I’ll start with two quasi-technical issues, then move on to issues of broader philosophical interest.

1. **Proving generalizations about truth.** Tarski pointed out long ago that we do not get an adequate theory of truth by simply taking as axioms all instances of the truth schema: this doesn’t give the generalizations we need (such as that whenever modus ponens is applied to true premises, the conclusion is true), but only gives their instances. In Section 3 I made a proposal for how to get around the problem, but what I said could have used more explanation.

   What I suggested takes off from an idea of Feferman (1991) and others, of building schematic variables into the language and subjecting them to a rule of substitution. This allows their scope to be automatically extended as the language expands. Feferman used the idea in connection with schematic theories like number theory and set theory. It is a bit more complicated to apply the idea in connection with the theory of truth, since the schematic letters appear both inside and outside quotes. My proposal was to handle this (in a language with quotation mark names) by using not only a substitution rule, but also a rule that allows you to pass from a schematic theorem in which all occurrences of a certain schematic letter appear inside quotes to a generalization about all sentences.

   One thing I should have said is that the syntactic theory needs to be done with schematic variables as well, for we need a schematic version of syntax in the suggested approach to truth theory. Consider the passage from the schema
(TF) ‘q or r’ is true if and only if ‘q’ is true or ‘r’ is true

(derived from three applications of the truth schema, using a substitutivity principle) to the generalized version

(TFG) For all sentences S₁ and S₂ of our language, [S₁ or S₂] is true iff S₁ is true or S₂ is true.

The required derivation uses schematic syntax to go from (TF) to

(TF*) ‘q’ ^ ‘or’ ^ ‘r’ is true if and only if ‘q’ is true or ‘r’ is true,

which by rule (ii) of Section 3 yields

(TFG*) For all sentences S₁ and S₂ of our language, S₁ ^ ‘or’ ^ S₂ is true iff S₁ is true or S₂ is true;

(TFG) is simply a rewrite of this using the corner-quote notation.

How does this schematic variable approach compare with Tarski’s? (Or rather, with the variant of Tarski’s approach that gives up the idea of defining truth, but takes over the compositional clauses that Tarski built into his definition as axioms. My reasons for favoring the axiomatic version of Tarski over the definitional are given in Section 6 of the Chapter.) I’m not sure that the difference between the schematic variable approach and the axiomatic variant of Tarski is all that significant, but I do see a few advantages to the schematic variable approach. First, we have the same advantage that Feferman stresses for schematic number theory and schematic set theory: the theory of truth doesn’t need to be revised as the language is expanded. Second, the use of “‘p’ is true if and only if p” (understood now as a schematic formula that is part of the language) in the
axiomatization seems to more directly capture the core of the notion of truth. (That’s very vague, and I wouldn’t put a great deal of weight on it.) Third and most significant, the theory doesn’t depend on our having a compositional account of the functioning of the other devices in the language. An axiomatic Tarskian truth theory for certain constructions, for instance, belief sentences, is notoriously hard to provide; the schematic variable approach works without such an account.

This last reason may well be thought to backfire: surely the compositional structure built into Tarski’s axiomatization is important, and a theory of truth that leaves it out would be deficient. But the schematic variable approach doesn’t leave it out, for I just sketched how a typical compositional axiom (the disjunction rule) can be derived. The derivation was based on a prior schematic derivation of (TF):

\['p \text{ or } q\] is true iff \(p \text{ or } q\)

\['p\] is true iff \(p\)

\['q\] is true iff \(q\)

So \(p \text{ or } q\) is true iff \(p\) is true or \(q\) is true.

The last line of this prior derivation follows from the earlier three only by a substitutivity principle that has no analog for problematic constructions like belief sentences, explaining why compositional principles are not generally available. The fact that truth is compositional in some cases but not others isn’t fundamentally a fact about truth, but rather about the underlying logic of the domains to which it is applied.

In summary, it is this last point that provides the main advantage of the schematic
variable approach over the axiomatic Tarskian approach. But in most respects, the axiomatic Tarskian approach is itself quite “deflationist”; for instance, the deflationistic “inversion” of the theory of reference discussed in Section 4 of Chapter 2 would work just as well on a Tarskian approach as on a schematic variables approach.

2. Semantic paradoxes. The paper was slightly cavalier in its treatment of the paradoxes, mostly because I have vacillated as to how best to avoid them. I think that our ordinary concept of truth involves principles that in classical logic lead to contradiction; any way of avoiding the contradiction within classical logic can only be a recommendation for how to improve our concept. (Generally I find the distinction between revising our concept and revising our views while keeping the concept fixed to be a murky one at best, but this is one case that clearly seems to fall under the former heading.)

Actually I should be more cautious: any reasonable candidate for our ordinary concept of truth involves principles that in classical logic lead to contradiction. The reason for the retreat to the more cautious statement is that it is not entirely implausible that “our ordinary concept” bifurcates over whether certain instances of the schema “if p then ‘p’ is true” are correct (e.g., instances where ‘p’ crucially contains vague terms, or evaluatives: see the distinction between weak and strong truth in Field 1994). Even so, it seems almost beyond doubt that on our ordinary concept, the converse schema holds, and that when one has produced a solid argument for ‘p’ one is entitled to conclude “‘p’ is true”. These are all that are needed to derive a contradiction involving Liar sentences in classical logic.

We can view the disquotational theory put forth in this paper either as an account of our ordinary concept, or as an account of a concept that improves on our ordinary
concept in certain ways. (As I’ve said, I don’t think the distinction between these alternatives to be entirely clear.) But because of the paradoxes, further improvements in the concept will be needed if classical logic is assumed to apply to it. For a very useful survey of some of the revised theories of truth that are available in classical logic, see Friedman and Sheard 1987.

Will the revisions needed for a consistent theory of truth in classical logic undermine the “deflationary” nature of the resulting conception? Certainly not, if we count a pure Tarskian theory (the kind of axiomatic theory discussed in Section 6 of Chapter 2) as “deflationary”. For all the revisions would do is complicate the axioms; they wouldn’t, for instance, undermine the deflationistic “inversion” of the theory of reference discussed in Section 4 of Chapter 2. On the other hand, to the extent that one builds into deflationism a special role for the schemas, deflationism would be somewhat undermined.

I’m not entirely content to leave the matter there, because the revisions in the theory of truth required to restore classical consistency also tend to undermine central ways in which we use the notion of truth. (That’s why the restoration of consistency seems such a clear case of alteration of the concept.) For this reason, I am tempted by the idea that we ought to keep the standard truth schemas, but weaken the underlying logic so that these schemas will no longer lead to contradiction. It is not clear that this requires a general weakening of classical logic: we might think of the weakened logic as simply a special system of conventions designed to restore coherence to truth and related concepts. At any rate, if we can keep the schemas by weakening the logic then we don’t need to qualify the deflationary nature of the truth theory in the least.
As a result of Kripke's influential work (Kripke 1975), the first thing one thinks of under the heading of a revision in classical logic to deal with the paradoxes is the use of a logic based on the strong Kleene truth tables: more particularly, the logic obtained by taking the valid inferences to be those that are guaranteed to preserve “correct assertion”, and taking an assertion to be correct if and only if it has the “highest” of the three semantic values. Unfortunately, this "Kleene logic" as it stands does not fully support deflationism as we’ve been understanding it, because the Tarski biconditionals do not all come out correct. The inference from "p" to "‘p’ is true" and its converse can be maintained (as it can in classical logic); more generally (and unlike the situation with classical logic), one can maintain that "p" and "‘p’ is true" are everywhere intersubstitutable in extensional contexts. Maybe the latter preserves enough of deflationism, but it would certainly be desirable to be able to assert the Tarski biconditionals.3 (Moreover, the Kleene logic seems unsatisfyingly weak, in that ‘if A then A’ is not generally assertable.)

One initially attractive response is to add a new kind of conditional A→B, not equivalent to ¬A ∨ B, together with the corresponding biconditional.4 This may require that in the semantics, the underlying space of semantic values be enlarged, in a way that doesn’t alter the valid principles for the other connectives. An especially appealing proposal along these lines is to use the Lukasiewicz continuum-valued semantics; the principle that every continuous function from the unit interval into itself has a fixed point yields an easy argument that any ordinary paradoxical sentence can consistently be assigned a semantic value in this semantics, and in a very appealing way. ("Ordinary" paradoxical sentences include all those in which the only role of the quantifiers is to achieve self-reference. In particular, any Curry paradox sentence ‘If this sentence is true
then \( p' \) gets a semantic value, even when ‘\( p \)’ has a value other than 0 or 1: its value is always \((p+1)/2\).) Unfortunately, certain non-ordinary paradoxical constructions give rise to a serious problem: though the truth schemas don’t produce outright inconsistency when applied to them, they do produce an \( \omega \)-inconsistency in number theory and hence in protosyntax: see Restall 1992.5

So despite its initial appeal, the Lukasiewicz approach doesn’t work. But I’ve recently shown (Field 2003, 2004a and 2004b) that something similar in spirit to it will; and Yablo (2004) suggests several alternatives of a similar nature. The logic I use, like the Lukasiewicz logic, is based on an infinite-valued semantics, though this time the values are not linearly ordered; and the logical relations among sentences not containing ‘\( \not\)’ are the same as in Kleene logic. (Also, the conditional reduces to the material conditional when excluded middle is assumed to hold of its antecedent and consequent.) In this logic, one can consistently assume the Tarski biconditionals; moreover, \( \text{“}p\text{” is true} \) comes out everywhere intersubstitutable with ‘\( p \)’, even inside the scope of a conditional.

An alternative approach has been advocated by Graham Priest (e.g. Priest 1998): on Priest’s view, the Liar sentence and its negation are each correctly assertable (as is their conjunction), and the rules of logic are weakened so that this does not allow the assertion of just anything. The most popular version of this (Priest’s LP) is also based on the strong Kleene 3-valued truth tables, and again takes the valid inferences to be those that are guaranteed to preserve “correct assertion”; but it takes an assertion to be correct if and only if it has one of the two highest of the three semantic values. The truth schema can be consistently asserted, as can the intersubstitutivity of ‘\( p \)’ and \( \text{“}p\text{” is true} \) even in
The Curry paradox is blocked, since the only conditional in the language is the material conditional, i.e. the one defined in the usual way from \( \neg \) and \( \lor \), and this conditional does not fully validate modus ponens in this logic. There is something counterintuitive about the absence of a stronger conditional in LP; it is probably more satisfactory to extend LP by adding one.\(^7\) (To avoid the Curry paradox, any such conditional must fail to obey either modus ponens or the deduction theorem; this is so for new conditionals added to Kleene logic also, and in my view it is the deduction theorem that should be abandoned.)

The distinction between the Kleene logic and the Priest variant is commonly described as being that Kleene’s takes “deviant” sentences (e.g. the Liar) to have truth value gaps, i.e. to be neither true nor false, but takes no sentences to have truth value gluts, i.e. to be both true and false; whereas Priest’s takes deviant sentences to have gluts and takes no sentences to have gaps. But for anyone who wants to maintain an equivalence between “‘p’ is true” and “p” (even in the interior of sentences), this is wholly misleading. For to assert a truth value gap in a sentence ‘A’ would be to assert “\( \neg [\text{True}(‘A’) \lor \text{True}(\neg ‘A’)] \)”, which should be equivalent to “\( \neg (A \lor \neg A) \)”; but no sentence of that form can ever be a legitimate assertion to an advocate of Kleene logic, whereas when ‘A’ is deviant this is a correct assertion according to Priest’s logic!

Similarly, to assert the absence of a truth value glut in sentence ‘A’ would be to assert “\( \neg [(\text{True}(‘A’) \& \text{True}(\neg ‘A’)] \)”, which should be equivalent to “\( \neg (A \& \neg A) \)”; and an advocate of Kleene logic cannot assert this absence for deviant sentences, whereas an advocate of Priest’s logic can.

If one wants to state the distinction between the logics in terms of a deflationary
notion of truth, the right way to do so is this: (i) on the Kleene logic one can’t assert of a “deviant” sentence either that it has a truth-value gap or that it doesn’t, nor can one assert either that it has a glut or that it doesn’t; (ii) in Priest’s logic, one can assert both that it has a truth-value gap and that it doesn’t, and one can also assert both that it has a glut and that it doesn’t.8 There is no distinction between gaps and gluts in either logic! What differentiates the logics is simply the “threshold of assertability”: deviant sentences are always assertable in Priest’s, never in Kleene’s. Since sentences attributing truth or non-truth or falsity or non-falsity to deviant sentences are themselves deviant, this means that the advocate of Kleene logic can say nothing whatever about the truth value of deviant sentences, whereas the advocate of Priest logic can say anything she pleases about them.9

I have occasionally been tempted by the Priest approach. But in the version with only the material conditional, its retention of the Tarski biconditionals seems a bit cheap: it asserts not only the Tarski biconditionals, but also the negations of the Tarski biconditionals that involve paradoxical sentences. Since the "paradoxical" Tarski biconditionals come out no better than their negations, this seems to be at best a minor gain over the Kleene approach with only the material conditional. Perhaps that problem could be solved in an extension of LP with a new conditional, but I am unaware of one that licenses the assertion of the Tarski biconditionals without asserting any of their negations. (Indeed, there are obstacles to getting an extension of LP of this sort–see Field 2004c–though they may not be insuperable.) For this reason, I currently favor the approach mentioned four paragraphs back.

To summarize this section: (1) Even assuming that classical logic is sacrosanct, the paradoxes do not undermine deflationism as an account of our ordinary truth-theoretic
concepts; they merely show that those ordinary concepts are inconsistent and need improvement. (2) The paradoxes also don’t undermine a deflationist attitude toward our improved concepts, if ‘deflationism’ is used in the weak sense on which a Tarskian theory is deflationist. And (3) if one is willing to contemplate a weakening of classical logic, then there are several ways of maintaining even a strong form of deflationism, in which the Tarski biconditionals and the equivalence of ‘"p" is true’ to ‘p’ hold unrestrictedly, so that no improvement on the ordinary concept is required.

3. The generalizing role of ‘refers’. In chapter 2 I stressed (following Quine and Leeds) that the term ‘true’ increases the expressive power of our language, by serving as a device for expressing infinite conjunctions and disjunctions (which could have been expressed by more obviously logical means like substitutional quantifiers). The idea obviously works equally well for ‘true of’. But one might doubt that the same holds for ‘refers’ as applied to singular terms: we already have quantifiers that play a generalizing role in the name position, so it might seem that no expressive power is gained.

I think this argument incorrect: ordinary quantifiers don’t allow for generalizations of names that appear both inside and outside quotation marks, but that is what we would need to express what is said by "Every name that came up in Department X’s discussion of who to hire referred to a male" without use of ‘refers’.

4. Untranslatable utterances (1). I now think that my restriction of an agent’s disquotational truth predicate to the sentences the agent currently understands wasn’t quite right, and is in considerable tension with my remarks on our commitment to extending the truth schema as we expand the language. Those latter remarks suggest that once we come to regard an expression S as an acceptable declarative sentence, then even
if we have no understanding of it (or virtually none) still we ought to accept the corresponding instance of the truth schema. Admittedly, the cognitive equivalence thesis implies that to the extent that we don’t yet understand S, we don’t understand the attributions of truth to it; this may seem to make the extension of the truth schema to it idle. But it is not completely idle, for we tie any future improved understanding of ‘true’ as applied to S to the future understanding of S. One way to think of the matter is in terms of indeterminacy: a sentence I don’t understand at all is maximally indeterminate in content for me, and so is an attribution of truth to it; but the indeterminacies are "tied together" in such a way that the truth schema holds.

The main point of modifying my view in this way is to better accommodate talk of truth for untranslatable utterances, utterances that are in no sense equivalent to anything in our current language. When I wrote Chapter 2 I was tempted to bite the bullet and say that we simply can’t apply the notion of truth to such utterances; I wasn’t happy with this, but could see no good alternative. Stewart Shapiro (unpublished) presented a nice problem case for this position. In Shapiro’s example, we have a guru who makes pronouncements about set theory; a disciple who thinks that everything the guru says is true, but who doesn’t understand set theory; and a logician who distrusts the guru’s set-theoretic pronouncements, but likes to draw their number-theoretic consequences, which the disciple does understand. Shapiro argues that if in addition the disciple trusts the logician’s acumen about logic (though not about set theory), then the disciple ought to be able to reason that since everything the guru says is true, and consequence preserves truth, and the logician truly says that the number-theoretic sentence ‘p’ is a consequence of what the guru says, then ‘p’ must be true, so p. But though the disciple understands this conclusion, the reasoning is blocked if we can’t meaningfully apply ‘true’ to the
There are ways of handling this example without modifying the view of Chapter 2, but they aren’t entirely natural, and I’m not sure that they would extend to all alterations of the example. But the modified view suggested two paragraphs back handles the example very neatly: the disciple regards the guru’s sentences as in a potential expansion of the disciple’s language, so that there is no difficulty in his carrying out the reasoning. It may seem mysterious that the reasoning can be carried out in this framework, since according to this framework the guru’s utterances and the notion of truth for these utterances have quite an indeterminate content for the disciple. But the joint acceptance of the claims (i) that the guru’s utterances are true, (ii) that they logically imply the number-theoretic claim ‘p’, (iii) that logical consequence preserves truth, and (iv) the truth schema, give just enough determinacy to enable the reasoning to be carried out. (The reasoning could be carried out even if the guru’s language were disjoint from the disciple’s, rather than an expansion of it; though in that case some might prefer to appeal to expansions of the disciple’s language that translate the guru’s utterances rather than that include them.)

Incidentally, the guru problem is a prima facie problem not only for the version of deflationism in the paper, but for versions of deflationism that take truth to be attributed primarily to propositions. The latter, if they are to have any claims to the title of deflationism, must explain what it is for a sentence to express the proposition that p in terms of equivalence in content to our sentence ‘p’; but this explanation is unavailable for the guru’s sentences. For those views as for mine, the idea that the attributions have a quite indeterminate content, constrained only by such beliefs as that they have specific
number-theoretic implications, provides a way out.

5. Untranslatable utterances (2). The incorporation model for dealing with foreign utterances has wider application than my remarks in Section 8 of Chapter 2 suggest. Indeed, in the next to last paragraph of Section 8 I explicitly ruled out using it for sentences that we don’t understand. As the discussion of Shapiro’s example should make clear, I recant.

Consider our reaction when on a certain occasion Z we hear another speaker, Mary, use a proper name such as ‘George’, and we take her to be "talking about someone we have never heard of before" (as we colloquially put it). If we are interested in what she says, we may pick up the name ‘George’ from her and start speculating with it ourself. ("I wonder if George lives in California.") When I say that we pick it up from her, I don’t just mean that we start using the sound ‘George’: we probably used that sound already, e.g. in connection with various U. S. Presidents. But we pick it up with a new use, not equivalent to these prior uses, and a use that is in some way deferential to Mary’s: we base our own beliefs on hers (more accurately, on what we take her beliefs to be). If you like, you can put this by saying that we mentally introduce a new term of form ‘George_j’ where ‘j’ is an "inner subscript" that we haven’t previously used in connection with ‘George’, and we use ‘George_j’ to translate those of Mary’s assertions that we regard Mary as having "mentally subscripted" by the same subscript that she mentally attaches on occasion Z.

I do not see any reason to think that understanding this incorporation of Mary’s term requires an antecedent grasp of the notion of reference. Of course, once we have a disquotational notion of reference, we can use it in this connection: I can say "‘George’
(as I’m using it now) refers to George”, using the term ‘George’ to incorporate Mary’s use of it on the occasion in question. Or on the "hidden subscript" picture: "‘George_j’ refers to George_j (and to nothing else)”. Since we translate Mary’s use of ‘George’ on occasion Z with ‘George_j’, we can also say that Mary’s use of ‘George’ on occasion Z refers to George_j (and to nothing else). In other words, I can conclude:

George_j is identical to the referent of Mary’s term ‘George’ on occasion Z.

This allows me to use the phrase "what Mary was referring to by ‘George’ on occasion Z" instead of ‘George_j’. Since the inner subscripts are silent (and other people don’t know my inner subscripts even were I to vocalize them!), doing so is sometimes useful: it is a very good way to make clear which way I’m using ‘George’ on this occasion. Of course, I could have introduced other means for doing the disambiguating: I could have just said "I’m going to introduce the new term ‘George_{Mary, Z}' (or ‘Shrdlu’) as an equivalent of Mary’s use of ‘George’ on occasion Z; but since we have a notion of disquotational reference already, there is no need for such new terms.

The approach carries over straightforwardly to demonstratives. Suppose that on occasion Z Mary uses the term ‘that’, and suppose that I have no idea how to pick out in English (either demonstratively or descriptively) what I take her to be talking about; alternatively, suppose I have little or no idea what to take her as talking about. Here it might be misleading to incorporate her use of ‘that’ into my language with my own use of ‘that’: normal conventions about the use of ‘that’ might suggest that my term stands for something quite different. But I could always introduce some special term, say ‘Oscar’ or ‘THAT_{Mary, Z}', as a device for incorporating her use of ‘that’ on this occasion. But if I did so, I could argue just as above that in a disquotational sense of ‘refers’ (extended by
translation), Oscar is identical to the referent of Mary’s term ‘that’ on occasion Z; so
given that we have a notion of disquotational reference already, there is no need for these
special terms.

The approach just suggested is quite similar to, and perhaps identical to, the
proposal of Brandom 1984. What I’ve said is that we should view the phrase "the
referent of Mary’s ‘that’ on occasion Z" as functioning as a device for incorporating her
use of ‘that’ on occasion Z into our language. What Brandom says is that it functions as
a device of cross-speaker anaphora. I don’t entirely like Brandom’s way of putting things,
because taken by itself it seems to suggest that there is a non-deflationary notion of
reference applicable to Mary’s pronoun, and that our anaphorically dependent phrase gets
its referent from it. However, Brandom says things incompatible with that understanding
of anaphora, and in support of some sort of deflationary story, and I think the
incorporation model is at least very close to what he intended.

The point that I’ve been making for reference can be made for truth as well. There
are many reasons why a direct incorporation of another person’s utterance into my
language could be inconvenient (e.g. if I can’t pronounce it) or misleading (e.g. if it
contains indexicals that would shift their reference were I to use them). In such cases, we
can use ‘Her utterance is true’ as a means of incorporating her utterance. This indeed is
what Grover, Camp and Belnap (1975) see as the main function of ‘true’. I prefer to view
the main function of ‘true’ as the disquotational function. But as with reference, I can
argue from the disquotational use of ‘true’ to the incorporation use. For if I were to
introduce a new sentence, say ‘$\text{UTT}_{\text{Guru}, Z}$’, to incorporate the guru’s unpronounceable
utterance on occasion Z, then the disquotational properties of truth give an equivalence
between

(1) ‘UTT\textsubscript{Guru, Z}’ is true

and

(2) UTT\textsubscript{Guru, Z};

but since ‘UTT\textsubscript{Guru, Z}’ is my translation of the guru’s utterance, (1) is equivalent to

(1') The guru’s utterance is true,

and so (1’) must be equivalent to (2). That is, (1’) is legitimized as a way of incorporating the guru’s utterance, so we don’t need the special sentence ‘UTT\textsubscript{Guru, Z}’ any more.

The discussion in this section (and in Brandom, and in Grover, Camp and Belnap) shows another way, besides as a device of infinite conjunction and disjunction, in which ‘true’ and ‘refers’ can increase expressive power: or at least, can increase the range of what can be expressed easily and in a way not subject to confusions of ambiguity. Again, this expressive function could have been achieved by other means (‘George\textsubscript{Mary, Z}’ or ‘Shrdlu’ in the case of ‘refers’); but ‘refers’ and ‘true of’ are a convenient way to do it, and I have argued that if we did use another device for doing it we would have no need for the other device once we had the disquotational use of ‘refers’ and ‘true’.

6. "Pure Disquotational" and "Quasi-Disquotational" Truth. The central feature of "pure disquotational truth" is that "‘p’ is true" is cognitively equivalent to "p" (at least, modulo the existence of the sentence). I say in Chapter 2 that ‘true’ in the pure disquotational sense is to be understood as something like ‘true-as-I-understand-it’; what I really meant was ‘true-as-I-actually-understand-it’. As I emphasized, this is just a
heuristic, and I claim that the notion does not require that we take "ways of understanding" to be understood in terms of propositional content. I have a bit more to say about this in Chapter 5 of Field 2001 (in which the present chapter originally appeared).

In Chapter 2 I recommended using this notion of pure disquotational truth where possible, but I of course recognized that we use the notion of truth in connection with other people’s utterances and in connection with our own utterances in counterfactual circumstances, and I suggested ways of accommodating this fact. One way is to use what I call "quasi-disquotational truth"; I downgraded it because it assumes a stand on Quinean issues about interpersonal synonymy that I prefer to remain neutral about. But the general idea behind quasi-disquotational truth could have been put in a way that is more neutral about interpersonal synonymy:

\[(*) \quad S \text{ is } \text{true}_{qd} \text{ (at possible world } v) \iff \text{ S is to be translated by a sentence of mine (in the actual world) that is purely disquotationally true (at } v).\]

My method of "explaining away" modal intuitions about the truth of our sentences at the end of Section 9 really amounts to suggesting that it is true_{qd} in the sense of (*) rather than pure disquotational truth that we are operating with in many contexts.

I have come to think that it is unnecessary to use two distinct truth predicates, the purely disquotational and the quasi-disquotational. We can use a single truth predicate as long as we take the entities in its extension not as orthographic types but as computational types: equivalence classes of (potential) tokens under the relation of computational equivalence. The relation of computational equivalence is explained in Chapter 5 of Field 2001. An important feature of it is that it is defined only within an individual } X \text{ in a}
given possible world $u$: it doesn’t make sense to ask if one of my tokens is computationally equivalent to a token of yours, or to a token of a counterpart of me in another possible world. So we can rewrite (*) with the individual and the world made explicit, as follows:

(\text{**}) \quad S_{X,u} \text{ is true}_{qd} \text{ (at possible world } v) \leftrightarrow S_{X,u} \text{ is to be translated by a sentence of mine (in the actual world) that is purely disquotationally true (at } v).

But then “pure disquotational truth” can be viewed as just the special case where the individual is me and the world is the actual world (by assuming that in that case the “translation” is just the identity function); (**) is merely a way to generalize from the special case of me in the actual world to others and to my counterparts. (Of course the heuristic ‘true-as-I actually-understand-it’ is appropriate only for the special case where translation drops out. It is worth emphasizing, though, that this special case plays a central role in the account: the other cases result from the special case by translation.) If sentences are typed orthographically, the same sentence can be used in distinct ways in distinct worlds and we need to distinguish between a truth predicate that “holds meaning constant” and one which doesn’t; but if sentences only exist within a single world then no such distinction is required.

If ordinary or quasi-disquotational truth is defined either as in the text or by (*) or (**), it has the curious feature that utterances not translatable into our language are not true. It seems more natural to suppose it indeterminate (or maybe just undetermined) whether they are true. The matter may be somewhat academic: if the points made in the two preceding sections of this postscript are correct, the notion of an untranslatable utterance does not have entirely clear application. Still, I would now prefer to do things
in the more natural way. The way to do so is clear: instead of defining ordinary or quasi-
disquotational truth in terms of purely disquotational truth, we explain it by a schema:

(***) If S\textsubscript{X,u} is translatable as ‘p’ then □(S\textsubscript{X,u} is true iff p).

[And if S\textsubscript{X,u} is translatable as ‘p’ then □(S\textsubscript{X,u} has the truth conditions that p).]

In Chapter 2 I attached some importance to the idea that we ought to define quasi-
disquotational truth, in terms of the more-or-less logical notion of purely disquotational
truth together with translation. I must have thought that the motivation for demanding a
substantive theory of truth and truth conditions and reference, like that advocated in Field
1972 and 1978, could only be undercut if these notions are explicitly defined in terms of
more-or-less logical notions plus translation. But this seems a mistake: the argument for
a substantive theory of truth conditions and reference depends on taking truth conditions
and reference as having a certain kind of "causal explanatory" role. (For a discussion of
this, see the Postscript to Field 1972, in Field 2001.) Introducing a notion of truth
conditions by means of the schema (***) does nothing to make the notion "causal
explanatory". Of course, using (***) to introduce the notion of truth conditions doesn’t
guarantee against putting the notion to the relevant sort of "causal explanatory" use, but
as long as one doesn’t do this—and also doesn’t make unwarranted determinacy claims for
these notions (see footnote to that Postscript)—then one can maintain a deflationist
position.

7. Bigger issues. There are two big areas in which much more needed to be said than I
said in Chapter 2. First, there is little discussion of the aspects of meaning that go beyond
truth, reference, and related notions. Don’t we need a deflationary account of those
aspects of meaning too? The answer is that we do need a deflationary account of
meaning generally, but that there is no special difficulty for getting one once we have a
deflationary account of truth and reference; I hinted as much at the beginning of Chapter 2, and say more about it in Chapter 5 of Field 2001.

The other area in which a lot more needed to be said was the explanatory role of the assignment of truth conditions to mental states, both in the explanation of behavior and in the explanation of the extent to which behavior is successful in achieving various results. In a paper written some years earlier (Field 1986), I had flirted with deflationism but ended up tentatively arguing against it, on the ground that such explanations required a role for the assignment of truth conditions to mental states that the deflationist could not legitimize. The argument was so abstract and convoluted it couldn’t possibly have convinced anyone; I was skeptical myself, but couldn’t pinpoint what was wrong with it. I now think it made a number of mistakes that collectively deprived the argument of all force. Or as Jimmy Carter said of his attempt to rescue the American hostages in Iran (an attempt that had to be aborted with eight deaths before the rescue helicopters even reached Iran): it was "an incomplete success". It isn’t worth going through the argument in detail here, but I’ll sketch the sort of considerations it involved and what I now think a deflationist should say about them. One key point (which was recognized in the ’86 paper, but not made sufficiently salient) will be that there is nothing in deflationism that prevents the use of ‘true’ in explanations as long as its only role there is as a device of generalization.

It is perfectly obvious that in explaining how a pilot manages to land a plane safely with some regularity, one will appeal to the fact that she has a good many true beliefs: beliefs about her airspeed at any moment, about whether she is above or below the
glideslope, about her altitude with respect to the ground, about which runway is in use, and so forth. (None of these beliefs need be based very directly on observation; she might be flying in bad weather with some of her instruments not working, so that she must rely on complicated cues.) The deflationist obviously needs to grant this explanatory role for the truth of her beliefs, and will have to say that it is somehow licensed by the generalizing role of ‘true’.

Can this deflationist strategy be maintained? A detailed explanation of how the pilot functions would involve the existence of some class C of internal representations, intuitively representing airspeed, such that when she believes a representation in C that represents too high an airspeed she slows the plane and when she believes one that represents too low a speed she speeds it up; and the representations in C that she believes tend to be true. The last part can of course be rephrased as a generalization: she tends to believe a representation in C that represents too low a speed when the plane is in fact too slow, etc. But is this restatement of the claim that she has true beliefs about airspeed "innocent" from the deflationist viewpoint? What may well seem a problem is that it is put in terms of what the pilot’s internal representations "represent". The basic idea behind the tortuous argument of the ’86 paper was that the deflationist is forced to understand the sense in which the representations "represent" airspeed in terms of their translation into the explainer’s language, but that this is inappropriate since this should be an objective explanation in which the explainer plays no causal role.13

An obvious strategy for responding to this argument is to say that talk of representation is serving a merely heuristic role in the explanation of the pilot’s ability; put without the heuristic, the explanation involves the existence of some class C of
internal representations in the pilot and two subclasses $C_1$ and $C_2$ of $C$ such that (i) when
she believes a representation in $C_1$ she slows the plane and when she believes one in $C_2$
she speeds it up, (ii) there is a 1-1 function $f$ from $C$ to a certain set of real numbers such
that (a) $C_1$ is that subclass of $C$ that is mapped into numbers above a certain threshold and
$C_2$ is that subclass of $C$ that is mapped into numbers below a certain (slightly lower)
threshold, and (b) she tends to believe a representation $r$ in $C$ when the airspeed in knots
is approximately $f(r)$. This makes clear that we have a perfectly objective explanation,
expressible without translating the pilot’s representations or talking of their truth
conditions or of what they represent. Of course part (ii)(b) of the explanation uses an
indication relation, but no deflationist could object to that. (This is discussed further in
the Postscript to Field 1978, in Field 2001.) The function mapping internal
representations into airspeeds needn’t even give the intuitive truth conditions of those
representations in all cases: one could tell a story in which the pilot’s beliefs about what
she was doing were so weird that it would be natural to assign quite different truth
conditions to her representations. (Perhaps she believes she isn’t in an airplane at all, but
is using the controls to direct US ground forces on a foreign mission.) The
representations would be reliably correlated with airspeed but would represent something
quite different that they were not reliably correlated with.

It is also important to realize that in normal cases (as opposed to cases where the
pilot has very weird beliefs about what she is doing) we can explain the pilot’s
competence in less detail, by simply saying that she tends to believe truths about the
airspeed. (We would have to rest content with this if we had no idea that the pilot needed
to maintain airspeed within a certain range to safely land, but knew only that she needed
to base her actions somehow on her airspeed.) The objective core of such an explanation
is clear: it is that a functional correspondence between internal representations and airspeeds like that mentioned above is playing some sort of role that we may not be in a position to specify. But when we are not in a position to specify that role with enough precision, it is inevitable that we use the notion of truth in one way or another. One way to do so would be to just say: some explanation roughly like the one in the previous paragraph is true; here it is transparent that truth is playing simply a role in generalizing, and so clearly this way of proceeding is available to the deflationist. But this way of proceeding requires that we have a sample objective explanation, however crude, to use as a model, and it is not always convenient to provide one. But the alternative way of proceeding mentioned at the start of this paragraph does not require this: if we just say "She tends to believe truths about the airspeed, and this somehow enters into her landing successfully", we can be vague on the details without even sketching a sample precise model.

Is this latter way of proceeding available to the deflationist?14 When we give this sort of explanation, we say in effect (i) that she has certain internal representations that are similar enough in their role to our representations of form ‘The airspeed is n’ to count as "saying the same thing", (ii) that they tend to be disquotationally true under the translation scheme just alluded to, and (iii) that the fact that their disquotational truth conditions under this mapping tend to be satisfied plays an important role in explaining the safe landing. This is a "projectivist" or "second class" explanation, in that we make use of a similarity between the agent and ourselves in order to avoid the need to fill in explanatorily crucial details. Admittedly, it is not a projectivist explanation in the narrow sense of the term, i.e. an explanation that proceeds by vaguely specifying a fuller explanation in the agent’s computational psychology. Rather, the projectivist explanation
now under consideration proceeds by vaguely specifying a fuller explanation in the agent’s extended psychology; where the extended psychology is the computational psychology together with the assumptions about correlations between the agent’s internal states and the external world. But the central point is the same: assignment of truth-conditional content is something we need only when we don’t know how to fill in the details of the explanation.\textsuperscript{15}

The same basic point holds in cases where the agent is less reliable in his beliefs about the external world. Suppose I explain Kennedy’s having avoided nuclear war over Cuba in terms of a lucky guess as to what Khrushchev would do. Again, the objective core of the explanation is that Kennedy’s action tends not to lead to war if Khrushchev acts as he did but to lead to war otherwise, and that the action was crucially based on his having believed a certain representation; if we knew a lot of detail about Kennedy and his decision making process, we could say a great deal about this representation and the role in his psychology, but typically we don’t know enough to do this. So here we specify the representation by its truth conditions; or what amounts to the same thing, by a translation into our own idiolect. That is, our explanation is that Kennedy believed (or anyway, assumed as a basis for action) some representation which had for him roughly the role that ‘Khrushchev will blink’ has for us, and his believing or assuming this was a salient cause of his actions, and that those actions didn’t lead to war only because Khrushchev blinked. Again, it is a "second class" explanation, in that a comparison to our psychology is used in place of specifying the relevant details of the agent’s psychology. As such, it is nothing that causes any problem for a deflationist.
1. The schematic variable approach can also be used if what we take as basic isn’t the Tarski schema, but the equivalence between $A[\text{True(‘p’)}]$ and $A[\text{p}]$ where the context $A[...]$ isn’t quotational or intentional or anything like that. Strictly speaking, it is that rather than the Tarski schema that I recommend taking as basic in Section 1 of Chapter 2.

2. Shapiro (1998) and Ketland (1999) argue that the schematic variable approach undermines deflationism, at least if one assumes that logic is first order. For a reply, see Field (1999).

3. However, a proponent of Adams’ (1975) thesis about conditionals might argue that what really ought to be involved in a strong deflationism has nothing to do with the acceptance of sentences involving a biconditional connective. Rather, what ought to be involved is *conditional degrees of belief*: $\text{Cr}(p \mid \text{‘p’ is true})$ and $\text{Cr}(\text{‘p’ is true} \mid p)$ should both be 1. I don’t know if the details of this position could be satisfactorily worked out, but if so it would be another way of defending a deflationism somewhat stronger than one that relies on compositional axioms. It would also remove the parenthetical difficulty that follows in the text.

4. Typically such a conditional will not have the monotonicity property that the Kleene connectives have, in which case Kripke’s fixed point constructions will obviously be inapplicable.

5. Consider a sentence $P$ that we can informally write as ‘There are natural numbers $n$ for which my $n$-fold jump is not true’; where the *jump* of a sentence $A$ is $\neg(A \rightarrow \neg A)$ (which is a strengthening of $A$ in this logic), and where the *$n$-fold jump* is the result of iterating the jump $n$ times. (In effect $P$ says ‘I am not super-true’, where to be super-true is to have
all of ones n-fold jumps true, which on any reasonable way of relating truth to the semantic values is to have semantic value 1.) Even if we enlarge the value range from [0,1] in the reals to [0,1] in a non-Archimedean field and generalize the Lukasiewicz semantics accordingly, it is clear that the only value that can be assigned to P consistently with the truth schema in the semantics is within an infinitesimal of 1. But then for any standard natural number n, the n-fold jump of P also has values within an infinitesimal of 1; so the truth schema demands that for standard n, ‘the n-fold jump of P isn’t true’ has a value within an infinitesimal of 0. Clearly the only consistent resolution involves taking the ‘there are natural numbers’ in P to have nonstandard numbers in its range; in other words, the only consistent resolution involves a protosyntax with sentences (in particular, jumps of P) that are not genuinely finite in length.

6. Restricting to transparent contexts, of course.

7. Alternatively, one could try to argue that the counterintuitiveness isn’t so great, because conditional probability judgements can be used as a surrogate for conditionals, as in note 3.

8. In Priest’s logic one can also assert of any deviant sentence that it is "solely true" (true and not false), and that it is "solely false" (false and not true).

9. Priest often expresses his view as the claim that contradictions can be true. It might be better to say that sentences of form ‘p and not-p’ aren’t, or aren’t always, genuinely contradictory. In any case, the claim that some such sentences are true is a bit misleading as a statement of the view, since it tends to make one overlook the fact that on his theory these true "contradictions" are also not true!

10. For instance, in the example as it stands, where the guru and the disciple share an
underlying logic, the disciple has a notion of consequence that applies to the guru’s language; one might argue that the disciple’s faith in the guru is inadequately represented by the claim that everything the guru says is true, it should be represented by the claim that all consequences of what the guru says (indeed, all consequences of those together with other truths) are true. In that case, we would get the desired conclusion without the need to reason as in the previous paragraph.

11. It isn’t essential to suppose that the disciple accepts the application of the truth schema to the guru’s sentence prior to his knowing that an extraordinarily reliable guru uttered it and prior to knowing that it has the number-theoretic consequence. These pieces of information do give the disciple some very minimal understanding of the guru’s sentence, so it isn’t really necessary to hold that we can apply ‘true’ to sentences of which we have no understanding at all.

12. This is so even if we know Mary to be unreliable. For instance, if we know her to have exaggerated beliefs, we might merely accept ‘George is over 6’ 4” when Mary accepts ‘George is over 6’ 7”. Or if Mary gushes over George, we might accept ‘Mary is infatuated with George’ instead of ‘George is a wonderful guy’.

13. As noted in Leeds 1995, pp. 28-9, arguing in such a manner seems prima facie to be misguided: of course it is the correlation between the agent’s states and external conditions that is of primary explanatory importance, but why should the fact that we use a correlation between an agent’s states and our sentences as a means of setting up the correlation be thought to undermine this? Why should we think that there must be some means of specifying the relevant correlation that doesn’t go via translation into our language but is given by a substantive theory of truth conditions? I would rather not press this line, however, because doing so leads quickly to complicated issues. (Indeed,
the most confusing parts of the ’86 paper concern precisely this.) Instead of arguing that
the basic line of argument against the deflationist is misguided in very conception, I will
argue merely that it cannot be carried out in detail. I believe that this way of proceeding
has the additional advantage of making clearer the role that talk of representation plays in
the explanation.

14. In probably the most crucial passage in my 1986 paper (p. 97 middle to 99 top), I
assumed not, without any clear argument. [In the immediately preceding pages I had
discussed avoiding mention of truth conditions in explanations, along the lines mentioned
two paragraphs above. I then raised the question about what to do when we didn’t know
the full explanatory details. The discussion is obscure: it recognizes that it is legitimate
for a deflationist to use a disquotational truth predicate as a way of giving an explanation
sketch in absence of full knowledge of the details (and that even an inflationist would
need to resort to this), but concludes for reasons that it doesn’t make at all clear that we
would need something more "inflationist" in addition.] Stephen Leeds rightly focuses
much of his critique of my paper on this crucial passage (Leeds 1995, pp. 17-20), and is
understandably baffled as to what my argument could have been. For what it’s worth, my
own diagnosis is given in the next footnote.

15. I think the key problem in the 1986 paper was an ambiguity about ‘projectivist
explanation’. In some passages I used it in the more inclusive sense, as when I argued
that the deflationist could appeal to the truth conditions of another agent’s states only in
projectivist explanations. But in other passages I used it in the narrower sense, as when I
argued that the role to which we put the assignment of truth conditions in explanations of
"success phenomena" like landing the plane with regularity can’t be projectivist. The
confusing structure of the paper disguised the ambiguity.