Truth and the Unprovability of Consistency

Hartry Field

Abstract: It might be thought that we could argue for the consistency of a mathematical theory T within T, by giving an inductive argument that all theorems of T are true and inferring consistency. By Gödel's second incompleteness theorem any such argument must break down, but *just how* it breaks down depends on the kind of theory of truth that is built into T. The paper surveys the possibilities, and suggests that some theories of truth give far more intuitive diagnoses of the breakdown than do others. The paper concludes with some morals about the nature of validity and about a possible alternative to the idea that mathematical theories are indefinitely extensible.

Gödel's second incompleteness theorem says, very roughly, that no reasonably powerful recursively axiomatized mathematical theory in classical logic can prove its own consistency. This is rough in various ways—e.g. it slurs over issues about how exactly the notion of consistency is to be formulated—but it will do for present purposes.

A natural initial thought is that the theorem is slightly puzzling: we ought to be able to prove the consistency of a mathematical theory T within T by

(i) inductively proving within T that all its theorems are true,

and

(ii) inferring from the truth of all theorems of T that T is consistent.

Of course, Gödel's result shows that this "Consistency Argument" must break down, but where?

The answer to this question depends on what kind of theory of truth is built into T. It's worth looking at how the breakdown occurs in specific theories of truth. I think that the breakdown is far less counterintuitive in some theories of truth than others, and that this provides some reason for preferring those theories of truth. But the purpose of this paper is less to argue for one theory of truth over another than to clarify their different diagnoses of how the

Consistency Argument breaks down.

1. The Consistency Argument. To spell out the Consistency Argument a bit further, let us for simplicity confine attention to theories that are formalized in a formulation of logic where reasoning is done only at the level of sentences: formulas with free variables are never axioms, and the rules of inference don't apply to them.[1] In that case, the reasoning of Step (i) is:

(1) Each axiom of T is true

(2) Each rule of inference of T preserves truth (that is, whenever the premises of the rule are true, so is the conclusion)

Since a theorem of T is just a sentence that results from the axioms by successive application of rules of inference, a simple mathematical induction yields

(3) All theorems of T are true.

(This argument can easily be recast for theories formalized in more familiar formulations of logic where free-variable reasoning is allowed: the direct induction is then that all theorems of T are satisfied by everything, from which the truth of all those theorems that are sentences follows.)

The reasoning of Step (ii) is that by elementary properties of truth, no sentence and its negation can both be true. It follows that no sentence and its negation can both be theorems of T, which is to say that T is consistent. A variant form of the reasoning of Step (ii) will be useful later: if T is not consistent, then every sentence is a theorem, which in conjunction with (3) would imply that every sentence is true, which is ruled out by elementary properties of truth.

2. The Tarskian diagnosis. If T is a theory like Zermelo-Fraenkel set theory (ZF), the diagnosis of the breakdown of the Consistency Argument is simple: theories like this don't contain a general truth (or satisfaction) predicate, so the argument can't even be formulated in the theory.

Such theories do contain restricted truth predicates, e.g. '$\text{true}_{n\text{-quant}}$' $=_{\text{df}}$ 'true and contains

at most *n* quantifiers' for specific *n*. Or rather, there is an uncontroversially acceptable way to define such restricted predicates in the theory. But such predicates are not enough to run the inductive argument for (3), or for obvious weakenings of (3) like

($3_{n\text{-quant}}$) All theorems of T with at most *n* quantifiers are $\text{true}_{n\text{-quant}}$.

For each *n*, we can inductively prove that every derivation in which every sentence has at most *n* quantifiers has a conclusion that is $\text{true}_{n\text{-quant}}$. But that wouldn't suffice for (3_n): a theorem with at most *n* quantifiers might only have proofs that involve sentences with more than *n* quantifiers.

What about if we expand ZF by adding the predicate 'is a true sentence in the language of ZF' plus appropriate axioms governing it? Call this ZF*. The problem is still the same: the truth predicate we've added is a general truth predicate for the language of ZF, but not for the full language of ZF*.

In the standard Tarskian picture there is an infinite hierarchy of ever-more-inclusive truth predicates: a predicate 'true_0' that has in its extension only sentences not containing any truth predicates; a predicate 'true_1' that has in its extension only sentences not containing any truth predicates other than 'true_0'; a predicate 'true_2' that has in its extension only sentences not containing any truth predicates other than 'true_0' and 'true_1'; and so on (where the hierarchy can be extended a good way into the countable ordinals, and has no last member). Again, the Consistency Argument cannot be formulated; and there is no even prima facie inductive argument for

(3_α) All theorems of T with no predicate 'true_β' for $\beta \geq \alpha$ are true_α,[2]

since a sentence that doesn't contain 'true_β' for large β might nonetheless be proved using sentences that do.

But there are strong reasons to be dissatisfied with the Tarskian hierarchy: see Kripke

1975. On the usual alternatives, we do have a unified truth predicate. Of course, Tarski proved an important negative result about theories with unified truth predicates: he proved that no theory of truth (in a sufficiently rich metalanguage that permits self-reference) whose logic is classical can have a general truth predicate that obeys the truth schema

(T) <p> is true if and only if p.

But this leaves two possibilities. First, it allows for theories of truth in classical logic that employ general "truth predicates" that restrict the truth schema. There are quite a few possibilities for theories of this sort (especially if we are generous about what counts as being in classical logic—see below); Friedman and Sheard 1987 surveys the possibilities that meet certain natural constraints. Second, it allows for theories that leave the truth schema unrestricted, but weaken classical logic a bit to accommodate it.[3] How does the Consistency Argument fare in theories of these types?

3. Two dubious diagnoses. A *conceivable* diagnosis of the failure of the Consistency Argument for theories with a unified truth predicate is that the problem arises from the extension of the induction schema to formulas containing 'true': mathematical induction, it could conceivably be held, works fine for ordinary formulas, but is suspect for formulas containing 'true'.

I think this diagnosis borders on the incredible: induction ought to be valid for any meaningful formula.[4] And as we will see, almost every standard approach to the theory of truth gives a diagnosis of the breakdown of the Consistency Argument that does not depend on restricting mathematical induction (when that is formulated in accordance with the previous footnote); the only exception is one form of dialetheism (to be mentioned in Section 11) that may have no advocates.

An alternative diagnosis of what's wrong with the Consistency Argument might be that a

theory T with a unified truth predicate and that includes a powerful mathematics like ZF will have infinitely many axioms. This may seem plausible since ZF itself has infinitely many axioms. The thought is that in such a T we will be able to prove of each axiom that it is true, but not to prove the universal generalization of this. In other words, (1) will not be provable in the theory.

But this alternative diagnosis can't be correct in general (or even, for all theories of sufficient strength). The reason is that in most theories with a truth predicate, that predicate can be used to finitely axiomatize. Given a theory T with infinitely many axioms, we can replace it by a theory $T^{\#}$ with the single axiom "All the axioms of T are true".[5] As long as the theory contains the single rule

(T-Elim) $\quad$ True(<A>) $\vdash$ A,

this will have all the consequences that the original theory has; and since it has only a single axiom, the diagnosis above can't hold for it. (Of course, $T^{\#}$ might be more powerful than T, and might be able to prove the consistency of T. This doesn't undermine my point, which is that if the diagnosis were to hold for all sufficiently strong theories it would have to hold for $T^{\#}$; but it doesn't, since $T^{\#}$ has only a single axiom and yet can't prove *its own* consistency.)

If the theory were to contain infinitely many rules, one might consider an analogous diagnosis: that though the theory implies the assertion that R is truth-preserving, for each rule R that it employs, it doesn't imply the universal generalization (2). But this diagnosis isn't correct either, for any theory of truth I know of: they all contain only finitely many rules.

If the above diagnoses of how the Consistency Argument fails are incorrect, what's left? Let's start by considering theories in which the logic is classical, in the sense that all arguments that are valid classically are taken as legitimate. (There might be additional validities beyond the classical ones: for instance, rules that essentially involve the notion of truth or satisfaction.) Then

aside from one totally unattractive classical theory ("hyper-dialetheism") that I'll mention in Section 5, the breakdown of the Consistency Argument in classical theories always occurs either because of a failure of *an individual instance of* (1), or because of a failure of *an individual instance of* (2). And when I say here that there's a failure, I mean not just that the theory doesn't contain the claim, I mean that it contains its negation. That is, it is always the case either that

(A) The theory employs an axiom that the theory implies is not true,

or that

(B) The theory employs a rule of inference that it implies is not truth-preserving.[6]

Details follow in Sections 4-7. Starting in Section 8 I move to theories that weaken the logic; a *prima facie* advantage of some such theories is that they simultaneously avoid (A) and (B).

4. A popular classical approach. Perhaps the most popular view among non-specialists is that we should accept one half of the schema (T) but not the other: more specifically, we should accept all instances of

(T-OUT) If True(<A>) then A,

but not all instances of the converse

(T-IN) If A then True(<A>).

(The basic theory of this sort is often called KF, for Kripke and Feferman.) Given the existence of "Liar sentences" that directly or indirectly assert their own untruth—that is, sentences Q for which Q is equivalent to ¬True(<Q>)—we can easily derive both Q and ¬True(<Q>).[7] That is, Q is a theorem of the theory T, but so is the claim about that theorem that it is untrue. But it isn't just certain *theorems* whose untruth is implied: the theory implies the untruth of certain of its *axioms*. For instance, the sentence

If True(<Q>) then Q

is an instance of (T-Out), hence an axiom; but the theory implies

(*) ¬True(<If True(<Q>) then Q>).

(The theory takes

(**) If True(<Q>) then Q

to be equivalent to

Either not True(<Q>), or Q.

This in turn is equivalent to Q, since by the Liar property, the untruth of Q is equivalent to Q. Since the theory takes (**) to be equivalent to Q, and takes Q not to be true, it is not surprising that it takes (**) not to be true, i.e. that it accepts (*).)

To my mind, a theory like KF that declares some of its axioms untrue is unsatisfactory. To those who share this view, one reaction might be to try to weaken KF to a theory KF_w without these "problematic" instances of (T-OUT).

A first point to be made about this suggestion is that it seems totally against the spirit of KF. An immediate lesson of the paradoxes is that if you are to keep classical logic then one or both of (T-OUT) and (T-IN) must be restricted, and the whole point of KF was to insist that restrictions are only required in the second—not in the first too, as with KF_w.

A second point to be made about this suggestion is that without some clear proposal about *how* (T-OUT) is to be restricted, the suggestion is almost useless. Until one is told precisely *which* instances of (T-OUT) are axioms, KF_w simply hasn't been specified.

A third point is that it is doubtful that any proposal for KF_w that is recursively axiomatized could be remotely satisfactory. Let f be any function that is definable in the language of T and that takes natural numbers to sentences in that language, and consider sentences of the form

(S_n) The result of applying ___ to n is not true

where the blank is filled with some definition of f. For each S_n, we have a corresponding instance

of (T-OUT):

(U_n) If <The result of applying ___ to n is not true> is true then the result of applying ___ to n is not true.

I take the spirit of the above suggestion on restricting (T-OUT) to be

Constraint I: If the result of applying f to n is an "unproblematic" sentence like 'Snow is white', then U_n should be part of the theory (probably an axiom, but at least a theorem).

And the following is clearly part of the proposal:

Constraint II: If the result of applying f to n is "pathological" (for instance, if it is S_n itself), then U_n should not be part of the theory (i.e. not even a theorem of the theory).

For without Constraint II, the unintuitive feature of KF would recur. But now let Z be a set of natural numbers that isn't recursively enumerable; if we consider (definable) functions that take "pathological" values for all and only those n that are not in Z, we see that the above constraints require that T not have a recursively enumerable set of theorems and hence not be recursively axiomatizable.

Of course, the problem could be avoided by weakening Constraint I, but this brings us back to the previous point: it isn't clear just how to weaken it in a satisfactory way.

There is a closely related point to be made, about sentences for which it is an empirical question whether they are pathological: e.g. "The first sentence uttered by a member of the NYU Philosophy Department in 2007 will not be true". A theory of truth must tell us whether the corresponding instance of (T-OUT) is part of the theory. To say that we can't tell whether that instance is part of the theory can't be settled until 2007 (at earliest) would seem most unsatisfactory.

5. Dialetheic theories in classical logic. I know of no one who has advocated the "reverse" of

the theory KF considered early in the last section: the theory that keeps (T-IN) but restricts (T-OUT). Despite its unpopularity I think it's worth briefly considering how such a view would treat the Consistency Argument.

Any classical theory that keeps (T-IN) is *dialetheic* in the sense that it takes certain sentences to be both true and false, where 'false' means 'has a true negation'. In particular, if Q is a Liar sentence, it will take both Q and its negation to both be true. (Proof in next paragraph.) But unlike the more familiar dialetheic views to be considered in Sections 11 and 12, which involve non-classical logics, these dialetheic views are classically consistent: they do not accept any contradictions. It might seem that they must accept the contradictory pair {Q, ¬Q}, given that they accept that each of its conjuncts is true. Not so! The views accept both True(<Q>) and True(<¬Q>), but these are not contradictory (neither is the negation of the other). If one had (T-OUT), or even the rule (T-Elim) from Section 3, then one could conclude to Q and to ¬Q, which is a contradictory pair; but one does not have (T-Elim) in this theory.

The argument that (T-IN) leads to both True(<Q>) and True(<¬Q>) in classical logic is a dual of the reasoning involving (T-OUT) in note 7. We have that if Q then True(<Q>), by (T-IN), and that if ¬Q then True(<Q>), by the meaning of Q, so True(<Q>) either way. But True(<Q>) is equivalent to ¬Q, so we have ¬Q; and by (T-IN) we get True(<¬Q>). So we have True(<Q>), True(<¬Q>), and ¬Q; but there's no way to get Q.

How does this "classical-logic dialetheism" diagnose the failure of the Consistency Argument? To answer this generally, we must subdivide: there are two possible versions of classical-logic dialetheism, though the first isn't very interesting; and they give very different diagnoses of how the Consistency Argument fails.

The uninteresting version might be called *hyper-dialetheism*: it is the view that every

sentence is true. This implies that every sentence is also false, given the identification of falsehood with truth of negation. Again, this is a perfectly consistent view: for though it regards both 'The earth is flat' and its negation as true, it disallows inferring from this that the earth is flat (or inferring its negation). On this view, 'true sentence' is just a long-winded way of saying 'sentence'. (We might call this the redundancy theory of truth, were that name not already taken for a somewhat more sensible doctrine.)

Clearly the problem with the Consistency Argument, on the hyper-dialetheic view, isn't in the inference to (3): that's a legitimate induction with an acceptable conclusion. The problem, rather, is that hyper-dialetheism blocks the inference from the truth of the theory to its consistency: it takes even inconsistent sentences to be true.

One might think that any version of dialetheism could block the Consistency Argument in the same way. This is not so. Consider theories in classical logic that accept (T-IN) and hence are dialetheic, but that are not hyper-dialetheic. Indeed, let's focus on classical theories that accept (T-IN) for which there is at least one sentence $\perp$ that the theory implies not to be true (perhaps '$0=1 \wedge \neg(0=1)$'). Then (3) implies that $\perp$ can't be a theorem of T, and hence implies that T is consistent. (This last step relies on the fact that in classical logic, anything follows from an inconsistency; in effect, I've used the "variant form of the reasoning in step (ii)" that was given in Section 1.) Consequently, the incompleteness theorem shows that in a dialetheic theory of this kind, the inductive argument for (3) must somehow be blocked.

And it is blocked, at step (2). Indeed, a theory of this kind must entail that modus ponens is not truth-preserving. It is still a classical theory in the sense defined above: it employs modus ponens. But it must declare its own rule not to be truth-preserving.

Why is this? We've seen that on such a view, True(<Q>), True(<$\neg$Q>), and $\neg$True($\perp$).

What about the conditional $\neg Q \rightarrow \bot$? Very likely, the view will regard it as not true: after all, it has a true antecedent and untrue consequent. If so, then the view will take the conclusion of the following instance of modus ponens to be untrue:

Q

$Q \rightarrow (\neg Q \rightarrow \bot)$

$\therefore \neg Q \rightarrow \bot$

But we've seen that the view takes the first premise to be true. And it takes the second premise to be true too, given that that is a truth of classical logic and (T-IN) holds. So we have an instance of modus ponens that isn't truth-preserving.

It would seem rather desperate for a non-hyper-dialetheic adherent of (T-IN) to declare $\neg Q \rightarrow \bot$ true despite its having a true antecedent and untrue consequent, but anyway, such desperation would be of no help: in that case, the following instance of modus ponens would have true premises and a false conclusion (according to the view):

$\neg Q$

$\neg Q \rightarrow \bot$

$\therefore \bot$

The situation then is that the sentence $\neg Q \rightarrow \bot$ is either true or not true (given the assumption of classical logic), and either way, modus ponens fails to be truth-preserving. (One might have a view that was agnostic as to the truth of $\neg Q \rightarrow \bot$ and hence agnostic as to which instance of modus ponens is problematic, but it's hard to see any advantage in that.)

6. Weakly classical approaches. The theories I've considered so far either (i) imply certain sentences that they also imply not to be true (Section 4) or (ii) imply certain sentences to be true while also implying their negations (Section 5). Indeed it may seem at first as if any classical theory (with a truth predicate and the elementary syntax required to talk about its own sentences)

must have this characteristic. For by standard Liar reasoning, any such classical theory will imply the disjunction

(D) Either Q and ¬True(<Q>), or ¬Q and True(<Q>).

And whichever disjunct one picks, we have either situation (i) or situation (ii).

But this ignores a third option: we can disavow Yogi Berra's advice "When you come to a fork in the road, take it". That is, we can accept the disjunction (D) without accepting either disjunct. If this just amounted to standard agnosticism (being undecided whether to adopt a theory of sort (i) or a theory of sort (ii)), it would be uninteresting: agnosticism as to which of two apparently unsatisfactory views to adopt isn't an interesting third possibility. But what I have in mind here—and what is embodied in many "classical logic" theories of truth, e.g. "rule of revision theories" such as Gupta 1982 and supervaluational theories such as McGee 1991—is not agnosticism of a standard sort. Rather, the idea of these theories is that it would be *absurd* to accept either disjunct of (D): the acceptance of either disjunct would commit one to a contradiction. But though it would be absurd to accept either disjunct, it is not absurd to accept the disjunction!

More fully, the view under consideration takes the following principles governing truth to be legitimate:

(T-Elim) True(<A>) ⊢ A

(T-Introd) A ⊢ True(<A>)

That is, the inference from True(<A>) to A is in some sense "valid", and so is its converse. By "valid" I mean that it is perfectly legitimate to infer from premise to conclusion: if you've established True(<A>) you can regard yourself as having established A, and vice versa. Given this, it is clear why accepting either disjunct of (D) would be absurd: accepting the first would lead to immediate contradiction via (T-Introd), and accepting the second would lead to immediate

contradiction via (T-Elim). But it turns out that the disjunction can be consistently maintained in what is, in a sense, a classical theory.

Of course we know that no classical theory can consistently maintain both (T-OUT) and (T-IN); so (T-Elim) and (T-Introd) must not imply (T-OUT) and (T-IN) in this theory. And that means that one of the standard meta-rules of classical logic, ¬-Introduction, must be somehow restricted when applied to sentences containing 'True'. But ¬-Introduction is a *meta*-rule, a rule allowing you to pass from validities to validities; giving it up is still compatible with being a classical theory *in the sense defined in Section 3*, namely taking all the classically valid inferences to be legitimate.

Another classical meta-rule which on this view can't apply without restriction is disjunction elimination (reasoning by cases): the view that if we can validly infer a claim C from either of two sentences then we can validly infer C from their disjunction. For the whole point of the view is that we can validly infer the contradiction $Q \wedge \neg Q$ from the Liar sentence Q and also from ¬Q, but we can't validly infer it from the logical truth $Q \vee \neg Q$. To my mind, abandoning reasoning by cases is highly counterintuitive, and does violence to the notion of disjunction; but it is not my purpose here to argue the matter.

There's no need to discuss the verbal question of whether theories that accept the validities of classical logic but restrict such meta-rules as ¬-introduction and reasoning by cases deserve the honorific "classical". But it's useful to have clear labels, so let's call these ones "weakly classical" and those that keep the meta-rules "strongly classical". (N.B.: "strongly classical" is taken not to imply "weakly classical"; rather, the two are exclusive varieties of classical.) The views discussed in Sections 4 and 5 were strongly classical.

Having said what these weakly classical views of truth are, I now turn to the question of

what they have to say about the Consistency Argument. And the answer is that (as for the main dialetheic views discussed in Section 5) they postulate that some of their own rules are not truth-preserving. Indeed, any reasonably detailed such theory implies a disjunction of at most two specific counterinstances to truth-preservation.

There is a basic result here, which is essentially contained in Section 5 of Friedman and Sheard 1987: If a weakly classical theory contains the rules (T-Elim) and (T-Introd) and declares each classical validity and each theorem of elementary syntax as true, then it implies that one of its rules (either (T-Elim) or (T-Introd) or modus ponens) fails to preserve truth.

Instead of reproducing their proof of this general claim, I will confine myself to illustrating how it works out for the most typical such theories: e.g. all the standard revision theories (e.g. Gupta 1982) and strong supervaluational theories (McGee 1991). For *some* of these theories there is no problem with modus ponens: the theories not only accept reasoning by modus ponens, they also declare modus ponens to be truth-preserving. (That is not so for some supervaluational theories weaker than McGee's. It is also not so for Gupta's theory, as noted by McGee p. 137; but as McGee also notes, it is so for stronger revision theories such as Herzberger 1982.) For (T-Elim) and (T-Introd), however, the situation is different: the theories accept these rules but declare them not to be truth-preserving. Indeed, they declare that the rules fail to preserve truth either in the case of the Liar sentence Q or in the case of its negation ¬Q (though they don't say which).

The reason is clear: in these theories, True(<Q>) is equivalent to ¬Q and True(<¬Q>) is equivalent to Q; using minimal assumptions accepted by these theories, it follows that True(<True(<Q>)>) is equivalent to Q and that True(<True(<¬Q>)>) is equivalent to ¬Q. Given this, the disjunction (D) from a few paragraphs back yields

(D*) Either True(<True(<Q>)>) and ¬True(<Q>), or True(<True(<¬Q>)>) and True(<Q>).

And these theories imply that no sentence and its negation are both true, so we get

(D**) Either True(<True(<Q>)>) and ¬True(<Q>), or True(<True(<¬Q>)>) and ¬True(<¬Q>).

In other words, (T-Elim) fails to preserve truth, either when applied to Q or when applied to ¬Q. (Of course, one couldn't say for which of these two the failure occurs, without committing to Q or to ¬Q, and hence without breeding inconsistency.) The argument for the failure of (T-Introd) to preserve truth is analogous. So Step (2) of the Consistency Argument is blocked twice over.

7. Restricted v. unrestricted truth preservation in weakly classical theories. Is the fact that weakly classical theories declare their own rules not to be truth-preserving a serious defect of those theories? While in some sense I think it is, it isn't obvious that it is a defect over and above other defects of the theory.

Initially, it may seem as if there is something very odd about employing a logical rule when we know it fails to preserve truth. But perhaps this isn't so. After all, it might fail to preserve truth *generally*, but nonetheless preserve truth *in the restricted circumstances where we will apply it*. And there is reason to think that that is exactly what happens in the case of the weakly classical theories:

(i) The failures of truth-preservation seem to arise only for pathological sentences like Q and ¬Q.

(ii) Rules like (T-Elim) and (T-Introd) aren't to be applied to arbitrary sentences, they are to be applied only in passing from theorems to theorems. And we don't expect such pathological sentences to be theorems; so the failure of truth-preservation won't matter in the situations where we apply the rules.

In short, even if the rules don't preserve truth generally, they may preserve it where it matters, and this seems enough to legitimize their employment.

Indeed, it might be thought misleading to say that a rule like (T-Elim) fails to preserve truth

in weakly classical theories. It is undeniable that when the premise is a true sentence that isn't a theorem, the conclusion needn't be true; but, it could reasonably be said, this is irrelevant, since the rule is only *properly applied* when the premise is a theorem. (Compare the necessitation rule in modal logic, which also preserves truth as applied to theorems though not as applied to non-theorems.) Let us not get hung up in a debate over the meaning of 'truth-preserving': let's just introduce a distinction between unrestricted and restricted truth preservation. To say that the rule (T-Elim) is unrestrictedly truth-preserving is to say that for all sentences x, if *the claim that x is true* is true then *x itself* is true. To say that it is restrictedly truth preserving is to say that this holds when x is a theorem (or more generally, when x can legitimately be asserted). From now on I'll mostly avoid using the unadorned term 'truth-preserving' (but when I do, it will mean unrestrictedly).

The distinction between restricted and unrestricted truth-preservation does not undermine the diagnosis of where the Consistency Argument breaks down in weakly classical theories: it still breaks down at Step (2). An advocate of a weakly classical theory of truth can't restore the Consistency Argument by saying that (T-Elim) and (T-Introd) and the other rules of the theory preserve truth when applied to theorems; for this presupposes that pathological claims like Q and ¬Q aren't theorems, which in turn presupposes the consistency of the theory. In other words, the claim that these rules are at least restrictedly truth-preserving may be plausible, but it presupposes consistency and can't be used in a non-question begging argument for it. In the sense of truth-preservation *that is ascertainable independently of what the theorems of the system are* ("unrestricted truth-preservation"), these rules fail to be truth-preserving.

I think these points do remove some of the prima facie oddity of a theory declaring that (in an important sense) its rules are not truth-preserving. To my mind, though, there is still a strong discomfort in the fact that theories like this accept (T-Elim) but at the same time accept

[True<True<Q>> ∧ ¬True<Q>] ∨ [True<True<¬Q>> ∧ ¬True<¬Q>].

For to accept this is to accept the disjunction of two claims each of which the theory (rightly) says is absurd. This, I think, is intuitively a problem, but it is not a new problem: it is simply another instance where the theory thinks that the disjunction of two absurdities needn't be absurd.

8. Classical logic v. the Intersubstitutivity Principle and truth schema. I have argued that except for hyper-dialetheism, each strongly or weakly classical theory with a truth predicate either denies the truth of one of its axioms or denies the (unrestricted) truth-preservingness of one of its own rules. The former is decidedly odd; perhaps the latter is not so clearly odd, though the particular form it takes in these theories seems counterintuitive.

There is also (what I take to be) a more serious difficulty with all the classical theories, including the weakly classical ones: they are incompatible with the truth predicate serving its standard role. Consider the claim

If everything Joe said yesterday is true then we are in trouble.

On the assumption that what Jones said yesterday was p_1, ..., p_n, then this ought to be equivalent to

If p_1 and ... and p_n then we are in trouble.

But clearly this can only be so in general if "<p> is true" is intersubstitutable with "p" even within a conditional. And such intersubstitutivity will not hold in any (strongly or weakly) classical theory. Indeed, the intersubstitutivity of "<p> is true" with "p" in the logical truth

If p, then p

would lead both to

(T-OUT) If <p> is true, then p

and to

(T-IN) If p, then <p> is true;

that is, it would lead to the full truth schema, which we know is classically inconsistent. The

classical theories must thus restrict intersubstitutivity, which precludes giving 'true' its standard role.

I would now like to look at theories that restore the standard role for truth by a weakening of the logic. Obviously there can be debate about whether weakening the logic is too high a price to pay for restoring the standard role for truth. Such a debate can't be intelligently conducted without looking in great detail at what the possibilities are for truth theory in a weakened logic, and that is not something I will undertake here. My goal here is limited to surveying some of the options for preserving the standard role for truth in a weakened logic, and seeing how they diagnose the failure of the Consistency Argument. Still, some remarks are necessary on what an acceptable theory must say about truth.

The standard role of truth requires that "<p> is true" be intersubstitutable with "p" not only within a conditional but in all transparent (non-quotational, non-intentional etc.) contexts.

> **Intersubstitutivity Principle:** If A and B are alike except that (in some transparent context) one has "p" where the other has "<p> is true", then one can legitimately infer B from A and A from B.

(Obviously this extends to multiple substitutions, by transitivity of legitimate inference.) Even a restricted form of this Principle implies all instances of the truth schema

(T) <p> is true if and only if p

in any logic meeting the very modest requirement that "if p then p" holds generally. (But this modest requirement is violated by the Kleene logic used in one version of Kripke's theory of truth). Conversely, the Intersubstitutivity Principle is implied by the Tarski schema, as long as the logic of the conditional satisfies certain very natural laws. I will restrict attention (except for a couple of remarks about Priest's theory) to logics strong enough for (T) and the Intersubstitutivity Principle to be equivalent.

I will also confine my attention to theories that keep one feature of classical logic that "weakly classical" theories abandon: the classical meta-rule of reasoning by cases will be assumed to hold without restriction. In *some* ways, then, the revisions of logic to be considered are less drastic than in the "weakly classical" theories: there is no tolerance for adhering to a disjunction if one rejects each of its disjuncts as absurd.

I will be concerned with diagnosing the failure of the Consistency Argument in such non-classical theories; but first, is it so clear that the argument does fail? The worry is that the second incompleteness theorem is a theorem of classical mathematics (more precisely, classical arithmetic); how do we know that it holds in the non-classical context? The answer is that the non-classical theories that will be under consideration are "effectively classical" as regards arithmetic: their non-classical aspects come out only as regards sentences containing 'true'.[8] The incompleteness theorem, as applied to theories containing 'true', certainly *mentions* the word 'true'; but it doesn't use it, so classical logic applies and the incompleteness theorem holds. So the Consistency Argument must indeed fail.

9. Paracomplete theories: restricting excluded middle without dialetheism. Non-classical theories that maintain the Intersubstitutivity Principle and/or the truth schema fall into two main types, the dialetheic and non-dialetheic. This section and the next will be concerned with the non-dialetheic ones and their response to the Consistency Argument. JC Beall (in conversation) has suggested the term 'paracomplete' for such theories. In the dialetheic context I'll mention theories that satisfy the truth schema without the Intersubstitutivity Principle, but in discussing paracomplete theories I will confine my attention to theories that satisfy both (and also allow reasoning by cases). The acceptance of the truth schema rules out theories whose underlying logic is Kleene logic, for instance the Kripkean theory designated KFS in Reinhardt 1986 and advocated in Soames 1999.

To motivate the paracomplete approach, consider a simple argument for the inconsistency of the Intersubstitutivity Principle. We know that in the presence of Intersubstitutivity, the Liar sentence Q implies its own negation ¬Q, and hence implies the contradiction Q ∧ ¬Q. Similarly, ¬Q implies Q, and hence also implies Q ∧ ¬Q. So if we allow reasoning by cases, Q ∨ ¬Q together with Intersubstitutivity implies Q ∧ ¬Q. But then by the law of excluded middle, according to which every sentence of form A ∨ ¬A is valid, we then get the contradiction Q ∧ ¬Q from Intersubstitutivity.

"Weakly classical" theories take reasoning by cases to be illegitimate, and thus block this particular argument (though they still get the inconsistency of Intersubstitutivity by other routes). But if you want to keep Intersubstitutivity together with reasoning by cases, and you reject contradictions, then it is clear that the law of excluded middle has to be restricted.[9] (There is no violation here of Yogi Berra's advice from Section 6: if we don't accept Q ∨ ¬Q we don't recognize a fork in the road, so there's no reason to take it.)

Can excluded middle be taken to be the *only* culprit in the paradoxes? There's a sense in which the answer is 'yes' and a sense in which it's 'no'. The sense in which it's 'no' is obvious: intuitionist logic gives up excluded middle, but is inconsistent with the Intersubstitutivity Principle. However, intuitionist logic accepts certain forms of reasoning which can plausibly be argued to depend on excluded middle for their motivation. An example is the intuitionist reductio rule: if Γ and A together imply ¬A then Γ alone implies ¬A. The most obvious argument for this rule, I think, is: if Γ and A together imply ¬A, then since Γ and ¬A together certainly imply ¬A, reasoning by cases yields that Γ and A∨¬A together imply ¬A; so by excluded middle, Γ alone implies ¬A. Thus the reductio rule can be argued to rest on excluded middle. (Of course, the intuitionist will try to argue for the reductio rule in a different manner; I will not discuss that here.) The application of the Intersubstitutivity Principle to the Liar sentence leads to inconsistency given the

reductio rule; but to the extent that that rule rests on excluded middle then there's a sense in which we can still regard excluded middle as the only culprit in the Liar paradox.

Of course, there are other paradoxes than the Liar. In particular, reasoning with the conditional requires some not immediately obvious restrictions, if the Intersubstitutivity Principle is to be preserved. The simplest paradox of the conditional is the Curry Paradox. This involves a sentence K which asserts (directly or indirectly) that if it itself is true then 0=1. (You can replace '0=1' by any other absurdity, e.g. 'The earth is flat'.) From this we seem to be able to argue that 0=1 (or that the earth is flat), by any of several routes; the most familiar is in two steps:

> Step One argues that *on the assumption True(<K>)*, 0=1 follows. The argument: from True(<K>) we get K (by Intersubstitutivity, or even just (T-Elim)), which is equivalent to True(<K>)→0=1; and True(<K>) and True(<K>)→0=1 then yield 0=1 by Modus Ponens.
>
> Step Two: since by the first step 0=1 follows from True(<K>), we get the conditional True(<K>)→0=1 by →-introduction. But that's equivalent to K, so we get True(<K>) by Intersubstitutivity (or even just (T-Introd)). So we can now conclude 0=1 by modus ponens, this time not in the scope of an assumption.

That's the paradox. If we insist on keeping Intersubstitutivity and Modus Ponens, as I shall, then it is clear that →-Introduction cannot be accepted without restriction.[10] Several other classical principles involving the → (for instance, the equivalence between A→(B→C) and A∧B→C) likewise can be shown to generate paradox and hence must be restricted.

But the restrictions on the laws of the conditional can also be taken to depend on restrictions on excluded middle. More fully: for any sentences B and C, the conditional B→C can be taken to be equivalent to ¬B ∨ C *on the assumption of the two instances of excluded middle B∨¬B and C∨¬C*. In classical logic of course B→C is always equivalent to ¬B∨C; that would

make A→A equivalent to an instance of excluded middle, so it can't possibly hold in a logic without excluded middle but with the law A→A. But the above says that though the equivalence doesn't hold generally, it does on the assumption of excluded middle for the antecedent and consequent of the conditional. So in a clear sense, any non-classicality in the conditional results from restrictions on excluded middle. (More generally, we can argue that if Γ classically implies A then Γ together with relevant instances of excluded middle imply A in this logic; where the relevant instances of excluded middle are confined to A and the members of Γ.)[11]

There is more than one way to fill in the details of a logic of the sort just adumbrated which is consistent with the Intersubstitutivity Principle. For present purposes the details won't matter much. Let me just say that my preferred versions have the following features: reasoning by cases is legitimate (in contrast to "weakly classical" theories), and the properties of the conditional required for the interdeducibility of the Intersubstitutivity Principle and the truth schema all hold. Modus ponens holds too. And unlike intuitionist logic, all the deMorgan laws hold and double negation is redundant. In the versions I prefer, we have contraposition in the strong conditional form (⊢(A→B)→(¬B→¬A)); and as in classical and intuitionist logic, contradictions "explode", i.e. imply everything. Moreover, the weakening of classical logic needn't ultimately affect reasoning with sentences not containing 'true'. In mathematics, physics, etc., logic is "effectively classical", because all 'true'-free instances of excluded middle can be taken as (non-logical) axioms. I will not give further details of such a logic, but perhaps the least technical introduction to a logic of this sort that has been shown consistent with "naive truth" is Field 2003b. (The consistency proof is in Field 2003a.)

Given such a paracomplete approach to truth, where does the reasoning of the Consistency Argument break down? The step from (3) to consistency is unproblematic: it follows from the truth theory that no arithmetical sentence is both true and false, so (3) implies that no arithmetical

sentence and its negation can both be theorems, and so (given that contradictions explode) the theory is consistent.[12] So the problem must be in the Inductive Argument for (3). But where does that inductive argument break down?

Is it in induction itself? I have already dismissed that possibility, but I should now add a qualification: *certain formulations* of induction are indeed suspect in absence of excluded middle: in particular, if induction is put as a sentence schema or as a least number rule, it requires some form of excluded middle premise. (I spare you the details.) Still, it is valid in either of the following rule forms:

(Simple) $A(0) \wedge \forall n(A(n) \rightarrow A(n+1)) \vdash \forall n A(n)$

(Course of values) $\forall n[\forall m(m<n \rightarrow A(m)) \rightarrow A(n)] \vdash \forall n A(n)$

And these are all we would need to derive (3) from (1) and (2).

It's also clear that the problem can't in general be in (1): for each axiom certainly implies its own truth given the truth schema, and we saw before that the possibility of using the truth predicate to finitely axiomatize a theory shows that the problem can't ultimately be due to getting from the instances of (1) to the universal generalization. As in the case of weakly classical theories, the breakdown of the Consistency Argument must be that premise (2) is unavailable: indeed, here too it must be that there are specific rules of inference that we employ but which we cannot assert to be unrestrictedly truth preserving (and which we can't even assert to preserve truth when applied to theorems, except by presupposing the consistency of our overall theory).

So it may appear superficially that the situation as regards the breakdown of the Consistency Argument in paracomplete theories is just like the situation in weakly classical theories. But I will now argue that there is a fairly substantial difference.

10. Truth-preservation for paracomplete theories. In paracomplete theories as in weakly

classical theories, we cannot assert, for each of the rules of inference we employ, that that rule is (unrestrictedly) truth-preserving: illustrations of this phenomenon will follow. But there is an important respect in which the situation for paracomplete theories is different from, and more palatable than, the situation with weakly classical theories. (This is in addition to the advantage noted at the start of Section 8, that by retaining naive truth theory the paracomplete theories can fully capture the generalizing role of 'true'.)

To simplify the notation, let's confine our attention to rules with a single premise.[13] Such a rule says that if x and y are any sentences that stand in a certain syntactic relation H (e.g., the relation of x being a conjunction whose second conjunct is y) then the inference from x to y is valid in the sense mentioned in Section 6: it is legitimate to infer from premise x to conclusion y. The issue for unrestricted truth-preservation is whether the theory implies that for any x and y that stand in that relation, if x is true then so is y. I take it that any reasonable theory here will accept that generalization if it accepts all its instances. (If not, the weakness could be remedied by adding the truth-preservation claim to the theory; as long as the original theory was ω-consistent, the strengthening would be consistent.) So the issue is whether whenever A and B are specific sentences standing in the syntactic relation H, the theory accepts True(<A>) → True(<B>). I'll call this "instance-wise unrestricted truth-preservation".

In the case of some rules, when B follows from A by that rule then the corresponding conditional A→B will be accepted. In that case, there is no question that a paracomplete theory of the sort considered in the previous section will yield instance-wise unrestricted truth-preservation: such theories satisfy the Intersubstitutivity Principle, so we can immediately infer from A→B to True(<A>) → True(<B>). One might wonder whether weakly classical theories might have a problem even with some rules of this form, since they don't satisfy Intersubstitutivity; but this is easily seen not to be a worry for those that assert that modus ponens preserves truth, and the best

weakly classical theories do assert that.

The interesting issue concerns rules for which we don't accept the corresponding conditional. We know that there are such rules, both in weakly classical theories and in paracomplete theories: that is what the failure of ¬→-Introduction in those theories amounts to. But we shouldn't expect the theory to declare *such* rules unrestrictedly truth-preserving, even instance-wise: to say True(<A>) → True(<B>) ought to be equivalent to saying A→B, which we don't have. In short: *the inability to assert unrestricted truth-preservation is what we would expect*, whenever we accept a rule without accepting the corresponding conditional.

As an example, consider the explosion rule: C ∧ ¬C ⊢ 0=1 (for arbitrary C). Paracomplete theories of the kind considered in the previous section contain this rule, but they don't contain the corresponding conditional C ∧ ¬C → 0=1 except for those C for which excluded middle can be assumed. (The reason is that C ∧ ¬C → 0=1 is equivalent to ¬(0=1) → ¬(C ∧ ¬C), which by modus ponens implies ¬(C ∧ ¬C), which is equivalent to the instance of excluded middle ¬C ∨ C.)[14] Since we don't have C ∧ ¬C → 0=1, *we shouldn't expect* True(<C ∧ ¬C>) → True(<0=1>).

Intuitively, the situation is that C ∧ ¬C ⊢ 0=1 is just a rule of conditional assertion whose antecedent is never fulfillable: you are never in a position to assert C ∧ ¬C, though for some C you also aren't in a position to assert its negation. For those C, you can't assert True(<C ∧ ¬C>) → True(<0=1>), since this would be tantamount to asserting ¬True(<C ∧ ¬C>), i.e. ¬(C ∧ ¬C), i.e. ¬C ∨ C.

As another example, consider modus ponens, the rule A ∧ (A→B) ⊢ B. This will be a rule in the sort of paracomplete theories I've considered, but the corresponding conditional [A ∧ (A→B)] → B won't be valid. (Reason: let B be an absurdity like 0=1, and let A be the corresponding Curry sentence K. Then A ∧ (A→B) is equivalent to K. So the rule A ∧ (A→B) ⊢ B

gives K ⊢ 0=1, which says truly that K leads to inconsistency; but the corresponding conditional [A ∧ (A→B)] → B gives K→0=1, which is equivalent to K, and thus is objectionable since it leads to inconsistency.) Since we don't have [A ∧ (A→B)] → B, *we shouldn't expect* True(<A>) ∧ True(<A→B>) → True(<B>).

Again, modus ponens is just a rule of conditional assertion: when you're in a position to assert both A and A→B (or equivalently, when you're in a position to assert their conjunction), then you're in a position to assert B; but this says nothing about what's the case when you aren't in a position to assert A and A→B, as when these are, say, K and K→0=1. In the vocabulary of Section 7, we have no reason to expect the rule to be unrestrictedly truth-preserving, but only to be truth-preserving *as applied to assumptions we are in a position to assert*. But even this restricted truth-preservation claim isn't one we can expect to be *demonstrable in our theory*: demonstrating it would require a prior proof that K and K → 0=1 aren't both theorems of our theory, and hence would require a prior proof of the consistency of our theory. We thus have a clear explanation of why the Consistency Argument breaks down.

The upshot is two-fold:

(I) Without assuming the consistency of our theory T, we could have no grounds for assuming that rules of T like explosion and modus ponens are even restrictedly truth-preserving;

and

(II) Even assuming the consistency of T, we have no grounds for assuming that explosion and modus ponens are *unrestrictedly* truth preserving: no grounds, for instance, for assuming for each C that True(<C ∧ ¬C>) → True(<0=1>), or for assuming for each A and B that True(<A>) ∧ True(<A→B>) → True(<B>).

While this situation is somewhat analogous to the situation for weakly classical theories

(there too the failure to assert truth-preservation arose for rules that we accept without accepting all the corresponding conditionals), there is a big difference.

For recall that in weakly classical theories, there are cases where we accept a rule but actually *deny* that it is unrestrictedly truth-preserving (understanding denial as assertion of the negation). Indeed, there are cases where we accept $A \vdash B$ whenever $\langle A,B\rangle$ stand in syntactic relation H, but nonetheless accept

$$\neg[\mathrm{True}(\langle A_1\rangle) \rightarrow \mathrm{True}(\langle B_1\rangle)] \vee \neg[\mathrm{True}(\langle A_2\rangle) \rightarrow \mathrm{True}(\langle B_2\rangle)],$$

or equivalently,

$$[\mathrm{True}(\langle A_1\rangle) \wedge \neg\mathrm{True}(\langle B_1\rangle)] \vee [\mathrm{True}(\langle A_2\rangle) \wedge \neg\mathrm{True}(\langle B_2\rangle)],$$

for two particular pairs $\langle A_1,B_1\rangle$ and $\langle A_2,B_2\rangle$ that stand in that syntactic relation. For instance, in revision-theoretic and McGee-type supervaluational theories, which are arguably the best versions of weakly classical theories, we accept the rule

$$\mathrm{True}(\langle A\rangle) \vdash A$$

but accept also its failure to unrestrictedly preserve truth in the following sharp form:

(*) $$[\mathrm{True}\langle \mathrm{True}\langle Q\rangle\rangle \wedge \neg\mathrm{True}\langle Q\rangle] \vee [\mathrm{True}\langle \mathrm{True}\langle \neg Q\rangle\rangle \wedge \neg\mathrm{True}\langle \neg Q\rangle].$$

It was this feature of the weakly classical theorist's explanation of the failure of the Consistency Argument that I declared somewhat disquieting; and nothing like this happens in the paracomplete case.

Obviously it doesn't happen in the paracomplete case *for the particular rule (T-Elim)*: indeed, in the paracomplete theories $\mathrm{True}(\langle B\rangle)$ is always equivalent to B (i.e. intersubstitutable with B), and hence $\mathrm{True}(\langle \mathrm{True}(\langle A\rangle)\rangle)$ is always equivalent to $\mathrm{True}(\langle A\rangle)$. But my point is more general: in the paracomplete case one never denies that one's rules are unrestrictedly truth-preserving; rather, one *neither asserts nor denies* that they are unrestrictedly truth-preserving.

Moreover, the inability to assert or deny of a paracomplete theory that certain of its rules preserve truth is not diagnosed as due to "ignorance" in any ordinary sense, or to an "incompleteness" of any ordinary sort in the theory. For whenever we recognize a rule but not all instances of the corresponding conditional, this is due to a belief that it would be inappropriate to assume excluded middle for that conditional (and also, inappropriate to assume excluded middle for its antecedent).

Consider our examples. (1) We accept $C \wedge \neg C \vdash 0=1$ (for arbitrary C), but don't accept $True(\langle Q \wedge \neg Q \rangle) \rightarrow True(\langle 0=1 \rangle)$ since we don't accept $Q \wedge \neg Q \rightarrow 0=1$. We don't accept the negations of these claims either. But that isn't because our theory is "incomplete" in any normal sense, or because we are "ignorant" as to whether $True(\langle Q \wedge \neg Q \rangle) \rightarrow True(\langle 0=1 \rangle)$. Rather, it is because we don't accept

$$[True(\langle Q \wedge \neg Q \rangle) \rightarrow True(\langle 0=1 \rangle)] \vee \neg[True(\langle Q \wedge \neg Q \rangle) \rightarrow True(\langle 0=1 \rangle)].$$

Indeed, it is easy to argue that we couldn't coherently accept this instance of excluded middle: accepting it is easily seen to breed contradictions.[15] The claim that this rule preserves truth thus has the same status as the Liar sentence itself has: intuitively speaking, it is "pathological" or "indeterminate".[16]

(2) The situation with modus ponens is similar: the failure to accept

(**) $$True(\langle K \rangle) \wedge True(\langle K \rightarrow 0=1 \rangle) \rightarrow True(\langle 0=1 \rangle)$$

isn't do either to acceptance of its negation or to ignorance as to which of it and its negation is true; indeed, the disjunction of (**) and its negation leads to contradiction. Intuitively, (**) isn't false but indeterminate, on this theory.

And not only is (**) intuitively indeterminate on this theory, so is the generalization that modus ponens unrestrictedly preserves truth. I'm inclined to think that this explanation of where

the Consistency Argument breaks down is quite intuitive. A weakly classical theory, on the other hand, regards the claim that its rules unrestrictedly preserve truth as *false*, and indeed we get this result in the strong form (*). This strikes me as far less intuitive.

I must admit, though, that this judgement against the weakly classical theorist's diagnosis of the failure of the Consistency Argument depends to some extent on a prior distaste for theories that accept disjunctions while simultaneously regarding each of the disjuncts as absurd. Perhaps those who do not share this distaste will find no advantage in the paracomplete theorists' diagnosis of where the Consistency Argument goes wrong over their own. (I repeat, though, that there is still one clear advantage of paracomplete theories over weakly classical theories: as argued at the start of Section 8, the weakly classical theories must seriously cripple the generalizing role of truth, whereas the paracomplete theories can keep this generalizing role intact.)

11. Standard dialetheic theories. I turn finally to dialetheic theories that keep the Intersubstitutivity Principle and/or the truth schema. Examples of the theories I have in mind include Priest 1987, 1998, and 2002; and Beall 2005. These might be thought to come out better than paracomplete and weakly classical theories in their diagnosis of the failure of the Consistency Argument; but I'll argue that this is not so, and that indeed the extant versions of dialetheism do worse.

Recall that dialetheic theories are ones according to which some sentences are both true and false, where 'false' means 'has a true negation'. In the presence of the Intersubstitutivity Principle (or the truth schema and modus ponens, or even just the rule (T-Elim)), this requires the acceptance of classical contradictions, and that is the kind of dialetheism that will be under consideration in this section and the next. (Recall that the dialetheic views of Section 5 avoided classical contradiction by abandoning (T-Elim).) So if we take classical logic to prohibit classical

contradictions, the dialetheic views now under consideration (which accept (T-Elim)) violate classical logic. Some might object to taking classical logic to include such a prohibition rule: they might regard it as including only rules of inference, in which case it can't prohibit anything. Even so, it would obviously be idiotic to accept everything; but contradictions classically imply everything, so any non-idiotic dialetheist that accepts (T-Elim) will have to reject the classical inference A, ¬A ⊢ B according to which contradictions explode. I'll use the term 'standard dialetheism' for dialetheic views that reject explosion and accept at least (T-Elim). From here on out, these will be the only kind of dialetheic views under discussion, so I won't always bother to explicitly say 'standard'.

Priest has claimed (1987 Ch. 3) that non-trivial dialetheic theories can prove their own soundness, i.e. prove claim (3) of the Consistency Argument:

(3) All theorems of T are true.

Some subsequent literature (e.g. Shapiro 2002) takes Priest at his word on this point. I'll be arguing that this is incorrect. But if it were correct, could the dialetheist block the Consistency Argument by blocking the inference from (3) to the consistency of T? In Section 1 I gave two forms of that inference. The second relied on the explosion principle, which the (standard) dialetheist doesn't have. The first relied on the assumption that no sentence and its negation can both be true, which the dialetheist denies. It seems, then, that dialetheist has a simple diagnosis of where the Consistency Argument fails: in the inference from (3) to the consistency of T.

But there is a problem with this easy response to the Consistency Argument. The inference from (3) to the consistency of T (in its first form) does indeed depend on the claim that no sentence and its negation are both true. And the dialetheist does deny this *in the sense of accepting its negation*. But accepting its negation doesn't preclude accepting the claim, on a dialetheist view; can we be sure that the dialetheist doesn't accept that no sentence is both true and

not true, as well as denying it?

I will not explore this in full detail, but there is strong prima facie reason to worry. Consider any specific claim p that the dialetheist regards as both true and false: he accepts

(X) $\mathrm{True}(\langle p\rangle) \wedge \mathrm{True}(\langle \neg p\rangle)$.

But the second conjuct implies ¬p, by the Intersubstitutivity Principle of the previous section; and by the Intersubstitutivity Principle again, this implies $\neg\mathrm{True}(\langle p\rangle)$, from which it follows that

(X*) $\neg[\mathrm{True}(\langle p\rangle) \wedge \mathrm{True}(\langle \neg p\rangle)]$.

In other words, X implies its own negation, given the principles of truth we are trying to preserve! If we can assume of each sentence that *it is either a dialetheia or not a dialetheia* (where a dialetheia is a sentence that is both true and false), then we have a proof of (X*) for any sentence p whatever. In particular, if we assume the law of excluded middle—as many dialetheists, including Priest, do—then we have a general proof of (X*). This isn't quite as strong as we need to get from (3) to consistency: for that, we need the universal generalization of (X*), i.e. that no sentence is both true and has a true negation. But it is hard to believe that many dialetheists would want the kind of ω-incompleteness that arises from accepting each instance of that generalization but not the generalization. And if the dialetheist does accept the generalization, then the inference from (3) to the consistency of T goes through; in which case the inductive argument for (3) must fail.

There are several ways in which one might block this argument.

(A) (The most important.) The argument depends on excluded middle (at least as applied to the predicate 'is a dialetheia'), and while many dialetheists (e.g. Priest) accept excluded middle, not all do: e.g. Beall 2005 doesn't. It's worth remarking, though, that once one has given up excluded middle, the *prima facie* advantages of going dialetheic are dramatically lessened: we've already seen that there are *non*-dialetheic theories without excluded middle that are compatible with the Intersubstitutivity Principle and the truth schema.

(B) The argument depends on the full Intersubstitutivity Principle, and in Chapter 5 of Priest 1987 he argues that Intersubstitutivity in the scope of a negation sometimes fails, which would be enough to block the argument. Priest accepts the truth schema, but in a logic for the conditional that does not accept the contraposition rule ($A \rightarrow B \vdash \neg B \rightarrow \neg A$); and without that rule we cannot derive Intersubstitutivity from the truth schema. (In some ways Priest's theory violates Intersubstitutivity more severely than most weakly classical theories do: most of those theories accept the inference from ¬p to ¬True(<p>), whereas Priest 1987 doesn't.)[17] To my mind, giving up on Intersubstitutivity destroys the main motivation for dialetheism: I think the truth schema is of little interest without the Intersubstitutivity Principle, for it is too weak to yield the generalizing role of 'true'. I will not argue the point (but I have at least one dialetheist, Beall 2005, on my side).

(C) Priest occasionally contemplates the idea that even in the arithmetic language some contradictions might be true; in that case, the denial of the incompleteness theorem might be true as well as the incompleteness theorem: our overall theory might prove its own consistency as well as not prove it, and then there would be no reason to block the Consistency Argument. Well, call me conservative and old-fashioned, but this is not a possibility I'm willing to take seriously.[18]

I've argued that a dialetheist can block the argument from (3) to consistency only by throwing away some of the *prima facie* advantages of the dialetheist position (e.g. throwing away Intersubstitutivity or excluded middle). But there is no need to block the argument from (3) to consistency: as I'll soon show, the inductive argument for (3) is blocked in any case.

Note that I'm not merely saying that a dialetheist is in a position to declare some of the theorems of his system false; that is completely obvious, for the essence of the view is that some are both true and false. The claim is that in addition, he is not in a position to declare them all true:

the inductive argument that one might expect breaks down.

Just where the inductive argument breaks down depends on the particulars of the dialetheic logic. For the simplest and best known dialetheic logic, Priest's LP (Priest 1998), it turns out that the problem is a breakdown of the induction rule, traceable to the fact that modus ponens isn't valid in the logic. But this is not a form of dialetheism that Priest seriously advocates—or that anyone else does either, as far as I know. From now on let's restrict to forms of dialetheism that keep modus ponens and induction.

A dialetheist of this sort who accepts Intersubstitutivity will clearly accept the instances of (1), and for reasons discussed above cannot in general diagnose the problem with the Consistency Argument as in the passage from the instances to the generalization. It seems, then, that the problem must lie, as for the paracomplete theorist and the weakly classical theorist, in (the instances of) (2). But it's worth looking at the details: as we'll see, for many versions of dialetheism the situation is in several respects more counterintuitive than it is even for the weakly classical theorist, let alone the paracomplete theorist.

12. Truth-Preservation for Standard Dialetheic Theories. First I return to the Curry Paradox. Though dialetheic rhetoric sometimes suggests that all the prima facie paradoxes should be treated as involving dialetheia (sentences that are both true and false), this is not so: no dialetheic view on which modus ponens and (T-Elim) hold can treat the Curry sentence as both true and false. For if K is both true and false, it is true. And this (True(<K>)) is enough to breed paradox even without assuming full intersubstitutivity: by (T-Elim) we get K, which is equivalent to True(<K>)→0=1, and we already have True(<K>, so modus ponens yields 0=1. (And we get any other conclusion, e.g. that the Earth is flat, by an analogous paradox involving a modified Curry sentence.) A dialetheist who accepts excluded middle (and reasoning by cases) must accept ¬K without

accepting K: K must be regarded as "solely false", i.e. false and not true. (A dialetheist who is not committed to excluded middle can follow the paracomplete theories in rejecting ¬K as well as rejecting K.)

This is relevant to the breakdown of the Consistency Argument in standard dialetheic theories that accept modus ponens. One place the breakdown must arise is that the theory cannot accept that modus ponens unrestrictedly preserves truth.[19] For an instance of that would be:

$$\text{True}(\langle K\rangle) \wedge \text{True}(\langle K \rightarrow 0=1\rangle) \rightarrow \text{True}(\langle 0=1\rangle),$$

which by Intersubstitutivity is equivalent to

$$K \wedge (K \rightarrow 0=1) \rightarrow 0=1.$$

(This step indeed requires only Intersubstitutivity in negation-free contexts, which Priest accepts.) But since the second conjunct of the antecedent is equivalent to K, this is just equivalent to K→0=1, and hence to K; and from K and K→0=1 we get 0=1 by modus ponens. So even though we accept modus ponens, we can't accept that it unrestrictedly preserves truth, on this sort of dialetheic theory.[20] (Priest 1987 appears to make a mistake about this (p. 63): he reads the truth-preservingness of modus ponens as requiring only the rule $\text{True}(\langle A \rightarrow B\rangle) \vdash \text{True}(\langle A\rangle) \rightarrow \text{True}(\langle B\rangle)$, but that is obviously not enough to formulate the inductive step in an inductive argument for the generalization (3).)

In the case of dialetheic theories that accept excluded middle (or that accept the negation of K for any other reason), the situation is more dramatic: in addition to not accepting that modus ponens preserves truth, they accept that it doesn't. Indeed they accept:

$$\neg[\text{True}(\langle K\rangle) \wedge \text{True}(\langle K \rightarrow 0=1\rangle) \rightarrow \text{True}(\langle 0=1\rangle)].$$[21]

The general claim of truth preservation thus has a specific instance that is "solely false". In one way this is more counterintuitive than the situation even for weakly classical theories (let alone for paracomplete theories): weakly classical theories don't accept any *specific* counterexample to the

unrestricted truth-preservingness of its rules, but only a disjunction of counterexamples. Perhaps this is one reason to prefer dialetheic theories that reject excluded middle as well as explosion: these cannot accept that modus ponens preserves truth in this example, but at least they don't deny that it does (i.e. accept the negation).

But there is a second and more important point of comparison between dialetheic theories on the one hand and weakly classical and paracomplete theories on the other. We saw that a paracomplete or weakly classical theorist, though unable to view her rules as unrestrictedly truth preserving, could at least profess faith that they preserve truth *when applied to theorems* (though this faith would have to rest on a faith in the consistency of her overall theory). Can something similar be said of the dialetheist? (Obviously the dialetheist's faith couldn't rest on faith in the *consistency* of his overall theory, since he has no such faith; but it might rest on faith in its non-triviality (i.e. its not implying everything), or its arithmetic consistency, i.e. its not leading to contradictions within arithmetic.) The example above does not undermine the claim that modus ponens preserves truth in this restricted sense, for the dialetheist does not accept K or K→0=1 (barring an unexpected triviality in his theory). But are there other examples that show that even restricted truth-preservation can't be maintained?

Here there's good news, bad news, worse news, and worser news.

The good news is that for any dialetheic theory that accepts Intersubstitutivity[22] and validates the inference from A∧B to A→B, there can be no such examples (examples undermining restricted truth-preservation). Reason: if the dialetheist accepts A, and A implies B, then he is committed to A∧B and hence, by that inference rule, to A→B; and Intersubstitutivity then would give True(<A>)→True(<B>).

The bad news is that this is irrelevant to the usual dialetheic logics: none of the dialetheic

conditionals of which I'm aware validates the inference from A∧B to A→B. (Most of them give the conditional a possible world semantics that blocks the inference.) So the argument for there being no problem is unavailable to most dialetheic theorists.

The worse news is that on these theories, the inference from A∧B to A→B fails in a way that itself prevents recognition of restricted truth-preservation of all the inferences employed in the theory. Let B be any axiom that has a conditional as its main connective, say C→C. Since B is an axiom, the inference from A to B is valid for any A, including any obvious truth not involving a conditional: say, ¬(0=1). But ¬(0=1)→B won't be true according to these theories (and in most of them, it will be false): the reason is that in the modal semantics used for the conditional in such theories, axioms like B will fail at certain "non-normal worlds", but ¬(0=1) is true at all worlds, and as a result A→B fails to be true. So by Intersubstitutivity, True(<A>)→True(<B>) must fail; and yet the inference from A to B is not only a valid inference, but one whose premise is assertable!

This leaves open the possibility of developing a dialetheic theory with a different kind of conditional not open to this problem. But the "worser" news is that there are serious limitations on the possibilities: for instance, the theory couldn't contain both excluded middle and a contraposition rule for the conditional (in addition to the usual double negation rules, which are common to all standard dialetheic theories, and Intersubstitutivity even limited to negation-free contexts). Reason: if it contains excluded middle, ¬K is a theorem, and hence follows from ¬(0=1). Since ¬(0=1) is a theorem, even *restricted* truth preservation would require True(<¬(0=1)>)→True(<¬K>), which by Intersubstitutivity is equivalent to ¬(0=1)→¬K. But contraposition (and double negation rules) yields K→0=1, which is equivalent to K and which we know we can't have. So we can't have even *restricted* truth-preservation in any such theory.

Indeed, using excluded middle again, the theory must regard the claim True(<¬(0=1)>)→True(<¬K>) as "solely false", since it is equivalent to K. So the theory says that we have a valid inference, from a premise which is true (indeed, solely true); and yet the claim that it preserves truth is solely false. This borders on the counterintuitive.

In partial summary: if we want *restricted* truth-preservation, as presumably we should, then we can't have both a contraposition rule and excluded middle; and even then, the kinds of possible worlds conditionals common to dialetheic logics must be ruled out. Indeed, I know of no dialetheic logic that employs a conditional that allows for restricted truth-preservation. I have no reason to doubt that one might be developed, though again, it couldn't obey both a contraposition rule and excluded middle. (We also saw earlier in this section that in dialetheic theories without excluded middle, the claim that modus ponens unrestrictedly preserves truth comes out solely false; if one prefers to avoid that, that is a reason to think that excluded middle rather than contraposition is the culprit as regards restricted truth-preservation.)

13. Validity and truth preservation. I think that the theories that have come out best with regard to the Consistency Argument are the paracomplete ones; but the weakly classical theories, and perhaps certain possible dialetheic theories even if no extant ones, aren't too far behind in their treatment of this Argument. (Weakly classical theories have independent problems, though: notably, their failure to accord with the generalizing role of 'true'.) All these theories diagnose the failure of the Consistency Argument as arising from our employing a theory that includes rules that the theory does not take to be *unrestrictedly* truth-preserving.

This raises an issue about the meaning of 'valid'. Many people think it true by definition that an inference is valid if and only if it is logically necessary that it preserves truth unrestrictedly. Call this validity$_T$. Others think it true by definition that the valid rules are the ones we should

reason with. Call this validity$_N$. Of course it makes little difference how you define 'valid': the substantive claim, which is rather widely accepted, is that validity$_T$ and validity$_N$ coincide: the rules one should reason with in logic are those that, of logical necessity, unrestrictedly preserve truth.

I doubt that either definition above coincides with the usual concept of validity. (Validity$_N$ strikes me as coming closer to being extensionally correct; but the idea that validity should be *defined* in terms of what we ought to believe strikes me as repugnant.) And I think that the question of how 'valid' is to be defined is unhelpful: it is best to view it not as a defined term, but as a primitive notion that governs our epistemic practices.[23] (Part of how it governs them is this: when we discover that the inference from *p* and *q* to *r* is valid, then we should ensure that our degree of belief in *r* is no lower than our degree of belief in the conjunction of *p* and *q*.) From this, we can construct an *argument* that validity coincides with necessary truth preservation:

(1) Assuming standard truth rules and the usual introduction and elimination rules for conjunction, the validity of the inference from $p_1,...,p_n$ to *q* is equivalent to the validity of the inference from *True(<*p_1*>) and ... and True(<*p_n*>)* to *True(<q>)*.

(2) And now assuming the usual introduction and elimination rules for the conditional, this is equivalent to the validity of the conditional *If True(<*p_1*>) and ... and True(<*p_n*>) then True(<q>)*.

Validity of a single claim is presumably equivalent to the logical necessity of that claim, so the natural-deduction rules of classical logic plus the truth predicate *show*, for any particular sentences $p_1,...,p_n$ and *q*, that if the inference is valid then of logical necessity it preserves truth, and conversely.[24]

But note: the derivation of the claim that valid inferences necessarily preserve truth is inapplicable to any logic that doesn't contain the introduction rule for the conditional, and the derivation of the converse is inapplicable to any logic that doesn't contain the elimination rule for

the conditional (modus ponens).[25] Nor should logics without these rules accept *the result* of the alleged derivation: any example where one would, e.g., accept that $\Gamma, A \models B$ without accepting $\Gamma \models A \neg B$ would be a case of accepting $\Gamma, A \models B$ without believing it to be necessarily truth preserving. So on this picture, advocates of logics that don't accept the classical rules for the conditional thus should not accept that validity coincides with validity_T. (Validity on this picture is obviously rather in the spirit of validity_N, even though the connection to what we ought to believe isn't taken as definitional; so if you prefer you could put the matter by saying that in logics like these, we should not accept the extensional equivalence of validity_N with validity_T.)

It looks at first blush, then, that the principle that the valid (or valid_N) inferences are those that necessarily preserve truth is a principle that we can argue for in strongly classical logic, or in non-classical logics like intuitionism that keep the classical rules for the conditional (and for conjunction), but for which we should give up for paracomplete, dialetheic, and weakly classical logics.

But this understates the difficulty with the identification of validity (or if you like, validity_N) with validity_T. For the derivation above relied not only on the standard introduction and elimination rules for the conditional (and for conjunction); it also relied on the principle that True(<p>) is intersubstitutable with p, or at least, on the rules (T-Introd) and (T-Elim). But those rules cannot be consistently maintained in strongly classical theories, or in intuitionist theories! Indeed, Curry's Paradox (Section 9) shows that any logic that accepts the standard introduction and elimination rules for the conditional and the introduction and elimination rules for truth is completely trivial: it implies anything whatever. Thus however compelling the argument that validity coincides with necessary truth-preservation may have seemed, it relies on assumptions that cannot be jointly accepted.

Admittedly, the equation of validity with validity$_T$ is fine for the strongly classical and intuitionistic logic of the sentential connectives and quantifiers; indeed, the above proof that validity coincides with validity$_T$ in this limited domain is perfectly acceptable. It is only in the case of rules for truth that the equation breaks down for these logics. But if we include rules of truth within our logic, we can't coherently maintain all of the principles on which the above proof rests, and so the motivation for equating validity with unrestricted truth-preservation breaks down.

Of course, nothing can stop one from simply defining 'valid' to mean 'valid$_T$'. We have seen, though, that if one does that there is no remotely acceptable way to avoid employing logical rules that one takes not to be "valid". In the case of classical (and intuitionist) theories, such rules will involve the truth predicate; examples will typically include the rules of (T-Elim) and (T-Introd) discussed earlier. In the case of paracomplete theories the only specific truth rule that's needed is the Tarski schema (T) (from which Intersubstitutivity follows in the logic), and the theory regards it as valid$_T$; but the cost is that certain rules of the underlying sentential logic (e.g. explosion and modus ponens) won't be regarded as valid$_T$. But this is certainly no more counterintuitive for these theories than the corresponding situation for classical theories: in both cases, the theories must employ rules that they don't regard as *unrestrictedly* truth-preserving. (And in fact it is somewhat less counterintuitive for paracomplete theories, since they don't declare their rules *not* to be truth-preserving.)

14. Must our mathematics be indefinitely extensible? The discussion of this paper has relevance to the popular idea that there could be no such thing as a person's (or community's) "overall mathematical theory" that is consistent and recursively axiomatized. According to this idea, it is a central feature of mathematics that it be "indefinitely extensible": in accepting a theory T we implicitly commit ourselves to the claim that it is consistent, which goes beyond T.

I will not attempt a serious discussion of this idea here. (Doing so would, among other things, require a fuller discussion of the Tarskian hierarchical approach to truth, within which the indefinite extensibility view is usually framed.) I will simply say that a common defense of the idea that when we accept a theory we implicitly commit ourselves to the claim that it is consistent is that we must believe the axioms to be true and the rules to preserve truth. I take it that the present paper undermines that defense: we shouldn't even believe that our rules unrestrictedly preserve truth, and the claim that they preserve truth as applied to theorems can have no justification independent of a belief that our theory is consistent.

By undermining this defense, we at least open ourselves to the possibility that there is indeed such a thing as "our overall mathematical theory". I hope it is clear that the advocate of the existence of such an overall theory need not and should not claim that there is no possible advancement in mathematics other than the deduction of new consequences from this theory: there is no bar to the introduction of conceptual innovations that would make it rational to advocate a more powerful theory, and doubtless such conceptual innovations have often been made. What the idea of an overall theory does require is that such non-deductive expansion of the theory would require conceptual innovation: it wouldn't result automatically by reflecting on the truth of the axioms and the character of the rules of the current theory.[26]

Notes

1. Such a formalization can be found in Quine 1940.

2. The gross impropriety about use and mention here could easily be remedied, but at cost of readability.

3. And of course a theory might restrict both the truth schema and the logic. (One well-known theory does this: the Kripkean theory KFS, which will be mentioned in Section 9.)

4. At least, this is so when induction is stated "in positive rule form"; see Section 9.

5. This assumes that ‘axiom of T’ is definable in T, but in theories like set theory or even arithmetic that is certainly true of all recursively axiomatized theories, and they are the only ones of interest here.

6. Strictly speaking, it *needn’t* quite do either: it could remain agnostic as to which axiom isn’t true, or agnostic as to which rule of inference fails to preserve truth, or even agnostic as to whether it’s the axioms or the rules that are in this way problematic. But the possibility of agnosticism is of little interest. (And in practice, the theories in question tend not to be agnostic over this.)

7. We have both

If True(<Q>), then Q

(by (T-OUT)) and

If ¬True(<Q>), then Q;

(by the equivalence between Q and ¬True(<Q>)). Q follows; and by the equivalence, ¬True(<Q>) does too.

8. In the non-classical logic I’ll be mainly interested in, all failures of classicality result from failures of excluded middle. We get full classicality within arithmetic by the simple device of adding all instances of excluded middle in the arithmetical language as (perhaps "nonlogical") axioms.

9. This diagnosis of where the Liar reasoning breaks down is also the one that would be given by the non-classical Kripkean theory KFS based on Kleene logic. But as I’ve remarked, that theory doesn’t have the truth schema (T) (though it has Intersubstitutivity); indeed, the truth schema is inconsistent in KFS.

10. Indeed, the argument didn’t require full Intersubstitutivity, just (T-Elim) and (T-Introd); so it provides another illustration of why even weakly classical theories can’t accept →-Introduction.

11. It seems independently natural to diagnose the Curry paradox as depending on excluded middle. For K leads to paradox, assuming only Intersubstitutivity and Modus Ponens, by Step One of the argument in the text. And ¬K implies ¬(0=1)→¬K, assuming the rule B ⊢ A→B or even the weaker rule A∧B ⊢ A→B. (And at least the weaker rule *must* hold if → reduces to ⊃ when excluded middle is assumed for antecedent and consequent.) But assuming a contraposition rule for the conditional, ¬(0=1)→¬K in turn implies K→0=1, which is equivalent to K. Since we saw above that K leads to paradox, and ¬K implies K, then ¬K leads to paradox too. Reasoning by cases, we have that K∨¬K leads to paradox.

12. Why do I rely on explosion and the restricted principle that no sentence *of arithmetic* can be both true and false, rather than using the unrestricted principle with 'of arithmetic' dropped? After all, any theory of this sort will take the claim that some sentences are both true and false to imply a contradiction. The answer is that in absence of excluded middle (and of intuitionist principles governing negation, which themselves lead to paradox), one can't infer from this that no sentence is both true and false: A can be contradictory without ¬A being a theorem.

13. This is really no limitation: any other (finitary) rule can be reduced to a single-premise one if we assume the single multi-premise rule of ∧-Introduction together with ∧-Elimination, and there is no problem about the truth-preservingness of these ∧-rules on any of the theories under consideration.

14. Incidentally, intuitionist logic does accept C ∧ ¬C → 0=1 despite not accepting excluded middle; it resists the above argument because it does not accept the deMorgan law that takes you from ¬(C ∧ ¬C) to ¬C ∨ ¬¬C. (Intuitionism rejects not only full excluded middle, but also the special case ¬C ∨ ¬¬C; so though it also doesn't accept the reduction of the latter to ¬C ∨ C, it is the failure of the deMorgan step that is crucial.) But intuitionist logic is of no particular relevance in the current context, since it is no better than classical logic in dealing with the paradoxes.

15. If it's incoherent to recognize a fork in the road, we can hardly be criticized for not

recognizing the fork, let alone for not taking it!

16. The question of how exactly to make sense of attributions of pathologicality or indeterminacy without breeding further paradoxes is complicated. For a discussion see Field forthcoming.

17. Priest 2002, on the other hand, suggests a truth theory in a logic with a rule form of contraposition; in it, the derivation of (X*) is acceptable. But he informs me that he actually prefers a modified version of this logic in which, as in Priest 1987, contraposition fails.

18. Shapiro 2002 shares this reaction. He thinks it cuts more centrally against dialetheism than I do, because he takes for granted that dialetheic theories can prove (3).

19. This point is also made in Section 6 of Beall forthcoming.

20. I'm taking the truth-preservation claim to be that for any x, y, and z such that y is an $\neg\!\rightarrow$-statement with x as antecedent and z as consequent,

$$\text{True}(x) \wedge \text{True}(y) \rightarrow \text{True}(z).$$

JC Beall suggested to me that I also consider truth-preservation in the material conditional form

$$\text{True}(x) \wedge \text{True}(y) \supset \text{True}(z)$$

(where y is still an $\neg\!\rightarrow$-statement rather than a $\supset$-statement); he noted that modus ponens does preserve truth in *this* sense, in Priest's theory. But he also correctly observed that this doesn't affect my main point, given that in Priest's theory, modus ponens fails for $\supset$ and the induction rule fails when the induction step is formulated via $\supset$. In particular, truth-preservation in this sense wouldn't validate the inference to (3) in the Consistency Argument.

21. Though they do not regard this as implying

$$\text{True}(\langle K \rangle) \wedge \text{True}(\langle K \neg\!\rightarrow 0=1 \rangle) \wedge \neg\text{True}(\langle 0=1 \rangle),$$

which they don't accept. See previous footnote for why truth-preservation is explained in terms of $\neg\!\rightarrow$ rather than $\supset$.

22. Or even Intersubstitutivity in negation-free contexts.

23. This viewpoint was advocated in Kreisel 1967; see also Field 1991.

24. Admittedly, there is an issue of how to universally generalize this. One approach would be to use the idea of Field 2006, but there are others.

25. The given derivation is also inapplicable to logics without conjunction-introduction, and the converse to logics without conjunction-elimination; I omit mention of these rules in the text since I find logics without them to be of little interest.

26. Thanks to JC Beall for comments on an earlier draft.

References

Beall, JC 2005. "Transparent Disquotationalism". In Beall and Armour-Garb, eds., *Deflationism and Paradox*, (OUP, Oxford), pp. 7-22.

Beall, JC forthcoming. "Truth and Paradox". In Dale Jacquette, ed., *Philosophy of Logic, volume II* (North-Holland, Amsterdam).

Field, Hartry 1991. "Metalogic and Modality", *Philosophical Studies* 62: 1-22.

Field, Hartry 2003a. "A Revenge-Immune Solution to the Semantic Paradoxes", *Journal of Philosophical Logic* 32: 139-177.

Field, Hartry 2003b. "The Semantic Paradoxes and the Paradoxes of Vagueness". In JC Beall, ed., *Liars and Heaps* (OUP, Oxford), pp. 262-311.

Field, Hartry 2006. "Compositional Principles versus Schematic Reasoning", *The Monist*.

Field, Hartry forthcoming. "Solving the Paradoxes, Escaping Revenge",

Friedman, Harvey and Michael Sheard 1987. "An Axiomatic Approach to Self-Referential Truth", *Annals of Pure and Applied Logic* 33: 1-21.

Gupta, Anil 1982. "Truth and Paradox", *Journal of Philosophical Logic* 11: 1-60.

Herzberger, Hans 1982. "Notes on Naive Semantics", *Journal of Philosophical Logic* 11: 61-102.

Kreisel, Georg 1967. "Informal Rigour and Completeness Proofs". In I. Lakatos, *Problems in the Philosophy of Mathematics* (North-Holland, Amsterdam).

Kripke, Saul 1975. "Outline of a Theory of Truth", *Journal of Philosophy* 72: 690-716.

McGee, Vann 1991. *Truth, Vagueness, and Paradox* (Hackett, Indianapolis).

Priest, Graham 1987. *In Contradiction* (Martinus Nijhoff, Dordrecht).

Priest, Graham 1998. "What is So Bad About Contradictions?", *Journal of Philosophy* 95: 410-426.

Priest, Graham 2002. 'Paraconsistent Logic'. In D.Gabbay and R.Guenthner, eds., *Handbook of Philosophical Logic*, 2nd ed., vol. 6 (Reidel, Dordrecht), pp. 287-393.

Quine, Willard 1940. *Mathematical Logic* (Harvard University, Cambridge).

Reinhardt, William 1986. "Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth", *Journal of Philosophical Logic* 15: 219-251.

Shapiro, Stewart 2002. "Incompleteness and Inconsistency", *Mind* 111: 817-832.

Soames, Scott 1999. *Understanding Truth* (OUP, Oxford).