manner rather similar to that in which the problem posed by the liar can be dissolved. One important problem of personal identity, widely discussed in the literature, is that of *fission*. For every relation Q otherwise seen as plausibly necessary and sufficient for personal identity, it appears to be metaphysically possible that Q obtains between a person at a time and two different persons at a different time.[22] But then Q cannot be identified with personal identity, because of the transitivity of identity. Fission cases are easily accounted for, however, in the general framework sketched here. In spite of the fission cases, it could be part of the folk theory of persons that Q is necessary and sufficient for personal identity in spite of the fission cases: the fission cases show only that the folk theory of persons is (necessarily) false. I shall illustrate and justify this claim at some length.

The two most popular kinds of criteria of personal identity are the *physical criterion*, according to which

> A at t is the same person as B at t' just in case B at t' is physically continuous with A at t (or, the intention behind the criterion is, B's body is the same as A's body),

and the *psychological criterion* according to which

[21] Kripke, I should note, hedges quite a bit and does not commit himself to this philosophical justification of his theory. On the basis of their philosophy of truth, Gupta and Belnap employ not a simple classification of sentences as true and false (and possibly also 'gappy' or 'neuter'), but instead a tripartite classification of sentences as *stably true*, *stably false* and *unstable*. This might make it seem that their *theory* of truth is hard to compare with that of Kripke, which employs a more standard conception of the semantic values of sentences. However, Gupta and Belnap's theory can be separated from their philosophy if we simply interpret 'stably true' as *true*, 'stably false' as *false*, and 'unstable' as *gappy*.

[22] As in chapter 2, I use 'Q' rather than 'R' because the latter is appropriated by Parfit.

> A at t is the same person as B at t' just in case B at t' is psychologically continuous with A at t (shares enough of A's psychological features and does so because of being suitably causally linked to A at t).

Both of these criteria face problems raised by the fission cases. Suppose lightning strikes and splits a person down the middle, and – miraculously – both the separated halves survive, and go on living separated. Or suppose one of my hemispheres is transplanted into one body and my other hemisphere is transplanted into another body. Both of these scenarios seem to pose problems for the psychological criterion, and at least the former poses problems for the physical criterion. (Whether the latter poses a problem for the physical criterion depends on whether we take the continuation of the brain or the continuation of rest of the body to be what matters, or to be what matter most in personal identity over time.) To see this, consider the case where lightning strikes and only one resulting half of the person or her body, let us say the left, exists and survives afterward; or the case where only one hemisphere (say, the left) survives and is transplanted into a living body. In both cases, a criterion of personal identity ought, to capture our intuitions, to yield the verdict that the person survives. But then the fission cases cause trouble. For, intuitively, the relation that in these latter cases obtains between the person before the loss of the right half, or the right hemisphere, and what remains of the person afterward, the left half or the left hemisphere, respectively, obtains between these relata also in the fission cases.

Theorists discussing personal identity have often, in response to fission cases, added so-called 'non-branching clauses' to their analyses. Thus, for example, a psychological continuity theorist would hold that A at t is the same person as B at t' just in case the relation of psychological continuity obtains between them and this relation does not take *branching form*: there are no other persons at t' psychologically continuous with A at t.

But a problem with this proposal is that this means that the identity of B at t' with A at t depends on what *other* entities exist. But this would appear to be ruled out by our conception of persons, or personal identity.[23]

If the Jackson-Lewis account of reference-determination is correct, then the alleged problem posed by fission cases for physical and psychological criteria can be turned on its head. It is implicit in the above characterization of the fission problem that it is not merely

[23] This claim is developed in more detail in chapter 2.

a problem regarding particular criteria of personal identity, but regarding our conception of persons, or personal identity, itself. For the fission problem shows that certain claims that all belong to that conception (or following rather immediately from claims which belong to it) are jointly inconsistent: (a) A person can survive losing half of her brain or half of her body. (b) In (what intuitively are) fission cases the original person does not survive. (c) Whether a person survives losing (say) half of her brain or half of her body survives does not depend on what else exists after the event of the loss.[24]

Given the Jackson-Lewis account, we can say that the principles that jointly lead to inconsistency all follow from claims that are part of the folk theory of persons. They are thus all reference-determining. (This provides an explication of the intuitive idea that they are part of our notion of personal identity.) Their joint inconsistency means that our folk theory of persons is (necessarily) false; and the semantic values of person-talk are whatever come closest to making this false theory true. The problem of personal identity then divides into two. On the one hand, there is the problem of laying out our folk theory of personal identity. On the other hand, there is the problem of giving an account of what the semantic values of person-talk are, given that this folk theory is false.

Returning to the physical and psychological criteria of personal identity, we can now see that although they lead to absurdity in fission cases, they thereby merely represent accurately our concept, or folk theory, of personal identity: for our folk theory itself leads to absurdity when applied to such cases. This is not a full-scale defense of either of these criteria (they face other problems); the point is only that the fact that these criteria lead to absurdity with respect to fission cases is only a point in their *favor*, if we are concerned with the *former* of the two questions about personal identity (namely, that of what our folk theory of personal identity is). For our folk theory appears to run into exactly the same problems as those faced by the physical and the psychological criteria.

It is not enough for matters of personal identity to be FTF-indeterminate that the folk theory characterizing personal identity be false. It must also be that there is no unique way of maximizing the correctness of this folk theory. But it seems likely that this further condition is satisfied. There are several ways of maximizing the correctness of the folk

[24] In Johnston (1989), there is a more careful discussion of exactly which principles apparently fundamental to our conception of person are jointly inconsistent. See footnote 14 of chapter 2 for the principles Johnston finds.

theory that both seem maximally good and on a par. For example, assignments under which one resultant of the fission is identical to the original person, assignments under which the other resultant of the fission is identical to the original person, and assignments under which neither resultant is identical to the original person (and the original person accordingly does not survive the fission).

One feature the liar paradox and the fission problem for personal identity have in common is that in both cases we have principles which seem clearly to be part of the concepts involved (to use a suitably intuitive phrase) - in the case of the liar, the disquotation schema, and in the case of personal identity, the principle that the survival of a person does not hinge metaphysically on the existence of other persons - which are not true. This feature calls for explanation. The Jackson-Lewis account of how reference (often) is determined provides such an explanation. The explanation is that these theses are all reference-determining, in spite of the fact that some of them are untrue.

The general pattern exhibited by the liar paradox and the problem of personal identity is exhibited by a large variety of philosophical problems. Many problems stem from the fact that we have a collection of principles, all of which appear not only obvious but even in some sense part of the concepts involved, but which taken together have an entirely unacceptable consequence. For all such problems, one kind of solution, or dissolution, that suggests itself, given the folk-theoretic account of how reference is determined, is that all of the principles that lead to the unacceptable conclusion are reference-determining.

Examples of problems that arguably are of this kind are: material constitution (see Rea (1995)), free will and determinism, belief attribution (see especially Kripke (1979)), moral dilemmas, rationality (e.g. Newcomb's paradox), skepticism (see the early pages of Lewis (1996)), and so on and so forth. All these problems can be stated on a form that highlights that here we have a set of highly plausible assumptions that jointly have a highly implausible conclusion. Of course, not all of these problems need be such as I have argued that the problem of the liar paradox and the problem of fission are. But for all of these problems, the kind of dissolution I have presented for the liar paradox and the problem of fission should seem prima facie attractive.

Recognition of the fact that a particular philosophical problem arises from the falsity of the folk theory defining expressions centrally employed in the statement of the problem does not solve the problem concerned. Realizing that the liar problem arises because the folk theory defining 'true' is false does not, for example, immediately tell us what the semantic value of 'true' is. What it does do, is to show why assigning an intuitively satisfactory semantic value to 'true' is so difficult; even impossible. And it strongly suggests that it is a highly indeterminate matter what the semantic value of 'true' is, and so that a complete answer to the question of the semantic value would have to specify all of the acceptable assignments. The task of developing such an answer is likely to be intractable, and, even should one be able to pull it off, it is not clear how much light would be shed on either language or the world. Diagnosing the liar problem as stemming from the falsity of the folk theory defining 'true' does not *solve* the problem; it only *dissolves* it. But the dissolution of the problem may be more illuminating than a solution would be.

3.5 Vagueness

In this section, I shall present an alternative analysis of vagueness, based on the idea that there is such a thing as FTF-indeterminacy. First, I shall compare FTF-indeterminacy with the recognized forms of indeterminacy. Then I shall use it to develop an analysis of vagueness.

In the literature, three kinds of indeterminacy are distinguished: *ontological*, *semantic* and *epistemic* indeterminacy. A sentence is ontologically indeterminate just in case it is indeterminate by virtue of indeterminacy *in the world.* Ontological indeterminacy is the result of the world being an inherently fuzzy place. A sentence is semantically indeterminate just in case it is indeterminate because of semantic features of the expressions employed. Partial definition is a relatively uncontroversial source of semantic indeterminacy. A sentence is epistemically indeterminate just in case it is neither ontologically nor semantically indeterminate, but for some principled reason we cannot find out what its semantic value is.

I have not listed *vagueness* as a *separate* kind of indeterminacy. It is a matter of dispute which kind of indeterminacy vagueness is: for all three kinds of indeterminacy I have mentioned, there are proponents of the thesis that vagueness – or *some* vagueness – is that kind of indeterminacy.

Timothy Williamson and Roy Sorensen have defended the view that vagueness is epistemic indeterminacy.[25] Michael Tye, among others, has defended the view that vagueness is (at least sometimes and partly) a matter of ontological indeterminacy.[26] By far the most popular proposal, however, is that vagueness is a matter of semantic indeterminacy, in particular, that vagueness is a matter of expressions being (as if) partially defined.

Vagueness seems neither to fit the model of epistemic indeterminacy (it's not just that we cannot find out where the boundaries are) nor the model of ontological indeterminacy (if it is indeterminate whether a man is bald that is not because the man himself is somehow indefinite). This may be the reason why vagueness is often thought to result from semantic underdetermination. For this thesis really does not seem to be intrinsically plausible. If vagueness really were a matter of semantic underdetermination – if borderline cases of a predicate 'F' were such because we had not made semantic decisions as to whether 'F' applies to cases of that kind – then intuitions about such cases would be akin to intuitions about division by zero. But we do have intuitions about the borderline cases, it is just that we do not have the intuition that a borderline case of redness looks red, or the intuition that it does not look red.

Those who regard vagueness as a matter of semantic indeterminacy tend to favor a *supervaluationist* theory of vagueness. For present purposes, the most important features of supervaluationism, as standardly conceived, are these: A sentence is true just in case it is true under each acceptable assignment. An assignment is acceptable provided all sentences are assigned determinate truth-values and compatible with the world and with the meanings of the expressions of the language, in so far as these meanings are determined. What gives rise to there being a manifold of different assignments is that the meanings of expressions are underdetermined. Vagueness, as it gives rise to there being a manifold of different assignments, arises from semantic indeterminacy, where this is conceived as underdetermination. Supervaluationists typically hold, moreover, that the laws of classical logic are true under every admissible assignment. Supervaluationists thus typically hold that the law of excluded middle is valid, but that the principle of bivalence is not.

It is clear that FTF-indeterminacy is a species of semantic indeterminacy. But it is different from generally recognized semantic indeterminacy in a fundamental way. Semantic indeterminacy is normally understood on the model of partial definition. If a term is seman-

[25] See e.g. Williamson (1994) and Sorensen (1991).

[26] See e.g. Tye (1990).

tically indeterminate, its meaning can be *extended* or *completed* without being *changed.*[27] Semantic indeterminacy results from *underdetermination* of meaning. Not so with FTF-indeterminacy.

The firmly entrenched beliefs that are part of folk theories are at least in certain respects part of the meanings of the terms implicitly defined by the folk theories. They play one of the roles traditionally ascribed to meanings: that of supplying a description such that the reference, or semantic value, of the term implicitly defined by the folk theory meets that description (or near enough). But then FTF-indeterminacy arises because the meanings are inconsistent, or inconsistent with empirical facts about the world. FTF-indeterminacy is or arises from *overdetermination* rather than underdetermination of meaning. The mere fact that semantic indeterminacy is generally regarded as having as its sole source underdetermination of meaning is an indication that FTF-indeterminacy is unknown or otherwise neglected.[28]

We must distinguish between two different kinds of indeterminacy that FTF-indeterminacy is or gives rise to: *first-level indeterminacy* and *second-level indeterminacy.* A sentence is second-level indeterminate just in case it has *different* semantic values under different assignments. It is *weakly* first-level indeterminate if it comes out truth-valueless under *some* acceptable assignment, and *strongly* first-level indeterminate if it comes out truth-valueless under *every* acceptable assignment. Second-level indeterminacy and strong first-level indeterminacy are incompatible. Weak first-level indeterminacy is compatible both with second-level indeterminacy and with strong first-level indeterminacy.

FTF-indeterminacy is second-level indeterminacy. It is compatible with the complete

[27] Fine (1975).

[28] The folk-theoretic philosophers are not committed to taking the claims of the folk theories to be part of the *meanings* of the expressions defined, as opposed to being only *reference-fixing*, in the sense of Kripke. However, the distinction between what is part of meaning and what is merely reference-fixing, although important for other purposes, is irrelevant here. Substitute, if you like, "overdetermination of reference-fixers" for "overdetermination of meaning"; the basic point stands. The problem isn't that we said *too little* to single out a unique referent; the problem is rather that we said *too much*, so not even one object manages to satisfy all conditions laid down.

In the text, I keep talking about folk theories as *reference-determining*. This is intended to be neutral as between taking folk theories to be reference-fixing and taking the claims of folk theories to be part of the meanings of expressions in a more substantial way.

Corresponding remarks to those in the main text should be made about the analysis of vagueness as semantic underdetermination. The proponents of this analysis are not either committed to the reference-determining claims being genuinely part of meaning. If something like the causal theory of reference is true, this does not touch the core of their analysis; for they can say that the *reference-fixing* claims are not sufficient to determine a unique referent (instead of saying that the meaning is not sufficient to determine a unique referent).

absence of first-level indeterminacy. It can be that within each acceptable assignment, all sentences in which an FTF-indeterminate term 'τ' occurs are either true or false. Does the FTF-indeterminacy of 'τ' then entail that there are sentences which are truth-valueless? That depends on whether we equate truth with truth under all acceptable assignments or not. If we do, then yes. If not, then no. And whether truth should be equated with truth under all assignments is not something that is settled simply by consideration of the nature of FTF-indeterminacy.

Having related FTF-indeterminacy to other kinds of indeterminacy, let me now turn to the analysis of vagueness.

The vagueness of a predicate 'F' is neatly explained if we assume that a *tolerance principle* for 'F' is part of its folk theory. A tolerance principle for 'F' says that, whereas large enough differences in its parameter of application (e.g. length for 'tall' and age for 'old') matter to the justice with which it is applied, some sufficiently small differences never matter to the justice with which it is applied.[29] Tolerance principles are always false, as shown by the sorites reasoning. Hence, the folk theories associated with vague predicates are always false.

Here, then, is an analysis of vagueness. First, the vagueness of a predicate consists in a tolerance principle being part of its associated folk theory. Second, this gives rise to FTF-indeterminacy, in the way described here; viz. second-level indeterminacy. Vague predicates have FTF-indeterminacy in common with many other expressions. What gives vague expressions their special flavor are the tolerance principles.[30]

The notion of tolerance was introduced in Wright (1975). Wright notes that no predicate

[29] Strictly, this would need a slightly more careful formulation. See chapter 1, footnote 14.

[30] In (1964), William Alston distinguishes between what he called *degree* vagueness and *combination of conditions* vagueness. Degree vagueness is more widely discussed: degree vagueness is the kind of vagueness which involves, in Alston's words, "the lack of a precise cutoff point along some dimension – age, number of inhabitants, or strength of opposition" (Alston 1964, p. 87). What Alston means by 'dimension' is what I earlier called the "parameter of application". It is degree vagueness that is analyzed here.

Combination of conditions vagueness arises when a word has "a number of independent conditions of application" Alston's most persuasive example of an expression exhibiting combination of conditions vagueness is 'religion'. There is a set of features such that the possession of all of them ensures that something is a religion; but such that it typically is not clear, regarding what has only most but not all of these features, whether it is a religion or not. Examples of such features are: "Beliefs in supernatural beings (gods)...Ritual acts focused around sacred objects...A moral code believed to be sanctioned by the gods...Prayer and other forms of communication with the gods..." (Alston 1964, p. 88).

The analysis of combination of conditions vagueness falls outside the scope of the present discussion. (Combination of conditions vagueness should not, of course, be *identified* with FTF-indeterminacy. It is not, for example, widely believed that all and only what possesses all and only the features associated with being a religion is a religion.)

can be tolerant, on pain of contradiction, and concludes that we must ***explain away*** the intuitions that they are. A different approach is suggested here: though vague predicates are not in fact tolerant, it is part of their folk theories that they are, and this constrains the assignment of semantic values to vague predicates.[31]

The way we think of the extension of a vague predicate 'F' is that some objects are clearly F, some are clearly not F, and there are gradual differences between what is F and what is not F; differences so smooth that small enough differences in 'F''s parameter of application do not matter at all to the justice with which 'F' is applied to an object. As all tolerance principles are false, the extension of 'F' cannot actually be as described. What properties does the extension of 'F' then have?

All that strictly follows from the falsity of tolerance principles, and hence of the folk theories defining vague predicates, is that the semantic values of vague predicates cannot be such as to make these folk theories true. In principle, this leaves open the very radical possibility that it is the second conjunct, that some sufficiently small difference in the parameter of application never matters, and not the first conjunct, that large enough differences in the parameter of application sometimes matter, that should come out true. But the semantic values of expressions defined by false folk theories are such as to make these theories come out as nearly true as possible, where the claims of the theory are appropriately weighed.

[31] The present analysis of vagueness as tolerance shares certain features with the analysis of vagueness that Mark Sainsbury defends (see e.g. Sainsbury (1991)}. A brief comparison with Sainsbury's proposal is in order.

According to what Sainsbury calls the "classical conception" of concepts, the function of a concept is to draw boundaries; concepts *classify*, in the sense of assigning things to classes (extensions and anti-extensions) (p. 167). This conception is what governs the thinking not only of those who defend bivalence in the face of vagueness but also of those who defend three-valued accounts (p. 167ff). Sainsbury argues that vagueness shows that we must abandon this conception. Vague concepts, Sainsbury says, draw no boundaries between their positive and negative cases. They classify "without drawing boundaries"; they are "boundaryless" (p. 179).

The similarity between Sainsbury's analysis and mine consists in precisely this: Sainsbury's talk of vague concepts being "boundaryless" could (for all of Sainsbury's view that has been presented so far) be explicated in terms of my analysis of vagueness as tolerance. One can say that a concept is boundaryless just in case it is part of its folk theory that it is tolerant. The idea would be that if a tolerance principle is part of the folk theory for predicate 'F', then 'F' is boundaryless in the sense that the meaning (set of reference-determining claims) for 'F' *says* that there is no boundary between the positive and the negative cases.

Sainsbury, however. takes a different route. He interprets lack of boundaries as a consistent feature of concepts, in the sense that it not merely is a principle that is reference-determining or part of the meaning of vague predicates, but moreover is a claim that the extensions of vague predicates actually can and do satisfy. He refers to Michael Tye's semantics for vagueness (in Tye 1990), for a semantics which is "congenial" to his approach. This is, offhand, puzzling, for Tye's semantics is three-valued; and Sainsbury has earlier argued that three-valued accounts of vagueness share the failings of bivalent accounts. The main difference between between Tye's semantics and other three-valued semantics is that according to Tye's semantics, the tripartite division effected by vague predicates is itself vague. But a concept's function can be to draw boundaries, as the classical conception has it, even if the boundaries it draws are vague.

And it would appear that, for 'F' a vague predicate, it is more essential that paradigmatic Fs be in the extension of 'F' than that small enough differences in 'F''s parameter of application never matter to the justice with which 'F' is applied. A vague predicate would still effect roughly the classification we take it to effect if the boundary between its instances and its counterinstances were sharp. But it would not effect any classification at all if small enough differences in its parameter of application never mattered at all to its applicability. An even more radical possibility is that all assignments of semantic values to vague predicates are so far from making the associated folk theories true that vague predicates lack semantic values altogether. As against this suggestion, one can again appeal to the greater centrality of the idea that intuitively paradigmatic Fs (or non-Fs) should be in the (anti-)extension of 'F', compared with the idea that the boundary between the Fs and the non-Fs is vague.

When these radical suggestions are ruled out, all the currently popular proposals about the semantic values of vague predicates remain in competition. But one of the remaining possibilities, the one I want to single out for consideration, is that under every acceptable assignment of an extension to 'F', 'F' partitions the domain into two classes; it is only that the extension of 'F' varies between acceptable assignments. Under a particular acceptable assignment of a semantic value to 'F', the kind of extension of this predicate is no different from that of precise predicates. Though tolerance principles belong to the folk theories for vague predicates, and thus in principle constrain the semantic values of these predicates, they do not have any real effect - or rather, the *second part* of tolerance principles, the part which says that some sufficiently small difference in the parameter of application never matters, does not have any real impact on the semantic values of vague predicates.

Vagueness thus need not give rise to *first-level* indeterminacy. It may well be that every acceptable assignment is bivalent. A sentence's borderline status is reflected, truth-valuewise, only in the fact that it has different classical truth-values under different acceptable assignments; it is not reflected in its lacking a classical truth-value under some or all assignments. If so, then whether vagueness is compatible with bivalence or not will depend on our semantics for the truth predicate. As far as vagueness itself goes, the complete diagnosis is that it gives rise to FTF-indeterminacy as indicated.

That all acceptable assignments are bivalent is, to repeat, only one of many live possibilities. But the reason for singling out this possibility for special attention is that it does seem that, given the distinction between first- and second-level indeterminacy, we could square classical logic and bivalence with the existence of vagueness while avoiding having to treat vagueness as merely an epistemic phenomenon. Moreover, if the folk-theoretic philosophers are right, then the semantic values also of logical expressions and of the truth predicate are determined by folk theories. And one reasonable suggestion concerning what the folk theories for these expressions are like, is that they say that the logical expressions are such that the inferences jointly characterizing classical logic are all valid, and the truth predicate is such as to make the disquotation schema valid. If so, we have a further reason for taking all acceptable assignments of semantic values to vague expressions to be classical and bivalent. For such assignments then have the distinctive advantage that they at least render true the folk theories characterizing the logical expressions and the truth predicate.[32]

3.6 Comparison with supervaluationism

Both the present analysis of vagueness and supervaluationism crucially invokes talk of different assignments, which in some sense are 'acceptable'. Both in light of this superficial similarity, and in light of the popularity of supervaluationism, it is worth briefly comparing the two analyses of vagueness.

One difference between supervaluationism and the analysis of vagueness here presented is of course that supervaluationists identify vagueness as stemming from semantic underdetermination, whereas I identify it as stemming from semantic overdetermination and, specifically, the fact that tolerance principles are part of the folk theories of vague predicates.

More interestingly, there is a way of thinking of supervaluationism according to which

[32] Of course, the liar paradox presents problems for rendering true the folk theories associated with the truth predicate and the logical expressions; but that is another matter.

The distinction between first- and second-level indeterminacy is potentially useful also in discussions of personal identity. Sometimes in discussions of personal identity it is claimed that something about our concept *person* determines that it never can be indeterminate whether A is the same person as B - even if other identity statements can be indeterminate. (See e.g. Williams (1970), pp. 174ff.) I shall not here try to explain why this would be so (I am not sure that the claim presented is correct); but there clearly is something intuitively extremely odd about the idea that personal identity could be indeterminate. At the same time, powerful arguments seem to show that personal identity sometimes is indeterminate (cf. cases where two persons gradually exchange bodies). Given the distinction between first- and second-order indeterminacy, there is room for reconciliation. It can be part of our folk theory of persons that personal identity always is determinate, so that all acceptable assignments of semantic values to our expressions of our language make all statements of personal identity determinate, while some statements about personal identity still are second-level indeterminate.

it does not conflict with the analysis of vagueness presented here. Supervaluationism can be separated into two components: First, it is an account of the *distribution of truth-values* to which vagueness gives rise: specifically, that vagueness gives rise to sentences that are neither true nor false, but that the laws of classical logic still are valid. Second, there is a kind of *philosophical justification* for this account of the distribution of truth-values: the diagnosis of vagueness as semantic underdetermination. In accordance with the terminological conventions introduced in the discussion of the liar, let us call the first component the supervaluationist's *theory* of vagueness, and the second the supervaluationist's *philosophy* of vagueness. The theory can be separated from the philosophy, and is so, by supervaluationists who emphasize supervaluationism primarily as a formal framework.[33] Only the supervaluationist philosophy is incompatible with the present analysis of vagueness. The present analysis of vagueness is compatible with the possibility that some, or even all, acceptable assignments of semantic values to expressions of our language give rise to the distributions of truth-values postulated by the supervaluationist theory. The general moral is this: *if* you are attracted to the supervaluationist *theory*, don't think you ought thereby to accept the supervaluationist philosophy.

Thirdly, the assignments that the supervaluationists talk about are not *per se* the assignments that (to put it loosely) best respect the meanings of the expressions of the language. All of the particular assignments that the supervaluationist consider are classical and bivalent, *not* because taking the laws of classical logic and the principle of bivalence to be valid best respect the meanings of our expressions (they do *not* – for the typical supervaluationist, the principle of bivalence is not valid) but because on every way of maximally extending the meanings of expressions, these principles come out valid.

3.7 Further remarks on the analysis of vagueness

The ideas about vagueness here defended are usefully divided into four. (1) Vague predicates are defined by folk theories. (2) In these folk theories are tolerance principles, so these folk theories are false. (3) The falsity of these folk theories can be expected to give rise to FTF-indeterminacy. (4) Every acceptable assignment of a semantic value to a vague predicate is

[33] See e.g. Vann McGee (1998). McGee, however, holds a view slightly different from that of the standard supervaluationist view presented.

bivalent.

Each of (1)-(4) builds on the preceding. The core of the proposal is in (1) and (2). A few remarks on these theses and their relations are in order.

First, (4) is rather speculative. But the reason for finding it attractive, given (1)-(3), is that since we can respect the idea that vague predicates are indeterminate by ascribing second-level indeterminacy to vague sentences, we need not take them to also give rise to first-level indeterminacy.

Second, thesis (3) does not *follow* from (1) and (2). The falsity of the folk theories defining vague predicates by no means entails that these predicates are FTF-indeterminate, or even indeterminate at all. It could still be that there is a unique best way of making the folk theories for vague predicates, or some of these folk theories, come out as near true as possible. It is just that it seems very likely that borderline cases of 'F' will be classed differently under different equally acceptable assignments.

Third, let an expression be ***FT-indeterminate*** (folk-theory indeterminate) if defined by a folk theory and indeterminate for reasons having to do with the folk theory defining it. It may thus be FT-indeterminate because it is FTF-indeterminate, or because it is a vague matter what claims are part of this folk theory, or because the folk theory does not say *enough* to single out a determinate semantic value for the term, or because the claims that are part of the folk theory are themselves vague. Then, even if vague predicates are FT-indeterminate, this need not depend on thesis (2). My thesis about vagueness is that a folk theory for a vague predicate 'F' contains a tolerance principle for 'F' and principles saying that paradigmatic Fs are in the extension of 'F' and paradigmatic non-Fs are in the anti-extension of 'F'. But consider the alternative hypothesis that the folk theory for a vague predicate 'F' does not contain a tolerance principle, but rather statements saying that anything sufficiently similar to a paradigmatic F is in the extension of 'F', and anything sufficiently different from a paradigmatic F is in the anti-extension of 'F'. Vague predicates would still be *FT-indeterminate*; but, the idea is, their FT-indeterminacy would result from the vagueness of the folk theories defining them (in particular the vagueness of "sufficiently similar") and not from these folk theories being false. Vague predicates would not be *FTF-indeterminate.*

This alternative analysis of vagueness would share some virtues with the one I have defended. In particular, FT-indeterminacy always necessitates distinguishing between first-

and second-level indeterminacy, and so one can argue that vagueness is compatible with classical logic and bivalence, given this alternative analysis of vagueness, analogously to the way this was argued above.

But even if vague predicates would exhibit FT-indeterminacy also under this alternative hypothesis about their folk theories, there is reason to hold on to, and emphasize, (2). For even though it is not preferable to this alternative hypothesis about the folk theories of vague predicates, as far as what it entails regarding semantic values under assignments, (2) explains the special role played by tolerance principles in our thinking involving vague expressions.[34]

Fourth, note that were (2) rejected but (1) retained, the important point would remain that we must distinguish between first- and second-level indeterminacy. This could still support (4), the idea that the indeterminacy to which vague predicates give rise can be squared with classical logic and semantics. For the indeterminacy vague predicates give rise to need only be second-level.

3.8 Reference-determination

Much of the discussion up to this point has proceeded on the assumption that the folk-theoretic account of how reference is determined is correct. This ought to make us somewhat skeptical; for this assumption is certainly not obviously correct. In this section I shall show that we do not need the full strength of the folk-theoretic analysis.

First, the folk-theoretic philosophers appear to hold that the reference of our expressions is determined *exclusively* by the folk theories associated with them. Even if this claim is not true, the theses for which I have argued (or only slightly modified versions of them) can hold. For my purposes it is enough that the statements of folk theories be among the determinants of reference. It could be, for example, that the referent of an expression (in the optimal case) is what makes the folk-theoretic statements associated with it true, *and* satisfies, e.g., certain causal constraints. In the cases where these conditions are not both

[34] One may also argue that the supposedly alternative hypothesis about vagueness just collapses into the original analysis. For according to the alternative hypothesis, the folk theory of a vague predicate ‘F’ says that anything sufficiently similar to a paradigmatic F is an F, and that anything sufficiently dissimilar from a paradigmatic F is a non-F. But “sufficiently similar [in respect of ‘F’’s parameter of application]” is itself vague. And one may think that it is part of its associated folk theory that any two things which are differ by less than some specified degree in ‘F’’s parameter of application are “sufficiently similar”. But then the alternative hypothesis collapses into my preferred hypothesis.

satisfied, the referent of an expression is what comes closest to satisfying both of these conditions.[35]

This said, however, I should remark that it is not clear that the putative alternative outlined really is an alternative to the conception of the folk-theoretic philosophers. Suppose that, as the causal theory of reference in its most usual form has it, 'water' refers to H_2O. The folk-theoretic philosopher can say that this is because one of the folk-theoretic beliefs about water is that it is the "stuff" causally linked, in the appropriate way, to our utterances of 'water'. Thus, the fact that 'water' refers to H_2O is explained along the lines of the folk-theoretic conception, unmodified. If our expressions refer to what stands in the appropriate causal relations to them, that may be because they should do so according to our folk theories.

We could have spoken a language where expressions corresponding to our ordinary expressions 'water', 'gold', etc. refer to the things that have the superficial characteristics of 'water', 'gold', and so forth. XYZ on Twin Earth would then have been in the extension of our expression 'water'. What makes it the case that we do not speak such a language? Here we cannot again appeal to causal considerations: for the question is precisely why our ordinary expressions refer to what they are appropriately causally linked to. It seems rather that the answer to this question would have to be provided by something like appeal to *what we take to be determinative of reference*, and from there it is a small step to saying that it is our folk theories are that determine that reference ordinarily has a causal element.

This leads us to a second point about how much of the folk-theoretic conception is necessary for the conclusions I have drawn here to follow. As I have characterized the folk-theoretic conception, the elements of folk theories are the *beliefs* that we draw upon when judging whether an expression applies in a given scenario. One may be skeptical that there really are beliefs that are sufficient to single out the referent of the expression; more specifically, that the broadly cognitive states that we draw upon when making judgments about the applicability of a particular expression are not always beliefs (and in particular are not even tacit beliefs). One may think, for example, that although the fact that 'water' refers to H_2O and not to whatever has the superficial characteristics of our water is determined

[35] Here is another reason why I might not want to take folk theories to be the only determinants of the reference of vague expressions. If folk theories were the only such determinants, vague expressions should be possible to eliminate in favor of non-vague expressions; but this does not seem plausible. (Thanks to Alex Byrne for emphasizing this to me.)

by facts about how we treat the term, we do not, properly speaking, have any *belief* to the effect that water is the stuff that is causally linked to our utterances of 'water'.

I am not sure to what extent the main proponents of the folk-theoretic conception are wedded to the idea that only beliefs can count as being part of folk theories; but at any rate nothing in the arguments I have presented here demands that the elements of folk theories be beliefs rather than the objects of some other appropriate cognitive state. This is however a very defensive move, and leads to a worry about what these other cognitive states might be. The response to the next issue that I will consider provides the answer also to this question.

I am concerned here with what elements of the folk-theoretic conception are necessary for the conclusions I have drawn from it to follow. So far I have concluded that I, as opposed to the folk-theoretic philosophers, need not commit myself to the thesis that reference is determined exclusively by folk theories, and I have urged that the elements of folk theories need not be beliefs. Now, thirdly, for a theoretically more interesting point, which I shall illustrate by considering the proposed analysis of vagueness (though the point could be made in connection with the other analyses as well). One potential problem with this analysis of vagueness is the following. The analysis says that tolerance principles belong to the folk theories of vague predicates. But some of us – some of the folk – have of course realized that these tolerance principles are not true. We can even suppose that all the folk could come to know that no tolerance principle is true (in primary school kids are taught about the sorites paradox), so that the folk theories associated with vague predicates change. It seems this would not, or at least need not, cause the meanings and extensions of vague predicates to change. Does this not refute our analysis of vagueness?

One might respond to this problem in two rather different ways.

First, one might take the statements of a folk theory to have an exclusively reference-fixing function, in the sense of Kripke. Then, so long as the references of vague predicates do not change, it does not matter that our opinions about the propositions that used to belong to the folk theories associated with these predicates change. (One problem about this suggestion, however, is that whereas it explains how old vague predicates do not change their meanings when tolerance principles are no longer believed, it does not explain another thing that seems possible: that new vague predicates can be introduced after tolerance principles are no longer believed.)

Second, perhaps more interestingly, we might respond to the problem by modifying our analysis of vagueness slightly. Suppose that the tolerance principles are so ingrained in our thinking about vague predicates (perhaps because they once were so firmly believed - but this causal explanation of their being so ingrained is not essential) that even when we are convinced of their falsity, we treat them as reference-determining, for example in that we take the extensions of vague predicates to be constrained by the fact that tolerance principles ought to come out as nearly true as possible. Then, arguably, the tolerance principles still are reference-determining, in the sense that they constrain reference in the way that statements of folk theories constrain reference. In order to be reference-determining, a principle need only be treated as reference-determining. The reason why the meanings of vague predicates need not change when the falsity of tolerance principles is generally recognized, is that the tolerance principles still might be treated as reference-determining.[36] (And in order to be treated as reference-determining, a principle need not be explicitly thought of as such: it is sufficient that speakers' judgments actually be constrained by their in effect trying to assign semantic values in such a way as to make the principle come out correct.)

I shall back up this general suggestion by two kinds of considerations: first I will consider the role of stipulations and then I will exploit the distinction drawn early on, between our explicit beliefs concerning a given concept and the (typically tacit) beliefs upon which we actually draw when making judgments about the application of this concept.

As is well known, stipulations may not be satisfiable. (This is illustrated by, for example, Arthur Prior's connective 'tonk', introduced by means of the stipulation that it make valid the inferences: from A, infer A tonk B; and from A tonk B, infer B.) And what is more important, stipulations can succeed in laying down meanings for expressions even when not believed to be true. I can introduce a new predicate into my language by means of some stipulation, reasoning as follows: "I have no idea whether these stipulations actually are consistent. But if they are, the expression introduced provides a valuable enrichment of the expressive resources of my language. I will hence assume, until proven wrong, that these stipulations are consistent." I do not here believe the stipulations to be consistent;

[36] We may describe the modification in two ways We can say either that, contrary to what was claimed earlier, a statement need not actually be believed to be part of a folk theory in order to be reference-determining, or that a claim can be reference-determining without being part of the folk theory. This is only a matter of terminology. But I find it more natural to say that a claim can be reference-determining without being part of a folk theory

but still the stipulations may well suffice to render the expression meaningful. In fact, the stipulations seem to work in much the same way as do statements of folk theories according to Jackson and Lewis: they determine reference, in the sense that if they are satisfiable, the expression introduced makes them true (and, presumably, if they are not satisfiable, the expression introduced refers to what makes them as close to true as possible).

More clearly relevant to the issue of tolerance principles, is that it may be that even when we are convinced that tolerance principles are all untrue, we cannot help but think of the world as if tolerance principles are true – in much the same way that (I suppose) even someone who denies the existence of ordinary middle-sized dry material objects cannot help but think of the world as containing ordinary middle-sized dry material objects. Perhaps we should *not* say, in the scenario envisaged, that we still *tacitly believe* that some tolerance principles are true (or that the ontological nihilist still tacitly believes that there are ordinary middle-sized dry material objects). But still, whatever cognitive attitude we continue to have toward tolerance principles may be sufficient for tolerance principles to be reference-determining for vague predicates.

The modified analysis of vagueness is that tolerance principles, whether parts of folk theories or not, are reference-determining for vague predicates. They are among the claims such that the semantic values of vague predicates are constrained by the fact that, if these claims are true, then the semantic values of vague predicates are what satisfies all of these claims, and if these claims are not all true, then the semantic values of vague predicates (if any) are what comes closest to satisfying these claims. It can, but need not, be by virtue of being part of a folk theory that a claim is in this way reference-determining.

This point extends beyond the analysis of vagueness. Generally, the diagnoses that we have given of philosophical problems do not require that there be statements that are part of folk theories in the sense of Jackson and Lewis, but instead only that there be statements that are treated as reference-determining. This is a significantly weaker claim.

3.9 Concluding remarks

In this chapter, I have shown that considerations of how reference is determined can yield the same kind of conclusions regarding the problems we have considered the paradoxes and the fission problems as was reached in other ways in the first two chapters.

In the first two chapters, I outlined a particular thesis about semantic competence – that sometimes we accept untrue claims by virtue of our semantic competence – and argued for this thesis mainly through showing how it could be used to resolve our puzzles. But now we have seen that certain theses about reference-determination apparently yield these conclusions on their own. What, then, is the status of the issue regarding semantic competence?

There are two points to be made here. The first (weaker) point is that the theses about reference-determination do not themselves account for all the data appealed to in the first two chapters. There is a sense of puzzlement about the paradoxes and the fission cases which the theses about reference-determination seem to do nothing to explain. The second (stronger) point concerns the content of the talk of "treating as reference-determining". I concluded that, in order to be reference-determining, a claim need not be believed (and, as noted early on, neither is it sufficient that the belief that it is true is firmly held and widely shared). A second point is that, as highlighted especially in the last section, there is an issue of what determines a claim as reference-determining. For most of chapter I considered folk theories, and how they determine reference. But as emphasized in the last section, there may be many other sources of a principle's having the status of reference-determining. One such source may be precisely that it is regarded as part of semantic competence that one is prepared to accept this principle, unless one has, for example, what one takes to be defeating evidence against it. It should be clear how this ties in with the claims about semantic competence in the previous two chapters.

Chapter 4

Truth, the Liar and Universality

4.1 Introduction

The discussion in the first chapter concerned one theme that arises from Tarski's discussion of the liar paradox: the issue of whether natural language is in some sense 'inconsistent'. A different theme from this discussion is that of whether natural language is 'universal'. As will be demonstrated later in this chapter, it is hard to say exactly what universality is supposed to be; but *very* roughly, a language L is universal if everything that can be expressed can be expressed in L.

I shall end up defending a thesis reasonably stated thus: *truth is inexpressible*. In fact, I shall argue that fairly *conservative* views on truth and the liar have this (surprising) consequence. (Or rather, I shall defend a disjunction, where this is one of several disjuncts. But the other disjuncts are very similar to this one in character.)

Let us start, however, with issues more mundane.

Suppose that the conditions of Tarski's theorem hold for English, in that (a) some expressions of English express the classical truth functions, and (b) all expressions of English can be named in English. What Tarski's theorem then shows is that the *disquotation schema*,

$\ulcorner \mathrm{p} \urcorner$ is true if and only if p,

(where for 'p' we can insert any declarative sentence of English[1]) cannot be valid. More generally, no predicate can satisfy the *T-schema*

[1] I ignore indexicality, context-sensitivity, etc. except where indicated.

$\ulcorner p \urcorner$ is T if and only if p,

where a predicate ***satisfies*** the T-schema if a valid schema results when this predicate is substituted for 'T'.[2] A ***property*** satisfies a schema just in case a necessary condition for a predicate to express it is that the predicate satisfy the schema. A property is expressible in L just in case some predicate of L expresses it. A property is absolutely inexpressible just in case no predicate of any possible language expresses it.

Now recall that many theses of philosophy of language appear to assume that the disquotation schema is valid. For example, all meaningful sentences are assumed to have truth-conditions, statable by true sentences that are instances of the disquotation schema. More controversially, truth is regarded as that at which assertions 'aim'; a sentence or proposition (in some sense) says that its truth condition is satisfied; and the truth conditions of a sentence are regarded as its content. All these theses appear to, and are generally taken to, require that a sentence S and a sentence which says that S is true have the same truth-value; and hence, in a classical framework, the disquotation schema should come out valid.

So if the conditions of Tarski's theorem hold for English -- if, as we may put it, English is a *Tarskian language* - then philosophy of language is in trouble. (Of course, many philosophers of language would deny the assumptions I have said are made in philosophy of language; but I shall disregard them and simply say that it is philosophy of language, as a whole, that is in trouble. It is an overstatement, used for effect.)

Some people may want to protest here that it is for more straightforward reasons rather clear that the disquotation schema cannot be valid. Phenomena like presupposition failure and reference failure give rise to sentences being neither true nor false and in most many-valued systems, when a sentence is neither true nor false, any biconditional in which it occurs will fail to be true. However, even should this be granted, the problem to which I want to draw attention to still arises, albeit in a slightly generalized form. Whether we hold that in a many-valued framework, a truth predicate takes all truth-values other than truth into falsity, or we hold that the sentence resulting from predicating truth of a sentence S always should have the same truth value as S, we can state conditions such that if English satisfies these conditions, English cannot express truth. Let us say that if English satisfies

[2] See e.g. McGee (1991), p 25f

these conditions, English is ***pseudo-Tarskian.***

The problem the liar poses for philosophy of language will be the main focus of the paper. I shall argue neither for the assumption that English is a Tarskian or pseudo-Tarskian language, nor for the assumption that truth will have to satisfy the T-schema (or one of the other conditions mentioned) to do its theoretical job. The reasons for omitting such arguments are the following. First, the problems that I will address are likely to arise in some form even if these assumptions are not true. If truth, conceived as satisfying the condition that predicating truth of a sentence S should result in a sentence equivalent in truth-value to S, is expressible, then it is not generally expressible that a sentence has one of the truth-values other than truth; that a sentence is ***untrue.*** So the problems I shall consider would arise with respect to untruth. Second, even if English is not a Tarskian or pseudo-Tarskian language, it certainly *could* have been. But fundamental assumptions in the philosophy of language should not be dependent on particular assumptions about our actual language.[3]

4.2 Universality

If, as incautiously stated in the introduction, the philosopher of language really is committed to the disquotation schema being valid, then the possibility of English being Tarskian really does refute her.

But consider: it should not or need not be essential to the philosophical theories of language that the property fulfilling such-and-such a function in the theory be expressed by a predicate spelled t-r-u-e. It would appear to be a thesis of a rather subsidiary character that our predicate 'true' of ordinary language – or *any* predicate of ordinary language – expresses (for example) the property that assertions aim at. Insofar as philosophers of language subscribe to that thesis, that ought to be clearly separated from their general concerns. (Dummett remarks somewhere that the real question of truth in ethics is not whether we predicate truth and falsity of ethical sentences, but rather whether, if we do, such predication could fill the same purpose in ethics as elsewhere.)[4]

[3] Since introducing the possibility of English being pseudo-Tarskian does not fundamentally affect the situation, I shall often henceforth consider only the possibility that English is Tarskian

[4] It is sometimes said that although natural languages are not universal (they may not be able to express all there is to be expressed) they are at least semantically universal: they can express all of their semantic properties. The main argument for the claim appears to be that natural language contains the means to

Generally, scientific theories (in a broad sense) should be expected to employ homonyms of expressions of ordinary language, related in meaning to their counterparts in ordinary language, but by no means coextensive with them. German is one language and Swedish and Norwegian two, given how we commonsensically individuate languages. The notion of 'language' employed in linguistics does not respect such judgments; and still it seems wrong to begrudge the linguist the word 'language'. Similarly, when a philosopher of language employs the word 'true', it would be unnecessary and unfair to immediately assume that the word 'true' in her mouth is coextensive with the word 'true' of the vernacular.

So when philosophers of language say, for example, that truth is that at which assertions aim, and hold further that any property at which assertions aim will have to satisfy the T-schema, then the core content of their claims should be taken to be that there is some property to which assertions aim, that this property satisfies other theoretical constraints truth is taken to satisfy, and that this property satisfies the T-schema. This core content may be true even if the subsidiary thesis that 'true' as actually used satisfies the T-schema is not. (Philosophers of language may have wanted to defend the so-called subsidiary thesis as well. My point is only that this can be separated from what should be regarded as the main concern.)

These points are, I believe, important. For it is all too common in the liar literature to assume that describing the ordinary property of truth and describing a property of truth fit to do theoretical work will have to come to the same thing.

But arguably, these points do not speak to the *main* problem about maintaining the view that truth satisfies the T-schema even if English is a Tarskian language. For if English is a Tarskian language, then no predicate of English satisfies the T-schema, not as a *contingent* matter but as a matter of *necessity*: not only is there no such predicate of English but such a predicate could not be added without affecting the semantic values of other expressions of English. This point can be developed so as to spell trouble for philosophy of language, in two ways.

First, one may hold that all properties are in principle expressible in English. Since a would-be property satisfying the T-schema is not in principle expressible in English, there

name all of its expressions and it contains truth and reference predicates (see e.g. Simmons (1993), p. 15f) But this assumes that the semantic properties of English are all specifiable using only the truth and reference predicates of English.

is then no such property.

Second, one may hold that however matters stand with the stronger claim about all properties being in principle expressible in English, surely any property of which we have a concept is in principle expressible in English; and the philosopher of language would be uncomfortable denying that we have a concept of the property of truth.

I shall consider these problems in turn.

Claims about every property being expressible in English are unclear in a very important respect. First, they can be interpreted *weakly*, as follows. Given the way we individuate natural languages, a natural language can survive rather drastic changes: we can, for example, add new expressions to English, and still it would be called English. If, when one says that every property is in principle expressible in English, one relies only on the fact that for every property, a predicate expressing that property could be added to English and the resulting language still be called English, then the thesis that everything can be expressed in English is rather trivial. It is, for example, compatible with it being the case that for every time *t* at which English exists, I can add to English as of that time a predicate satisfying the T-schema for English as of that time (it is only that I then necessarily *expand* the language).

So for the claim about every property being in principle expressible in English to pose trouble for the philosopher of language, it must be interpreted *strongly*, as being about what is expressible in English as it exists today. Let us accordingly interpret it that way.

Notice that the claim under consideration – the claim that every property is expressible in English, interpreted strongly – reasonably can be regarded as an explication of the much discussed thesis that natural languages are 'universal'.[5]

I shall evaluate this universality claim with special reference to the issue of whether English thought of as Tarskian, not containing any predicate that satisfies the T-schema, can be expanded with such a predicate (with the resulting language being distinct from English).[6]

Suppose that the philosopher of language suggests that when we do semantic theorizing

[5] The classic discussion of universality is in Tarski (1983) (first published in 1935). Other discussions are in e.g. Herzberger (1970 and 1981), Martin (1976), McGee (1997) and Simmons (1993).

[6] It is occasionally assumed to be trivial that natural languages can be so expanded. But though I will end up defending the claim that natural languages can be expanded as indicated, another intended lesson of the discussion to follow is that this is by no means *trivial*.

about English, we can and must do so in a metalanguage stronger in expressive resources than English. In particular, we can and should add to the language in which we state the theory a predicate, '**true**', say, satisfying the T-schema restricted to English. (By the qualification "restricted to English", I mean: when for 'p' we can insert only sentences of English.) What could be the trouble with such a suggestion? After all, the predicate '**true**' is applicable only to sentences of English, and thus no liar sentence for this predicate can be constructed.

Here is one potential problem. It is obviously essential to the proposal on the board that we should be able, in principle, to speak the expanded language. Suppose we speak it in 2005, and then consider the predicate of English "falls under the predicate '**true**' of the language we speak in 2005". This predicate would seem to be coextensive with '**true**' (under the circumstances envisaged). But no predicate of English can satisfy the T-schema; hence '**true**' cannot satisfy the T-schema either.

As the reasoning just presented can be generalized to constitute an argument for universality, let us dub it the *universality argument.*

If the universality argument is sound, it can be used to refute the vast majority of all proposed accounts of the liar paradox. For the vast majority of such accounts employ notions which, on pain of contradiction, cannot be expressed in the object language, but only in a richer metalanguage. As Kripke (1975) himself points out, the semantic notions he employs in the metalanguage, 'grounded', 'paradoxical', etc., cannot be expressed in the object language, on pain of contradiction. Kripke tries to justify this state of affairs by saying that

> ...in contrast to the notion of truth, none of these notions is to be found in natural language in its pristine purity, before philosophers reflect on its semantics (in particular, the semantic paradoxes). If we give up the goal of a universal language, models of the type presented in this paper are plausible as models of natural language at a stage before we reflect on the generation process associated with the concept of truth, the stage which continues in the daily life of nonphilosophical speakers.[7]

If sound, the universality argument shows that these considerations are simply irrelevant.

[7] Kripke (1975), in Martin (1984), p. 80n34.

English can be shown to be universal independently of considerations concerning the conceptual sophistication of speakers. Just substitute 'grounded' or 'paradoxical' for '**true**'.

In the theory of Gupta and Belnap – the ***revision theory*** of truth – the sentences of our language are classified as ***stably true***, ***stably false***, and ***unstable***. (And the paradoxical sentences come out as unstable.) Like Kripke, these theorists note that these semantic notions of the metalanguage of their theory cannot be expressed in the object language itself, on pain of contradiction. Just substitute "not stably true" for '**true**' and you land in trouble. In their (1993), Gupta and Belnap discuss whether it is a problem that some notions employed in their metalanguage cannot be expressed in the object language, on pain of contradiction. They conclude that it is not, on the ground that a universal language is an unattainable ideal.[8]

The intuition that I in effect exploited when briefly setting out the motivation for the claim that '**true**' can satisfy the T-schema restricted to English, is that we can *first* settle the truth-values of all sentences of English and *then*, as it were, tag on an expression that satisfies the T-schema restricted to English. What spells trouble for this idea, however, is that some English expressions, like 'true', apparently are meaningfully predicable of all sentences or utterances; in particular, also those of the extended language. This seems to entail that we cannot determine the truth-values of all sentences of English with respect to all contexts, independently of having determined the truth-values of the sentences of the extended language. The point bears directly on the question of the universality of natural languages. Semantic predicates of formal languages are meaningfully predicable only of a restricted range of sentences: those in the domain of the theory. Presumably, the domain can in principle be all-inclusive: but in the typical case, it is not. Semantic predicates of natural languages are, however, typically meaningfully predicable of expressions of all languages, as evidenced by the meaningfulness of the predicate "falls under the predicate '**true**' of the language we speak 2005". This makes for a prima facie difference between natural language and (typical) formal languages. Anything we can say in some language L′ we can say (in a rather roundabout way) in L, when L is a natural language, merely by appending L's counterpart of 'true' to what we say in L′. This is not always so when L is a formal language. For it may be that no predicate of the formal language is meaningfully predicable of sentences of L′. This seems to be a basis for the judgment that natural

[8] Gupta and Belnap (1993), pp. 256ff.

languages are universal.

The universality argument is unsound, however, and given the history of the notion of universality, it is so in a rather ironic way. Tarski claimed that it is the universality of natural language that is responsible for the liar paradox to arise (and hence also for natural language to be 'inconsistent'). But provided our language satisfies the conditions for being a Tarskian language, it is actually the liar reasoning that *prevents* it from being universal. For in order for the universality argument to be sound, it must be the case that for every predicate 'F', 'F' is coextensive with "falls under the predicate 'F'". But under the assumption that our language is Tarskian, the liar reasoning blocks that.

Though the universality argument should in the end be rejected, due to the liar reasoning, it must be emphasized how much the reasoning of the argument still demonstrates outside liar contexts. In (1993), Keith Simmons argues that the idea that natural language is universal (in the sense outlined here) should be rejected. He says, "Consider, for example, the sets in the ZF hierarchy. For each set, there is a distinct concept – say, *being a member of that particular set.* Given certain assumptions about natural languages (in particular, about upper limits on the size of vocabularies and on the length of sentences), these concepts would outrun the expressive capacity of any natural language".[9]

Even though I reject the universality argument, because of the liar reasoning, the reasoning of the argument still demonstrates that each of these concepts can be expressed in English. We can get at each of these concepts by indirect means: consider "falls under the concept chosen by God to be the most divine". Since God in principle could choose any one of the concepts described by Simmons as the most divine, each one of these concepts is expressible in English.

Simmons might object that when he talks about the possibility of expressing a concept, he means the possibility of there being a predicate that, as uttered in any possible situation, expresses this concept. But this is not the sense of expressibility most centrally at issue in the liar: for example, if our language is Tarskian, then truth conceived as satisfying the T-schema is inexpressible, in the sense of not being expressible by any predicate as uttered in any possible situation.[10]

[9] Simmons (1993), p. 15.

[10] Furthermore, after having rejected the claim that natural language is universal in the sense outlined, Simmons considers the ostensibly weaker claim, ascribed to Tarski, that "*if* a concept is expressible in *some* language, then it is expressible in any natural language" (p. 15). Though Simmons doubts the truth of this

4.3 How the property of truth is employed in philosophy of language

So it is possible to add to English a predicate '**true**' satisfying the T-schema. Hence one need not despair about the existence of a property satisfying the T-schema. Of course, we should not assume that every predicate expresses a property; but supposing there to have been a presumption in favor of 'true' (as used in theoretical discourses) expressing a property, we can at least conclude that there is no reason to abandon that presumption because of the liar paradox.

What about the second worry mentioned above, that in order to do its theoretical work in a philosophical account of language and meaning, truth will have to be grasped by speakers of natural language?

I will consider two versions of the idea that meaning is truth conditions; one of them I will associate with the early Donald Davidson (the author of "Truth and Meaning"), the other with Michael Dummett. As I shall argue, neither of these two versions of the thesis that meaning is truth conditions requires that the property of being true be expressible or that it must be the truth predicate of the object language that occurs in the statement of the meaning-theory. I shall not go into Davidson- and Dummett-exegesis, but only present enough of their conceptions of truth-conditional meaning-theories to show that under these influential conceptions of what such theories are supposed to do, it is not required for the workability of such a theory that the property of being true be expressible.

Let us first discuss the Davidsonian conception. Davidson does not assume that speakers are in any sense governed by knowledge of any relation between truth and meaning, or between truth and assertion. All that is assumed concerning truth is that it should be possible to have, in the language employed by the meaning-theorist, a predicate that satisfies the T-schema restricted to the object language, the language of those being 'interpreted'. The central feature of the Davidsonian approach is that the meaning-theorist introduces into the language she employs a *new* predicate, '$\delta(x)$', and constructs a meaning-theory, the (relevant) theorems of which are of the form:

claim, he does not take his argument from the ZF hierarchy to pose problems for it. But he should. For what is the scope of "any language" in the quote presented? It certainly sounds as if Simmons means all possible (natural) languages. But for each concept *being a member of [that particular set]*, there surely is some *possible* natural language where this concept is expressed; and by the claim attributed to Tarski, all these concepts are then expressible in all natural languages.

$\delta(s)$ if and only if p,

where s is a structural-descriptive name of a sentence of the object language and p is intended to be the translation of this sentence into the metalanguage. The idea is that we shall obtain a theorem of this form for every sentence of the object language. What the theory says is then determined by what conditions are laid down on when a sentence falls in the extension of '$\delta(x)$'. One must not *presuppose* that 'p' is the translation into the metalanguage of s, but it is by laying down the right conditions on '$\delta(x)$' that we obtain that. The idea behind the construction, as Davidson presents it, is that '$\delta(x)$' is to turn out to be coextensional with the predicate 'true'. But Davidson here simply disregards the liar problem, and assumes 'true' to satisfy the T-schema.[11] The core of Davidson's claim is that '$\delta(x)$' should satisfy the T-schema restricted to the object language, and as '$\delta(x)$' is not itself part of the object language it can do so.[12]

It may reasonably be held that the reason why the notion of truth need not be fully expressible for Davidson's purposes is that Davidson takes the notion to be only a tool of the meaning-theorist: the meaning-theory is not in any sense meant to represent knowledge speakers actually have, but instead is meant to state information knowledge of which would suffice for knowledge of the language that the theory is for.[13] It may therefore be of some interest to note that even when the notion of truth occurs in a more substantive way in theorizing about meaning, the notion need not be fully expressible.

Michael Dummett (as opposed to the early Davidson) assumes that the notion of truth in a sense plays a role in a speaker's understanding and use of her language.[14] I will show that the notion can play also the role Dummett assigns to it without being expressible. Dummett assumes that to assert is to present as true, and that this goes some way toward explaining what is distinctive about assertoric practice. Truth is then conceived as the property such that speakers within assertoric practice seek to utter only sentences with this property. The idea behind this account of assertoric practice is of course not that speakers, consciously or unconsciously, internally apply the predicate 'true', or the concept *true*, to

[11] Of course, in other contexts, Davidson does take explicit note of the liar problem.

[12] When the metalanguage, or rather the metalanguage minus '$\delta(x)$', is identical with the object language, English as it may be, '$\delta(x)$' has a better claim than 'true' to satisfy the T-schema restricted to the object language, and thus to express truth for the object language in a theoretically interesting sense.

[13] Remember, it is the early Davidson we are talking about.

[14] In the end, Dummett thinks that meaning-theories should not be truth-conditional. But Dummett still has one of the most elaborate conceptions of what is essential to a truth-conditional account of meaning, and we can legitimately talk about what a truth-conditional meaning-theory would look like for Dummett

every sentence they utter with assertoric force. Instead the idea is that in order successfully to participate in assertoric practice – and, especially given the central role accorded to that practice, to participate in linguistic practice in general – one must act in accordance with the norm of asserting only what has the property of being true. In a sense, speakers do have an understanding of the notion; however, this understanding manifests itself not in a capacity to form judgments in which the concept occurs, but it manifests itself precisely in successful participation in assertoric practice.

What I say here is *independent* of the thesis that truth in English is is not expressible in English. Sometimes the objection is raised against truth-conditional meaning-theories that, for example, *small children*, who (allegedly) obviously do not possess the concept of truth, are capable of participating in assertoric practice or, generally, linguistic practice. But then, it is claimed, knowledge of meaning cannot be analyzed in terms of truth-conditions: for here we have knowledge of meaning but not knowledge of truth-conditions.[15] This objection seems to me utterly misguided. The basis for the claim that little children do not possess the concept of truth appears to be that little children do not make judgments of which the concept true, or something expressing it, is a constituent, and do not possess a predicate 'true' expressing this concept. But this is not what is being presupposed by the proponent of the thesis that knowledge of meaning is knowledge of truth-conditions. Rather, insofar as it is presupposed that every speaker of a language possesses a concept of truth, possession of this concept is supposed to be completely manifested by the speaker's ability to participate successfully in assertoric practice. But this is consistent with truth's not being fully expressible by any predicate.[16]

[15] See for example Scott Soames (1989), p. 578. Soames says,

> Knowledge of truth conditions, as I have described it, presupposes possession of a metalinguistic concept of truth. Thus, the claim that such knowledge is necessary for understanding meaning entails that no one can learn or understand a language without first having such a concept. But this consequence seems false. Certainly, young children and unsophisticated adults can understand lots of sentences without understanding 'true' or any corresponding predicate.
>
> Must they, nevertheless, possess a metalinguistic concept of truth, even though they have no word for it? I don't see why.

Oddly, Soames then goes on to say that

> The child will get along fine so long as she knows that 'Momma is working' is to be assertively uttered only if Momma is working; 'Daddy is asleep' is to be assertively uttered only if Daddy is asleep; and so on.

Soames is apparently happy to ascribe to the child the concept expressed by "is to be assertively uttered"; but the reasons Soames hints at for thinking that little children don't possess the concept of truth surely carry over to the concept of assertive utterance, provided they are sound in the first place.

[16] I should add that Dummett's view on the liar appears to be precisely that it forces us to give the truth

4.4 Burge on meaning-theories and universality

In (1979), Tyler Burge addresses the problem that if we think that truth satisfies the T-schema, or some variant, and that language L is classical, then truth for L cannot be expressed in L itself. Burge says, presenting the problem:

> Some writers seeking to apply Tarski's theory have argued that natural languages are not universal [in the sense that "any word in another language can be translated"], holding for example that our predicate 'true in Urdu' cannot be translated into Urdu.
>
> The reasoning seems to go somewhat as follows. If Tarski's theory is to be applied to natural language, one must take a semantical system like his (including semantical postulates) as standing for or representing a natural language. A truth predicate in a natural language (e.g., 'true in English' or 'true in Urdu') should be represented by a predicate constant, with a fixed extension (e.g. all the true sentences of English) determined by this predicate's form and meaning. If Tarski's theory is to be applied, this constant must be governed by the usual semantic postulates. But, by Tarski's theorem, any such predicate for evaluating all the sentences of a semantical system cannot be introduced and used in the semantical system (with the usual semantical postulates) on pain of contradiction. So if Tarski's theory is to be applied, a truth predicate like 'true in English' cannot be allowed or cannot occur in English itself.[17]

According to Burge's own theory of truth, 'true' is indexical, with context supplying a particular index for each occurrence of 'true'. Thus, there is in a sense no univocal truth predicate for the whole language. This dissolves the problem, Burge says, for then truth in English can be expressed in English. (There is no univocal concept of truth in English, but all the various concepts $true_i$ can be expressed in English.)

However, as presented and justified earlier in Burge's paper, Burge's contextualist theory of truth is a theory of how our ordinary expression 'true' works. It is a significant claim that the ordinary notion of truth is also the notion fit to do theoretical work in semantics and the philosophy of language. Independent arguments would be needed against the idea

theory for a language in a richer metalanguage. See Dummett (1991), p. 72.

[17] Burge (1979), p. 196.

that it is a Tarski-style predicate constant satisfying the T-schema that is fit to do such work. Burge does not supply such argument; and neither, apparently, does he appreciate the need for them.

Burge also has a more specific criticism of the reasoning presented in the quote. He does not think that natural language is the sort of thing that can be said to be consistent or inconsistent, and so he thinks that consistency restrictions cannot play a role for what can and cannot be expressed in a particular language.

Burge's reason for thinking that languages cannot be inconsistent is that it is theories, not languages, that assert things; and to believe differently is to unjustifiably assimilate languages to theories. There are two different points here that must be separated. Burge could be making the simple ordinary-language point that we do not ordinarily speak of languages as being, or having as part, a body of claims which may be consistent or inconsistent. This point is correct, but would not suffice for Burge's purposes. For compare two languages, one being Tarskian and the other not. Neither contains a predicate satisfying the T-schema. But intuitively, the latter *could* contain a predicate satisfying the T-schema and the former could not. This point holds independently of whether languages can be said to assert things or not. What Burge would need is that one cannot endow "language L is inconsistent" with meaning in a natural and reasonable way.[18]

4.5 Truth simpliciter

I have argued, first, that it is possible that our language is not universal, and a predicate expressing truth for it can exist only in a metalanguage, and, second, that even should this possibility be actualized, truth can play an important role in our accounts of meaning.

Now, however, I shall consider a much tougher problem for the suggestion that truth is expressible in a stronger metalanguage. So far I have tacitly restricted the discussion to the expressibility in English of truth-in-English; or generally, the expressibility in L of truth-in-L. But when theorizing about the nature of truth, we not only theorize about truth-in-L for specific languages L: we also theorize about truth *simpliciter*, truth for variable L. A philosopher of language is likely to (wish to) hold that there exist not only various properties true-in-L but also a property of being true, simpliciter. For example, when

[18] See chapter 1, footnote 13, and section six of this chapter.

Dummett describes assertions as those speech acts which aim at truth, he does not only wish to say that assertions-in-English aim at truth-in-English, assertions-in-German aim at truth-in-German, etc.: he wants to say that there is some unifying feature of assertions across languages.

The problem, roughly put, is that it may seem that if truth-in-English is expressible only in a metalanguage, truth for that metalanguage only can be expressed in a further metalanguage, etc., then *truth simpliciter cannot be expressed in any language.* (One apparent problem with the emphasized claim, a problem to which we shall return, is that if true, it cannot itself be expressed. But it appears that we have just now managed to express it.)

However, this is not a particularly precise statement of the nature of our problem. Truth-in-L fails to be expressible in L only under rather specific hypotheses regarding what else is expressible in L. For all that we have seen reason to believe, there could be a language L* sufficiently limited in expressive resources that truth-in-L* can be expressed in L*, and truth for all other languages can be expressed in L*. I have supposed, for the purposes of this chapter, that English is sufficiently rich in other expressive resources that truth-in-English cannot be expressed in English, given that truth satisfies the T-schema. Even should this supposition be true, this does not rule out the existence of languages where truth can be expressed and where there are instead some other limitations on expressive resources.

But even should there be a language L*, otherwise impoverished in expressive resources, which manages to express truth for all languages, including itself, the basic problem posed by the presumed inexpressibility of truth remains.

Suppose that the reason L* can contain a predicate expressing truth-in-L* is that the language does not contain any expression expressing strong negation (the strong negation of a sentence is true whenever the sentence is either false or else has some other semantic value distinct from truth) for its own sentences. Then there cannot (under natural assumptions) be an expansion L** of L* where strong negation for L* can be expressed.[19] Recall the universality argument, which said that if English cannot contain a predicate satisfying the T-schema restricted to English then neither can an expansion of English. That objection has been satisfactorily answered. But the corresponding objection to the existence of a language

[19] Thus, for example in Kripke's system, with strong Kleene three-valued logic as the underlying logic, truth can be expressed (under one reasonable assumption about what it is for truth to be expressed in a three-valued language: that predicating truth of a sentence that is neuter results in a sentence that is neuter), but strong negation cannot be expressed.

such as L** seems valid. Suppose that there is a predicate 'Neg(x)' of L** expressing strong negation for L*, and that we now speak L*, and that at some time t in the future we speak L**. Consider then the predicate of L* "falls under the predicate 'Neg(x)' of the language we speak at t". Under the unfavorable circumstances mentioned, this predicate expresses strong negation for sentences of L*, under the assumption that the semantics of 'falls under' is linked to semantics of 'true'. (And this assumption is, moreover, not necessary: one could simply speak of the truth of the sentences of the form "Neg(...)" of L**, where for '...' sentences of L* are inserted.) But by hypothesis, L* cannot contain such a predicate.

Similar remarks would apply to other ways of conceiving of the expressive limitations of L* which would allow it to contain a predicate that expresses truth for all languages. Accordingly, let us simply *assume* that it is *truth* that cannot be expressed. Even if this is not true, similar problems will arise with respect to some property (or whatever it is we should say that negation and like operators express), for it will be the case that some particular property or function cannot be expressed in any language. Some problem of inexpressibility is sure to arise. What I shall argue is rather that we can accept inexpressibility.[20]

First, note that the most popular theories of the liar paradox, for example the theories of the liar earlier considered – Kripke's theory and the revision theory – face the same problem we do, only in a different guise. For these theories too relegate certain notions to the metalanguage. For example, in the case of Kripke's theory, grounded-in-L cannot

[20] One may wonder whether not the problem noted in the last few paragraphs may be resolved by appeal to *context-sensitivity*. Thus, one may say that all properties and functions are expressible, though not in the same context.

But the same reason that was adduced for why strong negation (simpliciter) cannot be expressed provided truth (simpliciter) can be expressed remains if we assume truth (simpliciter) can be expressed only in certain contexts. For let C be one of these contexts, and suppose that in C, I use a predicate of the form "falls under the predicate *so-and-so* as used in context *such-and-such*", where contingent facts about the world make this predicate coextensive with the predicate that in context *such-and-such* would express strong negation. The only way to block this argument would be to say that when I use a predicate which, perhaps given contingent facts about the world, expresses strong negation simpliciter, I thereby make it the case that I am no longer in a context where truth simpliciter can be expressed. But I know of no non-*ad hoc* reason to believe this kind of context-sensitivity to obtain.

And more importantly, even should both truth simpliciter and strong negation simpliciter be expressible in some context in some possible language, the basic problem remains that there cannot be a language in which truth simpliciter is expressed by a constant and a language in which strong negation simpliciter is expressed by a constant -- though one of these notions may be expressed by a constant in some possible language provided the other is not. (I should emphasize that I am simplifying matters quite a lot when presenting the matter as if it were only a matter of whether it is truth or strong negation that is expressible. Other options are in principle available. It has been suggested, for example, that it is absolutely unrestricted quantification that cannot be expressed.)

generally be represented in L. But then a general property of groundedness cannot be expressed in any language (unless there is some other property that cannot be expressed in any language).

It might be thought that there is a significant difference between groundedness and truth, in that there is a rather clear intuition regarding truth that there is also a property of truth simpliciter, whereas we are in a position to simply deny that there is a property of *groundedness simpliciter*, since the notion of groundedness is a technical innovation, However, the notion of truth we are concerned with here is as much a technical notion as the notion of groundedness, so if it can reasonably denied that there is such a thing as groundedness simpliciter, it can presumably also be reasonably denied that there is such a thing as truth simpliciter. What is more, it can hardly be denied that from Kripke's explanation of what groundedness is we have enough information to apply this notion to any language. But then it appears that Kripke has explained groundedness simpliciter.[21]

Certain theorists writing about the liar paradox have, partly as a reaction to Kripke and other theorists who have to appeal to stronger metalanguages, sought to present solutions to the liar which eschew such appeal. A prominent example is Vann McGee.[22] It may be thought that these theorists avoid the problem we are now considering. For their solutions to the liar do not force them to countenance any properties that cannot be expressed in English (as they conceive it). However, this would be mistaken. For what would be needed to avoid the problem we are considering is not merely a semantic theory employing only expressions for properties that can be expressed already in the object language, but a semantic account given which there are no properties that can be expressed only in the metalanguage. (In the absence of such an account, it is unclear what is the significance of semantic theories employing only notions that can be expressed in the object language, or why such semantic theories should be preferable to the alternatives. If there can be languages that do not

[21] When characterizing groundedness, Kripke says

> ...if a sentence....asserts that (all, some, most, etc.) of the sentences of a certain class C are true, its truth value can be ascertained if the truth values of the sentences in the class C are ascertained. If some of these sentences themselves involve the notion of truth, their truth value in turn must be ascertained by looking at *other* sentences, and so on. If ultimately this process terminates in sentences not mentioning the concept of truth, so that the truth value of the original statement can be ascertained, we call the original sentence *grounded*; otherwise, *ungrounded*. (Kripke 1975, p. 693f)

If you understand truth simpliciter, this characterization will certainly enable you to understand groundedness simpliciter.

[22] See McGee (1991).

contain the resources to express all the semantic properties of their expressions, then we need a special reason to think that our language is not among them.)

Second, notice that the soundness of the objection from inexpressibility does not in any way depend on identifying the property that satisfies the T-schema (for every language) as truth. So to avoid distraction, let us temporarily call this property 'Tr' rather than truth. By our reasoning, 'Tr' cannot be expressed in any language.

Earlier, when discussing only truth in English, I considered the theories of Davidson and Dummett, and asked whether they required that truth in English be expressible in English in order for truth to do its theoretical job. The answer was no. Now a similar, but perhaps tougher, problem arises: is it not necessary for truth (simpliciter) to be expressible in some language in order for it to do the theoretical job assigned to it?

Most of the points made earlier carry the same force now. The one new problem is that Dummett cannot be partly defended by appeal to the possibility of speakers to add to their language (at t) a predicate expressing truth for their language (at t). Speakers can in that way manifest grasp of truth-in-L, for all specific L, but they cannot in that way manifest their grasp of truth simpliciter.

One could say, at this point, that a speaker understands the property of truth simpliciter, if she is willing, regardless of any details of the language, to take as expressing truth in some specific language L a predicate satisfying the T-schema restricted to L.

An alternative approach is as follows. In chapter 1, I argued that the paradoxes like the liar and the sorites bring to light that our language is inconsistent in the following way: our competence with expressions of our language itself disposes us to accept principles (sentences and inference forms) which jointly lead to inconsistency. In particular, the liar may arise because our competence with the logical particles disposes us to take these particles to satisfy the proof-theoretic rules that characterize the classical logical operations, and our competence with 'true' disposes us to accept that 'true' satisfies the T-schema. The principles such that a speaker's competence with an expression requires her to be disposed to accept them may be called constitutive of meaning. When a language is inconsistent as outlined, the semantic values of expressions of the language cannot be such as to make all principles constitutive of meaning true and valid, respectively. Rather, I suggest, the semantic values are such as to make these principles as close to correct as possible.[23]

[23] For details, see chapter 1.

If this is accepted, we can say that a speaker has a concept of the property of truth simpliciter, if she is competent with a predicate, 'TRUE', say, a meaning-constitutive principle for which is that, whichever sentence S it is applied to, TRUE(S) is equivalent in truth-value to S.

The apparatus motivated by talk of inconsistent languages can also be used to introduce a weaker notion of expressibility, which we may call *quasi-expressibility.* I will explain it with special reference to the notion of truth. The property of truth in L is *expressible* in L only if some predicate of L satisfies the T-schema restricted to L. It is *quasi-expressible* in L only if for some predicate of L, competence with this predicate involves being disposed to accept that it satisfies the T-schema restricted to L. (And if L is a Tarskian language, truth in L may be quasi-expressible though not expressible in L.)

Earlier I discussed the counterintuitiveness of saying of a familiar property that it is inexpressible. This counterintuitiveness can be limited by appeal to the fact that what is not expressible may still be quasi-expressible. What I attempt to express what is not really expressible, for example when using the predicate 'true' when doing philosophy of language, my audience can still recognize my semantic intention of expressing a property satisfying the T-schema, provided such a property is quasi-expressible.

Another problem with the claim that truth simpliciter is not expressible is that this claim, if true, cannot say what it is intended to say. For if it is true, then "truth simpliciter", as it occurs in the claim, cannot express truth simpliciter. (And of course, this claim in turn cannot say what it is intended to say either.)

But if the thesis that truth simpliciter is not expressible is not statable, it may still be quasi-statable or quasi-expressible. This can explain how we can grasp the inexpressibility thesis even if it is true. We know what it aims to say, even if it cannot actually say it.

Even should the idea of quasi-expressibility be rejected, however, I am not very worried about the non-statability of the claim that truth simpliciter is not expressible. For, as noted earlier, truth simpliciter (or perhaps, e.g., strong negation simpliciter) is at any rate not *available*, and already this presents the same potential problem that the inexpressibility thesis faces.

Appeal to quasi-expressibility can also help resolve a related problem; one that I have tried to sweep under the rug so far. Ever since the first pages of this paper, I have been talking about "predicates satisfying the T-schema". And though my terminology differs in its detail from that employed by other authors, certainly other authors employ equivalent locutions. Now, to say that a predicate 'F' satisfies the T-schema is to say that all instances of

$\ulcorner \text{p} \urcorner$ is F if and only if p,

are true. This is problematic in the present case. Vann McGee has proven that

> If S is an extension of Robinson's arithmetic, we can for every sentence ϕ in the language of S effectively find a sentence ψ such that $\ulcorner \psi \urcorner$ is equivalent in the theory to $\ulcorner Tr(\ulcorner \psi \urcorner) \leftrightarrow \psi \urcorner$, regardless of how we choose the truth predicate '$Tr(x)$'.[24]

The proof is fairly simple. Let '$A(x)$' abbreviate '$Tr(x) \leftrightarrow \phi$'. Then, by the diagonal lemma, there is a sentence ψ such that $\ulcorner \psi \leftrightarrow A(\ulcorner \psi \urcorner) \urcorner$, i.e. $\ulcorner \psi \leftrightarrow (Tr(\ulcorner \psi \urcorner) \leftrightarrow \phi) \urcorner$. By propositional logic, $\ulcorner (\psi \leftrightarrow Tr(\ulcorner \psi \urcorner)) \leftrightarrow \phi \urcorner$.

But now consider what "satisfying the T-schema" can mean. As noted, to say that a particular predicate 'F' satisfies the T-schema is to say that every instance of

$\ulcorner \text{p} \urcorner$ is F if and only if p

is true. But how do we descend, i.e. get rid of the truth predicate? In light of McGee's theorem, not every instance of

$$Tr(\ulcorner Tr(\ulcorner \phi \urcorner) \leftrightarrow \phi \urcorner) \leftrightarrow (Tr(\ulcorner \phi \urcorner) \leftrightarrow \phi)$$

is true. In particular, for the liar sentence we can effectively find an equivalent sentence of the form $\ulcorner Tr(\ulcorner \psi \urcorner) \leftrightarrow \psi \urcorner$. But for this sentence the corresponding instance of the disquotation schema is not true, as it is not true for the liar sentence. (Of course, in this argument, I have employed yet another truth predicate in discussing how to understand talk about the validity of the T-schema; and the employment of this truth predicate raises further problems. But this is no cause for comfort.)

"Satisfies the T-schema" cannot mean what we intuitively thought it did. But if there are notions that are quasi-expressible, this problem is if not eliminated then at least considerably diminished. For then we can at least quasi-express what "satisfies the T-schema" is meant

[24] McGee (1992).

to express. (And to the extent that this problem arises also for other liar theorists, they too would be considerably helped by appeal to the notion of quasi-expressibility.)

I distinguished early on between the predicate 'true' of ordinary language and a predicate expressing a theoretically significant property of truth. This was in order to emphasize that even if the property expressed by our ordinary predicate 'true' cannot satisfy the T-schema, a truth predicate introduced for theoretical purposes might. I am somewhat unhappy to risk blurring the distinction drawn. as it is really important in its own right. But it should be noted that granted the notion of quasi-expressibility, and its potential divergence from regular expressibility, the following possibility opens up: A meaning-constitutive principle for 'true' of ordinary language is that this predicate satisfies the T-schema. But the predicate 'true' does not actually make true this principle, for the meaning-constitutive principles for the *logical particles* require that these particles satisfy the inference rules which jointly characterize classical logic – and the way to make the meaning-constitutive principles of our language come out as nearly true as possible *on the whole* is to render the meaning-constitutive principles for the *logical particles*, rather than those for the predicate 'true', true. Even if this is so, however, a predicate of an *expanded* language can satisfy the T-schema restricted to English, and thus can succeeds in expressing the property that the predicate 'true' of English aims to express. This might justify taking our predicate 'true' to be in an important sense identical in *meaning* to the truth predicate of the expanded language, in spite of the fact that their extensions are different. The meaning-constitutive principles for both predicates are the same.

Let me sum up where we have gotten to so far in this section. If what I have said earlier about truth-in-L being expressible only in a metalanguage (provided L satisfies certain conditions) is correct, then the problem arises of what to say about truth simpliciter. Under certain assumptions, truth simpliciter is not expressed by any predicate of any possible language. (And, as noted before, if these assumptions do not hold true, similar problems are sure to arise with respect to some other notion, for example negation.) The property of truth simpliciter then either simply does not exist, or, if it does, it cannot be expressed in any possible language. This seems quite implausible; and it also poses problems for philosophy of language, which appeals to such a property. I have now presented a number of arguments designed to make the consequence that truth simpliciter cannot be expressed seem more acceptable.

So far I have not said anything that distinguishes between the claim that truth simpliciter does not exist and the claim that the property, though it exists, is not expressible. To some extent, it is not important for me to make a decision between these options. What is important is only that appeal to truth in semantic and metasemantic theories is justified. However, it seems to me most natural to say that truth is inexpressible rather than that it does not exist, given that we, by appeal to the notion of quasi-expressibility, can explain away the apparent absurdity of such a claim. Here is one way of putting it: *if* a property of truth must exist in order for appeal to such a property to do its job in philosophy of language, then I have removed the obstacle to taking it to exist but be inexpressible.[25]

So I have ended up where I said I would end up. My thesis is fairly reasonably stated as: truth is inexpressible. However, 'truth' as it occurs in the statement of this thesis should not be taken to be the property expressed by our ordinary predicate 'true', but instead the property that does the theoretical work traditionally ascribed to truth. Second, truth is inexpressible only if other relevant notions are expressible. If, however, it is not truth but one of these other notions, for example strong negation, that fails to be expressible, the issues I have been concerned with here still arise: it is only that they arise with respect to strong negation rather than with respect to truth. Third, truth for English is not *absolutely* inexpressible: it can be expressed in a metalanguage. What is absolutely inexpressible is only truth simpliciter. Fourth, the thesis cannot, of course, be *stated*: it is only *quasi-statable.*

4.6 Concluding remarks

The path to the desired conclusion has been thorny. Let me end by briefly restating the main points that were made along the way.

I first argued that if English is Tarskian, then truth-in-English (where 'truth' here is to be thought of as the notion fit to do the theoretical work traditionally ascribed to truth) is not expressible in English but only in a richer metalanguage. Moreover, I argued, with reference to the accounts of meaning of Davidson and Dummett, that the notion of truth can still be useful in theorizing about English under the assumptions mentioned.

Then I considered truth simpliciter, truth for variable language L. If what I argued

[25] Toward the end of section four of chapter six I will however discuss another problem for this thesis.

up until that point is correct, then truth simpliciter is, under natural assumptions, not expressible in any language. The "natural assumptions" in question are that certain other properties are expressible in some language or other. A corollary is that some property we appear to understand or have a concept of is not expressible in any language, if it exists at all. I argued that this consequence regarding absolute inexpressibility, surprising and disturbing though it is, ought not to lead us to think that something went wrong in the arguments leading up to it. Among other things, I introduced the notion of quasi-expressibility, and attempted to limit the counterintuitiveness of the conclusion that there are inexpressible properties by arguing that these properties still are quasi-expressible.

I must emphasize that nothing I have argued here amounts in any way to a *solution* to the liar paradox. I have merely investigated the consequences of a particular class of possible solutions to the liar paradox: the solutions according to which English is Tarskian or pseudo-Tarskian. The one potentially important consequence of the present discussion for the project of finding a solution to the liar paradox is: a solution need not be rejected even if it entails that there are absolutely inexpressible properties.

At the end of the next chapter, however,I shall use the conclusions from this and the preceding chapters (mainly chapter 1) to argue that the liar paradox is in effect solved. The account of chapter 1 apparently requires ascending into a richer metalanguage. In order to justify ascending to a richer metalanguage I will use the considerations that I have presented in this chapter.

Chapter 5

The Liar Paradox (Dis)solved

5.1 Introduction

The problem of the liar has loomed large in the earlier chapters. However, I have neither discussed the existing accounts of the liar nor attempted to provide a solution. In this chapter, I will do both. (Though I shall not strictly provide a solution, but rather argue that there already exist several acceptable solutions.)

Let me begin by briefly restating my account of the liar, as set out in chapter 1.

Call the untrue premises and invalid steps in the liar reasoning the *culprits*, and call the premises and steps that are reasonable candidates for being culprits *suspects.* Let us use the label *intuition* to cover pretheoretical beliefs generally; and let the intuitions we have by virtue of our semantic competence be called *competence intuitions.*

It is uncontroversial that the liar arises because our intuitions are inconsistent. However, my more daring claim is that the jointly inconsistent intuitions that the suspects are true and valid are all competence intuitions. Competence intuitions hence need not all be veridical (the objects of these intuitions need not all be true). There are principles that are jointly inconsistent but such that we are disposed by our semantic competence to accept them all. Let an argument *exert pull* just in case it is unsound but such that we are disposed to accept the culprits in it by virtue of our semantic competence. Then what I claim is that the liar paradox exerts pull. Principles we are disposed to accept by virtue of our semantic competence I call *constitutive of meaning.* For our purposes, meanings can be thought of as such meaning-constitutive principles.[1] When the meaning-constitutive principles of the expressions of a language are inconsistent, I call the language inconsistent.

My first thesis about the liar is that the paradox arises because meaning-constitutive principles are jointly inconsistent. This thesis by itself says nothing about how the meanings (meaning-constitutive principles) of expressions are related to their semantic values - contributions to truth conditions. My second thesis is that the semantic values of expressions of a language are such as to make the meaning-constitutive principles come out as nearly correct as possible.

A consequence of the first and the second thesis is that one must distinguish between two projects of giving the meaning or semantics for an expression (or a language fragment): there is on the one hand the project of stating the meaning-constitutive principles for the expression, and on the other hand the project of saying what the semantic value of the expression is and what principles are actually rendered true and valid by the semantic value of the expression.

It is very likely that given my first two theses, it is highly indeterminate both what the semantic values of the expressions crucially employed in the liar are, and what the culprit is.

When comparing my account with other accounts in the literature, it is useful to distinguish between two kinds of accounts. First, there are *standard accounts*, according to which, roughly, the liar straightforwardly presents a problem as to which premise or step is invalid; at some specific point in the reasoning we have simply made a mistake. Second, there are the accounts according to which, as on my account, the liar arises because our language or conceptual system is in some sense or other 'incoherent' or 'inconsistent'. These I shall call *inconsistency views*. Among accounts of the former kind are, among many others, Kripke, Burge, Barwise and Etchemendy, Gupta and Belnap, Gaifman, McGee, and Simmons.[2] Among the latter are, for example, the accounts of Tarski, Chihara, Priest, and Yablo.[3]

[1] Two remarks on this identification. First, there are many kinds of things that can be called meanings. I focus on meanings conceived as objects of semantic competence, since my present concern is semantic competence. Second, even when meanings are conceived thus, there may of course be more to meanings than meaning-constitutive principles; but I shall abstract away from such other features.

[2] See e.g. Barwise and Etchemendy (1987), Burge (1979), Gaifman (1992), Gupta and Belnap (1993), Kripke (1975), McGee (1991) and Simmons (1993).

[3] See e.g. Tarski (1983), Chihara (1979), Priest (1979, 1984b and 1987) and Yablo (1993 and 1993b).

I shall discuss the inconsistency theorists individually, but the standard accounts will be discussed together, at a more general level. Standard theorists seldom raise explicitly the kinds of issue I will discuss here. In some cases it would be possible, through textual interpretation, to figure out how a standard theorist comes down on the issues discussed. But I think a discussion at a more general level will be more useful.

5.2 Standard accounts

In this section I shall discuss the shortcomings of standard accounts of the liar. However, I shall begin by discussing one particular feature of my account of the liar, and two other possible accounts that have the same feature. The purpose of this discussion is to highlight a particular contrast between my account and standard accounts.

One notable feature of my account of the liar is that all the suspects have roughly the same status with respect to being part of meaning, or *meaning status* as we may call it. I claim that all the suspects are meaning-constitutive.

There are other possible accounts of the liar that share this feature.

Consider Quineans, who hold that there is no distinction between what is and is not part of the meaning of an expression, and replace this distinction with Quine's picture of our belief system as a network, with some beliefs closer to the 'core' and some beliefs closer to the 'periphery'.

The Quinean, like everyone else, must face the question of how the semantic values are determined. (For example, she must be able to provide an answer to questions like: how come 'cow' is true of cows but not of horses?) The most reasonable way of doing so would seem to be to say that the semantic values of our expressions are such as to make our belief system, as a whole, as correct as possible (given the way the world is); and to this end, it is more important that core beliefs be true than that beliefs near the periphery be true.

For the Quinean, the liar ought not to pose a big, or even theoretically interesting, problem. The liar arises because beliefs all close to the core are jointly inconsistent. All of these beliefs are therefore among those that the semantic values of our expressions ought to make true. As this is impossible, it is indeterminate – to a rather considerable extent – what the semantic values are, and what the culprit in the liar reasoning is.

This Quinean account shares some important features with mine. The beliefs that are reference-determining (*all* of our beliefs, according to her) are jointly inconsistent. The inconsistency that comes to light in the liar gives rise to considerable indeterminacy. And the suspects in the liar all have the same kind of meaning status: they are all core beliefs.

The suspects in the liar all have the same meaning status also on a third kind of view. One might hold, as against the Quinean, that there is a distinction between what is part of meaning and what is not, while holding also that it is very often a *vague* matter whether a given principle is part of the meaning of an expression. One might then hold that the suspects in the liar reasoning are *all* borderline cases of being part of the meanings of the expressions employed. The suspects then all have the same meaning-status, as on the other views just considered, my view and the Quinean's. And again we have the same widespread indeterminacy as on my account and on the Quinean view. For it will again be indeterminate which suspects are made true by the semantic values of expressions.[4]

However, on standard accounts of the liar – the accounts listed above – the suspects can hardly all have the same meaning status. If they did, the semantic values of the expressions employed, and the diagnosis of the culprit, would likely be considerably more indeterminate than on standard views. Standard accounts are typically centered around a fairly specific diagnosis of what the culprit is; or, to put it differently, standard accounts are typically justified and motivated by arguments designed to show that they have specific consequences regarding how the liar reasoning goes wrong. All the suspects appear to be meaning-constitutive, but one or more – the culprits – can on closer examination be seen not to be meaning-constitutive. [5]

Perhaps the most explicit statement of what a standard account involves is in Martin (1970). Martin says that to solve the liar paradox is to "uncover" some "rulelike features of our language" such that they "have the effect of blocking at least one of the assumptions of the argument"; what is common to the standard accounts is that they attempt to uncover just such features of our language. There are, it is assumed, one or more specific assumptions

[4] Why would one hold the view outlined in this paragraph? Well, if one is not a Quinean, and one thinks that semantic competence can never dispose us to accept what is not true, and still cannot see how the suspects could have different meaning status, then the position outlined should seem compelling.

[5] There is a possible – and possibly not entirely unreasonable – view that I shall simply disregard here. It is the view that combines an anti-Quinean position with saying that more suspects than the culprit fail to be even borderline meaning-constitutive. What determines which of these suspects is the culprit is reality itself. The reason I shall not discuss such views is that I simply do not see how even to outline an answer to the question of how, in this case, we could *know* which of the suspects is the culprit. I believe, however, that some of the considerations against standard views apply also to this version of a standard view.

of the argument such that when sufficient attention is paid to the "rulelike features" of our language, we shall see that these assumptions should be doubted.[6]

I shall now argue, as against standard accounts, that it is very likely that the suspects all have the same meaning status.[7]

Consider a language English* just like English, except for the difference that in English* there are expressions introduced by the stipulations that (have the consequence that) the suspects in the liar reasoning – or, strictly speaking, its counterpart liar* reasoning – all come out true. It is easy to see that in the liar* reasoning in English*, all the suspects have the same meaning status. I would say that they are all part of the meanings of the expressions employed; the Quinean might say that the objects of the stipulations are core beliefs; and the borderline case theorist might say that the stipulations are all borderline cases of being part of the meaning of the expressions employed.

Why believe that English is any different from English*, in respect of the suspects having the same meaning status?

English *could* of course be different from English* in this respect: it could be, for example, that the truth predicate of English was generally taken to satisfy only principles much weaker than the disquotation schema; or that the truth predicate is context-sensitive, and that the liar reasoning derives its apparent plausibility from our overlooking this fact; or it could be that the fact that we believe the disquotation schema to be valid stems merely from our overgeneralizing from its true instances. But it is not necessarily the case that English differs from the hypothetical language English* in respect of the suspects having the same meaning status, as is shown by the possibility of other views.

There are different ways in which a standard theorist might argue that English is different from English* in the relevant respect. She could argue, for example, that whereas some suspects are part of the meanings of the expressions employed, close consideration of the behavior of the expressions employed shows that some suspects are not thus meaning-

[6] Martin (1970), p. 91. Martin states this as a condition on what a solution to the liar should achieve. Saying that to count as a solution something must satisfy this condition goes beyond saying that a solution to the liar satisfy this condition. Not all standard theorists need agree on the stronger claim. It will be discussed in the last sections of this chapter.

[7] I shall not argue, either now or later, against the two accounts according to which, as according to mine, all the suspects have the same meaning status. Arguing against Quineanism would take us too far afield; and at any rate the Quinean and I have roughly the same take on the liar. As regards the suggestion that, though there is a distinction between what is part of the meaning of an expression and what is not, all the suspects are borderline cases of being part of the meanings of the expressions employed, my only remark is that this suggestion seems unmotivated given the theoretical possibility of meanings being inconsistent.

constitutive. But, given the popularity of standard accounts, there is surprisingly little in the way of arguments to this effect. And the arguments are not convincing. I shall illustrate this by consideration of one argument of a kind not unfamiliar in the literature. There are other arguments. But the reasons why the kind of argument I consider does not work are general.

The first premise of the argument is that the liar sentence should be expected to have a semantic value different from truth and falsity – the third truth-value "neither true nor false", or lacking a truth-value, or being 'undecided', etc. I shall call this status *neuter*. This name fits better the view that the liar sentence is neither true nor false: but I shall use the label to cover any view of the kind mentioned.

To justify the first premise, one may for example appeal to the notion of groundedness, explained as follows in Kripke (1975):

> ...if a sentence....asserts that (all, some, most, etc.) of the sentences of a certain class C are true, its truth value can be ascertained if the truth values of the sentences in the class C are ascertained. If some of these sentences themselves involve the notion of truth, their truth value in turn must be ascertained by looking at *other* sentences, and so on. If ultimately this process terminates in sentences not mentioning the concept of truth, so that the truth value of the original statement can be ascertained, we call the original sentence *grounded*; otherwise, *ungrounded.*[8]

One may hold that a sentence must be grounded in order for it to have a truth-value, reasoning as follows: Semantic facts supervene on non-semantic facts. Facts about truth and falsity are paradigmatic examples of semantic facts. Hence they supervene on non-semantic facts. But then for a sentence to be true or false it must be grounded. For a sentence is grounded if and only if at the end of the chain of facts upon which its truth-value depends, there are non-semantic facts that determine its truth-value. But this, the argument goes, is just an explication of what it is for semantic facts to supervene on non-semantic facts.[9]

[8] Kripke 1975, p. 693f.

[9] This argument, as stated, is much too crude; it is presumably virtually a caricature of what proponents of similar arguments have had in mind. Suppose all Smith says is that everything Jones says is untrue, and all Jones says is that everything Smith says is untrue. Reflecting on their utterances I say: "At most one of Smith and Jones speaks the truth". Intuitively, what I say is true. But if the argument just presented is correct it is not.

The simplicity of the argument given in the text is however ideal for present, illustrative purposes.

The liar sentence is obviously non-grounded. Hence, by the above argument, it is neuter.

The second premise is that negation should be expected to take neuter into neuter (i.e. if a sentence is neuter, then so is its negation). One might have thought that if a sentence is neuter, the negation of the sentence is true; but what the second premise asserts is that the negation of the sentence is in this case neuter.

The third premise is that truth too takes neuter into neuter. One might have thought that when a sentence S is neuter, it is *not true*, and that accordingly, the sentence that expresses that S is true is false. But this is what is being denied by the third premise.

If these three premises hold, the liar sentence can be ascribed a semantic value without contradiction. For we can, without untoward consequences, then ascribe the liar sentence the value neuter. (If the second or third premise did not hold, we would still have trouble with a liar sentence that says of itself that it is not true. For then if we ascribe such a sentence the value neuter, we can derive that it has a classical truth-value after all.)

Merely asserting that these three premises are true does not, however, make you a standard theorist. To be a standard theorist you must also hold that these premises are the ones that reflect the meanings of the expressions involved; and that to think differently is to be mistaken about the meanings of one or other of these expressions.

But this further claim seems to me problematic, even if, upon reflection, we should accept the three premises mentioned. For does not precisely the fact that we have all the intuitions leading up to contradiction entail that the three premises stated, in so far as they serve to avoid contradiction, fail to adequately describe our intuitions?

There is no doubt that there are *possible* expressions that would work just as the expressions used in the liar reasoning according to the argument presented do work. Neither is there any doubt that there are possible expressions such that the three premises accurately describe the meanings they have.

What is at issue is rather whether the expressions we actually use are endowed with such meanings. Think of the matter this way. Would we really – *the liar reasoning aside* – prefer the assumptions embodied in the three premises to the assumptions that jointly, by the liar reasoning, lead to contradiction? If not, why should we think that the three premises better describe the meaning of 'true' and other expressions employed in the paradoxical reasoning than do inconsistency accounts?

The proponent of a standard account of the kind considered (which is broadly Kripkean[10] is likely to respond by saying that we are demanding too much of a purported solution. It is uncontroversial that the liar reasoning is problematic because we have contradictory intuitions, and trivial that any attempted solution to the liar (except solutions that involve accepting true contradictions) will have to deny some of the intuitions that lead to contradiction; so the mere fact that the Kripkean account violates some intuitions ought not to count against it.

But this response should not be accepted. It is true, as the response says, that any account will have to deny some intuitions. But, first, there are very many potential solutions that involve denying some of the intuitions that jointly lead to contradiction; what the Kripkean would need to adduce is a special reason for denying exactly those intuitions that she denies. Second, while it is true that any account will have to deny some intuitions, this (truism) does not entail that it is not the case that all the suspects have the same meaning status. If the suspects have the same meaning status – and this status is that of being part of meaning, or being a core belief, or being a borderline part of meaning – then the problem of the liar splits into two. One is that of saying what the claims that jointly lead to contradiction are and what their meaning status is; another is that of saying which of these claims actually are true. Some intuitions that we have must be denied, in the sense that on every acceptable assignment of semantic values to our expressions, some intuitions are deemed false. But this by itself means neither that some of our intuitions are not competence intuitions, nor that there is some intuition employed in the liar reasoning that comes out false under every acceptable assignment of semantic values. It could be that the semantic values of the expressions employed in the liar are so indeterminate that every suspect is rendered false by at least some acceptable assignment of semantic values.

[10] I say that the account is broadly Kripkean because the argument, if sound, serves to justify a hypothesis about the semantic values of the expressions in the liar reasoning that is very much like the main hypothesis considered in Kripke (1975).

Kripke himself does not adduce a philosophical justification of this kind for the account he presents. He instead indicates three different kinds of justification for it. First, he notes that the theory is mathematically interesting, regardless of how well it captures the concept of truth. Second, he says that it accords with many of our intuitions about truth (without indicating that it accords with all of them and without indicating that because it accords with so many intuitions about truth, conflicting intuitions should be disregarded or explained away). Third, he indicates that the minimal fixed point is what a speaker would arrive at given instructions we would naturally give when explaining to someone the concept of truth (Kripke 1975, p. 701ff).

Moreover, Kripke officially remains neutral on which fixed point (best) represents the extension of 'true'. However, the minimal fixed point receives most of the attention, and it is the extension of 'true' in the minimal fixed point that is the genuine extension of 'true' given the justification of the third kind listed.

The Kripkean may suggest, in response to the first of the problems mentioned in the previous paragraph, that her account is preferable to other accounts because although her account violates some of our intuitions, it *on the whole* respects our intuitions better than other consistent accounts.

But first, and most importantly, this seems to be *surrendering*, as regards the question of the suspects not having the same meaning status. For it seems to be agreeing that all the suspects have a (roughly) equal claim to be respected by an assignment of semantic values.

Second, on a more speculative note, the Kripkean's new claim is rather difficult to evaluate. How can we reasonably compare the Kripkean theory, with its acceptance of what is in effect a many-valued logic, with, for example, the revision theory of truth, which respects classical logic at the cost of making the semantics of the truth predicate more exotic?[11] I do not say that this is in principle impossible – it could of course be that one theory succeeded determinately better than the other – but what I do not see is how we can reasonably assume that there will be facts which determine which of these two theories on the whole better respects our intuitions.[12]

Notice, however, that for all I have argued it may well be that a Kripkean account (of the kind motivated by the argument considered) is among the acceptable assignments of semantic values – and it could even be that every acceptable assignment of semantic values is Kripkean. What I have been mainly concerned to argue is that even if this should be the case, that is not because the meaning status of the assumption the Kripkean diagnoses as the culprit is different from the meaning status of other suspects.

There is also a different kind of worry about standard accounts. Suppose a standard account, for example a Kripkean account of the kind described, is true. Then let us ask: could there be a predicate which is like 'true' except that, with this predicate substituted for 'true', all the suspects have the same status? If so, then it seems that the version of the liar reasoning we get with this other, hypothetical truth predicate presents an as yet unsolved problem. This problem is presumably deeper than that posed by the original liar, provided a standard account is correct, since the modified liar reasoning forces us to ask questions

[11] For the revision theory of truth, see Gupta and Belnap (1993).

[12] In chapter 3 I distinguished between the theories and philosophies of different accounts of truth and the liar. The theory is the proposed account of semantic values. The philosophy is the justification given for thinking that this account is the correct one. What I am talking about here is the *theories* of Kripke and of Gupta and Belnap.

not only about the behavior of a specific predicate but about the nature of meaning itself.

5.3 Negation

In the last section, I presented general reasons for rejecting standard accounts. In this section I shall revisit the assumption about negation in a many-valued setting made by certain standard theorists. I shall discuss and criticize the philosophy behind attempts to solve the liar by appealing to this assumption. By and large I shall cover the same ground as was covered in the last section; but the argumentative roads traveled will be different.

Terence Parsons (1984) and Jamie Tappenden (1999) argue that, partly for reasons unrelated to the liar reasoning, negation in a many-valued setting is best construed as weak negation rather than strong negation. (Weak negation takes neuter, the third truth-value, into neuter; strong negation takes every truth-value apart from truth into truth.) If negation should be understood thus, then the possibility opens up that a liar sentence can be assigned the value neuter, with no untoward consequences.

One might quarrel with the reasons that Parsons and Tappenden adduce for thinking that negation should always be understood as weak negation. However, I shall not even discuss what those reasons are, but instead simply grant Parsons and Tappenden their point about how 'not' and "it is not the case that" are being used. My worries are different.[13]

First, for the liar problem to arise it is sufficient that there be *some* expression that expresses strong negation. That 'not' does not would then be beside the point.

Second, there are in fact specific reasons for thinking that, even granting Parsons and Tappenden their point about 'not', some expressions of English actually do express external negation. First, I take it that the above explanation of the notion of external negation *succeeded*; but then the explanans must have succeeded in expressing external negation. 'Neuter', as it occurred there, can be replaced, if you like, with "the truth-value distinct from truth and falsity". (Of course, my own positive view, as developed in chapter 4, is that the explanation of external negation could convey what it is intended to convey even if, say, "has the truth-value distinct from truth and falsity" could not, appearances to the contrary, be true of exactly the sentences of which it was intended to be true. But the

[13] In what follows I shall focus exclusively on Tappenden's discussion, which is more recent. But the same remarks as apply to Tappenden's discussion apply to Parsons' discussion.

argument here is intended to be *ad hominem.*[14]) Second, *arguably*, if S is neuter, then the statement that S is true will be false. But then the predicate "is not true" will express external negation even if 'not' expresses weak negation.[15]

The point made in the last paragraph can be made more forcefully. Tappenden's main point concerns a class of sentences that have certain properties yet fail to be true; but then for Tappenden's main point to be made, it has to be possible to express that these sentences are not in fact true.[16]

Third, let us for argument's sake *grant* Parsons and Tappenden that *no* expression of English can express external negation. Can we not introduce by stipulation an expression that expresses external negation? Tappenden discusses this objection at some length, and it will be instructive to consider his discussion. In Tappenden's formulation, the objection is: "...the reader might sigh: Again those darn native speakers have failed to talk as they ought to. But that's easy enough to remedy, isn't it? Can't we just add it to the language now?".[17]

Tappenden's first point in response is that a stipulation may fail, as in the case of 'tonk'. It may be, Tappenden says, that "'H()' [this is Tappenden's would-be external negation], like 'tonk', has a clear definition isolating no real operator".[18] He then goes on to discuss "the normative principles dividing prospective new logical vocabulary into acceptable and not"; the question of "what restrictions can be placed on proposed new pieces of logical vocabulary to ensure that they are not bad ones like 'Tonk'".[19]

What Tappenden claims is that the very fact that a contradiction would result if the operator 'H()' were introduced shows that the operator is 'illegitimate'. The condition Tappenden proposes on the legitimacy of new vocabulary is that it not disrupt the meanings of expressions previously in the language.[20]

Now, there may well be a worthwhile project, prescriptive in nature, of determining what the conditions on the legitimacy of new vocabulary are. But the problem is, it may

[14] In other works (1993 and 1993b) Tappenden explores ideas which may be deployed in defense of the view that some things can be conveyed without being strictly expressible. But Tappenden does not refer to those ideas in his (1999); and at any rate I shall argue against these ideas in the next section.

[15] I am ignoring the distinctions between predicates and sentential operators, since for my present destructive purposes it does not matter whether it is a predicate or an operator that expresses external negation.

[16] This echoes Donnellan's (1970) comments on Martin (1970).

[17] Tappenden (1999), p. 282.

[18] Ibid., p. 283.

[19] Ibid.

[20] Ibid.

be that not all new definitions satisfy the proposed conditions of legitimacy. Those 'darn native speakers' may fail to introduce new vocabulary as they ought to.

An interesting question is, what would happen *then*? A hard-liner might say that the definition *simply fails*: that these illegitimate definitions cannot affect the meanings of expressions previously in the language, and that the new expression fails to be meaningful at all. But first, this does not seem very plausible. (Suppose the speakers who introduce the external negation operator do not know of and do not discover the liar problem. Then the use and meaning of the newly introduced operator will soon become as entrenched as that of the expressions previously in the language.) Second, Tappenden himself seems not to take this hard line. If he did, his condition on the legitimacy of definitions would be satisfied by all stipulations that succeed in rendering meaningful the expression introduced; but it does not seem as if he holds this view.[21]

Having criticized Tappenden's treatment of the objection that external negation may be introduced into the language by stipulation, I should add that it is not entirely clear that it is an objection at all. Or rather: either it is not an objection, or Tappenden's response is for very general reasons inadequate. Let me explain.

What kind of problem does the liar paradox pose? One can think of it as posing a problem concerning the semantics of actual natural languages such as English. The problem would be (roughly): what are the actual meanings of expressions such that the liar reasoning is not sound? If this is how Tappenden conceives of the liar problem, the possibility of introducing external negation via stipulation ought not to trouble him at all. For his solution to the liar might still correctly describe actually existing English. Alternatively, one can think of the liar as posing a problem concerning all possible languages: solving the liar is solving it not only as it arises for English as it actually exists but also as it, or variants of it, arise for all kinds of possible languages. But this had better not be how Tappenden conceives of the problem posed by the liar, for then his solution is inadequate. For his response to the objection that external negation *can* be introduced into English by stipulation obviously relies on the assumption that the language, English, into which external negation is supposed to be introduced, already contains sufficient expressive

[21] The question Tappenden seems to be addressing is: what conditions must a definition satisfy in order that the language we use be the same after the new expression is introduced. (Or, as Tappenden takes this question: what conditions must a definition satisfy in order that it not disrupt the meanings of expressions already in the language.

resources so that external negation *cannot* be consistently introduced.

Perhaps a charitable interpretation of Tappenden is that he is really concerned with two distinct questions. One is that of why the liar reasoning is not sound in actual English; the other is that of why the liar paradox, or a counterpart, cannot arise in any language. Tappenden's story of negation is intended to be an answer only to the first question; the answer to the second question is rather a general point about definitions: definitions do not succeed when they conflict with the semantics of expressions already in the language.

This interpretation would also explain Tappenden's distinguishing between a 'narrow' and a 'wide' issue raised in his work, the narrow issue being specifically about (a version of) the liar paradox as it arises in actual English, the wide issue being about possible languages, more specifically, the "ways the potential for acceptable language change is constrained by linguistic meaning".[22] However, even given this more charitable interpretation of what Tappenden's project is, I am skeptical. Tappenden's "wide issue" concerns only definitions incompatible with the meanings of expressions already in the language. As I have argued, the same incompatibility that can exist between new definitions and what is already in the language can exist also between the meanings of expressions all 'already' in the language. And Tappenden's account of the former kind of incompatibility does not extend to the latter, since his resolution of the former kind of incompatibility is to take the meanings of expressions already in the language to trump the new definitions, as it were.

5.4 Tappenden on pre-analyticity

In other work, Tappenden has presented views more congenial to mine.[23] He argues (in effect) that the liar and the sorites paradox arise because the *pre-analytic* sentences of our language are inconsistent, where a sentence S is pre-analytic just in case competence with it involves knowing that it cannot be false (though it can fail to be true). The instances of the disquotation schema, and the sorites premise, are among the pre-analytic sentences of our language. The key to understanding what goes on in the paradoxes, Tappenden argues, is to realize that a sentence may be pre-analytic without being analytic. A sentence can be as if true by convention while failing to be true.

[22] Tappenden (1999), p. 261.

[23] See Tappenden (1993) and (1993b).

Compare my class of meaning-constitutive principles with Tappenden's class of pre-analytic sentences. There are a number of salient differences. One is that the characterization of what it is for a sentence to be meaning-constitutive allows for the possibility that a meaning-constitutive sentence is outright *false.* Perhaps argument can show that acceptable assignments of semantic values always will fail to render any meaning-constitutive sentence false. But if so, that is a highly non-trivial result. Another salient difference is that there is a tighter link between meaning-constitutivity and semantic values than between pre-analyticity and semantic values (at least for all that Tappenden tells us about pre-analyticity). On my account, the semantic values of expressions of a language L are constrained by the fact that the meaning-constitutive principles of L should come out as nearly correct (true and truth-preserving, respectively) as possible. But for all that Tappenden says, it appears that the pre-analytic sentences constrain only the semantic values of their constituent expressions in that the pre-analytic sentences must not come out false. This is compatible with not a single one of them being true – even if they all consistently could be true.

Another important feature of Tappenden's account is that he postulates the existence of a speech act of *articulation*, where a sentence is articulated if "it is uttered for the purposes of inducing someone else to withdraw, or refrain from asserting, some sentence". Presumably, some utterances will be both assertions and articulations. What is important for Tappenden's purposes is only that some articulations are not assertions. In particular, pre-analytic sentences are examples of sentences which may justifiably be articulated but which cannot be correctly asserted.

One example Tappenden uses to back up his claims about a speech act of articulation concerns the way we normally present the liar paradox. In the following passage, 'k' is the liar sentence and '(3*)' is the biconditional "k is true iff k is not true":

> ...(3*) is self-contradictory and hence cannot correctly be said to be true, but it also seems to capture one feature of the sentence k in virtue of which k is paradoxical! Does this not sum up nicely what is funny about k: "k is true if and only if k is not true"?....Imagine that you are explaining the liar to someone who does not catch on right away. You might well say: "Here is what is funny about k. If k is true then it is not true, and if k is not true then it is true." You have apparently just contradicted yourself by uttering a variation on (3*). But

(3*) seems to be the right thing to say in the situation.[24]

To account for the correctness of uttering the self-contradictory sentence, Tappenden says that it is an articulation, not an assertion. I shall not either endorse or contest what Tappenden says about these utterances. What I shall be concerned to show is that *if* we agree with Tappenden that the sentences he considers have the special status of being pre-analytic and of being 'articulable' without being assertible, then a very good way of doing so is by adopting my account of the liar. I am not, however, convinced by Tappenden's argument for postulating a speech act of articulation. When explaining Russell's paradox of the barber to someone, is it not reasonable to say, "Here's what is funny about the village barber. If he shaves himself then he doesn't shave himself, and if he doesn't shave himself then he shaves himself."? Yet I take it we would not, on the strength of this, want either to postulate a speech act of articulation, or to say that the analogue of (3*) in the case of the barber cannot correctly be called false. But I will simply set aside these worries and assume, for the sake of the argument, that there is reason to postulate a special kind of speech act to account for the correctness of certain utterances of (3*).

Tappenden gives conditions for when a sentence is pre-analytic and also for when a sentence is fit to be articulated. But he does not give an explanation of how it comes about that there are sentences which satisfy these conditions and yet are not properly analytic and properly assertible. He doesn't say why we should expect there to be sentences which are pre-analytic without being analytic, or articulable without being assertible. A suggestion, based primarily on his discussion in (1993b), is that one reason he would adduce is that we know that stipulation cannot (always) make sentences true, for stipulations may be jointly inconsistent. But a sentence introduced by stipulation still can never be correctly called false.

However, this is unsatisfactory. Why is it reasonable to say that, although stipulations cannot guarantee truth, they can guarantee non-falsity? To make my worry more concrete, suppose that the truth predicate is 'strong', in that it takes neuter into falsity. If S is neuter then the statement that S is true is false. Now consider any inconsistent set of stipulations you like, $\{S_1,...,S_n\}$, where $S_1,...S_n$ are the sentences stipulated to be true. According to Tappenden, the stipulations cannot guarantee the truth of $S_1,...,S_n$, but they can guaran-

[24] Tappenden (1993), p. 552.

tee their non-falsity. However, the incautious stipulator could just as well stipulate that Tr(S_1),...,Tr(S_n) come out true; and by the semantics of the truth predicate, at least one of these sentences must be false if S_1,...,S_n are jointly inconsistent. In other words, the phenomenon Tappenden appears to appeal to in order to establish the existence of sentences which are pre-analytic and articulable, without being analytic and assertible, is not adequately described by Tappenden's notions of pre-analyticity and articulation.

The argument just given will not convince Tappenden, for he takes truth to be 'weak' – to take neuter into neuter – and he has theoretical reasons for that. But my point is more general. If a set of stipulations may be inconsistent, then it is given only some fairly non-obvious assumption that Tappenden is right that no stipulation can come out false. (Again it should be emphasized that for Tappenden's purposes it is not sufficient that *truth* is weak: *no* predicate must have the semantics of strong truth.)

If Tappenden cannot claim for the pre-analytic sentences that our semantic competence involves knowledge that they can never correctly be called false, the class of pre-analytic sentences cannot be picked out by our semantic knowledge of what truth-values they can and cannot have. But this does not mean that there is no reasonable way of picking out at least something like Tappenden's class of pre-analytic sentences. But suppose we consider Tappenden's account modified so that the 'pre-analytic' sentences are exactly what I have called the meaning-constitutive sentences. Then the data to which Tappenden points can still be satisfactorily accounted for. For example, if competence with 'true' involves being disposed to accept the disquotation schema, it is hardly surprising if the disquotation schema or an instance of it (even the contradictory ones) can sometimes be used to convey something about the content of claims about truth.

Tappenden, like me, attempts to make natural sense of the claim that natural language is 'inconsistent'. For Tappenden, a language is inconsistent if and only if it is inconsistent in its pre-analytic sentences. However, there is reason for dissatisfaction with this characterization of inconsistency of languages. Suppose there are pre-analytic sentences of our language such that our linguistic intuitions tell us that these sentences are neither true nor false (or that they are 'gappy', or however else you would like to describe this semantic status). Then a language may be Tappenden-inconsistent even though there is nothing amiss with our linguistic intuitions: all our linguistic intuitions can still be veridical in the sense that what they tell us about truth and falsity (and what is in between) may indeed be the case.

But this differs markedly from how inconsistency theorists have regarded the paradoxes: according to inconsistency theorists the paradoxes show that not all linguistic intuitions can be veridical (or else that there exist true contradictions). Tappenden-inconsistency is hence too weak a notion to capture an inconsistency view.

Tappenden is of course free to stipulate "language L is inconsistent" to mean whatever he likes. The present point is supposed to be deeper. It seems that two entirely distinct phenomena fall under Tappenden's heading "inconsistent language", or, better put, that two entirely distinct phenomena can give rise to a language being inconsistent in Tappenden's sense. On the one hand, the paradoxes which display an inconsistency of competence intuitions, and on the other hand, the cases where we have competence intuitions to the effect that certain sentence are 'gappy' and their being true is incompatible with the truth of sentences we take by our competence to be true.

Tappenden's account is motivated by the fact that sometimes we lay down stipulations of the form "P(x) if and only if Def(x)", where 'P' is the predicate we are defining and 'Def' simply is short for the definiens. If we assume that non-referring and referentially indeterminate names (sometimes) give rise to truth-value gaps and that in a framework where truth-value gaps are allowed, a biconditional is true only if flanked by two sentences with classical truth-values (as in the Kleene schemes), these stipulations are not true. But still these sentences have some special status. According to Tappenden, they are pre-analytic. However, even if one agrees with Tappenden that these stipulations have some distinguished status without being true, one may think that the situation here described is not very closely analogous to that presented by the paradoxes, for example because in the paradoxes not all competence intuitions are veridical, but in the case here described they may well all be.[25]

[25] Tappenden of course thinks of the definitions mentioned as being of the form "*For every* x, P(x) if and only if Def(x)". Once we keep this firmly in mind there is a very simple explanation of the phenomenon Tappenden points to. In a language with names that may be, in various ways, semantically deficient, one must distinguish between (a) it being the case that every object x is such that P(x) if and only if Def(x), and (b) it being the case that, no matter what name we substitute for 'x' in "P(x) if and only if Def(x)", a true sentence results. It is well known that, for independent reasons, (b) can be true without (a) being so (our language need not contain a name for every object). In languages with semantically deficient names, (a) can be true without (b) being so: for when names are semantically deficient they can fail to name unique objects.

5.5 Priest's dialetheism

Tappenden's view, discussed in the last section, may be called an inconsistency view on the liar. He does, after all, take the liar paradox to constitute one reason for thinking that natural language is inconsistent (in the sense he characterizes). In chapter 1, I discussed Tarski's inconsistency view on the liar, or rather Scott Soames' interpretation of this view. Other theorists who have defended inconsistency views on the liar are Charles Chihara, Graham Priest and Stephen Yablo. In this section I shall discuss Priest's view; in the next Chihara's; and thereafter Yablo's.

Graham Priest has used the liar paradox to argue for *dialetheism*, the view that there are true contradictions. The argument from the liar is not his only argument to that effect; but I shall consider the argument from the liar as a self-sufficient argument.

Priest's basic argument is in effect rather simple. He defends each principle in the liar reasoning on the basis that the meanings of the expressions employed require that the principles be true and valid, respectively. In particular, he assumes the meaning of 'true' to demand that this predicate satisfy the T-schema, and assumes also that the meanings of the logical expressions demand that they be such as to make the liar reasoning sound. Hence, there are, according to Priest, true contradictions.[26]

Now, a language can be, in some sense, 'inconsistent' in importantly different ways. Let a language L be *Priest-inconsistent* just in case the semantics of the expression performing the function of negation in L (where this is understood suitably loosely), '$\neg$', as we may symbolize it, is such as to allow for there to be sentences 'P' of L such that both 'P' and '$\neg$P' are true. Let a language L be *deeply inconsistent* just in case the semantic values of expressions of L cannot be such as to make all principles constitutive of meaning true and valid, respectively.

A language can be inconsistent in both ways at once; but importantly, it can also be inconsistent in one of these two ways without being so in the other. Priest's arguments for Priest-inconsistency suffice to establish their desired conclusion only if it is not a meaning-

[26] Priest's assumption about 'true' is often hidden under the assumption that English is "semantically closed". But Priest's talk of semantic closure is really about basic assumptions concerning the semantic values of truth and satisfaction predicates. Thus, in (1984), Priest says "A theory is semantically closed (with respect to its satisfaction relation) iff (i) for every formula with one free variable ϕ, there is a term a_ϕ, its name, (ii) there is a formula with two free variables $Sat(x, y)$ such that every instance of the scheme $Sat(t, a_\phi) \leftrightarrow \phi(v/t)$ is a theorem, where t is any term, ϕ any formula with one free variable v, and, $\phi(v/t)$ is ϕ [with] all occurrences of 'v' replaced by 't'" (Priest 1984, p. 118).

constitutive principle that contradictions cannot come out true. It is not sufficient that the meanings of the expressions are such that the principles leading up to the contradiction are demanded (to use a suitably intuitive expression) by the meanings to be true. This would not by itself establish Priest-inconsistency, but is equally consistent with deep inconsistency.

Standard, 'consistent' accounts of the paradox fail to explain why we are intuitively attracted to each of the premises and steps in the liar argument. Hence they fail to explain why the liar paradox exerts pull, where, recall, an unsound argument exerts pull just in case our competence with the expressions employed in the argument disposes us to be initially attracted to the culprits in it.

When pull is thus characterized, only unsound arguments can exert pull. But let an argument exert *generalized pull* just in case it either exerts pull in the sense above or is a sound argument such that our semantic competence disposes us to be disinclined to accept the conclusion. In chapter 1, I argued that the liar exerts pull. The arguments can be employed, without significant modifications, also for the claim that the liar exerts generalized pull.

Priest's account, if it says nothing more than that the solution to the liar paradox is that we should accept some contradictions as true, fails to explain generalized pull. This is just the same kind of problem as that which was discussed in chapter 1 for standard solutions. Generalized pull exertion implies deep inconsistency, and Priest-inconsistency is not deep.

Priest's account is accordingly reasonably classed with the standard accounts. For Priest, no intuitions leading up to the contradiction but rather the intuition *that the contradictory conclusion is paradoxical* is a seeming competence intuition rejected as not having been based on semantic competence in the first place. What is special about Priest's account is only that his view about *which* apparently obvious claim it is that has a different meaning status differs from that of most theorists, since he takes this claim to be the law of non-contradiction.

The argument so far in this section has aimed to show only that *if* the liar exerts pull, *then* Priest's version of dialetheism, according to which it does not, is not acceptable. So far nothing that has been said rules out any version of dialetheism that does allow for pull.

But once the possibility of deep inconsistency is recognized, much, if not all, of the motivation, for embracing the very radical doctrine of dialetheism should go away. For the main motivation for dialetheism is the apparent absurdity involved in denying any one of the assumptions that, by the liar reasoning, yield a contradiction. But, given the possibility of deep inconsistency, we can see this absurdity as indicative of *deep* inconsistency rather than Priest-inconsistency. It may be (and certainly seems seems to be) that accepting the contradictory conclusion is at least as absurd as is denying one of the premises. And if our language is deeply inconsistent, dialetheism is true only if the best way of maximizing the correctness of our jointly inconsistent meaning-constitutive principles involves making some contradictions true. But it certainly appears that the law of non-contradiction is more entrenched or more fundamental in our language and conceptual scheme than are some other principles that may be abandoned.

I have here focused on Priest's argument that the liar paradox provides a reason for accepting dialetheism. But I believe that the major points generalize to all of Priest's arguments for dialetheism and all the motivation for dialetheism.

In the very first paragraph of the first chapter of his (1987), Priest says:

> The paradoxes are all arguments starting with apparently analytic principles concerning truth, membership etc., and proceeding via apparently valid reasoning, to a conclusion of the form 'α and not-α'. *Prima facie*, therefore they show the existence of dialetheias [true contradictions]. Those who would deny dialetheism have to show what is wrong with the arguments – of [sic] every single argument, that is. For every single argument they must locate a premise that is untrue, or a step which is invalid. Of course, choosing a point at which to break each argument is not difficult: we can just choose one at random. The problem is to justify the choice. It is my contention that no choice has been satisfactorily justified and, moreover, that no choice can be.[27]

Priest is here talking about what he calls "the paradoxes of self reference" – the semantic and set-theoretic paradoxes. But what he says about these paradoxes extends to some of his other arguments for dialetheism, for example the argument from multi-criterial terms.[28] There are certain arguments that lead to contradictions and are such that we cannot give non-*ad hoc* reasons for rejecting any one of their premises and steps. Priest concludes from this that all the premises and steps of these arguments are in fact true.

[27] Priest (1987), p. 11.

[28] Priest (1987), p. 85ff.

There is however an important gap in the argument, even granting, as I of course am willing to do, the premise that there are arguments that lead to contradiction and are such that there are no non-*ad hoc* reasons to reject any of the premises or steps. One might say against Priest that there is likewise no non-*ad hoc* reason for abandoning the view that no contradictions are true. The claim that Priest would have us give up is itself in the same boat as the premises and steps of the paradoxical arguments.

The situation is in a way worse than Priest takes it to be. The paradoxical arguments are of the following form: apparently analytic claims (e.g. the premises and steps of a particular paradox) conflict with another apparently analytic claim (e.g. that no contradiction is true). Priest evidently thinks that there is always one apparently analytic claim that we can give non-*ad hoc* reasons for abandoning; in many cases, the claim that no contradictions are true. I think that we cannot give non-*ad hoc* reasons for abandoning this claim either.

5.6 Chihara's inconsistency view

Chihara's inconsistency view on the liar paradox is that "there are generally accepted conventions which give the meaning of 'true' and which are expressed by [a version of the disquotation schema]", and these conventions in conjunction with true and valid principles lead to contradiction.[29] Though I believe I am in basic agreement with Chihara, I should like to remark on some difficulties in interpreting his inconsistency view.

Before I do so, however, I want to say a few words about Chihara's celebrated distinction between the "diagnostic problem" and the "preventative problem" of the semantic paradoxes. Briefly, I think that Chihara is getting at a very important distinction; but as he draws the distinction it relies too heavily on his own positive view on the liar paradox. Chihara characterizes the distinction as follows:

> The problem of pinpointing what is deceiving us and, if possible, explaining how and why the deception was produced is what I wish to call '*the diagnostic problem of the paradox*'. The related problem of devising languages and logical systems which capture certain essential or useful features of the relevant semantical concepts, but within which the paradox cannot arise, I shall call '*the*

[29] Chihara (1979), p. 611.

preventative problem of the paradox'.[30]

Someone who does not agree that the inconsistency view is correct might well hold that to say that a paradox "arises within a language" is at best merely shorthand for: speakers of that language are for some reason or other prone to be taken in by the paradoxical reasoning. This means that she ought not, or at least need not, agree that the diagnostic problem, as Chihara describes it, is a problem for the philosopher or logician at all. She may reasonably say that it is instead a problem for the psychologist or psycholinguist. Compare: it is not for the philosopher to explain how particular visual illusions arise.[31] (For Chihara himself or for me, on the other hand, it is literally correct that the paradox *arises within* the language itself.[32])

Moreover, if it is not literally correct to say that the paradox arises within the language, there certainly need not be any reason to *reform* our language as a response to the paradox: all that needs to be done is to explain what the semantic features of expressions of our language were like all along. Even if it *is* literally correct that (somehow) the paradox "arises within" our language, it is hardly the case that the semantic values of expressions of our language make contradictions true.[33] If the paradox "arises within" our language it must be in some other sense (as Chihara certainly agrees). But then it is not clear why reform would be called for to avoid the paradox, even should the paradox "arise within" the language. But as Chihara characterizes the preventative problem, the liar paradox calls for reform, if we want to avoid paradox.

Chihara, to be sure, emphasizes that our language is fine as it is, even though paradox arises "within it". But what I want to emphasize is that since, as Chihara agrees, the semantic values of expressions of our language do not make contradictions true, an account of semantic values that avoids paradox, that is, which avoids contradiction, does not constitute

[30] Chihara (1979), p. 590f.

[31] It may be for the philosopher to give an analysis of what a visual illusion is; but that is a different matter entirely.

[32] Gupta and Belnap – standard theorists who are circularity theorists – also hold that the liar paradox "arises within" the language itself: but they interpret this differently. They interpret it not to mean that the language in some sense licenses the derivation of a contradiction, but that there is some other behavior (a certain kind of semantic instability) that is paradoxical, and whose explanation is semantic rather than non-semantic.

When we take "the paradox arises within the language" to mean that the language itself – somehow or in some sense – licenses the derivation of a contradiction, paradox no longer arises within the language given Gupta and Belnap's account.

[33] Priest would disagree of course.

a reform of our language.[34]

As for Chihara's statement of his positive view on the liar, the "inconsistency view", it just is not clear what he means by saying that the conventions giving the meaning of 'true' are inconsistent.

The claim would have a clear meaning if the use of 'true' were governed by explicitly given conventions, but of course it is not. What we have to make sense of is the claim that the use of 'true" is *implicitly* governed by inconsistent conventions. Chihara compares 'true' with a predicate 'true*' introduced by the stipulation that it satisfy the T-schema and argues (I believe plausibly) that the meanings of the two predicates will be essentially the same. But this still does not answer the question what it is for 'true' to be implicitly governed by inconsistent conventions. What the comparison can show is that 'true' is in some important respects like a predicate introduced by an inconsistent stipulation. But one would like to know in *what* respects.[35]

The straightforward understanding of the claim that a statement ϕ of English in which 'true' is used is a convention implicitly governing the use of 'true', is that this statement is

[34] There are structural similarities between (a) Chihara's distinction between problems posed by the semantic paradoxes, and (b) Strawson's distinction between "descriptive" and "revisionary" metaphysics. Descriptive metaphysics is the project of describing "the actual structure of our thought about the world"; revisionary metaphysics strives to develop a better structure. Strawson mentions Kant as a descriptive metaphysician and Descartes and Berkeley as revisionary metaphysicians. But as J. O. Urmson remarks in his (1961), Descartes and Berkeley would have regarded themselves as respecting 'our' metaphysics.

But I think it is clear what Strawson is getting at, and the distinction is better characterized as follows: "Descriptive metaphysics" seeks to describe the actual structure of our thought about the world, irrespective of whether it accurately represents the way the world is. "Revisionary metaphysics" seeks to present the actual structure of the world. When the distinction is characterized thus, it is a good deal more plausible to take Descartes and Berkeley as revisionary metaphysicians: even if they *also* believed that their metaphysics accorded with our thinking about metaphysical matters, their chief aim was to present true metaphysical theories.

Similarly, one can improve upon Chihara's distinction by distinguishing between (a) the problem of accounting for or solving the liar paradox as it arises in natural language (or if you prefer: in our conceptual scheme, or in our commonsense view of things) and (b) the problem of making our natural language or conceptual scheme or theory of things better in various respects. (However, as emphasized in the text, the liar paradox by no means necessitates any changes in our language or conceptual scheme.)

[35] Chihara's utilization of a predicate 'true*' introduced by the stipulation that it satisfy the T-schema is unclear also in other respects. Chihara contrasts his inconsistency view on the liar with the "consistency view" according to which "our reasons for accepting [the disquotation schema] are simply empirical in nature". The problem is that Chihara himself presents the "defined Liar", involving 'true*', and similar problems to arise because we (wrongly) take stipulations to be true by fiat: we take the culprit, the counterpart of the disquotation schema with 'true*' substituted for 'true', to be true by fiat. (See the discussion of parallel cases in (1979), pp. 593-4, and also Chihara's discussion of the same issues in (1984), where he says that the semantic paradoxes arise "because something appears to have been made true by fiat or convention".) But this is not incompatible with the consistency view as Chihara has characterized this view: for all that Chihara says, one can say that in cases like the defined liar, the false empirical statement believed to be true is that all stipulations are true by fiat.

an analytic truth about truth. But the disquotation schema is not valid, so this explanation will not do. If this is the way to explain what it is for a convention to govern implicitly the use of an expression, 'true' is not governed by the disquotation schema.

A second way of understanding the claim that 'true' is implicitly governed by the disquotation schema is to take it as saying that speakers invariably take ϕ to be true, and their use of 'true' is informed by it, in that when they use 'true' competently, they make only statements compatible with ϕ. But surely, speakers familiar with the liar reasoning can reasonably - and without manifesting lack of semantic competence - reject some instances of the disquotation schema as untrue.

Chihara's inconsistency view is accordingly objectionably unclear as it stands. However, I have on offer an explication of it under which it comes out both with a determinate content and plausible. Let ϕ be a convention governing the use of 'true' just in case speakers are disposed to accept ϕ by virtue of their competence with 'true' (where this disposition is defeasible); that is, ϕ is a convention governing 'true' just in case it is a meaning-constitutive principle for the expression.

5.7 Yablo on inconsistent definitions

Another inconsistency view that has been put forward is Stephen Yablo's. Yablo motivates the basic idea underlying the view as follows:

> "Defining a word is not *asserting* something but stating a rule or policy for the word's employment." Some such view of definitions is widely accepted; but what is the philosophical payoff if it is right? This is where I have a suggestion to make. Show me an inconsistent assertion, and I will show you a *false* assertion. But an inconsistent rule is *not* false; indeed it may be correct in the only sense that matters, that of according with speakers' semantic intentions.[36]

This opens up the possibility that definitions - or more generally, semantic rules or semantic obligations[37] - be inconsistent. Yablo's idea is that a definition, unlike an assertion, can do the job it is supposed to do, even if not true; and that a speaker can reasonably hold on to

[36] Yablo (1993), p. 147.

[37] Yablo talks about definitions introducing new expressions. But from Yablo's discussion it is quite clear that he takes the conclusions about definitions to apply to and shed light on semantic rules and semantic obligations generally.

a proposition *qua* definition or semantic rule even when she might reject the corresponding assertion (that is, she need not accept the proposition as true).[38]

On what I suppose is the traditional view on semantic rules, they are analytically true, and to obey a semantic rule is to accept it as analytically true.

On the inconsistency view I defend, semantic rules need not be true. For a sentence to be or to express a semantic rule, it need only be the case that semantic competence involves defeasibly holding it to be true, and that the sentence is reference-determining, in the sense that semantic values are constrained by the fact that it should come out as close to true as possible, given other constraints on semantic values.

Like me, Yablo holds that semantic rules need not be true. But he develops this idea a bit differently. His view, as I understand it, is that definitions (semantic rules) function as *instructions* for assignments of semantic values. The semantic values are not necessarily such as to make the definitions as close to true as possible; instead, the definitions determine the semantic values in that the latter are what we arrive at through following the instructions laid down. To illustrate the general idea, consider the following definition:

(1) x is G $=_{df}$ x is F or x is both H and non-G.

(The example is from Gupta and Belnap but serves nicely to illustrate Yablo's account.) According to Yablo, definitions are governed by three principles: *equivalence*, *forcing* and *grounding*. By equivalence, the definiendum and the definiens are materially equivalent. By forcing, something falls in the extension of the definiendum just in case it belongs to the definition's "least solution": just in case it is among the objects the definition *forces* us to treat as falling within the definiendum. By grounding, an object falls under the definiendum "when, and only when, that object has shown itself to satisfy the definiens".[39]

If all Hs are Fs, then definition (1) does not lead us into trouble. This case is not very interesting. But suppose some object a is both non-F and H. Then, according to Yablo, the

[38] One might partly agree with Yablo but hold the following view. Assertions, rules, commands, questions, etc. can all share the same propositional content. For example, an assertion that the disquotation schema is valid, a rule that states that the disquotation schema should come out valid, and a question as to whether the disquotation schema is valid, can all have the same content: the proposition that the disquotation schema is valid. And it is the content that is true or false, so the assertion and the definition and the question all have the same truth-conditions. However – and it is here that something like Yablo's point comes in – a definition, like a question and unlike an assertion, is not correct or incorrect according as the corresponding content is true or false. The correctness of a definition does not depend on the truth of the corresponding content.

[39] Yablo (1993), p. 155.

question whether *a* is G reduces to the question of whether *a* is non-G, to decide which, it appears, we would have to apply the definition again, and so on, ad infinitum. By ***grounding***, *a* is then non-G. But then, by ***equivalence***, *a* is G after all. And so on, ad infinitum.

The definition fails to yield a verdict about *a*. But it seems that it still manages to endow 'G' with meaning. This is seen most clearly from the fact that everything that is not both non-F and H is unproblematically classified as either a G or a non-G.

Yablo defines a rule to be inconsistent if and only if equivalence, forcing and grounding are not jointly satisfiable. As shown by the example of 'G', the consistency of a rule, thus defined, can depend on contingent features of the world (in the present case, whether all Hs are Fs).

The application of all this to truth is straightforward. Yablo outlines rules for truth given which no extension assigned to the truth predicate satisfies both equivalence, forcing and grounding. In particular, however we evaluate the liar sentence, we end up violating one or more of these principles. If we first hold that the liar is non-true, and apply the rules for 'true' and the general principles concerning definitions, we can conclude that the liar is, after all, true; but then, applying the same rules and principles, we can again conclude that the liar is non-true, etc.[40]

From the present standpoint, Yablo's account of the liar is unsatisfactory, for the reason that if it is right, the liar does not exert pull. On Yablo's account, as he presents it, the rules governing 'true' at no point license the derivation of a contradiction, for example that

[40] This is Yablo's definition, and a demonstration of its inconsistency. (I will follow Yablo (1993).)

ϕ is true $=_{df}$
(A1) $\phi = \ulcorner Ra \urcorner$ and a's referent belongs to R's extension,
(A2) or $\phi = \ulcorner \neg\psi \urcorner$ and ψ is false
(A3) or $\phi = \ulcorner \psi \wedge \chi \urcorner$ and both ψ and χ are true
(A4) or $\phi = \ulcorner \forall x \psi(x) \urcorner$ and all its instances are true
(A5) or $\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is true

ϕ is false $=_{df}$
(B1) $\phi = \ulcorner Ra \urcorner$ and a's referent doesn't belong to R's extension,
(B2) or $\phi = \ulcorner \neg\psi \urcorner$ and ψ is true
(B3) or $\phi = \ulcorner \psi \wedge \chi \urcorner$ and either ψ or χ is false
(B4) or $\phi = \ulcorner \forall x \psi(x) \urcorner$ and some of its instances are false
(B5) or $\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is not true

(Yablo (1993), p. 165. I have added the numbering of the lines.) Let the liar sentence, λ, be: '$\neg(\ulcorner \lambda \urcorner$ is true)'. We can assume that from the outset, the liar sentence is counted as not true, for by grounding a sentence must have shown itself to be true to be true. Then, by (B5), '$\ulcorner \lambda \urcorner$ is true' is false. By (A2), '$\neg(\ulcorner \lambda \urcorner$ is true)' is true; and since this is λ, λ is true. By (A5), '$\ulcorner \lambda \urcorner$ is true' is true; and by (B2), '$\neg(\ulcorner \lambda \urcorner$ is true)' is false. Since this is λ, λ is false. But then '$\ulcorner \lambda \urcorner$ is true' is false; and we are back where we started.

the liar is both true and non-true. Rather, the rules license first taking the liar sentence to have one truth-value, then taking it to have another truth-value, etc. This means that if a speaker concludes a contradiction, she has not applied the rules correctly.

This is my first point against Yablo: that his account fails to accommodate the fact that the liar exerts pull.[41]

A second problem I have with Yablo's account is the following. Take the case of 'G' in a situation where some H, a, is not an F. According to Yablo's account as presented, we can first apply the rule for 'G' and the principles concerning definitions to yield one truth-value for "a is G", then apply these principles to yield a different truth-value for this sentence, and then again apply these principles, again overturning the verdict given, etc. But Yablo never explains why, having concluded (say) first that "a is G" is not true and then that "a is G" is true, we should not conclude that this sentence is both true and non-true, but instead we should take the later verdict to overturn the former. (I am not saying that Yablo is not right to say that, in some sense, we should take the latter judgment to overturn the former: all I am saying is that no justification has been provided for it.)

How important is this omission on Yablo's part? Not very important at all, it may seem. For Yablo, the flip-flopping between taking "a is G" to be true and taking it to be non-true is only reflects the fact that the rules governing 'G' yield no stable verdict on the truth-value of "a is G". This latter claim is true even if we allow that a speaker who has concluded first that "a is G" is true and then that it is non-true can, following semantic rules, conclude that "a is G" is both true and non-true. Yablo's account can be modified as follows.

Having concluded first that "a is G" is true and then that it is non-true, a speaker can, following semantic rules, equally well conclude that the sentence is both true and non-true as to take the latter claim to override the former. For just as, in the latter case, semantic rules can again be appealed to in order to override the verdict, semantic rule can be appealed to in order to criticize the former verdict. For instance, one can take the rules for negation to rule out the possibility that there are true contradictions.[42]

[41] That the liar does exert pull is of course nothing I have argued in this chapter. See chapter 1 for arguments to this effect.

[42] I call this a modification of Yablo's account because it is not part of the account in Yablo (1993), nor is it part of the positive account he develops from section IV onwards in Yablo (1993b). However, I should note that in the early parts of (1993b), he does consider the possibility of assigning to the liar sentence the status "both true and false", and concludes regarding this suggestion, as regarding others, that it conflicts

If we modify Yablo's account as outlined, then, it may seem, we respect its basic motivation, which is that inconsistent definitions give rise to sentences concerning which we cannot give stable verdicts. Besides, if the account is modified as outlined, then the resulting account accommodates the fact that the liar exerts pull: for then the account allows that speakers who derive a contradiction follow semantic rules.

However, the modification is not as minor as it may seem. As Yablo presents his account, it sounds as if it constitutes a precise method for determining the semantic values of sentences, given the semantic rules. But this depends crucially on assuming that a speaker who has first concluded that ϕ is true and then that ϕ is non-true, should take the latter judgment to override the former. For if a speaker can conclude that ϕ is both true and non-true, it seems that there is nothing to prevent her from concluding from this, using *ex falso quodlibet*, that ψ is true, where ψ is any sentence. Of course, the judgment that ψ is true can later be overturned. But ψ would still come out unstable; and a guiding idea behind Yablo's account of inconsistent rules endowing sentences with meaning is that some sentences still receive stable verdicts. These are the sentences which are clearly true and clearly false.[43]

5.8 'Solutions' to paradoxes

Most of the literature on the paradoxes consists in attempts to 'solve' them. In light of this one may expect it to be at least somewhat clear what a solution should involve; but I very much doubt that it is. Here, however, are some general points about what a solution may be required to involve.

In (1993), Gupta and Belnap distinguish between the descriptive and the normative questions that may be raised by a paradox such as the liar. The normative questions

with the rules governing 'true' (pp. 373ff).

[43] The remarks in the last paragraph show that Yablo's account is not easily modified so as to allow for the exertion of pull. This brings us back to the question how Yablo can justify, in an intuitively reasonable way, taking the derivation of the non-truth "a is G" to override the derivation of the truth of this same sentence, rather than concluding that the sentence is both true and non-true. The suggestion that comes to mind is that one is allowed to draw further conclusions from a sentence only after having ascertained that this sentence has not been arrived at in a way that violates one of the constraints on definitions (equivalence, forcing and grounding). This suggestion is developed in Appendix 1 of Yablo (1993b).

I should explain, incidentally, why the same problem does not afflict a Kripkean theory of truth. On Kripke's theory of truth, once a sentence receives a (classical) truth-value, it keeps this truth-value. A sentence is at unevaluated at the outset, and, the though is, simply lacks a truth-value, and once it has received a truth-value it keeps it.

concern the *construction* of "paradox-free concepts of truth". The descriptive question concerns how to account for actual language use.

The same remarks that applied to Chihara's distinction between the diagnostic and the preventative problem apply to Gupta and Belnap's distinction. The distinction is characterized in a theoretically loaded way. But one may hold that for reasons independent of paradox, one must distinguish between, on the one hand, questions about the nature of expressions and concepts in use (these can be called *descriptive* questions), and on the other hand, questions about the properties of a notion of truth fit to do theoretical work (these may be called *normative* questions). The problem posed by the liar for questions of the latter kind is only this: the liar reasoning shows that there are certain theoretical virtues which cannot be possessed jointly. For example, a truth predicate cannot satisfy the T-schema if the logic is classical and the language contains sufficient means for self-reference. But if our concern is with a *solution* to the liar, it is fairly clear that we are, or ought to be, concerned with descriptive questions. For the liar, *qua* paradox, is a piece of reasoning employing a truth predicate, a negation sign, and means for self-reference that are *already in existence*, and the paradox is that it *appears* that these expressions, or what they stand for, are such that a contradiction is true. But no contradictions are true. Something has gone wrong; but what? So a first remark on solutions is that they should concern expressions or concepts actually in use.[44]

Moreover, it is or should be relatively uncontroversial that a solution should at least incorporate an account of *truth conditions*, given which one can show either (as on most views) that the paradoxical reasoning is not sound, or (as on Priest's view) that the paradoxical reasoning is sound, but that this can be accepted since contradictions do not entail every other proposition.

But first, if the view I have argued for is correct, then accounts of truth conditions are likely to fail, by themselves, to be very enlightening. When a paradox arises because the language is inconsistent, the acceptable assignments of semantic values to the expressions

[44] To say that a solution to the liar paradox constitutes an answer to a descriptive question is not to say that all, or even most, significant questions raised by the liar reasoning are descriptive. A question looming large in the previous chapter, and made problematic by the liar reasoning, was that of what properties a truth predicate introduced for theoretical purposes should have. This question may legitimately be held to be more interesting than that of what are the properties of the truth predicate actually used. All I am emphasizing is that, even so, an answer to this normative question does not constitute a solution to the liar paradox.

employed are those that come as close as possible to maximizing the correctness of the meaning-constitutive principles; no account can make all of these principles correct. It is offhand likely that the members of the class of acceptable assignments will be a very heterogeneous bunch. There will be very many of them, and very few interesting properties will be common to all of them. (This is not *inevitable*, of course, but it seems like a very live option.)

Second, on *any* view about the nature of the liar paradox, an account of truth conditions cannot by itself constitute the whole solution to the paradox.

Given my inconsistency view, an account of truth conditions is by itself powerless to explain why the liar paradox exerts pull. But such an explanation is evidently essential if we want to know what is going on in the liar reasoning. And on a standard account of the liar, an account of truth conditions must be supplemented by some kind of story about *why* the truth conditions are as they are. Such a story is perceived to be necessary not only dialectically, to convince one's opponents; but without it, it has not been explained what is going on in the paradoxical reasoning.

So far I have discussed only what I take to be common ground between proponents of different views on the paradoxes. I want now to go over to less uncontroversial territory. In (1970), Martin stated that "A solution [to a paradox] consists in convincing ourselves that one of the assumptions that led to the contradiction is after all not so plausible".[45] And he goes on to qualify this: "What is wanted, ideally, is the uncovering, the making explicit, of some rulelike features of our language which when considered carefully have the effect of blocking at least one of the assumptions of the argument".[46] It is hard to state precisely what assumption Martin makes, but roughly, it is that there be some reason, pertaining to the nature of truth or (say) the nature of negation or universal quantification, that we can point to when arguing that one of the principles leading to contradiction is false. Martin demands a particular kind of answer to the question of why such-and-such a principle employed in the liar reasoning fails to be true or valid.

According to Martin, moreover, we cannot reasonably reject one of the assumptions leading to contradiction simply because it conflicts with other plausible assumptions. A solution must provide an independent reason for taking the assumption singled out as the

[45] Martin (1970), p. 91.
[46] Ibid.

culprit to be the culprit.

If something like my account is correct, then no solution of the kind Martin seeks can be found. The rules of our language are themselves inconsistent. We can conclude either that Martin's proposed adequacy condition on a solution is too strong, or that the liar paradox is in this sense *unsolvable.*

5.9 The strengthened liar

Against the background of the discussion of reasonable conditions on an acceptable solution to the liar paradox, I can now state my thesis about the status of this paradox given my inconsistency view: if we accept Martin's adequacy condition on a solution, the liar paradox is *unsolvable*; if we do not accept his condition, then the liar paradox is already for all intents and purposes *solved.*

Before I argue that, however, I should consider the problem traditionally considered most serious for accounts of the liar paradox: the strengthened liar.

Suppose someone were to argue, "Given your account, with the apparatus of truth under an acceptable assignment, we can construct a strengthened liar sentence saying of itself that it is not true under all assignments. If it will not come out true under all assignments, it is true, and hence true under all assignments. If it comes out true under all assignments, it is not true, and hence not true under all assignments. This is in essence the same problem we have struggled with all along (or one of the problems we have struggled with all along). How can you then say that the liar is for all intents and purposes solved?"

This is the strengthened liar reasoning applied to my account. The objector uses the conceptual resources I employ – in particular, the notion of *truth under an assignment* – and presents a new version of the liar paradox different from, but modeled on, the original one.[47]

To explain how I deal with the strengthened liar, let me first back up and again explain, briefly, how I deal with the liar itself. My claim about the liar paradox is that it arises be-

[47] In fact, what I have throughout called a liar sentence, namely a sentence saying of itself that it is not true, is sometimes regarded as a strengthened liar sentence. A liar sentence is then regarded a sentence that says of itself that it is *false.*

A different use of the locution "strengthened liar", the one I use here, in the main text, is to take a strengthened liar problem to be a problem raised using the conceptual apparatus introduced to deal with one or more of the problems in the family of liar problems.

cause our semantic competence disposes us to accept principles that are jointly inconsistent. For example, semantic competence with 'true' involves being disposed to accept that the disquotation schema as valid, semantic competence with 'not' involves being disposed to accept that this expression satisfies the proof-theoretic rules characterizing classical negation, etc. Accordingly, our semantic competence disposes us to accept jointly inconsistent principles. As a result, one or more of the expressions employed in the liar reasoning *misfires*, in the sense that the principles it actually satisfies are not exactly the principles our competence disposes us to take it to satisfy.

It will come as no surprise that I say exactly the same thing about the strengthened liar. The strengthened liar arises because competence with the expressions employed in the strengthened liar reasoning involves being disposed to accept jointly inconsistent principles; as a result, one or more of the expressions employed misfires.

It may be argued, however, that it is a major problem for me if some expressions crucially employed in the strengthened liar misfires. For since the expressions employed in the strengthened liar reasoning are the ones I employ in stating my positive theory, it follows that my theory cannot be stated or expressed.

However, here I can refer back to chapter 4, where I argued that the expressive resources of the metalanguage can justifiably be taken to outstrip those of the object language. My semantic theory for English, involving talk of the semantic values of English expressions under different assignments, should be thought of as given in an essentially stronger metalanguage.

It may be thought that if I so crucially refer to the possibility of ascending to a stronger metalanguage, then that, and not the suggestion that our language is inconsistent in the sense characterized, is what is doing the real work when I talk of having the resources for solving the liar (an issue to which I will return shortly)

However, the suggestion that our language is inconsistent plays a role in making plausible the suggestion that we can ascend to a stronger metalanguage. If my suggestion, that the semantic theory I have outlined be given in a stronger metalanguage, is accepted, then we are faced with the difficult question of what to say about the semantics of the expressions of the *object language* which certainly *appear* to express the notions employed in the metalanguage. The suggestion that our language is inconsistent plays an obvious role in such an answer. There are in fact two slightly different issues here. One pertains to the relevant expressions of the object language themselves: what semantic properties do they have? The other concerns the *relation* between the predicate of English "is an English sentence true under every acceptable assignment", and the homophonic metalanguage predicate. As regards the second issue: These predicates certainly seem very similar in meaning – even synonymous – and yet their extensions are, according to my proposed answer to the question we are considering, different. How can this be? Given the suggestion that our language is inconsistent, there is an attractive, straightforward explanation. The competence dispositions associated with the two predicates are the same. As regards the first issue, the semantic properties of the object language expressions are whatever properties are such as to make the meaning-constitutive principles for expressions of the object language jointly come out as nearly correct as possible. It is of course a highly non-trivial task to arrive at the semantic properties from the jointly inconsistent principles; and it is moreover likely that it will be quite indeterminate what these semantic properties are.

5.10 The liar dissolved

I can now make good on the claim that the liar paradox is either (if Martin's adequacy condition on solutions is accepted) unsolvable or (if not) practically solved. I shall argue that, for all intents and purposes, we have at hand the resources for giving 'solutions' satisfying the other adequacy conditions on solutions.

First, as outlined, given my view on the liar paradox, standard objections to the accounts of truth conditions currently on offer can be met. It is acceptable for accounts of truth conditions to have certain counterintuitive consequences, which cannot be explained away. For since the meaning-constitutive principles for the expressions employed in the liar reasoning cannot all be true and valid, any acceptable assignment of semantic values to these expressions will have to render some such principle false. Accordingly, any acceptable assignment is bound to have counterintuitive consequences that cannot be explained away.

It is of course highly unlikely that we will have at hand all accounts of truth conditions that are acceptable (that maximize the correctness of our competence dispositions), but the most important accounts on offer seem to at least be among the acceptable accounts of truth conditions.

Second, we have at hand an account of *why* the liar paradox arises. It arises because of the joint inconsistency of the belief-forming dispositions bound up with competence with the relevant expressions.

Naturally, it is an open question exactly which accounts of truth-conditions currently on offer actually do maximize the correctness of our competence dispositions and which of the accounts currently on offer are among them; and it is likewise an open question exactly what our competence dispositions are. As regards the competence dispositions, I find it attractive to think that Tarski got matters essentially right. We are disposed by our semantic competence to take 'true' to satisfy the T-schema and to take the logical expressions to obey the laws of classical logic. But nothing I have argued hinges on taking this view to be correct.

When I say that the liar paradox is for all intents and purposes solved, I mean only that the principled obstacles that seemed to lie in the way of a solution no longer seem to present obstacles. A solution of this kind seems to me sufficient. For extremely few words of our language can we actually confidently state either the semantic values or the meaning-constitutive principles. There is no reason why 'true' should be any different. What we can reasonably ask for, when we ask for a solution to the liar paradox, is not the exact semantic values and meaning-constitutive principles of the expressions employed, but instead only general conditions satisfied by the semantic values and meaning-constitutive principles.

5.11 Concluding remarks

I have accomplished three main things in this chapter. First, I called attention to an assumption that standard accounts of the liar paradox have in common, which we have no good reason to believe: the assumption that the suspects do not have the same meaning status. Standard accounts should thus be rejected. It deserves emphasis that my rejection of standard accounts did not rely on any specific features of my inconsistency view. I outlined two views competing with mine, which, like mine, deny the assumption that the suspects have unequal meaning status. Second, I discussed and criticized "inconsistency views" (ostensibly) competing with mine. Common themes in this discussion were that crucial elements of the competing views have not been satisfactorily explained (but that the necessary explanations can be given if my account is adopted), and that whatever plausibility these views have is shared by my view. Third, I explained why, given my arguments regarding the liar in chapters 1 and 4, we have as much of a solution to the liar as we can reasonably hope for.

Chapter 6

The Significance of Questions About Truth and Logic

6.1 Introduction

In previous chapters, I have discussed the liar and the sorites paradoxes. But I have hardly addressed the questions that most theorists take to be the most central ones raised by these problems: the question of how these arguments go wrong and the question of which logical and semantic principles can be upheld in the face of these arguments. (The exception is the tentative defense of bivalence in the face of vagueness, which I presented in chapters 1 and 3.) In part, the reason why I have devoted so little attention to these questions is that, as argued in previous chapters, if my inconsistency view is correct, the answers to these questions are likely to be highly indeterminate. But there are also deeper reasons for my relative lack of concern with the questions normally discussed in connection with the paradoxes. In this chapter, I will go through these reasons.

Consider questions such as

> Is the principle of bivalence valid?
>
> Can contradictions be true?

which concern truth, and questions like

> Is the law of excluded middle valid?
>
> Is *ex falso quodlibet*, which says that from a contradiction everything follows, valid?

which concern logic. I shall here talk about "questions about truth and logic", and it will be questions like these that I will refer to. Naturally, there are also questions concerning truth and logic that are of a different sort, like "Is truth correspondence with the facts?" and "Is (so-called) higher-order logic really logic?".

The questions that I will address are about which principles about truth and the laws of logic are valid. They have been widely discussed in the philosophical literature. Here I shall consider how these questions must be understood in order for them to have the kind of significance generally accorded to them. In particular, I shall argue for a skeptical conclusion: for these questions to have the philosophical importance normally attached to them, certain theoretical and non-obvious claims will have to be true.[1]

Discussion of the semantic paradoxes and the sorites paradox often centers around their consequences regarding truth and logic. If skepticism about such questions is justified, then we should be skeptical also of the importance of traditional discussions of the paradoxes. If the paradoxes are philosophically significant at all, their significance may have to consist in their consequences for questions other than those about truth and logic. I shall sketch some alternatives.

In section two, I shall go through the most obvious suggestions concerning how to regard disputes about truth and logic, and conclude that if these disputes are understood in accordance with one of these suggestions, they do not have the importance usually attached to them. In section three, I shall cover roughly the same ground once again, but now with special reference to disputes about the consequences of the liar paradox with regard to truth and logic. Section four will have more of a positive character: I will there list some ways in which discussion of the paradoxes may be philosophically significant even if disputes about the validity of principles about truth and laws of logic is not. In section five I shall return to the questions of truth and logic, presenting a positive proposal concerning how to understand them. Section six will be devoted to spelling out the implications of understanding the questions about truth and logic as suggested for debates about those questions.

[1] George Boolos is credited with having said, "When I hear someone talk of philosophical significance I reach for my gun" – and it is hard not to feel somewhat sympathetic to that remark. Here, for skeptics, are a few words about what I mean by philosophical significance.

When saying that something is philosophically significant, I mean that it is *theoretically important* for some topic or other; and what makes this importance specifically *philosophical* is (i) that it has the appropriate *generality*, and (ii) that it can be, and is most reasonably, studied by (relatively) *a priori* means.

6.2 On some natural construals of questions about truth and logic

Let us focus on one particular set of questions about truth and logic: the questions of whether there are true contradictions and whether *ex falso quodlibet* is valid. The ***dialetheist*** claims (a) that there are true contradictions and (b) that *ex falso quodlibet* is not valid, whence a *paraconsistent logic*, in which this law is not valid, is the correct one.[2] Dialetheism sounds like an extremely radical thesis. It sounds as if the dialetheist is affirming something the denial of which the rest of us take for granted. This is what makes dialetheism an especially suitable thesis to discuss; for, as we shall see, it is very difficult to say what is at stake between the dialetheist and the rest of us.

It may seem very easy to characterize what is at stake between us and the dialetheist. It may even seem that I have already managed to do so, when I said that the issue concerns whether there are true contradictions: whether there are sentences such that both they and their negations are true.

But how should, for example, "negation", as it occurs in this statement, be understood?

We cannot, of course, identify a negation sign as any sign that satisfies such-and-such proof-theoretic rules. That would beg the question against one or the other of the participants to the debate. The issue concerns exactly which proof-theoretic rules negation satisfies. (We could avoid question-begging by choosing proof-theoretic rules such that all participants agree that negation satisfies them. But then the dispute would still be a non-dispute: all parties would be right.)

A better proposal is to suggest that negation is what is expressed by some suitable (presumed) logical expression of natural language, such as 'not' or 'it is not the case that'. Dialetheism is then understood as the thesis that for some sentence S of English, both S and 'it is not the case that S' are true. Call this proposal regarding how questions about truth and logic are to be understood *the natural language proposal.*

However, it is fairly clear that the natural language proposal is misguided, or, slightly more cautiously, that *if* this is how the issue is to be understood, then participants to the debate have seriously mistaken views about the significance of the issue.[3]

[2] See Priest, e.g. (1979), (1984b) and (1987).

[3] As Priest remarks, discussing in a recent article (1999) how to understand issues about negation, we

(1) Suppose it turns out that no sign of English expresses classical negation, that is, satisfies exactly the proof-theoretic rules satisfied by classical negation. The classical logician would, I presume, not necessarily be much perturbed by that: she would not conclude that the laws of classical logic are not valid but rather that no sign of English expresses real, logical negation. (Or suppose that contra Grice, sentences like "they got married and had a baby" and "they had a baby and got married" can differ in truth-value, and the believed-to-be logical particles of natural language are not truth-functional at all. *If* the present construal of debates about logical laws were correct, there would in that case seem to be only *losers* in the debate between classical logicians and dialetheists.)

The reasoning in the above paragraph relies on a generalization about the thinking of proponents of classical logic. But the next reason for rejecting the natural language proposal does not rely on any such hypothesis.

(2) If the natural language proposal is correct, the issue hinges on contingent features of English. It is then possible that, for example, classical logic is the logic of English, but *paraconsistent logic* is the logic of *French.* But what is the philosophical interest of finding out that classical logic is the logic of English, if other logics are the logics of other natural languages?[4]

Another very natural way of construing what is at stake between classical logician and the dialetheist is to construe it as an issue of *rational reconstruction.* Call this *the rational reconstruction proposal.* The question is then about what properties the logical particles

would not want to regard a theory of the English word 'not' as a theory of negation, for the reasons that (a) some uses of 'not' rather serve to reject connotations of what is said ("I am not his wife: he is my husband"), and (b) inserting 'not' does not always negate the original sentence, as witnessed by the fact that we would not call "some man is not mortal" the negation of "some man is mortal" (p. 103).

On the basis of these remarks, Priest takes a theory of negation to be a theory of the relation of contradiction. He does not, however, address the issue of whether such a theory primarily is a theory of some of the behavior of 'not' or "it is not the case that", singled out on the basis of linguistic considerations, or whether we should think of this theory as of something different altogether.. If the latter, Priest has not made clear just what his proposal is. If the former, at least the second objection to the natural language proposal applies to Priest's proposal.

[4] One response frequently made to this argument, when I present it in conversation, is that it relies on taking the truth or falsity of the claim that there can be true contradictions to be an *analytic* or *verbal* matter, and that it accordingly relies crucially on the analytic/synthetic distinction, widely regarded with skepticism. But suppose that, as a matter of fact, the law of non-contradiction is valid, not analytically so, but because the world conspires with the meaning of 'it is not the case that' to make every sentence of the form "S and it is not the case that S" false. And suppose further that in some other language actually used, French, say, there is some expression such that it satisfies the proof-theoretic rules characterizing paraconsistent negation, likewise not analytically, but as a synthetic matter. The question is what the philosophical interest is of the fact that it so happens that, in the language we use, some expression satisfies the rules characterizing classical negation, not the rules characterizing paraconsistent negation. See further the remarks on the logical realism construal, below.

used *ought* to have: what logic is most *useful* for various practical and theoretical purposes? Some problems for the rational reconstruction proposal are the following:

(1) There is little reason to suppose that any single logic will be the most useful for all theoretical purposes. Construing the issue this way would mean encouraging a pluralism, which it appears participants to the debate would not be happy with.

(2) Relatedly, it seems that proponents of classical logic can happily grant that working in other 'logics' is more useful than working in classical logic. Whatever one thinks of quantum logic, one can hardly deny that it could have turned out that quantum logic was of instrumental value – even though its acceptance as a useful tool would not have constituted an overthrow of classical logic.

(3) Recall that useful scientific theories have turned out to be inconsistent (and have been used also after the inconsistency has been discovered). If we work in such theories, paraconsistent logic is likely to be more useful than classical logic, in which *ex falso quodlibet* is valid. So if the rational reconstruction proposal is accepted, the widespread skepticism – even horror – that is the common reaction to dialetheism is scarcely justified. This could of course simply be a point in favor of the dialetheist. But a fairer assessment of the issue seems to be that the opponents of dialetheism do not view the matter as one of rational reconstruction.

On the strength of the above considerations, we should reject the rational reconstruction proposal.

In rejecting the natural language proposal and the rational reconstruction proposal, I do not mean to imply that the questions of the semantics of the truth predicate and the logical particles of natural language fail to be important, or that the project of constructing maximally useful logics and truth theories lacks significance. What I am suggesting is only that they lack the kind of philosophical significance usually attached to the questions of truth and logic.

The way (many) philosophers of logic reason about the questions about truth and logic would be explained and justified if we ascribed to them the view that the proper answers to these questions are the ones that *both* come closest to capturing the behavior of our actual logical particles *and* are most 'useful'. The former criterion (concerning capturing the behavior of logical particles) would justify (according to orthodoxy – which, for my two cents, seems to have things right in this case) ruling out dialetheism, for dialetheism is very *far* from capturing the behavior of our ordinary logical expressions; The latter criterion (usefulness) would justify setting aside as irrelevant the possibility that our actual logical particles aren't exactly classical (for, again according to orthodoxy, even if our actual logical particles are not *exactly* truth-functional and classical, they at least come close to being so). However, if this is how the questions about truth and logic are being tacitly understood, why bother about them at all? The question of how the actual logical particles behave, whether philosophically significant or not, is an interesting question. And the question of which logic would be most useful is likewise interesting. But why bother about the *synthesis* of the two?

Call the proposal on the table the *synthesis proposal.* To me it appears *obvious* that the synthesis cannot be interesting: two entirely different kinds of issues are run together in an arbitrary way. But let me anyway try out two possible ways of justifying interest in the synthesis.

The first is by appeal to a principle of charity. The idea would be that at bottom, we are interested in describing natural language; but the principle of charity, and hence considerations of usefulness, are crucially implicated in the interpretation of natural language.

But such appeal to charity will not serve to justify interest in the synthesis, even if charity principles are crucially used in interpretation. For when used in that way, charity principles play a role only in determining what natural language expressions actually mean: such appeal does not justify departures from the actual meanings of natural language expressions; but that is what is needed if interest in the synthesis is to be justified.

Second, one may take natural language expressions to be, to a large degree, indeterminate in meaning, and regard considerations of usefulness as relevant to the question of how to make the expressions more determinate or precise.

But even if, in general, meanings can be thus indeterminate and considerations of usefulness can enter in when we decide how to make meanings more determinate, this suggestion seems in the present case utterly misguided. It *compounds* the problems of the natural language proposal and the rational reconstruction proposal. The first problem regarding the natural language proposal, namely that the behavior of the logical particles of natural language is inconsistent with them expressing the classical truth-functions, applies equally here. And the same goes for the problem regarding the rational reconstruction proposal, namely that one and the same logical system is unlikely to be useful for all practical and theoretical purposes.

Having rejected the two most natural and straightforward ways of construing the debate between the classicist and the dialetheist, and their synthesis, I shall now consider some other suggestions that may come to mind.

First, Quine has proposed that there can be no alternative logics, as constraints on acceptable translation forbid us to translate someone as rejecting a statement we take to be a logical truth. If Quine's thesis is accepted, then it may seem that contrary to what I suggested above, the question of the logic of English is philosophically interesting, for the logic of English is also the logic of every other language possibly used.

Even setting aside misgivings concerning Quine's thesis about translation itself, *the Quinean translation proposal* (as we may call the proposal under scrutiny) is badly misguided. Quine does *not* say that we cannot translate any expression of a foreign language as (say) paraconsistent negation if we use classical negation: he only says that if we interpret someone as believing "P and not P" to be true, for some P, his "not" must not be translated into *our* "not". But if we can interpret some expression of a foreign language as paraconsistent negation, we can interpret some other people as employing paraconsistent negation and not classical negation.[5]

Second, another suggestion concerning how to construe the dispute that takes its cue from Quine. In "Truth by Convention", Quine argued against the logical positivists' contention that logically true statements are true by convention. Quine's argument was, very briefly put, that there are infinitely many logically true statements, and that we can lay and have laid down only finitely many conventions. To derive the infinitely many truths from these finitely many conventions, logic is needed. But this appeal to logic cannot itself be given a conventionalist foundation.

Someone who wishes to defend the significance of questions about truth and logic might try to exploit Quine's basic idea as follows: Such debates, it may be said, concern not what the contents of such and such statements are but what the logical consequences of statements with such-and-such contents are. The idea is that in principle, the classical and the paraconsistent logician can agree on the contents of all sentences of English, and that their disagreement concerns the logical relations between sentences of English (and other

[5] Also, even were it the case that, if classical logic is the logic of our language it is the logic of all languages, still our logic (and hence the logic of all languages) could for all that has been said have been different.

languages), *given* that the sentences have these contents. For lack of a better name, call this *the logical realism proposal.*

One potential problem with the logical realism proposal stems from the difficulties in ascertaining that the classical and the paraconsistent logician indeed agree on the contents of English sentences, given that they disagree on the logical relations between these sentences.

But let us set such worries aside. The problem for the logical realism proposal that I want to focus on is rather the following. Suppose the classical and the paraconsistent logician do agree on the contents of sentences of English. Suppose that their dispute then concerns whether a sentence of the form "P and not P" – understood as they both understand it – entails every other sentence or not. Assume further that, as it happens, the classical logician is right and the paraconsistent logician wrong. But now suppose there is some *other* class of contents, expressible by sentences of the form "P and not* P", where 'not*' is some actually or possibly used expression different in meaning from 'not', such that the contents expressed by sentences of this form bear the logical relations to other contents that the paraconsistent logician says are borne by the contents of sentences of the form "P and not P". The paraconsistent logician, though wrong about the logical properties of contents expressed by "P and not P", would be right about the contents expressed by the would-be sentences of the form "P and not* P". Given these assumptions, which logic is the right logic? Granted, there is a *temptation* to say that classical logic under these assumptions is the right logic, just because it is classical logic that gets negation in English right. But to succumb to this temptation is obviously, in effect, to interpret the question of what logic is the right logic as a question about the logical properties of expressions of English: but I have already argued that the issue had better not be thought of that way.

The last suggestion concerning how to understand questions about truth and logic that I shall consider, for now, is that the debate between the classicist and the paraconsistent logician concerns which logical operations *there are.* Thus, Graham Priest argues in (1990) that there is no such thing as classical, Boolean negation. (It is not just that classical negation is not expressed in this or that language: there just is no such thing.) I shall call this suggestion *the existence proposal.*

If the questions about truth and logic are understood in accordance with the existence proposal, they are non-trivial, they do not hinge on contingent features of our natural language or our conceptual scheme, and their answers are not interest-relative. Moreover, suppose that the classical and the paraconsistent logician were to agree about what logical operations there are, and disagree only about which logical operation is properly identified as negation. Then the dispute between them would be merely verbal. Arguably, a dispute is verbal if there is a language L such that (a) all participants to the dispute can agree that everything they want to express (with respect to the issue at hand) can be expressed in L, and (b) the disputants can agree on the truth-values of all sentences of L.

If the disputants disagree only about which logical operation is properly identified as negation, then these conditions are both satisfied. This is a further argument against the natural language proposal. But if, on the other hand, the disputants disagree about *what there is to be expressed*, as with the existence proposal, the issue is not merely verbal.

Before I go into the problems with this proposal, let me introduce some terminological conventions that will be useful later. Let us call properties and logical operations alike *notions*. What Priest argues is that some notion we thought existed does not actually exist. One could also use Priest's argument to reach a different conclusion: that some notion we thought we could express in the language we use cannot be expressed there, although it does exist. Let us say that such notions are *ineffable*. It will be useful to have a simple means of saying that a particular (presumed) notion *either* does not exist *or* is ineffable. Let us say that such notions are *unavailable*.

As Priest notes when arguing that classical negation is unavailable, what logical considerations show can only be things of the form: *if* these and these things are available, *then* those and those things are not available. Priest argues that some notions are available, given the availability of which classical negation is unavailable (and given the further assumption that there are no ineffable notions it follows that there is no such thing as classical, Boolean negation). The classical logician can respond by saying that the classical logical operations are available, and given the availability of *these* notions, some notion Priest takes to be available is not, in particular truth conceived as satisfying the T-schema,

$$\ulcorner p \urcorner \text{ is T if and only if } p.$$

This appears to lead to a stand-off. What kinds of considerations could settle the issue? The kinds of considerations that come to mind are (i) questions about the semantics of expressions we actually use, and (ii) questions about the usefulness of various notions.[6] But

[6] Priest seems to opt for the latter alternative.

this means that the present proposal concerning how to understand what is at issue between the dialetheist and us faces the same problems earlier proposals we have already rejected do. There are also some further problems, at least if considerations of type (i) are taken to settle the issue. For as noted, it appears possible that *we* use *classical* negation and *some other community* uses *paraconsistent* negation. (We shall return to these problems later, in our discussion of the liar paradox.)

I have gone through many different ways of understanding the questions about truth and logic and have found them all wanting. I have not shown that these questions really fail to be important. Later I shall present a more positive proposal concerning how to understand them. But for the time being, I shall continue in the skeptical mode.

6.3 The discussion of the liar paradox

Most discussions of the sorites and the liar paradoxes centrally concern the implications of these paradoxes for principles like bivalence and the law of excluded middle. Thus, some, presumably rather much, of the significance of the paradoxes has been taken to be their implications for truth and logic. If we must reevaluate the significance of the traditional questions concerning truth and logic, we must also reconsider the significance of the paradoxes.

My remarks in this section and the next will all concern the liar paradox: but most of them – enough for the main points to generalize – apply equally to other paradoxes discussed in the philosophical literature, such as the sorites paradox.

What kind of problem is the liar paradox? It *can* be thought of as merely a problem concerning the *expressive limitations* of various *possible languages.* Tarski's theorem, the classic result in this area, says that in no language which can talk freely about its own expressions are both truth conceived as satisfying the T-schema and the classical truth functions expressible. This theorem serves as a starting point for much work on the liar: much formal work on the liar has in effect been devoted to the question of how *close* we can get to expressing these things. For example: taking a language which can talk freely about its own expressions and in which the classical truth-functions can be expressed, how close can a predicate of this language come to satisfying the T-schema?

The expressibility question (or the *availability* question) raised by the liar can be stated in a more general way. It is widely held that the liar reasoning forces us to abandon bivalence and classical logic, for example because the liar sentence is regarded as neither true nor false. In various many-valued logics, it is rather trivial that the T-schema is not valid, for under standard ways of reading the biconditional in a many-valued logic, the biconditional comes out untrue when the sentences flanking it are neither true nor false. In such a setting, the question of expressibility, or, rather, availability, takes a different form. Consider the E(quivalence)-schema:

$\ulcorner \mathrm{T}(\ulcorner \phi \urcorner) \urcorner$ is equivalent in truth-value to $\ulcorner \phi \urcorner$.

This is the natural counterpart of the T-schema for a truth predicate in a non-bivalent setting. The relevant availability question in a non-bivalent setting is whether (a) a predicate (a would-be truth predicate) satisfying the E-schema can be expressed, or (b) strong negation (taking both 'gappiness' and every truth-value other than truth into truth) can be expressed.[7] In a classical framework, satisfying the E-schema amounts to satisfying the T-schema and expressing strong negation amounts to expressing negation. In what follows, I shall sometimes simply assume that the property truth – if such a thing there be – satisfies the E-schema. The assumption is made only for simplicity. (If you are skeptical about it, substitute for 'truth' either "truth conceived as satisfying the E-schema" or "a property satisfying the E-schema".)

The expressibility question raised by the liar reasoning the question of how close a language can come to both having a predicate that satisfies the E-schema, and having the means to express all of classical logic. This question is obviously answered by formal logical considerations.

What formal work cannot and is not meant to show is what actually *is* expressed in English: what can settle that, if anything, is attention to the details about how English is used. This point raises two issues.

First, it is an open question whether the thoughts, intentions, conventions and practices of speakers of English – *whatever* it is that determines the meanings of expressions of English – determine exactly which notions are expressed. In fact, it would seem rather miraculous if they did. It is not the case that both strong negation and truth, conceived as satisfying

[7] This is oversimplified. (One may for example think that it is absolutely unrestricted quantification that cannot be expressed.) Other options are in principle available. However, both here and later I shall restrict my attention to these two options.

the E-schema, can be expressed in English. What would determine whether it is strong negation or truth (or perhaps neither) that can be expressed in English are the thoughts and practices of English speakers. But it seems at the outset – before we encounter the liar problem – equally plausible that strong negation should be expressible as that 'true' satisfies the E-schema. The claim that truth satisfies the E-schema and the claim that some expression satisfies the rules that characterize strong negation have the same analytical feel to them, as it were. Speakers not aware of the liar reasoning would presumably both take strong negation to be expressible, and take 'true' to satisfy the E-schema. But then it surely cannot be assumed that the thoughts and practices of speakers of English also determine which assumption it is that should be rejected when it turns out that there is a clash.

Consider a hypothetical case. Suppose that both the logical particles and the predicate 'true' were introduced by stipulation, and that for hundreds of years, it was not discovered that these stipulations could not be jointly true, and it did not occur to anyone to doubt them. Then the liar reasoning is discovered, and people realize that these stipulations cannot all be true. Why suppose that, in this scenario, something determines which one of them it is that fails to be true? After all, the stipulations would seem to be on an equal footing. Of course, the English predicate 'true' and the logical expressions of English were not introduced by stipulation: but the hypothetical example of expressions introduced through stipulation brings out how our thoughts and practices *could* fail to determine which one of notions not jointly available is unavailable.

Some theorists may wish to defend the significance of the project of finding the correct truth conditions for the sentences employed in the liar reasoning by claiming that the puzzle of the liar is that we have not been able to find a single promising candidate for getting the truth conditions right. The problem in the case of the liar is not, they might say, that of looking for facts about English that allow us to settle which one of several promising candidates is the correct account of the truth conditions of the sentences employed. The problem is rather that of finding at least *one* promising candidate.

But consider again the hypothetical linguistic community where the logical particles and the predicate 'true' were introduced by stipulation. If semantic theorists in this community were to try to give the right truth conditions for sentences of that language, they would likely be in the same situation as those theorists who are concerned with the liar paradox as it arises for us, in the respect that they would be unable to come up with the right truth conditions, in the sense of coming up with an account of truth conditions which respects all speaker judgments that they seek to respect. In the hypothetical case it is easy to see this, for there it is easy to see there that the speaker judgments are inconsistent. So in the hypothetical case, the search for an account of truth conditions fully faithful to speaker judgments is hopeless. And it is hard to see why it should not be assumed that the search for such an account is not hopeless also in the case of actual English. And once we have abandoned the search for such an account, and have decided to settle for an account that fails to respect every speaker judgment, it is no longer so clear that there are no really good candidates out there.[8]

Second, suppose that something about the thoughts and practices of speakers of English *does* determine exactly which notions are available in English. *So what?* Suppose it turns out that speakers of English generally conclude that given that truth and strong negation cannot both be expressed, it is truth, conceived as satisfying the E-schema, rather than strong negation that is not available in English. Speakers give uniform answers to the question of which principles they abandon when they realize truth and strong negation cannot both be expressed; and hence there is something that determines what is and what is not available in English.[9] This would seem to be a highly *contingent* matter. The dispositions of speakers could easily have been different. Perhaps the French respond differently to puzzle cases. What is the philosophical interest of the fact that the thoughts and practices of English speakers are just the way they happen to be?

I have been concerned exclusively with the question of what kind of problem the liar paradox is. But see where we have ended up. We have in effect concluded that if the liar is merely posing a problem concerning the semantics of the logical particles of natural language and of the predicate 'true', it does not pose any philosophically significant problem.

Other authors seem to take the liar as posing a problem to be solved by rational recon-

[8] In chapter 1, I defended the view that the principles constitutive of meaning of the expressions used in stating the liar paradox are jointly inconsistent; and that this is what is responsible for the paradox to arise. The semantic values – contributions to truth conditions – of the expressions are whichever set of them come closest to making all of the meaning-constitutive principles for the expressions true. It is to be expected that there are many equally acceptable assignments of semantic values.

If my account in chapter 1 is correct, then what I say here in the main text is surely correct. There is no acceptable account completely faithful to speaker judgments. But even should my (controversial) account of chapter 1 fail to get matters right, the claims above stand. The claims made here in the main text do not require any heavy-duty theory such as the one I present in chapter 1.

[9] The details do not really matter. I am only presenting this as an *example* of what *might* determine which notions are available.

struction. The liar paradox shows, they reason, that we cannot have everything we want: for example, we cannot respect all our intuitive judgments about truth, and we cannot unrestrictedly use a truth predicate as a device for disquotation, and we cannot talk about all semantic features of our language using this very language. But with enough ingenuity, we can approximate those goals rather well. The correct account of the liar is then the account that comes closest to these goals, which primarily are pragmatic in nature.

This may be a worthwhile project, if recognized for what it is. But the same remarks apply here as applied to the "rational reconstruction" proposal discussed above. Specifically, it seems that different predicates can satisfy the desiderata to different extents: so we will have a plurality of equally acceptable truth predicates with importantly different properties.

6.4 The significance of the liar paradox

I have been trying to cast doubt on the view that the paradoxes tell us much of interest, *in a certain respect.* More specifically, I have been trying to cast doubt on the view that the questions about truth and logic often raised in connection with the paradoxes have the philosophical significance usually attached to them.

Even if such skepticism should prove entirely justified, the paradoxes need not lack philosophical interest altogether. I shall now enter the positive mode and go through some theses for which some people, including me (in previous chapters), have used the paradoxes to argue, and which if true would seem to be of considerable interest. I shall not here argue for these theses: I do not even endorse all of them. My aim is only to consider the *kinds* of issues among those raised by the paradoxes that would be interesting, even should questions about truth and logic lack the philosophical significance often accorded to them.

First, Anil Gupta and Nuel Belnap have used the liar paradox to argue that some concepts, in particular *truth*, are *circular* or *circularly defined.* These concepts are, Gupta and Belnap claim, associated with a rule of *revision* rather than a rule of *application.* Gupta says in (1989) that the meaning of a predicate expressing a non-circular concept "is a rule that gives the extension of the predicate in all possible situations....the meaning determines the conditions for the predicate's applicability". A circular definition, on the other hand, provides only "a rule that can be used to calculate what the extension would be *once we* *make a hypothesis concerning the extension of the definiendum*".[10] Such a definition "does not determine the conditions of the applicability of the definiendum absolutely, but only hypothetically".[11]

In his discussions of the liar paradox, Stephen Yablo (1993a and 1993b) makes use of what is very roughly the same formal idea but puts a different philosophical spin on it. Yablo's idea is that the rules of use for a given word do not only put conditions on objects in the sense that they state what conditions an object must meet for the word to apply to it: they also put conditions on *subjects* in that they yield conditions for the proper use of the word. Definitions, stating rules of use, can accordingly be inconsistent in two different ways. On the one hand, they can be inconsistent in the sense of putting impossible conditions on objects. On the other hand, they can be inconsistent in the sense of that they "impose irreconcilable obligations on *speakers*".[12] When a definition is inconsistent in the latter sense, no extension can be stably ascribed to the word defined. For every extension which a speaker concludes that the word has, she can apply the definition once again to yield a different extension of the word.

The theses of Gupta and Belnap and of Yablo are of considerable interest if true: for then some expressions or concepts are quite different from what we thought any expressions and concepts could be. Notice that it is in a sense incidental to the purpose of the theses considered that they indeed fit the actually used expression 'true'. If some expression used could have the meaning they ascribe to 'true', the general conclusion about meaning would already follow. That 'true' actually functions as they say, if indeed it does, is of importance only in so far as it cuts short any dispute about whether a predicate of the kind described really could be used meaningfully.

So one issue raised by the liar, the issue of whether tiuth is "circular" (or more generally, whether there may be circular concepts), is clearly of interest even should questions about truth and logic not be. Now let us proceed to a second such issue.

In recent work Michael Glanzberg has argued that the key to a solution of the liar paradox lies in recognizing a hitherto unappreciated kind of context-sensitivity.[13] In brief, the argument is as follows. Since the only way to avoid the contradiction that otherwise is

[10] Gupta (1989), p. 234; emphasis in the original.

[11] Ibid.

[12] Yablo (1993), p. 147; Yablo's emphasis.

[13] Glanzberg (forthcoming).

derived in the liar reasoning is to take context-sensitivity to enter in somewhere, context-sensitivity to enter in somewhere. But none of the actually existing models of context-sensitivity seems to fit. This, Glanzberg says, is the real puzzle raised by the liar.

This too is a thesis somewhat removed from the actual workings of natural language, in the sense that, although Glanzberg does claim that his thesis fits natural language, it is enough for its philosophical interest that it could fit natural language: that there could be linguistic phenomena that exhibit this uncharted kind of context-sensitivity.

A third issue raised by the paradoxes and which would retain its significance even if we were to conclude that questions about truth and logic lack significance is the following. In chapter 1, I used the sorites and the liar paradox to argue that, roughly, the meanings of some expressions are incoherent. The sorites and the liar paradoxes are unsound arguments, but they *exert pull* in the sense that competent speakers are required by their semantic competence to be initially attracted to the arguments' culprits – untrue or invalid premises and steps. More precisely, each culprit is such that speakers are required to accept it in the absence of defeating evidence (such as is provided by having seen that these principles lead to contradiction).

If this is right, then any acceptable account of semantic competence must allow that we can be disposed to accept, by virtue of our semantic competence, what is untrue or invalid. Most, perhaps all, popular accounts of semantic competence disallow that.[14]

The same remarks apply here as applied to the proposals previously considered. The interesting conclusion about language is independent of whether the account really fits actually existing languages: it is enough for the philosophical significance of the conclusion that some language we could have used is incoherent in the sense outlined. The importance of the proposal's fitting an actually used language is only that this constructively shows that a possibly used language has been described.

A fourth issue distinct from questions about truth and logic which is raised by the liar paradox concerns *ineffability*. This issue was discussed in chapter 4; let me here repeat the main points. To set up the issue, we need some definitions and stipulations. First, the term 'English' as I shall use it henceforth refers to English *as it exists today*. To say that something is not available for expression in English is then to say that English *of today* does not contain the means for expressing it. Second, when I say that a property can be

[14] See section 5 of chapter 1.

expressed in a language, I mean that there is some predicate of that language which, in some context of utterance, is true of all and only those things that have the property. Third, for a predicate to satisfy the E-schema *restricted to a set of sentences S* is for every instance of the E-schema to be true when we substitute this predicate for 'T' and a sentence of S for 'p'.

Suppose now that English satisfies the conditions of Tarski's theorem: that the logical particles of English are classical and that English contains the means for naming its own expressions. Then the liar reasoning shows that English cannot contain a predicate satisfying the E-schema restricted to English.

Does this show that there is no *property* satisfying the E-schema restricted to English? Both an affirmative and a negative answer to this question have interesting consequences. An affirmative answer has repercussions for the philosophy of language; for in much philosophy of language it is assumed that there is some semantically significant property satisfying the E-schema. A negative answer – the answer I favor – has the consequence that (under the hypothesis mentioned) there are things that are not expressible in English, as a matter of necessity.

An affirmative answer also has the consequence, under further natural assumptions, that there are some properties which are *absolutely* inexpressible: not expressible in *any* language. Suppose we call the property expressed by a predicate satisfying the E-schema restricted to English **truth**-in-English. For every language L, we may suppose, there is some property **truth**-in-L. Now, provided there also is such a property as **truth** simpliciter, **truth** for variable L, this property is not expressible in any language. Suppose it could be expressed in some language L*. Then in L*, **truth** in L* can be expressed; but that is impossible.[15]

Suppose I am right. Then the liar reasoning shows that English is *essentially limited in expressive resources*: not every property there is can be expressed in English, and this is so not as a contingent matter of fact but as a matter of logic. It is perhaps no surprise that as it happens, not everything that can be expressed can be expressed in English. For example, it may not be surprising if the notions of some future science cannot be expressed in present-day English. But if what I say is right, then it is for a priori reasons impossible for English to express everything that can be expressed: for it cannot express the property that satisfies the E-schema with respect to English. Moreover, if my thesis is true, there

[15] This is much too quick; a more careful discussion is to be found in chapter 4.

are properties which cannot be expressed in *any language*. These are claims about what it is in principle possible to express. I use the behavior of the actually used predicate 'true' to argue for these claims, but that is not really essential. Even it were not the case that English could talk freely about its expressions and has sufficient logical resources for it to be impossible that a predicate of English satisfy the E-schema, a philosopher could give my argument (more or less) for the claim that English is essentially limited in expressive resources. She could say: if *this* notion were available, then, of necessity, *that* notion would not be.

Fifth, the liar reasoning also presents some little discussed metaphysical problems. Virtually all discussions of the liar problem are centered around liar *sentences*, and, as a result, about what to say about (our) language in face of the liar. But accounts of the liar that are *purely linguistic*, accounts which consist merely of theses regarding language, are arguably *incomplete.*

Let us consider two examples of purely linguistic accounts of the liar. First, there are accounts, of which mine (just presented) is one, according to which the liar shows that our language is in some way or other essentially limited in expressive resources. For example, I say that the liar reasoning shows that if English satisfies the conditions of Tarski's theorem, properties satisfying the E-schema are inexpressible in English. Second, there are accounts of the liar which consist of appeals to context-sensitivity. An example is Charles Parsons' account, according to which the liar reasoning shows that we can never quantify over all propositions there are.[16]

Such accounts face the problem of what we may call *the metaphysical liar*: what to say about a *liar proposition.* Just as, intuitively, there is a liar sentence which says of itself that it is not true, there is, intuitively, a liar proposition which says of itself that it is not true. A liar sentence which says of itself that it is not true can exist only if 'not' or 'true' fails to mean what we naively take it to mean. The same goes, *mutatis mutandis*, for a liar proposition.[17] But a purely linguistic account of the liar deals only with the liar problem as it arises for sentences. A liar sentence intuitively says of itself that it is "not true",

[16] Parsons (1974). See also Burge (1979), Gaifman (1992), Glanzberg (forthcoming) and Simmons (1993).

[17] How exactly to spell out the "mutatis mutandis" will depend on how we think of propositions. In the main text I will try to remain neutral on such issues. But the account of propositions that I favor, and given which the problem discussed will arise very forcefully, is that propositions are structured entities made up of concepts.

where we take truth to satisfy the E-schema and take the negation to be classical. One way of avoiding the contradiction in the case of a liar sentence is to say that no predicate of the language employed can satisfy the E-schema for the language employed; *a fortiori*, 'true' cannot. But this does not go any distance toward blocking the existence of a liar proposition (understood analogously to a liar sentence), having the properties we would intuitively ascribe to such a proposition. That is, this proposition says of itself that it is not true, where the property truth satisfies the E-schema, and the truth function of negation is classical. But it is impossible that a proposition with these attributes should exist.

However, the metaphysical liar not only poses problems for purely linguistic accounts of the liar. Let us again consider English. It seemed clear that the question of whether it is strong negation or a predicate satisfying the E-schema that cannot be expressed in English could in principle be settled by attention to the thoughts and practices of speakers. But here is a puzzle.

As stressed, our thoughts and practices could have been different. It seemed that it could be the case that whereas in *our* language strong negation can be expressed, and there is no predicate 'T' such that $T(\phi)$ always coincides in truth-value with ϕ, in some other possible language where the reverse is the case. But it appears that it cannot be that both truth and strong negation exist. This leads to epistemological problems. We thought we could figure out whether 'true' satisfies the E-schema, or else negation is strong, by considering our speaker judgments. But the members of another linguistic community are likely to arrive at the opposite conclusion about their language by employing the same method. And it cannot be that both we and they are right: for it cannot be that both what we take ourselves to express and what they take themselves to express exist.

The metaphysical liar also poses problems concerning how meaning and reference are determined. We thought that what determines the meanings of expressions like the truth predicate and the logical constants was how these expressions are being used. But this is now called into doubt or at least significantly problematized. For it would appear both that an expression can be used to express truth simpliciter and that an expression, at least an expression in a different language, can express strong negation simpliciter; but that is impossible. Of course, the meanings of natural kind terms and proper names, for example, are widely agreed to depend not only on use but also on features of the environment. But what the meanings of such expressions depend on, apart from usage, are causal relations between language users and objects in the world they inhabit. And it is hard to see how something similar could be the case when it comes to the truth predicate and, especially, logical terms. Thus, there is a mystery about what determines the meanings of the logical expressions and the truth predicate.[18]

Perhaps the solution is to deny the assumption that any entities of the right types compose a proposition. That is, perhaps the solution is to say that although truth exists and strong negation exists, the liar proposition does not exist. Or perhaps propositions are not structured entities, contrary to what – it may be argued – is presupposed in the argument that troubles us. Or perhaps the solution is something more radical: that the very idea of propositions as objects (that are not mere mirror images of sentences or utterances) is misguided.

Or perhaps the solution is to be a little more precise about what it is for a property or logical operation to be expressed by some given expression. This may sound very moderate as compared with other suggestions, but I think that this suggestion, if accepted, has radical implications of its own. Compare two languages L and L′ such that one is considerably weaker in expressive resources than the other, with two expressions, 'and_L' (in L) and '$\text{and}_{L'}$' (in L′), such that both expressions satisfy the standard proof-theoretic rules characterizing conjunction. Do these expressions express the same thing? Incautiously, we would say 'yes'. But perhaps we should say that what an expression expresses is crucially dependent on what other expressive resources the language has. Then we can say that the solution to the epistemological problem noted above is that we should not say that (for example) we can express strong negation but they cannot: for since our language and theirs are different in other respects, what we can express and what they cannot express are different things. (Of course, they still cannot express our notion of strong negation, but now for a rather more trivial reason: since their language in other respects is different from ours, it follows immediately that they do not express the same property of strong negation as we do.)

However, in (forthcoming), Vann McGee argues that our acceptance of the inference rules for the logical expressions is *open-ended*, in the sense that our acceptance that these inferences are valid in no way relies on what is actually expressible in the language we speak,

[18] Some theorists on the liar do discuss the problem in in terms of liar propositions (Barwise and Etchemendy (1987) is a prominent example). Some of what they say can be taken to address the issues raised here.

but we would continue to regard these inferences as valid no matter how our language were enriched. If (i) this is true, and (ii) any two expressions which open-endedly satisfy a particular set of principles or inference rules express the same thing, then we have a conflict with the suggested response to the threat posed by the liar proposition.

Moreover, (i) and (ii) seem very reasonable. Our acceptance of the inference rules for the logical constants does appear to be open-ended. And it appears that (say) 'and' is different in meaning from a would-be expression different from 'and' in exactly the respect that our acceptance of the inference rules characterizing this expression – the same as for 'and' – is not open-ended. This difference in meaning appears to entail that the two expressions express different things. But, to repeat, to accept (i) and (ii) is to accept something that conflicts with the proposed solution to the metaphysical liar.

I have not here tried to present a solution to the metaphysical liar. All I have tried to show is that whatever we conclude from the metaphysical liar is likely to have significance beyond mere questions about what happens to be expressed in the language we actually speak.

6.5 The metaphysical liar

The metaphysical liar is interesting also because it poses problems for the overall conception of metaphysics that I am most attracted to. In this section, I shall first indicate what this conception of metaphysics is, and then say how the metaphysical liar poses problems for it.

In chapter 2, I showed how to resolve certain issues in metaphysics. I shall not here repeat this discussion, but call attention to a central feature of it: I took the central claims used as evidence in these pieces of metaphysical argumentation to be judgments that we make in virtue of our semantic or conceptual competence. Though this does not commit me to the further thesis that all claims thus used as evidence (and legitimately so) are judgments of this kind, I am attracted to this stronger thesis as well. The questions "What are persons? What is personal identity over time?", as raised in metaphysics, are questions to be answered by attention to our notion of person. Call this a *linguistic* approach to metaphysics. It contrasts with the *metaphysical* approach to metaphysics (as we may call it), according to which these questions are to be answered by appeal to what there is, rather than appeal to our *conception* of what there is. An especially clear example of a metaphysical approach is in van Inwagen (1991). The main concern in this book is with the question of when 'simples' (for example, the simplest existents as physics conceives them) together constitute a complex object. The radical answer van Inwagen gives is that they constitute such an object only when they constitute a *life* or an *organism.* Given this metaphysical view, the only object there is that could be the *person* X is the organism associated with the person; and thus, for van Inwagen, persons are organisms. What is important for present purposes is that van Inwagen arrives at this conclusion about the nature of persons not by consideration of our conception of persons, but by consideration of what objects there are for persons to be identified with.[19]

What I want to call attention to is this. van Inwagen's reasoning about the nature of persons depends on taking the universe to be relatively *sparse.* If, for every reasonable conception of personhood, there corresponded objects, metaphysical considerations could not determine the question of what persons are. Conversely, if the universe is sparsely populated – in particular, if there are not objects corresponding to each broadly reasonable conception of personhood – then attention to the precise nature of our conception of persons would not settle metaphysical issues.

To see this, consider the psychological and the physical criterion of personal identity. In chapters 2 and 3 I was concerned with the coherence of these criteria. But let us set aside those issues and pretend that we have reasonable versions of these criteria that do not face such problems. Then I take it that we all should agree that both criteria are in the running for getting personal identity right, in the following, rather precise sense: if the only person-like entities that existed were such that their persistence conditions were those stated by one of these criteria, then these person-like entities would be persons. Coherence considerations aside, both (some version of) the psychological and (some version of) the physical criterion of personal identity are sufficiently close to correct that if the only person-like entities that exist satisfy this criterion, these person-like entities are persons. Provided some entities exist

[19] This would need to be stated more carefully. Of course, if we meant by 'person' what we mean by 'computer', then van Inwagen's metaphysical conclusions would not have the consequence that persons are organisms. His metaphysical conclusions only justify van Inwagen's thesis about persons over other theses about persons that appear to capture how we think of persons just as well as his thesis does.

A terminological note. In this section, I will use 'conception", as it occurs in "our conception of persons", to mean *the way we think of persons.* 'Concept', as it occurs in "our concept *person*, is that which is expressed by our expression. Thus, our concept *person* is what is expressed by our expression 'person'. Not all elements of our conception of persons need be part of our concept *person*, and for broadly externalist reasons, the concept need not be determined by the conception.

that correspond to some version of, for example, the physical criterion, the only reason these entities would not be persons would be that there exist other person-like entities that fit our conception of persons even better.

If this is correct, then whether a linguistic or a metaphysical approach to metaphysical issues is correct will be tied to issues of whether reality is sparsely populated or not. Exactly how to characterize (relatively) sparse and (relatively) dense population of reality in terms that are not theory-laden is difficult; and I will not attempt to do so here. But, very roughly, if reality is extremely densely populated, then there exists as many entities as there can possibly exist; for example, any way of combining simples into complexes results in objects; and to every consistent mathematical theory there correspond objects etc.[20] If time-world slices are accepted as fundamental elements of our ontology, the difference between the views may, in the case of spatiotemporally located objects, be characterized as follows: on the view that the universe is densely populated, any sum of time-world slices constitutes an object; whereas if it is sparsely populated, a sum of time-world slices must meet some substantial constraints if it is to constitute an object. But this way of characterizing the issue relies on a controversial ontological claim about time-world slices.[21]

Just to *illustrate* the issue, let me briefly mention one argument that can be given for the linguistic conception of metaphysics; this argument quite clearly requires reality to be rather densely populated. Suppose there is a linguistic community with a language and conceptual scheme just like ours, except that the members of this community take persons – or rather, what they call 'persons' – to be somewhat more modally fragile than we take persons to be: there are certain adventures we take persons to survive that they do not take 'persons' to survive. If we were to deny that there exist entities corresponding to their conception of 'persons', but holding that there exist entities corresponding to our conception of 'persons', we would be faced with the question of why they, and not we, are thus ontologically unfortunate. The members of this linguistic community seem no less justified in their belief that 'persons' exist than we are in our belief that persons exist. So if we are convinced that there exist entities corresponding to our conception of persons, we

[20] One may want to treat abstract objects differently from the way one treats concrete objects; but a principal motivation for taking concrete reality to be densely populated - the desire to treat metaphysical issues as conceptual or linguistic – carries over to abstract reality.

[21] It is not necessary for the world to be extremely densely populated that there exist time-world slices. It may be necessary to refer to time-world slices in order to be able to effectively specify the exotic objects that exist according to the view that the world is densely populated.

should likewise take there to exist entities corresponding to their conception of 'persons'. And this of course generalizes to other kinds of entities.[22]

The kind of epistemological skepticism that might lead one to take there to be 'persons' extends to the case of abstract objects. One might for example think that there exist structures that make true all consistent alternative set theories, provided that there exists a structure that makes true at least one set theory. The debate about the continuum hypothesis provides an illustration. According to some, probably most, theorists, the question of whether the continuum hypothesis is true is a mathematical question. But here is a different picture. Both ZF plus the continuum hypothesis and ZF plus its negation are consistent theories – and they are theories to which there correspond structures that make them true. That the continuum hypothesis is not settled by the current axioms of set theory, nor any intuitively obvious extension of it, only reflects a semantic indeterminacy in our current concept of set.[23] This picture would be motivated by considering how else we could know the continuum hypothesis (or its negation) other than through conceptual knowledge and knowledge of consistency. This view on abstract objects, like the corresponding view about concrete objects, depends on the universe – in this case the universe of abstract objects – being densely populated.

It may seem odd that the linguistic conception of metaphysics should depend for its truth on a claim which is itself metaphysical: that the universe really is densely populated. What is the source of our knowledge of this claim? The linguistic conception may appear *self-defeating*, depending for its truth on a metaphysical claim, our knowledge of which cannot be explained simply by appeal to our linguistic or conceptual competence.

However, it does appear that the proponent of the linguistic conception can reasonably simply appeal again to linguistic competence: she can say, for example, that it is part of, or follows from, our concept of existence that the universe is densely populated; for example, that it follows from our concept of existence that no matter what kind of an object you are or purport to be, no more is required for your existence than that the entities which would constitute you do exist.

[22] This argument would be question-begging against someone who, like van Inwagen, takes seriously the possibility that our knowledge of metaphysical claims is not conceptual or linguistic in nature. It only serves to dramatize skepticism of the assumption that there are other sources of metaphysical knowledge.

[23] Field (2000). (Field, of course, is a nominalist. I am outlining how to accept this picture from a platonist viewpoint.)

I do not mean to say that this claim is unproblematic. I only want to emphasize that this claim is fully compatible with the linguistic conception. In fact, the reason why I have been discussing the linguistic conception of metaphysics at length here is that the metaphysical liar presents a problem for it – or rather for the existence of a justification for it of the kind just mentioned.[24]

As we have seen, the linguistic conception of metaphysics requires that reality be densely populated. Applied to the case of abstract objects, and in particular objects that exist necessarily if they exist at all, the view would appear to have to be (roughly) that everything that consistently could exist does exist. But the metaphysical liar presents a prima facie problem for these ideas. We could consistently use a language in which truth simpliciter is expressed; we could also consistently use a language in which strong negation (likewise 'simpliciter') is expressed. But what would be expressed in each of these languages cannot coexist.[25]

One may think that this is not a very deep problem. All it immediately shows is that regarding one very specialized matter, we cannot employ the methods we (according to the linguistic conception) ordinarily use to establish metaphysical claims. This does not imply either that these methods cannot be used elsewhere, or that there are other ways of establishing metaphysical claims.

But here is one reason to suspect that the problem may go deeper. Although the metaphysical liar shows only that there is one instance where the linguistic method in metaphysics yields absurdity, it entails that the linguistic method is not conceptually guaranteed to bring truth. It was noted above that the metaphysical claim upon which the linguistic conception of metaphysics rests must itself be linguistically or conceptually justified for the linguistic conception not to be self-defeating. The metaphysical liar poses a problem regarding the existence of this kind of justification.

[24] It should be pointed out that whereas for a particular metaphysical issue, for example that about the nature of persons, to be settled by attention to specifics of our concepts, it is sufficient that there exist entities corresponding to all reasonable and consistent theories of the relevant concepts, the suggested justification for a linguistic approach to metaphysics demands that reality be as densely populated as consistently possible. (See Shoemaker (1988) for criticism of the view that reality is extremely densely populated.)

[25] One may think that since a truth predicate and a negation sign belong to the wrong grammatical categories, there exist no objects corresponding to (in the sense of being the semantic values of) these expressions. But I could then simply assume that we have recourse to nominalization devices, and given the general metaphysical picture presupposed by the suggested justification for the linguistic conception of metaphysics, the nominalizations refer to objects, and it is the properties of these objects that are problematized.

6.6 Debates about truth and logic revisited

In sections two and three, I motivated skepticism about the philosophical significance of questions about truth and logic. In the fourth section, I showed how the paradoxes could still be of philosophical significance even though questions about truth and logic are not. In the last section, I took a detour over the consequences of the liar for metaphysics.

I shall now present a positive suggestion concerning how to understand questions about truth and logic. It seems to me that if certain theoretical claims hold true, then there is a way of allowing questions about truth and, by extension, logic to be significant.

However, two features of this vindication are worth noting. First, if this is how debates about truth and logic are vindicated, then, as we shall see, the ground rules of such debates are quite different from what is generally supposed. Second, the theoretical assumptions under which questions of truth and logic retain their significance are by no means obviously true. I suspect that many who would want to see debates about truth and logic as significant would find these theoretical assumptions difficult to swallow.

I shall start by considering how to render questions about truth – such as whether bivalence is valid and whether there can be true contradictions – significant. The general idea here is that if our predicates 'true' and 'false', or some would-be predicates relevantly like them, track (express, stand for) philosophically significant features, the so-called questions about truth become relevant when formulated in terms of these predicates tracking philosophically significant properties.

Here is an example. Suppose that when reflecting upon and studying the nature of language we find for example that: (i) Utterances are usefully divided up into different *speech acts.* (ii) Among these speech acts are *assertion* and *denial.* (iii) These speech acts are usefully characterized by appeal to the norms peculiar to them. Of course to be acceptable in every respect, an assertion must satisfy many norms: it must be relevant, polite, etc. But suppose there is a norm such that what distinguishes assertions from other speech acts is that this norm governs only the practice of assertion, and similarly for denial. (iv) Suppose now that we introduce two expressions 'true*' and 'false*' such that 'true*' is a predicate that applies to a sentence just in case an assertion of that sentence satisfies the norm peculiar to assertion; and *mutatis mutandis* for 'false*' and denial.

'True*' and 'false*' are technical counterparts of the truth and falsity predicates of ordinary language. Questions of, for example, bivalence become significant when formulated in terms of these technical notions, provided the theoretical assumptions are correct. The question of bivalence become significant when formulated in terms of these technical notions, provided the theoretical assumptions are correct.

The two expressions thus introduced are presumably more-or-less coextensive with their ordinary-language counterparts. Perhaps they even are coextensive with them. But that is incidental to their usefulness in reasoning about language: what is important is that they register philosophically important norms. If there did not already exist expressions expressing what 'true*' and 'false*' express, we would, for theoretical purposes, have to invent such expressions.

I should again stress that what is required for questions about truth to be significant is not exactly that *assertibility* be a philosophically important notion. I provide only one relatively straightforward example of how our truth predicate or a predicate rather similar to it in behavior may track some significant property. The central thesis, to repeat, is simply that there is some philosophically significant property of truth (and some philosophically significant property of falsity).

Thus phrased, the assumption under which questions about truth are philosophically significant may sound trivial, in two ways. First, it may seem obvious that the notion of truth is philosophically significant – it is, after all, the topic of much philosophical discussion. Second, the suggestion on the table sounds like the following: questions about truth are philosophically significant provided the notion of truth is philosophically significant. But that sounds embarrassingly trivial.

But, first, that truth is a topic of much philosophical discussion is not sufficient for its being philosophically significant, in the sense that is relevant here. This is illustrated by, for example, the fact that a substantial portion of the philosophical discussion of truth concerns precisely whether truth does track a theoretically significant property, or even tracks a property at all. The philosophical discussion I am alluding is that about *deflationism.*[26]

And as regards the second objection or worry: the proposal does indeed sound trivial under the restatement of it. But we must be careful about what is meant by saying that a particular notion is philosophically significant, as shown for example by the points in the last paragraph. On the one hand, a notion can be philosophically significant because

[26] See e.g. Horwich (1990) and Field (1994).

of, or in the sense of, tracking a philosophically significant property. On the other hand, a notion can be philosophically significant because of, or in the sense of, there being important philosophical questions raised concerning it. Among such questions may be: does the notion track a philosophically significant property?

It is important not to confuse my positive proposal regarding how to render questions about truth significant with the rational reconstruction proposal rejected earlier. The rational reconstruction proposal emphasizes usefulness – pragmatic virtues – and thus the expressive gains of employing such-and-such a truth predicate and logical particles. But a truth predicate tracking some significant property need not be the truth predicate which is most useful in respect of expressive power.

If the significance of questions about truth is vindicated as suggested, the significance of questions about logic is easily vindicated in a similar way. One can let the significance of questions about logic be parasitic on the significance of questions about truth. One can take the logical expressions the questions of logic are about to be *truth-functional*, where the property of truth in question is a (supposedly) philosophically significant property of truth.

But this strategy suggests several analogous strategies. Intuitionists have analyzed the meanings of logical expressions in terms of the *existence of proof*. One can take questions about logic to be questions about logical expressions thus understood. Then the significance of questions about logic is dependent on the philosophical significance of the notion of proof, or the existence of proof. Supervaluationists have analyzed the logical expressions in terms of the notion of *truth under an assignment*. And questions about logic can be taken to be questions about logical expressions understood as the supervaluationist understands them. Their significance then becomes dependent on the significance of the notion of truth under a valuation. And similarly for other notions in terms of which logical expressions may be explained.

It is important to see what this kind of vindication of the significance of questions about logic would *not* yield. Questions about logic are not, given this vindication of their significance, *autonomously* significant: their significance is *derivative*. That is, their significance is dependent on that of the notion in terms of which the meanings of the logical expressions are explained. Second, there is no significant question about which notion the logical constants are explained in terms of. There exist broadly logical expressions analyzable in terms of, for example, notions of truth, existence of proof, and truth under an assignment. There is no issue about which of these broadly logical expressions are *the* logical expressions, or what conjunction or disjunction or negation is really like; there is no question about which notion we *ought* to analyze logical expressions in terms of. The closest we come to such an issue is the question of which among the notions in terms of which logical expressions might be explained really are philosophically significant.[27,28]

6.7 The paradoxes and bivalence

I suspect some readers will feel that the proposed interpretation of questions about truth and logic is the one they – and everyone else – clearly have had in mind all the time; and so that it is unclear what the interest of the discussion of the proper interpretation of these questions has been. In this section I shall spell out some consequences of interpreting questions about truth and logic in accordance with the 'semantic' proposal above, and show that this is of some consequence for how to regard the implications of the paradoxes for the questions about truth and logic.

If questions about truth and logic are so interpreted, the paradoxes can again be seen to have important consequences for a theoretically significant issue of bivalence. The liar sentence, for example, might be an example of a sentence which can neither be correctly asserted nor correctly denied (when 'assertion' and 'denial' are understood as theoretical

[27] To explain these assertions about the nature of questions about logic, let me relate them to Dummett's views on logic. An important theme in Dummett's writings is whether intuitionist or classical logic is correct; where Dummett sees the intuitionist logical expressions as explained in terms of a notion of existence of proof and classical logical expressions as explained in terms of a notion of truth. When Dummett's views are set out thus, it should appear that there is an immediate clash between my views and those of Dummett. On my view on questions about logic, there is no genuine issue of whether the logical expressions ought to be explained in terms of proof or in terms of truth; and accordingly the dispute between intuitionist and classical logic, construed as Dummett construes it, is a non-issue.

However, this overlooks the fact that for Dummett, the notion of existence of proof and the notion of (classical, bivalent) truth are competitors for a particular theoretical role: that of *being central in a meaning theory*. A detailed excursion into Dummett's philosophy of language would be needed to elucidate this notion. But suffice it here to say that it is a notion that, in Dummett's philosophy, plays a particular, central theoretical role. Thus, for Dummett, the question of whether intuitionistic or classical logic is the right one becomes a question of whether intuitionistic or classical logic is the right one, given that the logical expressions are explained in terms of the notion that plays a central role in the theory of meaning. This in turn becomes merely a question of which notion plays this role, and what the properties of this notion are. (This justifies, *on my terms*, Dummett's approach to the question of whether intuitionistic or classical logic is the right logic. I do not mean to imply that Dummett thinks of the matter in just this way.)

[28] All I am saying here is that questions about logic are not autonomously significant given the present vindication of their significance. I can see no way of ruling out the possibility that there exists some other vindication of their significance, that would make questions about logic autonomously significant.

notions); and, of course, similarly for a large number of vague sentences. But importantly, the relevance is of a different *kind* than is generally assumed; and if the questions are interpreted as suggested, debates about classical logic and bivalence should take a different form. A corollary is that if theorists do indeed understand questions about truth and logic as I have here suggested that such questions should be understood, then they have at least failed to see the implications of so understanding them.

Suppose that the liar paradox shows that English cannot contain truth and falsity predicates such that for every declarative sentence (or utterance thereof) one of these two predicates applies to it. Given the the proposed way of understanding questions about bivalence, this does not settle the question of bivalence negatively. It *could* still be that every sentence can be either correctly asserted or correctly denied. What would be shown is only that English cannot contain predicates that track assertibility and deniability.[29] (I have already suggested that the liar sentence may be neither correctly assertible nor correctly deniable. My point here is just that the question of its assertibility or deniability must be separated from the question of whether the principle of bivalence as stated in the object language is valid.)

Correspondingly, focusing for a change on the sorites paradox and pretending that the liar does not exist, the following is a possible scenario. First, a supervaluationist approach to vagueness is correct, in that the vagueness of a language gives rise to there being a multitude of acceptable assignments of semantic values to the sentences of the language. Under every such assignment the sentences we would ordinarily take to express the laws of classical logic come out true. Thus, every sentence of the form "P or not P" comes out true, every sentence of the form "If P and Q then P" comes out true, etc. Second, every instance of the disquotation schema comes out true under every assignment, and every instance of the counterpart of the disquotation schema for falsity,

'P' is false if and only if it is not the case that P,

comes out true. Bivalence is then valid, in the relatively uninteresting sense that every sentence of the form "'P' is true or 'P' is false" comes out true. But the philosophically significant question of bivalence may still be answered in the negative. For example, it

[29] My arguments in chapter 4, "Truth, the Liar and Universality", can, but need not, be seen as developing this suggestion.

may be that a sentence is correctly assertible just in case it is true under every acceptable assignment, and that a sentence is correctly deniable just in case it is false under every acceptable assignment. If all this is correct, vagueness would show that the philosophically significant question of bivalence should be answered in the negative.[30]

These examples should suffice to show that if the questions about truth and logic are interpreted as suggested in the last section, the character of the implications of the paradoxes for questions about truth and logic are slightly different from what has been assumed. In particular, one must distinguish between the consequences of the paradoxes as regards the semantics of the expressions 'true', 'not', etc., of our language and, for example, the consequences they have as regards the nature of assertibility and deniability, etc. As noted, it is possible that 'true' applies to exactly the sentences that are correctly assertible in the technical sense. But this certainly is not something that can be presupposed.

6.8 Concluding remarks

These remarks conclude this chapter, and the thesis. Part of what I hope to have accomplished in this thesis is, briefly, to replace one set of concerns normally raised in connection with the sorites and liar paradoxes, namely those about which logical and semantic principles our expressions and concepts satisfy, and what are the semantic values of logical and semantic expressions and vague predicates, with questions that appear to me more theoretically interesting, such as those about inconsistency-by-virtue-of-competence and the limits of expressibility. In earlier chapters, primarily chapters 1, 3 and 4, I made a case for for my positive view about how the latter questions are raised by the paradoxes. (In fact, I presented answers to some such questions.) In this last chapter, I have been casting doubt on the significance of the questions traditionally raised.

[30] In chapters 1 and 3, I outlined how the principle of bivalence could be upheld in the face of vagueness. This, I should now confess, was a defense of bivalence in the *uninteresting* sense. The reason for still bothering about this defense of bivalence is that it illustrates, for example, the distinction between first-level and second-level indeterminacy, explained in these chapters (more extensively in chapter 3).

References

Alston, William: 1964, *Philosophy of Language*, Englewood Cliffs: Prentice-Hall.

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

Bedard, Katherine: 1993, "Partial Denotations of Theoretical Terms", *Noûs* 27: 499-511.

Boghossian, Paul: 1997, "Analyticity", in Bob Hale and Crispin Wright (eds.), *A Companion to the Philosophy of Language*, Oxford: Blackwell, pp. 331-68.

Burge, Tyler: 1979, "Semantical Paradox", *Journal of Philosophy* 76: 169-98.

Byrne, Alex: 1999, "Cosmic Hermeneutics", *Philosophical Perspectives* 13: 347-83.

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

Chihara, Charles: 1984, "The Semantic Paradoxes: Some Second Thoughts", *Philosophical Studies* 45: 223-9.

Chomsky, Noam: 1995, "Language and Nature", *Mind* 104: 1-61.

Churchland, Paul: 1981, "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy* 78: 67-90.

Davidson, Donald: 1967, "Truth and Meaning", *Synthèse* 17: 304-23.

Donnellan, Keith: 1970, "Categories, Negation, and the Liar Paradox", in Martin (1970b), pp. 113-120.

Dummett, Michael: 1975, "Wang's Paradox", *Synthèse* 30: 301-24.

Dummett, Michael: 1981, *Frege: Philosophy of Language*, 2nd edn., London: Duckworth.

Dummett, Michael: 1991, *The Logical Basis of Metaphysics*, London: Duckworth.

Edgington, Dorothy: 1992, "Validity, Uncertainty and Vagueness", *Analysis* 52: 193-204.

Field, Hartry: 1973, "Theory Change and Indeterminacy of Reference", *Journal of Philosophy* 70: 462-81.

Field, Hartry: 1994, "Deflationist Views of Meaning and Content", *Mind* 103: 249-85.

Field, Hartry: 2000, "Indeterminacy, Degree of Belief, and Excluded Middle", *Noûs* 34: 1-30.

Fine, Kit: 1975, "Vagueness, Truth and Logic", *Synthèse* 30: 265-300.

Fodor, Jerry and Ernest LePore: 1998, "The Emptiness of the Lexicon: Reflections on James Pustejovsky's *The Generative Lexicon*", *Linguistic Inquiry* 29: 269-88.

Gabbay, Dov and Heinrich Wansing (eds.): 1999, *What is Negation?*, Dordrecht, Boston and London: Kluwer Academic Publishers.

Gaifman, Haif: 1992, "Pointers to Truth", *Journal of Philosophy* 89: 223-61.

Glanzberg, Michael: forthcoming, "The Liar in Context", *Philosophical Studies.*

Gupta, Anil: 1988-89, "Remarks on Definitions and the Concept of Truth", *Proceedings of the Aristotelian Society* 89: 227-46.

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

Herzberger, Hans: 1967, "The Truth-Conditional Consistency of Natural Languages", *Journal of Philosophy* 64: 29-35.

Herzberger, Hans: 1970, "Paradoxes of Grounding in Semantics", *Journal of Philosophy* 67: 145-67.

Herzberger, Hans: 1981, "New Paradoxes for Old", *Proceedings of the Aristotelian Society* 81: 109-23.

Herzberger, Hans: 1982, "Naive Semantics and the Liar Paradox", *Journal of Philosophy* 79: 179-97.

Horgan, Terence: 1983, "Supervenience and Cosmic Hermeneutics", *Southern Journal of Philosophy* 22, Supplement: 19-38.

Horgan, Terence: 1995, "Transvaluationism: A Dionysian Approach to Vagueness", *Southern Journal of Philosophy* 33, Spindel Conference Supplement on Vagueness, pp. 97-126.

Horwich, Paul: 1990, *Truth*, Oxford: Blackwell.

Jackson, Frank: 1998, *From Metaphysics to Ethics: A Defense of Conceptual Analysis*, Oxford: Oxford University Press.

Johnston, Mark: 1987, "Human Beings", *Journal of Philosophy* 84: 59-83.

Johnston, Mark: 1989, "Fission and the Facts", *Philosophical Perspectives, Vol. 3, Philosophy of Mind and Action Theory*, pp. 369-97.

Kripke, Saul: 1975, "Outline of a Theory of Truth", *Journal of Philosophy* 72: 690-716. Reprinted in Martin (1984).

Kripke, Saul: 1979, "A Puzzle About Belief", in Avishai Margalit (ed.), *Meaning and Use*, Dordrecht: D. Reidel, pp. 239-83.

Lewis, David: 1970, "How to Define Theoretical Terms", *Journal of Philosophy* 67: 427-46.

Lewis, David: 1972, "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy* 50: 249-58.

Lewis, David: 1976, "Survival and Identity", in Amelie Rorty (ed.), *The Identities of Persons*, Berkeley: University of California Press, pp. 17-40.

Lewis, David: 1983, "New Work for a Theory of Universals", *Australasian Journal of Philosophy* 61: 343-77.

Lewis, David: 1997, "Naming the Colours", *Australasian Journal of Philosophy* 75: 325-42.

Lowe, E.J.: 1983, "On the Identity of Artifacts", *Journal of Philosophy* 80: 220-32.

Martin, Robert L.: 1970, "A Category Solution to the Liar", in Martin (1970b), pp. 91-112.

Martin, Robert L. (ed.): 1970b, *The Paradox of the Liar*, New Haven and London: Yale University Press.

Martin, Robert L.: 1976, "Are Natural Languages Universal?", *Synthèse* 32: 271-91.

Martin, Robert L. (ed.): 1984, *Recent Essays on Truth and the Liar Paradox*, Oxford: Oxford University Press.

McGee, Vann: 1991, *Truth, Vagueness and Paradox: An Essay on the Logic of Truth*, Indianapolis: Hackett Publishing Company.

McGee, Vann: 1992, "Maximal Consistent Sets of Sentences of Tarski's T-schema", *Journal of Philosophical Logic* 21: 235-41.

McGee, Vann: 1997: "Revision", in Enrique Villanueva (ed.), *Truth*, Atascadero: Ridgeview, pp. 387-406.

McGee, Vann: 1998, "Kilimanjaro", *Canadian Journal of Philosophy*, supp. vol. 23: 141-98.

McGee, Vann: forthcoming, ""Everything"", in Gila Sher and Richard Tieszen (eds.), *Between Logic and Intuition*, New York and Cambridge: Cambridge University Press.

McGee, Vann and Brian McLaughlin: 1995, "Distinctions Without a Difference", *Southern Journal of Philosophy* 33, Spindel Conference Supplement on Vagueness, pp. 203-51.

Nozick, Robert: 1981, *Philosophical Explanations*, Cambridge, Massachusetts: Belknap Press.

Parfit, Derek: 1984, *Reasons and Persons*, Oxford: Clarendon Press.

Parsons, Charles: 1974, "The Liar Paradox", *Journal of Philosophical Logic* 3: 381-412. Reprinted, with postscript, in *Mathematics in Philosophy*, Ithaca: Cornell University Press, 1983, pp. 221-67.

Parsons, Terence: 1984, "Assertion, Denial, and the Liar Paradox", *Journal of Philosophical Logic* 11: 117-52.

Peacocke, Christopher: 1992, *A Study of Concepts*, Cambridge, Mass.: MIT Press.

Perry, John: 1993, "Williams on the Self and Its Future", in John Perry and Michael Bratman (eds.), *Introduction to Philosophy: Classical and Contemporary Readings*, New York and Oxford: Oxford University Press, 2nd edn, pp. 427-35.

Priest, Graham: 1979, "The Liar Paradox", *Journal of Philosophical Logic* 8: 219-41.

Priest, Graham: 1984, "Semantic Closure", *Studia Logica* 43: 117-29.

Priest, Graham: 1984b, "The Liar Paradox Revisited", *Journal of Philosophical Logic* 13: 153-79.

Priest, Graham: 1987, *In Contradiction: A Study of the Transconsistent*, Dordrecht and Boston: Kluwer Academic Publishers.

Priest, Graham: 1990, "Boolean Negation and All That", *Journal of Philosophical Logic.*

Priest, Graham: 1999, "What Not? A Defence of Dialetheic Theory of Negation", in Gabbay and Wansing (1999), pp. 101-20.

Prior, Arthur: 1960, "The Runabout Inference-Ticket", *Analysis* 21: 38-9.

Pustejovsky, James: 1998, "Generativity and Explanation in Semantics: A Reply to Fodor and LePore", *Linguistic Inquiry* 29: 289-311.

Putnam, Hilary: 1983, "Vagueness and Alternative Logic", *Erkenntnis* 19: 297-314.

Quine, Willard van Orman: 1936, "Truth by Convention", in O. H. Lee (ed.), *Philosophical Essays for A.N. Whitehead*, New York: Longmans, pp. 90-124.

Quine, Willard van Orman: "Two Dogmas of Empiricism", in *From a Logical Point of View*, Cambridge, Mass.: Harvard University Press, pp. 20-46.

Quine, Willard van Orman: 1960, *Word and Object*, Cambridge, Mass.: MIT Press.

Quine, Willard van Orman: 1972, "Book Review: *Identity and Individuation.* Milton K. Munitz, editor", *Journal of Philosophy* 69: 488-97.

Quine, Willard van Orman: 1981, "What Price Bivalence?", *Journal of Philosophy* 78: 90-5.

Raffman, Diana: 1994, "Vagueness Without Paradox", *Philosophical Review* 103: 41-74.

Rea, Michael C.: 1995, "The Problem of Material Constitution", *Philosophical Review* 104: 525-52.

Rieber, Steven: 1998, "The Concept of Personal Identity", *Philosophy and Phenomenological Research* 58: 581-94.

Rolf, Bertil: 1981, *Topics on Vagueness*, Lund: Studentlitteratur.

Rovane, Carol: 1993, "Self-Reference: The Radicalization of Locke", *Journal of Philosophy*

90: 73-97.

Rovane, Carol: 1998, *The Bounds of Agency*, Princeton: Princeton University Press.

Sainsbury, Mark: 1991, "Is There Higher-Order Vagueness?", *Philosophical Quarterly* 41: 167-82.

Sainsbury, Mark: 1996, "Concepts Without Boundaries" in Rosanna Keefe and Peter Smith (eds.), *Vagueness: A Reader*, Cambridge, Mass.: MIT Press, pp. 251-64. Inaugural Lecture at King's College London, 6 November, 1990.

Shoemaker, Sydney: 1984, "Personal Identity: a Materialist's Account", in Sydney Shoemaker and Richard Swinburne, *Personal Identity*, Oxford: Basil Blackwell, pp. 67-132.

Shoemaker, Sydney: 1988, "On What There Are", *Philosophical Topics* 16: 201-24.

Simmons, Keith: 1993, *Universality and the Liar*, Cambridge: Cambridge University Press.

Smart, J.J.C.: 1961, "Colors", *Philosophy* 36: 128-42.

Soames, Scott: 1989, "Semantics and Semantic Competence", in *Philosophical Perspectives, Vol. 3, Philosophy of Mind and Action Theory*, pp. 575-96.

Soames, Scott: 1999, *Understanding Truth*, Oxford: Oxford University Press.

Sorensen, Roy: 1991, "Vagueness Within the Language of Thought", *Philosophical Quarterly* 41: 389-413.

Sosa, Ernest: 1990, "Surviving Matters", *Noûs* 24: 297-322.

Stich, Stephen: 1992, "What Is a Theory of Mental Representation?", *Mind* 101: 243-61.

Strawson, Peter: 1959, *Individuals*, London: Methuen.

Tappenden, Jamie: 1993, "The Liar and Sorites Paradoxes: Toward a Unified Treatment", *Journal of Philosophy* 90: 551-77.

Tappenden, Jamie: 1993b, "Analytic Truth – It's Worse (or Perhaps Better) than You Thought", *Philosophical Topics* 21: 233-61.

Tappenden, Jamie: 1999, "Negation, Denial and Language Change in Philosophical Logic", in Gabbay and Wansing (1999), pp. 261-98.

Tarski, Alfred: 1983, "The Concept of Truth in Formalized Languages", in J. H. Corcoran (ed.), *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, 2nd edn, Indianapolis: Hackett Publishing Company. English translation by J.H. Woodger of "Der Wahrheitsbegriff in den formalizierten Sprachen", *Studia Philosophica* 1 (1935).

Thomson, Judith: 1997, "People and their Bodies", in Jonathan Dancy (ed.), *Reading Parfit*, Oxford: Blackwell, pp. 202-29.

Tye, Michael: 1990, "Vague Objects", *Mind* 99: 535-57.

Tye, Michael: 1994, "Sorites Paradoxes and the Semantics of Vagueness", *Philosophical Perspectives* 8: 189-206.

Unger, Peter: 1979, "There Are No Ordinary Things", *Synthèse* 41: 117-54.

Urmson, J. O.: 1961, "*Individuals: an Essay in Descriptive Metaphysics.* By P. F. Strawson", *Mind* 70: 258-64.

van Fraassen, Bas: 1968, "Presupposition, Implication and Self-Reference", *Journal of Philosophy* 65: 136-52.

van Inwagen, Peter: 1991, *Material Beings*, Ithaca: Cornell University Press.

Varzi, Achille: 1995, "Vagueness, Indiscernibility, and Pragmatics: Comments on Burns", *Southern Journal of Philosophy* 33, Spindel Conference Supplement on Vagueness, pp. 49-62.

Warmbrod, Ken: 1991, "The Need for Charity in Semantics", *Philosophical Review* 100: 431-58.

Wiggins, David: 1967, *Identity and Spatio-Temporal Continuity*, Oxford: Basil Blackwell.

Wilkes, Kathleen: 1988, *Real People: Personal Identity Without Thought Experiments* New York and Oxford: Oxford University Press.

Williams, Bernard: 1970, "The Self and the Future", *Philosophical Review* 79: 161-80.

Williamson, Timothy: 1994, *Vagueness*, London: Routledge.

Wright, Crispin: 1975, "On the Coherence of Vague Predicates", *Synthèse* 30: 325-65.

Wright, Crispin: 1987, "Further Reflections on the Sorites Paradox", *Philosophical Topics* 15: 227-90.

Yablo, Stephen: 1993, "Definitions: Consistent and Inconsistent", *Philosophical Studies* 72: 147-75.

Yablo, Stephen: 1993b, "Hop, Skip and Jump", *Philosophical Perspectives* 7: 371-96.