Notre Dame Journal of Formal Logic
Volume 32, Number 2, Spring 1991

Intensional Paradoxes

GRAHAM PRIEST*

Abstract The topic of this paper is that class of paradoxes of self-reference whose members involve intensional notions such as *knowing that*, *saying that*, etc. The paper discusses a number of solutions that have been proposed by, e.g., Prior and several AI workers, and argues that they are inadequate. It argues, instead, for a dialetheic/paraconsistent resolution. A formal theory of propositions is given; this is based on arithmetic, and treats propositions as sentences. In the theory the paradoxes are accommodated in a satisfactory manner. An Appendix establishes that the contradictions in the theory do not spread to the underlying arithmetic machinery.

Introduction The paradoxes of self-reference are known and loved (or hated) by logicians. Attempts to solve them have provided the cornerstone of logic this century. The set-theoretic paradoxes were integral to the development of modern set theory; the semantic paradoxes to formal semantics. There is, however, a third clearly distinguishable group (though all three of these groups tend to merge into the others at the edges): the intensional paradoxes. Paradoxes in this group are just as venerable as paradoxes in the others (indeed, much more venerable than the set-theoretic paradoxes), since they are to be found in Buridan, if not in antiquity, in intensional versions of the liar paradox. Yet they have had little attention compared with their more illustrious cousins. Even after the brilliant papers of Prior [28] and Montague [13] on the subject, little heed was paid to them. The situation is now changing, however, due to the need to develop a formal theory of intensionality. The pressure for this comes largely from a, perhaps, somewhat unexpected direction: artificial intelligence. Workers in this area have confronted the problem of how an AI reasoning system should reason about what it and others know, believe, etc., and have run head-on into the problem of the intensional paradoxes (see Asher and Kamp [3], Morgenstern [15], des

*An earlier draft of this paper was read at a meeting of the Australasian Association of Logic in Perth, May 1988. I am grateful to a number of people there for comments. I am also grateful to Ross Brady for written comments and to an anonymous referee for many helpful comments.

Received July 5, 1989; revised December 18, 1989

Rivieres and Levesque [30], Perlis [16], and for a general view, Thomason [37]). The point of this paper is to review briefly the intensional paradoxes, the attempts to resolve them, particularly by AI workers, and to suggest a better approach.

1 Semantic paradoxes The solution I advocate will be no surprise to anyone familiar with recent debates concerning the logical paradoxes. I have argued elsewhere[1] that it is necessary to accept paradoxical sentences such as the liar as true contradictions, and hence to use a paraconsistent logic. I will not repeat this argument here. To someone who has taken this solution to heart, the solution to the intensional paradoxes will be a trivial corollary, since the structural similarity between all the paradoxes of self-reference cries out for a uniform approach. Indeed, this paper is a really quite trivial extension of well-known work. What makes it worth writing, I take it, is simply that it is well past time that someone said it, particularly for the benefit of AI workers.

In order to discuss the intensional paradoxes and their paraconsistent resolution, it will be useful to have a prolegomenon on the semantic paradoxes, and, in particular, on a semantically closed theory. Take a first-order language which has names for its own sentences and sufficient self-referential machinery. The details of this are not crucial; but for the sake of definiteness, let us take the language of first-order arithmetic. For the sake of simplicity, take this in a form where for each n-place recursive function (or just primitive recursive function), there is a corresponding term of the language with n free variables.[2] We will also suppose a standard Gödelization, a map from formulas to numbers. If φ is a formula, I will use $\langle\varphi\rangle$ as a (metalinguistic) expression for the numeral of its Gödel number. In addition, suppose that the language has a one-place predicate, T, which is going to function as the truth predicate of the language. The logic underlying the language is to be a paraconsistent logic in which standard quantifier inferences, modus ponens, and the substitutivity of equivalents (but not absorption[3]) hold. We will also assume that the logic contains the law of excluded middle (though this assumption is unnecessary for the production of some semantic paradoxes: see Priest [19] and Priest [22], Section 1.3). For the sake of definiteness, let us say that we are using a quantificational extension of the relevant logic E_{fdf} or some higher degree extension thereof, such as DK.[4]

The logic is also to contain identity. Since the properties of this will be crucial sometimes in what follows, it is necessary to be precise about them. The standard principles, Identity and Leibniz' Law:

$$\forall x(x = x)$$

$$t_1 = t_2 \vdash \varphi(t_1) \leftrightarrow \varphi(t_2)$$

(where t_1 and t_2 are any terms, and the usual qualifications concerning substitution are made) are both acceptable. However, the classical consequence of Leibniz' Law, $\varphi(t_1) \wedge \neg\varphi(t_2) \vdash t_1 \neq t_2$, is not correct (as standard relevant/paraconsistent semantics show: see [22], Section 5.3, and [31], Section 7). To see why, let a be any object that has paradoxical properties, i.e., such that for some predicate, P, both Pa and $\neg Pa$ are true. Then if this form were correct, it would follow that $a \neq a$, which it does not: there may be objects which are non-self-identical (as well as self-identical), but this is quite a different matter,

and should by no means follow merely from the fact that an object has *some* inconsistent properties.

In addition to the logical axioms we will take simply all equations and inequations true in the standard model of arithmetic and, also, the T-schema:

$$\mathrm{T}\langle\varphi\rangle \leftrightarrow \varphi$$

where φ is any sentence (closed formula).

I will call the theory just outlined the *extensional theory*. This is inconsistent. For we can prove the diagonal lemma: if $\alpha(v)$ is any formula of one free variable, v, there is a sentence ψ such that $\vdash \psi \leftrightarrow \alpha(\langle\psi\rangle)$.[5] Taking the formula $\neg \mathrm{T}v$ for $\alpha(v)$, we get a formula ψ such that $\psi \leftrightarrow \neg \mathrm{T}\langle\psi\rangle$. But $\mathrm{T}\langle\psi\rangle \leftrightarrow \psi$, by the T-schema. Whence by transitivity $\mathrm{T}\langle\psi\rangle \leftrightarrow \neg \mathrm{T}\langle\psi\rangle$, and by this and the law of excluded middle (LEM), $\mathrm{T}\langle\psi\rangle \wedge \neg \mathrm{T}\langle\psi\rangle$. Thus the liar paradox is provable in the theory. However, this does no harm since contradictions do not spread. I will formulate this claim more precisely in Section 6.

*2 **Propositions*** Let us now return to the subject of the intensional paradoxes. What characterizes these is that they all make use of the notion of a proposition. This may be done in one or both of two ways: first, they may quantify over propositions; second, propositions may be the argument of propositional functions, such as 'is identical to', 'John believes/knows/fears/says that', 'It is provable that', and so on. The paradoxes do not depend, by and large, on any specific properties of propositions, and so of any substantive theory thereof. However, I wish to give a formal treatment on the paradoxes, and thus the question of how to treat propositions must be faced. I have no very strong philosophical views on the matter; moreover, the substantive points I shall make would apply equally well if one were to take propositions and propositional quantification as primitive. But for the sake of definiteness I shall take propositions to be sentences. We may thus simply add to the extensional language of Section 1, monadic predicates such as B (It is believed that), K (It is known that), etc. Then to the axioms of the extensional theory may be added the logical principles specific to each predicate. What, exactly, these are we need not go into. It will suffice to note that most writers subscribe to some subset of the "S5" principles for K:

Ki	$\vdash K\langle\varphi\rangle \rightarrow \varphi$
Kii	$\vdash K\langle\varphi \rightarrow \psi\rangle \rightarrow (K\langle\varphi\rangle \rightarrow K\langle\psi\rangle)$
Kiii	$\vdash K\langle\varphi\rangle \wedge K\langle\psi\rangle \rightarrow K\langle\varphi \wedge \psi\rangle$
Kiv	If $\vdash\varphi$ then $\vdash K\langle\varphi\rangle$
Kv	$\vdash K\langle\neg\varphi\rangle \rightarrow \neg K\langle\varphi\rangle$
Kvi	$\vdash K\langle\varphi\rangle \rightarrow K\langle K\langle\varphi\rangle\rangle$
Kvii	$\vdash \neg K\langle\varphi\rangle \rightarrow K\langle\neg K\langle\varphi\rangle\rangle$

where φ and ψ are sentences. The same is true when we substitute 'B' uniformly for 'K'. (Of course, depending on the underlying logic, these principles may not be independent of each other.) I note for future reference that all of the above are provable in the extensional theory if the intensional operator is replaced by T. (I leave the proof of this as an exercise.) Other intensional notions, such as

fears that, may have no very interesting purely logical postulates, though of course for any concrete situation we may formulate nonlogical axioms. I shall call any augmentation of the extensional theory of the kind indicated, an *intensional theory*.

Taking propositions to be sentences is worth a few comments before we continue. First, it would obviously be incorrect to do this if we had a language with indexicals; but the intensional language above has none. Secondly, and ironically, the intensional paradoxes themselves have been held to refute the very possibility of taking intensional operators as syntactic predicates (see Thomason [34]). The present paper will therefore scotch this objection once and for all. Third, to obtain the effects of propositional quantifiers on this approach, it is necessary to use the truth predicate. Thus, for example, instead of writing $\forall p(p \vee \neg p)$, we must write $\forall x(Fx \rightarrow (\mathrm{T}x \vee \neg \mathrm{T}x))$ or $\forall x(Fx \rightarrow (\mathrm{T}x \vee \mathrm{TNeg}x))$ where F is a predicate true of (codes of) sentences and Neg is the primitive recursive function which maps a (code of a) formula to (that of) its negation. In virtue of the T--schema, these are equivalent. Finally, treating propositions as sentences has very clear computational advantages: processing claims and arguments about knowledge etc. can be done very simply and efficiently. In virtue of the applications of this subject in AI, this is a notable plus.

3 The paradoxes At last, then, here are the paradoxes, together with formal analyses. The list is certainly not a comprehensive one. In particular, I have not included those paradoxes which involve quantification over propositional functions.[6] It should be clear, however, how they can be handled in principle. The first two paradoxes are plain contradictions, and can simply be left to stand.

(i) The simplest paradox is Tarski's[7]: there is some proposition identical to this one which is false. Clearly, this is false iff it is true. To formalize it, we apply the diagonal lemma to get a sentence ψ such that:

(1) $\psi \leftrightarrow \exists x(x = \langle \psi \rangle \wedge \mathrm{TNeg}x)$.

Suppose	ψ	
then	$\exists x(x = \langle \psi \rangle \wedge \mathrm{TNeg}x$	by (1)
	$a = \langle \psi \rangle \wedge \mathrm{TNeg}a$	existential instantiation
	$\mathrm{TNeg}\langle \psi \rangle$	by identity rules
	$\mathrm{T}\langle \neg \psi \rangle$	since $\mathrm{Neg}\langle \psi \rangle = \langle \neg \psi \rangle$
	$\neg \psi$	by the T-schema.

The argument also runs in the other direction since $\langle \psi \rangle = \langle \psi \rangle$.

(ii) A slightly more sophisticated paradox is the knower paradox of Kaplan and Montague (see [9] and Montague [13]). It concerns the sentence: I know this is false. Suppose this is true. Then it is known to be false, and hence false. Hence it is false. Moreover we have just proved this. Hence it is known to be false, i.e., it is true. To formalize this, apply the diagonal lemma to get a sentences, ψ, such that:

(2) $\psi \leftrightarrow K\mathrm{Neg}\langle\psi\rangle$

Suppose	ψ	
then	$K\mathrm{Neg}\langle\psi\rangle$	by (2)
Hence	$K\langle\neg\psi\rangle$	
	$\neg\psi$	by Ki
Hence	$\neg\psi$	by the LEM, since $\{\neg\psi\} \vdash \neg\psi$
But then	$K\langle\neg\psi\rangle$	by Kiv
i.e.	ψ	by (2).

One way out of this paradox is to deny Kiv, the principle of logical omniscience, which is obviously false for real knowers. However, this will not do (see Anderson [2]). First, because although the rule is not in general correct, it is if we have explicitly recognized the proof of ψ, which we have in this case; secondly, because if we take K to mean 'is provable', Kiv (and Ki) *is* intuitively correct. We still, therefore, have an intensional paradox.[8]

(iii) A more sophisticated paradox yet is Prior's.[9] Suppose that during a certain time a Cretan says that everything said by a Cretan during that time is false. Suppose this is true; then everything said by a Cretan (during that time) is false, whence it is false. Thus it is false. Hence, something said by a Cretan (during that time) is true. Since it cannot be the thing uttered, some Cretan must have said something else. To formalize this, let the monadic predicate D be 'During a particular time, t, a Cretan said that'; let the sentence:

$$\forall x(Dx \rightarrow \mathrm{TNeg}x) \text{ be } \psi.$$

Then suppose that:

(3) $D\langle\psi\rangle$

and that:

	ψ	
then	$D\langle\psi\rangle \rightarrow \mathrm{TNeg}\langle\psi\rangle$	by instantiation
hence	$\mathrm{TNeg}\langle\psi\rangle$	by modus ponens (*)
	$\neg\psi$	by the T-schema
Hence	$\neg\psi$	by the LEM, since $\{\neg\psi\} \vdash \neg\psi$
i.e.	$\neg\forall x(Dx \rightarrow \mathrm{TNeg}x)$	
so	$\exists x(Dx \wedge \neg\mathrm{TNeg}x)$	(**)
	$Da \wedge \neg\mathrm{TNeg}a$	by instantiation
But	$\mathrm{TNeg}\langle\psi\rangle$	by the T-schema
thus	$a \neq \langle\psi\rangle$	by identity principles (***)
So	$\exists x(Dx \wedge x \neq \langle\psi\rangle)$.	

Prior's argument is not a paradox in the sense that it ends in a contradiction. The conclusion is, on its own, quite consistent. The paradox lies, rather, in the a priori deduction of something that is quite contingent, and which we can easily arrange to be false. Prior's solution to the paradox is that if nothing else is said during this time then (3) must be false, by reductio. Implausible as this may seem, it has at least this going for it: we may be sure that a Cretan uttered words during that time; however, that he stated a proposition with the use of those words

is not so obvious. Quite independently, since Prior, several writers have urged that whether or not an utterance has content may depend on quite contingent features not necessarily visible to the utterer (though for no reason that Prior could appeal to here).[10]

The real bite of Prior's paradox comes, as Prior observed, when we note that the argument makes *no* substantial assumptions about D at all. Hence exactly the same conclusion must be drawn if we read D as 'It is feared/thought etc. at a certain time that'. It is one thing to suppose that one did not say something though one thought one did; it is quite another to suppose that one does not fear something when one thinks one does. Indeed, suppose we run the argument with the intensional predicate 'It is thought that it is feared that etc.'. Then Prior's conclusion is to the effect that one cannot even think that one fears something. What in Heaven's name is one doing then? Even worse (as Dummett noted; see [28], p. 29) we can run the argument with D as 'John uttered a sentence which normally means that' to prove that John didn't even do that. This, surely, is totally absurd (and shows, incidentally, that there are some things that are much more absurd than some contradictions).

The solution is as follows. The argument above is invalid at line (**): the inference from $\neg(\alpha \rightarrow \beta)$ to $(\alpha \wedge \neg\beta)$ is invalid. This, however, hardly gets to the bottom of the matter. We can repair this step if we replace the $\rightarrow$ everywhere with a material conditional. Now, however, step (*) becomes an invalid detachment for material implication. But if this step is not truth-preserving, it must be because the antecedent is both true and false (see [22], Chapter 8, and Priest [24]). Hence we must have $D\langle\psi\rangle \wedge \neg D\langle\psi\rangle$. This at least saves $D\langle\psi\rangle$, but at the expense of endorsing its negation too, and it is not clear that this solution is much better than Prior's.

In any case, both this solution and the last hinge on the fact that the only obvious ways of translating the Aristotelian A form into quantifier plus connective fails to preserve its standard properties in relevant logic. Maybe this is inevitable; but maybe what is required (as many have noted) is a new theory of bounded quantification, suitable for relevant logic, but preserving (more) standard syllogistic forms. And if this is obtainable then maybe both (*) and (**) can be made valid. There are therefore at least two good reasons for supposing that we have not yet got to the heart of the problem. The heart, as will probably have been clear straight away, is the inference (***). As I noted in Section 1, this inference is quite invalid. In particular, it breaks down if the object in question has contradictory properties: in this case $\langle\psi\rangle$ $(=a)$ is both true and not true. But this is exactly what one would expect from standard dialetheic approaches to the logical paradoxes. The assertion made by the Cretan, or whoever, is simply a paradoxical one in the context. Isn't that obvious?

(iv) A variant of Prior's paradox, which Prior attributes to Buridan, is as follows. I say that exactly one of this pronouncement and your next one is true. You then say that apartheid is an evil, which is true. Then my saying is true iff it is false. To formalize this let D be the predicate 'is your next saying', $\veebar$ be exclusive disjunction; and suppose you say something true:

(4) $\exists x(Dx \wedge \mathrm{T}x)$.

Instantiating the quantifier we get:

$$Da \wedge \mathrm{T}a.$$

Now, by the diagonal lemma there is a ψ such that:

(5) $\psi \leftrightarrow (\mathrm{T}a \veebar \mathrm{T}\langle\psi\rangle)$.

Suppose	ψ	
Then	$\mathrm{T}\langle\psi\rangle$	by the T-schema
Then	$\neg(\mathrm{T}a \veebar \mathrm{T}\langle\psi\rangle$	by standard properties of $\veebar$
Hence	$\neg\psi$	by (5)
Hence	$\neg\psi$	by the LEM
But then	$\neg\mathrm{T}\langle\psi\rangle$	by the T-schema and contraposition
So	$\mathrm{T}a \veebar \mathrm{T}\langle\psi\rangle$	by properties of $\veebar$
i.e.	ψ	by (5).

Prior's solution to this is simply to take it as a reductio of (4). The counterintuitiveness of this hardly needs to be stressed. It shows that you cannot say anything true, even if the words you utter are as benign as you like. Even worse, as in the previous paradox, we have made no substantive assumptions about *D*. Hence by reinterpreting *D*, Prior's solution would show that you cannot even think something true, etc. It is much more simple and natural to take the paradox to show, what it appears to, that ψ is both true and false. (Notice also that the argument cannot then be run as a reductio of the assumption (4), since the assumption does not *on its own* entail the contradiction.)

4 Other solutions In the last section I discussed Prior's solutions to paradoxes iii and iv and argued that they are most implausible. If this were not enough, Prior's solution will not work anyway, just because it can be applied only to *some* of the paradoxes: those which have contingent premises. It gets no grip at all on paradoxes i and ii. I will now review other suggested solutions. They all fail, and for at least the same reason: they can be applied at best to only some of the intensional paradoxes.

(a) An extreme suggestion is to ban all propositional quantification. No doubt this would appeal to certain Quineans. However, it will not work: it makes nonsense of straightforward claims (such as that there is something I know that you don't), claims that are an integral part of much cognitive and AI reasoning. And apart from anything else, paradox ii, the knower paradox, makes no use of such quantification.

(b) Another reaction is to reject the interpretation of propositions as sentences and, more generally, any account of propositions according to which they are representational, i.e., are internally structured entities (see, e.g., [34]). Such an internal structure is exploited in proving the diagonal lemma, the argument for which therefore fails. Those who accept this solution face the difficult task of giving a nonrepresentational theory of propositions which is adequate to its task, and which does not reintroduce the paradoxes. This is no easy matter, for many such accounts do not avoid the paradoxes, or rather avoid them only be-

cause they cannot express perfectly legitimate notions. For example, the addition of a predicate expressing the perfectly legitimate relation *sentence x expresses proposition y* is sufficient to regenerate the paradoxes in many such theories (see [3]). But even if such a balancing trick can be performed, the solution still will not work, since it does not avoid those paradoxes which do not use diagonalization. Prior's paradox iii is such a paradox.

(c) Another possible solution is provided by an idea of Rivieres and Levesque [30],[11] which is as follows. Suppose we take some demonstratively consistent theory and find a deduction-preserving map, $+$, from the language of the theory into the intensional language, which, moreover, respects negation (i.e., such that $\{\varphi\} \vdash \psi$ iff $\{\varphi^+\} \vdash \psi^+$, and $(\neg\varphi)^+ = \neg(\varphi^+)$). If we now take as axioms only those formulas that are images of the axioms of the consistent theory under the map, we can be sure that these are consistent. To apply this strategy Rivieres and Levesque produce a suitable map from the language of a standard (consistent) epistemic logic. Certain instances of the principles *K*i–*K*vii are filtered out by this procedure, notably those where the formula of which *K* is predicated is the formula given by the diagonal lemma.

As a solution to the paradoxes, this obviously leaves a good deal to be desired. It establishes that there are consistent subtheories of our (classically) inconsistent theory; but this was never in doubt. Perhaps it establishes that there are reasonably strong subtheories; but unless it provides a rationale as to why translatability from the source language is a criterion of truth, and, more crucially, why failure to translate is a criterion of falsity, the construction is of little theoretical value. One is reminded of some older attempts to solve the semantic paradoxes (of a kind which have now disappeared, fortunately): any consistent theory in a language containing a predicate written as the upper case 15th letter of the alphabet was thought to be good enough. If no rationale is forthcoming, this kind of approach is at best an exercise in formal virtuosity. Moreover, in this and similar cases of "strategic withdrawing to a safe domain" (Lakatos [12], pp. 28ff), since there is no rationale, there is no guarantee that the axioms have not been decimated unduly: that there is not perfectly legitimate reasoning which is ruled out. For example, consider the sentence 'This sentence is known to be a well-formed formula' (or its arithmetic analogue, formed by diagonalizing on the predicate 'is the (code of) a well-formed formula'). This is quite unproblematically true. Yet its construction requires the diagonal lemma, and so the obvious argument for its truth is blocked on this approach. (I leave the verification of this as an exercise.)

Finally, though this approach ensures consistency, it does not guarantee the absence of other paradoxes and absurdities. Thus, Prior's paradoxes iii and iv do not use any instance of a modal axiom! Their proofs therefore go through in Rivieres and Levesque's approach. As Kant might (not) have put it: conceptions without intuitions are blind.

(d) A fourth possibility for resolving the paradoxes starts from the observation that the principle of bivalence, in the form of the law of excluded middle, is employed in the paradoxes. Hence, jettisoning this may provide a solution. It is worth noting that moving to intuitionistic logic will not solve the paradoxes. This is because the intuitionistically valid $(\alpha \to \neg\alpha) \to \neg\alpha$ (*consequentia*

mirabilis) would do instead of the law of excluded middle in a number of the cases. But suppose we think more classically and grant there to be truth-value gaps which invalidate both the law of excluded middle and *consequentia mirabilis*? It should be noted that this suggestion is quite unsatisfactory unless we have an independent argument that the paradoxical sentences *are* truth-valueless. But even if we had such an argument, the suggestion will still not solve the paradoxes. For, as in the alethic case, there are paradoxes which explicitly take the possibility of truth-value gaps into account, "extended paradoxes". (For similar comments, see [2].)

For example, consider the "extended knower paradox": This sentence is known to be false or truth-valueless. Suppose it is true. Then since what is known is true, it must be false or truth-valueless. Hence it is not true. It must therefore be false or truth-valueless; and we have just proved that. Hence we know it to be false or truth-valueless, i.e., it is true.

Similarly, there is an extended version of Prior's paradox iii. Suppose a Cretan says that everything a Cretan says is either false or truth-valueless. Suppose also that nothing else is said by the Cretan. If the pronouncement is true, then it is either false or truth-valueless. Contradiction. But if it is either false or truth-valueless, it is true. Contradiction. Hence if a Cretan says this, something else must be said by a Cretan.

Supposing there to be truth-value gaps does not, therefore, solve either the semantic or the intensional paradoxes.

5 Borrowing alethic technology Another thought as to how to solve the paradoxes is this. The notion of truth, and in particular, the T-schema, is prominent in a number of the paradoxes. There are, of course, many suggested solutions to the semantic paradoxes which reject this. Maybe if we employ one of these solutions, the intensional paradoxes can also be avoided. The thought will not work. For a start, the techniques concerning truth are all known to be unsatisfactory in their own right (see, e.g., Priest [20], Priest [23], Priest [25] and especially [22], Chapter 1); hence no generally satisfactory solution can be expected in this way. Secondly, the notion of truth was necessary in formulating the paradoxes only because I chose to represent propositions as sentences and eschewed propositional quantification in its own right. Had we used this, the notion of truth would not even have appeared in the paradoxes. Third, and conclusively, not all the paradoxes made use of the notion of truth. The knower paradox, iii, for example, does not.

In reply to the last point, one might suggest that although truth itself does not appear in the knower paradox, it might be possible to apply some of the devices familiar from accounts of truth aimed at solving the semantic paradoxes to notions which do occur in the knower (and similar) paradoxes, and thus solve them. This suggestion has, indeed, been made by various people, so it is worth commenting on. Since the techniques used are all known to be problematic in the alethic case it would be amazing if they worked any better for the intensional case; and, indeed, they do not, being susceptible to exactly the same problems. I will indicate this briefly. Those who do not want to tread these well-trodden paths again can skip on to the next section.

(a) The first technique that has been used is the Tarski hierarchy of metalanguages. It supposes that natural language is really a hierarchy of languages; each language has its own family of intensional predicates, gramatically applicable only to the sentences of the language below. Thus, self-reference is made impossible.[12]

The move has little independent rationale: there is no linguistic evidence that language has this structure. Thus the strategy is merely that of "withdrawal to a safe domain" again, and, as usual, such strategies withdraw too far. There is perfectly good self-referential reasoning concerning intensional notions, which cannot be handled this way. We have already met the sentence 'This sentence is known to be well formed' which is unproblematically true, but cannot be formulated, let alone have its properties established, on this view. Neither are the sentences involved necessarily of this artificial kind. The skeptic's claim, that he knows nothing, is self-refuting just because it applies to itself. Grice's famous definition of $\text{meaning}_{\text{N.N.}}$ involves intentions concerning, *inter alia*, the effects of those intentions. Many of us have beliefs that all our beliefs (including this one) are fallible. And so it goes on. The banning of intensional self-reference just cannot do justice to our rich and self-reflexive mental life. To put the cap on it, there isn't even any syntactic way of determining which sentences involve self-reference, since whether self-reference occurs may depend on quite contingent facts, such as the reference of certain noun phrases (as Kripke has pointed out in [11]). Solutions that work on syntactic regimentation are therefore hopelessly misguided.

Many of these objections are avoided by a slightly liberalized version of the hierarchy solution suggested by Anderson [2] and Burge [6]. According to them, the extension of the knowledge predicate of English, though normally fixed, may change (is "indexical"), particularly during certain forms of ratiocination. The extension of the predicate may contain sentences containing the knowledge predicate itself, and a certain amount of self-referential reasoning is therefore possible. Lo and behold, however, at just the point where a contradiction threatens, the extension surreptitiously switches to avoid it. The claim that the predicate is indexical is still *ad hoc*, however, in that it lacks any independent linguistic grounding. Anderson offers an argument, based on Gödel's Theorem; but in the end what it amounts to is: if it were not like this, inconsistency would arise; which is hardly an independent justification of why it is like this. Moreover, a second locus of *adhoc*ness enters the picture: for even given that the predicate may change its extension we need an independent reason as to why it changes "in time" to prevent inconsistency.[13]

In any case, all versions of the hierarchy view succumb to extended paradoxes, or at least, to avoid them we must maintain the inexpressibility of some perfectly legitimate notion. For call a sentence *known-in-some-sense* if it is in the union of all possible extensions of the K predicate. And consider the claim: this sentence is known-in-some-sense to be false. The familiar reasoning leads to contradiction.

(b) Another posited solution appeals to Kripke's work on truth, rather than Tarski's. This is pursued by Morgenstern [15]. The idea here is essentially as follows. Truth is defined *à la* Kripke. In particular, then, certain sentences are un-

grounded and hence neither true nor false. Let Bx be the predicate 'x is believed' (or 'x is believed with justification'). Then Kx is defined as $Bx \wedge Tx$. A simple exercise now shows that all of Ki–vii break down in general (with the possible exception of Kiv, depending on how one interprets it). A major problem with this approach is that it gets truth-values wrong, even on its own terms. Suppose that φ is some proposition (sentence) that we know to have no truth value at some fixed point. We therefore know that it is not true. It is therefore true that we do not know it, i.e. $\neg K\langle\varphi\rangle$ is true. But by construction $T\langle\varphi\rangle$ has no truth value (at the fixed point), as therefore does $B\langle\varphi\rangle \wedge T\langle\varphi\rangle$, as, therefore, does its negation $\neg K\langle\varphi\rangle$.

Another substantial objection to this suggestion is that it, again, maintains consistency only by certain perfectly legitimate notions failing to be expressible in the language. Crucially, for example, there is no predicate which defines the set of sentences that are neither true nor false. If there were, paradoxes of the extended variety, such as the extended knower paradox, would occur, as may be checked.

Similar comments apply to a solution proposed by Perlis [16]. He defines K in the same way, except that he takes the truth predicate to be a total predicate, applying truly to grounded true sentences and falsely to all others. His logic is, consequently, classical. This account, similarly, gets its truth values wrong. Let φ be any formula that is known to be true but not grounded (e.g., $T\langle\psi\rangle \vee \neg T\langle\psi\rangle$ where ψ is not grounded). Then $T\langle\varphi\rangle$ is false; whence $K\langle\varphi\rangle$ is false, contrary to the facts. Moreover, the approach, being classical, is subject to Tarski's theorem. There is no predicate which defines the set of genuinely true sentences. If there were, we could construct extended knower paradoxes in the familiar way.

(c) The final approach that borrows standard alethic technology borrows not from Tarski or Kripke but from Gupta and Herzberger. This is pursued by Asher and Kamp [3]. Like Kripke, Gupta and Herzberger use a single language with a truth predicate; like Kripke, they also define a transfinite sequence of models such that the extension of the truth predicate (at a successor ordinal) is the set of truths in the previous model. Unlike Kripke, however, they use classical logic. As a result of this, there is no fixed point to the construction, but it does obtain a certain stability, with many formulas (but not paradoxical ones like the liar and its ilk) reaching a point where they have a fixed truth value in all subsequent models. Asher and Kamp run a similar construction based on a frame for an epistemic logic (for which we may suppose the accessibility relation to be at least reflexive). A formula is in the extension of the knowledge predicate at any world at any (successor) ordinal if it is true at all epistemic alternatives of that world at the previous stage. (A limiting construction, whose exact details need not concern us, is used at limit ordinals.) The results concerning stability are similar to but more complex than those in the alethic case, the nature of the frame being of crucial importance. And although Kii–v are guaranteed to be validated by this construction, Ki, on which the knower paradox depends, is not. (Kvi–vii depend, of course, on other frame properties even in the standard case.)

This suggestion fares no better than the others, and, arguably, fares worse. For a start, since there are no fixed points, how this construction is supposed to relate to English and the assertions we make in that language is quite opaque.

Secondly, as do all possible-worlds constructions, these semantics suffer from the problem of logical omniscience.[14] Thirdly, and again, it avoids paradoxes only by being expressively incomplete. Call a sentence, φ, *absolutely known* if $K\langle\varphi\rangle$ is true at all worlds at all levels. Absolute knowledge is a world-invariant, level-invariant notion. Hence it should be specifiable in a world-invariant, level-invariant way. But there is no formula of one free variable, $\alpha(v)$ which defines the set of absolutely known formulas in all levels of all worlds. For suppose $\alpha(v)$ defined the absolutely known formulas at every level at every world. By the diagonal lemma we could construct a formula ψ such that $\psi \leftrightarrow \neg\alpha(\langle\psi\rangle)$ holds (at all worlds and levels). Suppose that $\alpha(\langle\psi\rangle)$ is true at level λ at world w. Then ψ is absolutely known. Hence $K\langle\psi\rangle$ is true at w at level $\lambda + 1$, and ψ is true at every world accessible to w at level λ. In particular, ψ is true at w at level λ (since the accessibility relation is reflexive). Whence, by definition of ψ, $\alpha(\langle\psi\rangle)$ is false at w at level λ. Contradiction. Alternatively, suppose that $\alpha(\langle\psi\rangle)$ is false at w at level λ. Then ψ is not absolutely known. There is therefore some world w' and some (successor) level λ' such that $K\langle\psi\rangle$ is false at w' at level λ'. Thus there is some alternative to w', w'', such that ψ is false as w'' at level $\lambda' - 1$. Hence, by definition, $\alpha(\langle\psi\rangle)$ is true at this level at w'', and we know that this leads to contradiction.

Thus, all these appropriations of alethic technology fail. Crucially, all the constructions are expressively impoverished. Indeed, this is exactly how consistency is maintained. This is exactly what happens in the alethic case too.[15] All such solutions move to a language which is expressively weaker than English, and hence fail to show that our ordinary notions, expressed in ordinary English, are, appearances notwithstanding, consistent. Indeed, they are not, as any honest person must confess.

6 The confinement of inconsistency We have now reviewed a number of mooted solutions to the intensional paradoxes. It is clear that they are implausible, frequently contrived, or just plain wrong. By contrast, the paraconsistent/dialetheic solution is simple and natural. The paradoxical sentences in question are, as the paradoxical arguments show, both true and false; and provided a paraconsistent logic is used, we may at least suppose (until proved otherwise) that the contradictions do not spread, and therefore do no harm.[16]

This raises the question of how far the contradictions do spread. A proof of nontriviality is always a welcome result; a result circumscribing the realm of contradiction is even more so. In this final section I will provide just such a result for certain intensional theories. The result is a corollary of the following theorem concerning the extensional theory.

Theorem 1 *For the extensional theory of Section* 1, *if* α *is any sentence not containing* T *then* α *is not provably inconsistent.*

In particular, therefore, contradictions do not spread into the underlying theory of arithmetic. I indicate the proof of this theorem in an appendix. To state the corollary, a definition is useful. Let us say that an intensional theory is *regular* if every axiom (or rule of inference) involving intensional operators would be a theorem (or an admissible rule) of the extensional theory if all the intensional

operators were replaced by T. In fact, most normal intensional theories are regular. Standard theories for knowledge and belief are regular, as I observed in Section 2. Many intensional operators do not even have proper logical axioms. Regular intensional theories are therefore most theories of logical interest. The corollary is:

Corollary *In any regular intensional theory, if α is any sentence not containing an intensional predicate or* T, *it is not provably inconsistent.*

Proof: The proof of this is trivial: if a theory is regular it can be interpreted in the extensional theory simply by mapping intensional predicates to T.

As I stated at the beginning of the paper, the correct solution to the intensional paradoxes is (quite literally) but a corollary of the correct solution to the semantic paradoxes.

NOTES

1. See Priest [18], [20], and particularly [22], Part One. In an AI context, I have argued the thesis in [25], where the present approach to the intensional paradoxes was mooted. The present paper can be profitably read as a sequel to that one.
2. This can be done in a finitary way with resources slightly exceeding those of first-order logic, by having function symbols for each basic recursive function, and functional symbols for substitution, primitive recursion (and minimization). See Boolos [4], Chapter 7.
3. $\alpha \to (\alpha \to \beta) \vdash \alpha \to \beta$: see Note 8.
4. For E_{fdf} see, e.g., [1], Section 19. For DK see [31], Section 6.
5. *Proof:* Let $\delta(v)$ represent the diagonal function. Consider $\alpha(\delta(v))$. Call this $\varphi(v)$. Its diagonalization is $\varphi(\langle\varphi\rangle)$. Since δ represents the diagonal function, $\vdash\delta(\langle\varphi\rangle) = \langle\varphi(\langle\varphi\rangle)\rangle$. Hence $\vdash\alpha(\delta(\langle\varphi\rangle)) \Leftrightarrow \alpha(\langle\varphi(\langle\varphi\rangle)\rangle)$, i.e. $\vdash\varphi(\langle\varphi\rangle) \leftrightarrow \alpha(\langle\varphi(\langle\varphi\rangle)\rangle)$. Thus, $\varphi(\langle\varphi\rangle)$ is the required formula.
6. For example, the propositional version of the heterological paradox. For a discussion of some others see Thomason [36]. The Surprise Exam paradox might also be thought of as a paradox of intensionality (see [9]). In its simple form, where the teacher's pronouncement is not self-referential, it seems to me that the most plausible solution is just to deny that the children know that the teacher is speaking the truth – however earnestly s/he speaks (thus refuting the definition of knowledge as justified true belief). In its more complex form, where the teacher's pronouncement is self-referential, the situation is more like that in the knower paradox, and a dialetheic solution may be more plausible.
7. Tarski [33], pp. 160 ff. I have reformulated it slightly.
8. There is a version of the knower pardox for belief, which does not use *B*i. Slightly different versions are given by Thomason [35] and Burge [5]; but essentially it goes as follows. By diagonalization we can find a formula φ such that:

(*)	$\varphi \leftrightarrow \neg B\langle\varphi\rangle$	
	$B\langle\varphi \to \neg B\psi\langle\varphi\rangle\rangle$	by *B*iv
So	$B\langle\varphi\rangle \to B\langle\neg B\langle\varphi\rangle\rangle$	by *B*ii

But	$B\langle \neg B\langle \varphi\rangle\rangle \rightarrow \neg B\langle B\langle \varphi\rangle\rangle$	by Bv
So	$B\langle \varphi\rangle \rightarrow \neg B\langle B\langle \varphi\rangle\rangle$	
But	$B\langle \varphi\rangle \rightarrow B\langle B\langle \varphi\rangle\rangle$	by Bvi
Thus	$\neg B\langle \varphi\rangle$	by contraposition and the LEM
Whence	φ	by (*)
and	$B\langle \varphi\rangle$	by Biv.

Interpreting B as belief, this argument has little to recommend it since it depends on the manifestly false Bv. If we interpret B as 'It is rational to believe that', then all the steps are much more plausible. If they are correct then we have a true contradiction. I would argue that Bv is incorrect even for rational belief, however. It is possible to believe a sentence and its negation rationally. See Priest [21] or [22], Chapter 7.

There is also a Curried form of the knower paradox: Given an arbitrary β, by diagonalization we find a sentence of the form:

(*)	$\varphi \leftrightarrow (K\langle \varphi\rangle \rightarrow \beta)$	
Thus	$K\langle \varphi\rangle \rightarrow (K\langle \varphi\rangle \rightarrow \beta)$	by Ki
(**)	$(K\langle \varphi\rangle \rightarrow \beta)$	by absorption
Hence	φ	by (*)
So	$K\langle \varphi\rangle$	by Kiv
Whence	β	by (**).

This argument fails if absorption fails, as it must if ordinary Curry paradoxes are to be avoided.

. For this and the next paradox see [28]. See also Prior [29], Chapter 6.

. This observation and a number of others in the following discussion of Prior's paradoxes come from the (still, unfortunately) unpublished [36], which paper first alerted me to the subject of intensional paradoxes. See also the excellent discussion in [6].

. To be fair to them, it is not clear that they think of themselves as trying to solve the paradoxes. Perhaps they are just showing that there are reasonably strong consistent sentential theories.

. Clearly, on this approach naming cannot be performed by Gödel numbering; some other device is used. The approach is mooted by Kaplan and Montague [9], and carried out in Priest [17]. It is also implemented by Konolige [10], though it is not clear that this is in response to the paradoxes.

. Burge has a sophisticated but somewhat tortuous argument aimed at justifying this for his version of the believer paradox (see Note 8). However, whether he takes this to apply to the knower paradox, and if so, how, is unclear to me.

. To be fair to Asher and Kamp, they recognize this problem and say that they would avoid it by applying the technique in a different way, but what this is is not explained.

. See [22], Chapter 1. Clearly, in the extensional theory, and, a fortiori, in intensional theories, both truth and falsity are expressible. It might be thought, however, that these theories, too, suffer from expressive incompleteness. For a discussion and rejection of this view see Priest [26].

. Naturally, it is possible to formulate objections to this approach to the paradoxes, too, though this is not the place to discuss them. For full details see [22].

. For the logic RM3 see Brady [7], Section 2, or Anderson and Belnap [1], p. 470.

REFERENCES

[1] Anderson, A. and N. Belnap, *Entailment*, Princeton University Press, Princeton, New Jersey, 1975.

[2] Anderson, C. A., "The paradox of the knower," *Journal of Philosophy*, vol. 80 (1983), pp. 338–355.

[3] Asher, N. and J. Kamp, "The knower's paradox and representational theories of attitudes," in [8].

[4] Boolos, G. and R. Jeffrey, *Logic and Computability*, Cambridge University Press, Cambridge, 1974.

[5] Burge, T., "Buridan and epistemic paradox," *Philosophical Studies*, vol. 34 (1978), pp. 21–35.

[6] Burge, T., "Epistemic paradox," *Journal of Philosophy*, vol. 81 (1984), pp. 5–29.

[7] Brady, R., "The non-triviality of dialectical set theory," in [27].

[8] Halpern, J. Y., *Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufman, Los Altos, California, 1986.

[9] Kaplan, D. and R. Montague, "A paradox regained," *Notre Dame Journal of Formal Logic*, vol. 1 (1960), pp. 79–90 (reprinted as Chapter 9 of [14]).

[10] Konolige, K., "A first order formalisation of knowledge and action for a multiagent planning system," *Machine Intelligence*, vol. 10 (1982), pp. 41–72.

[11] Kripke, S., "Outline of a theory of truth," *Journal of Philosophy*, vol. 72 (1976), pp. 690–716.

[12] Lakatos, I., *Proofs and Refutions*, Cambridge University Press, Cambridge, 1976.

[13] Montague, R., "Syntactic treatments of modality, with corollaries on reflection principles and finite axiomatizability," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 153–167 (reprinted as chapter 10 of [14]).

[14] Montague, R., *Formal Philosophy*, New Haven, Connecticut, Yale University Press, 1974.

[15] Morgenstern, L., "A first order theory of planning, knowledge and action," in [8].

[16] Perlis, D., "Languages with self-reference II: Knowledge, belief and modality," *Artificial Intelligence*, vol. 34 (1989), pp. 179–212.

[17] Priest, G., "A refoundation of modal logic," *Notre Dame Journal of Formal Logic*," vol. 18 (1977), pp. 340–354.

[18] Priest, G., "Logic of paradox," *Journal of Philosophical Logic*, vol. 8 (1979), pp. 219–241.

[19] Priest, G., "The logical paradoxes and the law of excluded middle," *Philosophical Quarterly*, vol. 33 (1983), pp. 160–165.

[20] Priest, G., "Semantic closure," *Studia Logica*, vol. 43 (1984), pp. 117–129.

[21] Priest, G., "Rationality, belief and contradiction," *Proceedings of the Aristotelian Society*, vol. 86 (1986), pp. 99–116.

[22] Priest, G., *In Contradiction*, Martin Nijhoff, Dordrecht, 1987.

[23] Priest, G., "Unstable solutions to the liar paradox," in S. Bartlett and P. Suber (eds.), *Self-Reference: Reflections on Reflexivity*, Dordrecht, Martin Nijhoff, 1987.

[24] Priest, G., *Reductio ad absurdum et modus tollendo ponens*, in [27].

[25] Priest, G., "Reasoning about truth," *Artificial Intelligence*, vol. 39 (1989), pp. 231-244.

[26] Priest, G., "Boolean negation and all that," *Journal of Philosophical Logic*, vol. 19 (1990), pp. 201-215.

[27] Priest, G., R. Routley, and J. Norman, *Paraconsistent Logic*, Philosophia Verlag, Munich, 1989.

[28] Prior, A., "On a family of paradoxes," *Notre Dame Journal of Formal Logic*, vol. 2 (1961), pp. 16-32.

[29] Prior, A., *Objects of Thought*, Clarendon Press, Oxford, 1971.

[30] des Rivieres, J. and H. Levesque, "The consistency of syntactic treatments of knowledge," in [8].

[31] Routley, R., "Ultralogic as universal?," *Relevance Logic Newsletter*, vol. 2 (1977), pp. 50-90 and 138-175 (reprinted as an Appendix in [32]).

[32] Routley, R., *Exploring Meinong's Jungle and Beyond*, Department of Philosophy, RSSS, Australian National University (1980).

[33] Tarski, A., "The concept of truth in formalised languages," chapter 8 of *Logic, Semantics and Meta-Mathematics*, Clarendon Press, Oxford, 1956.

[34] Thomason, R., "Indirect discourse is not quotational," *Monist*, vol. 60 (1977), pp. 341-352.

[35] Thomason, R., "A note on syntactic treatments of modality," *Synthese*, vol. 44 (1980), pp. 391-395.

[36] Thomason, R., "Paradoxes of Intensionality," unpublished manuscript, University of Pittsburgh (1982).

[37] Thomason, R., "Paradoxes and semantic representation," in [8].

Department of Philosophy
University of Queensland
St. Lucia, Queensland
Australia 4067

Appendix In this appendix I will sketch a proof of Theorem 1 of Section 6. The proof relies heavily on a construction of Brady [7], and I will not repeat a number of proofs to be found in that paper. The theorem is itself a corollary of a more general theorem. To explain what that is, it will help to have a couple of definitions. If α is any formula let $\alpha^{\supset}$ be α with all occurrences of $\rightarrow$ replaced by $\supset$. If X is a theory (i.e., set of sentences), let $X^{\supset} = \{\alpha^{\supset}; \alpha \in X\}$. The theorem is as follows:

Theorem 0 *Let X be a theory in a language not containing* T. *Let φ be a sentence in the same language such that $X^{\supset} \nvdash \varphi^{\supset}$ (where the $\vdash$ is classical*

consequence). *Then* φ *is not a theorem of the theory* (*in the extended language*) X + T-*schema.*

Proof of Theorem 1: Suppose that α is an arithmetic sentence, and that A is the set of arithmetic axioms of the extensional theory. Then clearly $A^{\supset} \nvdash (\alpha \wedge \neg\alpha)^{\supset}$ since $A^{\supset}$ is true in the standard model. By Theorem 0, $\alpha \wedge \neg\alpha$ is not a theorem of the extensional theory.

Proof of Theorem 0: The first step is to construct a theory X^+ in the language of X plus a denumerable number of individual constants, which contains X, is maximally consistent, saturated in the constants, but such that φ cannot be deduced from X^+. Since $X^{\supset} \nvdash \varphi^{\supset}$, $X^{\supset}$ may be extended to a consistent saturated theory, Y, such that $Y \nvdash \varphi^{\supset}$ by the usual Henkin construction. Now let $X^+ = \{\alpha; \alpha^{\supset} \in Y\}$. X^+ can easily be checked to have the right properties.

Let us call the language of $X^+ L$. Let L^T be the language L extended by a new monadic predicate T. We assume that to each formula, φ, of L is assigned a term in L, $\langle\varphi\rangle$, to function as its name. This is always possible since there are countably many constants. (In particular, if L contains the language of arithmetic—as it does in the case we are interested in—$\langle\varphi\rangle$ is just the numeral of its Gödel number under some standard Gödelization.)

An *initial sentence* is any sentence of L^T which is either atomic or is of the form $\alpha \rightarrow \beta$. I will call any map which assigns to each initial sentence one of the values t, b, and f, and then uses RM3 truth conditions and substitutional quantification to extend this to all formulas of L^T, an *evaluation.*[17] At one point below, RM3 implication will also make an appearance. I will write this as $\Rightarrow$.

Define an ordering on evaluations, $<$, as follows:

$\nu < \nu'$ iff for every initial sentence, α:

$$\text{if } \nu(\alpha) = \text{t or f then } \nu(\alpha) = \nu'(\alpha) \qquad (*)$$

Lemma 1 *If* $\nu < \nu'$ *then condition* (*) *holds for all sentences of* L^T.

Proof: See Brady [7], Lemma 1. The proof is straightforward.

The argument now employs two iterative constructions.

Construction One. Given any evaluation, ν, define a (transfinite) sequence of evaluations thus:

$\nu_0 = \nu$.

For $\lambda \neq 0$:

$$\begin{aligned} \nu_\lambda(\mathrm{T}\langle\varphi\rangle) &= t \text{ if } \exists\kappa\forall\mu(\kappa \leq \mu < \lambda \text{ implies } \nu_\mu(\varphi) = t) \\ &\quad\ f \text{ if } \exists\kappa\forall\mu(\kappa \leq \mu < \lambda \text{ implies } \nu_\mu(\nu) = f) \\ &\quad\ b \text{ otherwise} \\ \nu_\lambda(\alpha) &= \nu_0(\alpha) \\ &\quad\ \text{where } \alpha \text{ is any other initial sentence.} \end{aligned}$$

(Note that Brady's definition is slightly different, but equivalent.)

Lemma 2 *If* $\kappa < \mu$ *then* $\nu_\kappa < \nu_\mu$.

Proof: The proof is straightforward.

Now consider the sequence $\{\nu_\lambda; \lambda \text{ an ordinal}\}$. By Lemma 2 and some simple facts about cardinality, it follows that there must be a λ such that, for all $\kappa > \lambda$, $\nu_\lambda = \nu_\kappa$. Let λ be the least such ordinal; and let $\nu^* = \nu_\lambda$.

Observe the following about ν^*:

(i) $\nu^*(\alpha) = \nu^*(\mathrm{T}\langle\alpha\rangle)$.
(ii) $\nu < \nu^*$.

Construction One produces an evaluation which would validate the T-schema if it were formulated with $\Rightarrow$ but not as formulated with $\rightarrow$. To rectify this we employ Construction Two.

Construction Two. Given an evaluation ν, define a (transfinite) sequence of evaluations as follows:

$\nu_0 = \nu$.

For $\lambda \neq 0$:

$$\begin{aligned}\nu_\lambda(\alpha \rightarrow \beta) &= t \text{ if } \forall\mu < \lambda\ \nu_\mu^*(\alpha \Rightarrow \beta) = t\\ &\quad\ f \text{ if } \exists\mu < \lambda\ \nu_\mu^*(\alpha \Rightarrow \beta) = f\\ &\quad\ b \text{ otherwise}\\ &\quad\ \text{where } \alpha \text{ and } \beta \text{ are formulas of } L^T\\ \nu_\lambda(\gamma) &= t \text{ if } \exists\kappa\forall\mu(\kappa \leq \mu < \lambda \text{ implies } \nu_\mu^*(\gamma) = t)\\ &\quad\ f \text{ if } \exists\kappa\forall\mu(\kappa \leq \mu < \lambda \text{ implies } \nu_\mu^*(\gamma) = f)\\ &\quad\ b \text{ otherwise}\\ &\quad\ \text{where } \gamma \text{ is any other initial sentence.}\end{aligned}$$

(Again, Brady's definition is slightly different but equivalent.)

Lemma 3 *If* $\kappa < \mu$, *then* $\nu_\kappa < \nu_\mu$.

Proof: See Brady [7], Lemma 2.

Now consider the sequence $\{\nu_\lambda; \lambda \text{ an ordinal}\}$. By Lemma 3 and some simple facts about cardinality, it follows that there must be a λ such that, for all $\kappa > \lambda$, $\nu_\lambda = \nu_\kappa$. Let λ be the least such ordinal; and let $\nu^T = \nu_\lambda$.

Observe the following facts about ν^T:

(i) $\nu < \nu^T$
(ii) $\nu^T(T\langle\alpha\rangle \leftrightarrow \alpha) = t$.

Now define an evaluation on initial sentences thus:

$\nu(\alpha) = t$ if $\alpha \in X^+$
$\nu(\alpha) = f$ if $\neg\alpha \in X^+$
$\nu(\alpha) = b$ otherwise.

This is well defined since X^+ is consistent. Moreover, it is easy to check that if α is any sentence of L, if $\alpha \in X^+$ then $\nu(\alpha) = t$, and if $\alpha \notin X^+$ then $\nu(\alpha) = f$. By the above observations, ν^T verifies all members of X^+ (and hence X), the T-schema, but not φ. If ν^T is a model of the logic, we are home.

Whether or not this is so depends, of course, on the exact logic we are using, a matter I left partly indeterminate in Section 1. There are, however, many logics of the kind I indicated there for which this is true:

Lemma 4 *v^T is a model of many logics of the kind indicated in Section* 1.

Proof: The proof is tedious but straightforward. See Brady [7], Theorem 1.

A final comment: The proof actually proves something stronger than the stated theorem. Since the RM3 truth conditions for $\wedge$, $\vee$, and $\neg$ are essentially those of the strong Kleene three-valued logic, and Construction One is essentially the construction of a Kripke fixed-point, any grounded formula (in Kripke's sense) receives a classical truth value at v^T. Hence if α is grounded $\alpha \wedge \neg\alpha$ is not provable in the extensional theory. This gives correspondingly stronger forms of Theorem 1 and its Corollary.