RESEARCH NOTE

Reasoning about Truth

G. Priest
Department of Philosophy, University of Queensland, St. Lucia, Brisbane, Queensland, Australia 4047

Introduction

Any AI reasoning system with reasonable ambitions must have a way of describing, specifying or representing situations, states of affairs or wot not. Moreover, any AI reasoner that wants to perform cognitive reasoning, a central aspect of people's intelligence, must be able to express and reason about the cognitive attitudes that are taken, by the reasoner and others, to those representations: whether they are known, believed, true, provable, etc. (all quite distinct notions). Thus, for example, a person (or processor which is part of a distributed system), A, may reason "What B has told me in the past has usually been true. B now tells me that he has just heard from C. Hence I may reasonably believe that B has heard from C."

In the past several years, we have seen a number of studies concerning the AI handling of cognitive reasoning. (See, for example, any number of the papers in [9].) These studies have concentrated on only some cognitive attitudes, particularly, knowledge and belief. The notion of truth has been largely ignored; and this despite the fact that it is, arguably, one of the more central notions. For example, it is one of the distinguishing characteristics between knowledge and belief; it is a necessary condition of adequate proof; etc. The recent paper by Perlis [12], which does address the problem of reasoning about truth is therefore highly timely. Moreover, it creates yet another bond between AI and logic;[1] for reasoning about truth is something that has formed a central part of logical investigations this century (though, it should be said, this does not mean that there is a consensus in the area. Indeed, the area of the semantic paradoxes is perhaps the most contentious in modern logic).

[1] As will probably be clear from this paper, I write from the logicians' side of this partition; however, I hope to succeed, at least partly, in crossing it.

Artificial Intelligence **39** (1989) 231–244

The paper falls into two parts. In the first part I will describe Perlis' construction and argue that it has certain inadequacies. In the second part I will describe an approach to the problem of reasoning about truth which I think is preferable, and explain why. The approach draws on fairly recent work in a branch of logic called "paraconsistent logic." This part of the paper may therefore serve the function of introducing the reader to parts of logic, relevant to AI, of which they may not be aware.

1. Semantic Closure

The aim of an AI account of reasoning about truth (and related notions) is to produce some formally tractable way of representing the legitimate inferences that cognitive reasoners are wont to make about truth (and its associated notions). Obviously, the first prerequisite of such an account is to have a language with a predicate "is true" (which I will write as T). Syntactically, the predicate is to apply to the representations of states. We may conveniently take these to be sentences. This is not only simple and apparently adequate, but is also the dominant line that logicians have taken since Tarski, and so allows the application of any established logical technology.

What, however, makes this predicate a *truth* predicate? If we survey the inferences that characteristically involve the notion of truth, we find essentially two. We infer "$\underline{\alpha}$ is true" from α, and α from "$\underline{\alpha}$ is true" (where $\underline{\alpha}$ is a noun phrase which we can think of as a name for the sentence α). This suggests that the truth predicate, T, is characterized by the inference scheme:

$$T\underline{\alpha} \Leftrightarrow \alpha \ ,$$

which is called by logicians the T-scheme.[2]

It is perhaps rather surprising that such a trivial form of inference leads to trouble. Yet it does so. For all we need is a modicum of self-reference, obtainable in numerous ways, to find a state which claims that it, itself, is not true, i.e. a state, β, of the form $\neg T\underline{\beta}$. Applying the T-scheme to this, we obtain:

$$T\underline{\beta} \Leftrightarrow \neg T\underline{\beta}$$

and hence:

$$T\underline{\beta} \wedge \neg T\underline{\beta} \ .$$

[2] Since various notions of implication will play a role in this paper, let me comment briefly on my notation. I will use $\Rightarrow$ as a generic implication connective (where its precise properties are not at issue); $\supset$ as material implication (always defined using negation and disjunction); and $\rightarrow$ as a bona fide implication guaranteed to satisfy at least modus ponens. Their respective bi-implications are $\Leftrightarrow$, $\equiv$ and $\leftrightarrow$.

This contradiction, the liar paradox, in itself, might not be too much of a problem. After all, is it surprising that such counter-intuitive results follow from consideration of such a pathological state? Unfortunately, if we then throw in the principle of standard logic that anything can be deduced from a contradiction a real problem arises. For the reasoner who has gone through this process can now infer everything. Which is slightly too much.

A dodge that logicians have used since the 30s to avoid this problem is to separate out the system to whose entities truth is attributed (the object language) and the system which attributes truth (the metalanguage), and claim that these must be distinct. (The construction is due to Tarski, though it should be said, in fairness to him, that he did not think that ordinary language reasoning about truth worked in this way.) Thus, the claim that this very state is not true cannot be expressed at all. If it could be then, since it attributes truth, it must be in the metalanguage but, since it is that to which truth is attributed, it must be in the object language. Hence this is impossible.

In his paper Perlis argues, quite correctly, that this fix will not work. Cognitive representations are not intrinsically typed in this fashion, and any attempt to impose such a partition is not only artificial, but renders a great deal of perfectly correct and unproblematical reasoning impossible. This point is one of which logicians are now acutely aware, and most would agree that Perlis is quite right. As a result of this awareness, in recent years logicians have been investigating systems which aim at semantic closure, that is, systems that can talk about the truth of their own sentences. Characteristically, these approaches reject, or at least weaken the T-scheme. Such approaches all face well-known problems. (See [17; 18, Chapter 1] for references and a discussion of the various views.) Perlis produces a novel such approach, which is not only simple, but works within the framework of orthodox logic. To this I now turn.

2. The T^*-Scheme

Perlis' suggestion is simply to replace the T-scheme by what we will call the T^*-scheme:[3]

$$T\underline{\alpha} \equiv \alpha^* , \tag{1}$$

where α^* is the result of putting α into normal form (either conjunctive of disjunctive), and then replacing all occurrences of the form $\neg T\underline{\beta}$ by $T\underline{\neg\beta}$. Thus, in α^*, only atomic sentences are negated, and atomic sentences containing the truth predicate are never negated. Since a formula is logically equivalent to its normal form, this preserves the T-scheme for those sentences that do not contain T, and even for those that do, provided that the occurrences of T are not within the scope of a negation.

[3] In practice, Perlis also assumes other principles verified by the model construction to follow, such as (2) below, but which do not, as far as I can see, follow from T^*.

The T^*-scheme may look a little strange, and is certainly unlikely to occur to anyone a priori. The mystique may be removed, however, by considering its intuitive motivation.

As Perlis describes it, this builds on an idea of Kripke. First, we determine a class of sentences whose truth value may be fixed in a certain (transfinite) recursive fashion. These include (properly) all the sentences which do not contain T. Call these sentences *grounded*. The negation of a true grounded sentence is a false grounded sentence, and vice versa. On the Kripke construction, sentences that are not grounded are neither true nor false. The truth predicate applies truly to true sentences, falsely to false sentences, and neither-true-nor-falsely to ungrounded sentences. Consequently, $T\underline{\alpha}$ always has the same truth value (or lack of it) as α.

As an analysis of truth. Kripke's construction is problematic for a number of reasons. One is that it does not dispense with the object language/metalanguage distinction. This is because the fact that a sentence is not true cannot be correctly expressed in the language itself. For if α is neither true nor false then $\neg T\underline{\alpha}$ is neither true nor false, not true as required. Moreover, not only does the T-scheme fail (if α is neither true nor false, so is the instance of the T-scheme for it), but no reasonable approximation to it seems to be available.

Perlis' suggestion is, in effect, to define a new classical interpretation of the language, $\mathscr{I}$, such that the truth conditions of atomic sentences not containing T are the same as those in the Kripke interpretation, and those for sentences containing T are:

$T\underline{\alpha}$ is true in $\mathscr{I}$ iff α is Kripke-true
$T\underline{\alpha}$ is false in $\mathscr{I}$ otherwise.

It follows that all sentences are either true or false (in $\mathscr{I}$). Moreover, all Kripke-true sentences are true (in $\mathscr{I}$), and all Kripke-false sentences are false (as a simple induction shows). It therefore follows that

$$T\underline{\alpha} \supset \alpha$$

is valid in $\mathscr{I}$. The converse, however, is not. For if α is a true Kripke-neither sentence the consequent is true and the antecedent is false. Thus, the T-scheme, as is to be expected, fails in general. However, the T^*-scheme is valid. I leave the proof of this as an exercise for those familiar with Kripke's construction.[4] Thus, we see what semantics for the truth predicate really underlie the T^*-scheme.

[4] *Hint*. First show the result for α in normal form. The only nontrivial part of this argument concerns negated T-sentences; but here one can use the fact that $\neg T\underline{\beta}$ is Kripke-true iff $T\underline{\neg\beta}$ is Kripke-true. Next, observe that in the Kleene strong three-valued logic α is equivalent to its normal form α'. Hence, $T\underline{\alpha}$ is equivalent to $T\underline{\alpha}'$. Finally, observe that α'^* is just α^*.

To see what happens to the liar paradox in these semantics, note that

$$\neg(T\underline{\alpha} \wedge T\underline{\neg\alpha}) \tag{2}$$

is true (in $\mathcal{J}$) for every α. Next, note that if β is the liar sentence, the T^*-scheme gives:

$$T\underline{\beta} \equiv T\underline{\neg\beta} \quad (=\beta^*)\,.$$

Thus, by (2) and classical logic; $\neg T\underline{\beta}$, i.e., β. The inference to $T\underline{\beta}$ is, however, blocked. Thus, the contradiction does not arise; though we do have the rather odd $\beta \wedge \neg T\underline{\beta}$. Moreover, $\mathcal{J}$ provides a consistent interpretation of the T^*-scheme (and (2)), which establishes that no other contradictions arise in the theory.[5]

3. Criticisms of This Account

Though Perlis' solution to the problem of how to formalize reasoning concerning truth is neat, it will not work. It is wrong for both theoretical and practical reasons. Let us start with the theoretical reasons.

One objection to Perlis' construction is provided by the very fact that the T-scheme does not hold in general. There are a number of arguments to the effect that the T-scheme must hold for the truth predicate, that it, indeed, characterizes truth. Some of the arguments are as ancient as Aristotle, and some as modern as Frege. I will not rehearse them here, since I do not wish this to be a philosophical paper. (Some of these arguments can be found in [18, Sections 4.2–4.3].) Let us, therefore, move on to more technical objections.

One of the weaknesses of Kripke's construction is that it does not dispose of the object language/metalanguage distinction, as I noted above. But Perlis' construction is in exactly the same boat. For there is still no way in Perlis' construction of expressing the fact that a sentence is true (in $\mathcal{J}$)! The easiest way to see this is just to note that if the expressive power of the language is sufficiently strong then, since the logic is classical, we can apply Tarski's theorem to show that the set of true sentences cannot be defined by any formula with one free variable. Thus Perlis' own talk of truth (in interpretation $\mathcal{J}$) must be conceived of as occurring within a distinct metalanguage—which renders the content of his claim to have got rid of such a metalanguage (p. 312) unclear.

It follows, in particular, that the formula Tx does not express the claim that x is true. In fact, as the semantics make clear, $T\underline{\alpha}$ is true iff α is Kripke-true; but

[5] The model construction in Perlis' paper is somewhat different, but the final model is the same. Again, the proof is not difficult to find, and I leave it as an exercise. *Hint.* Show by induction that the extension of the truth predicate is the same at each level of the Kripke and Perlis hierarchies.

there are plenty of sentences that are true but not Kripke-true. As we noted, if β is the liar sentence, $\beta \wedge \neg T\underline{\beta}$ is true (in $\mathcal{I}$). Thus, β is one such formula.[6] Nor is T even a good approximation to truth (in $\mathcal{I}$), since some of the most fundamental facts about truth and T differ: For example, for every sentence either it or its negation is true; but there are α such that $T\underline{\alpha} \vee T\underline{\neg\alpha}$ is false. Similarly, if α is not true then its negation is true; but there are α for which $\neg T\underline{\alpha} \supset T\underline{\neg\alpha}$ fails. (For counter-examples to both, take α to be Kripke-neither.)

As we see, Perlis' account is theoretically flawed. It might be suggested, however, that this doesn't matter since the point of the construction is not a theoretical but a practical one. Specifically, the aim is to construct a formalization that can represent our ordinary reasoning concerning truth; and, it may be suggested, the T^*-scheme is, in fact, adequate for this. Indeed, Perlis provides some nice examples of inferences involving the T-scheme which are accounted for equally by the T^*-scheme.

Unfortunately, the theoretical inadequacies inevitably flow over into practical ones. Suppose, for example, that someone has the job of having destroyed all and only those books that contain some truth, i.e., they act on the command:

$$\exists x\,(Tx \;\&\; \text{book}(y) \;\&\; \text{occurs\_in}(x, y)) \equiv \text{destroy}(y).$$

They learn of book b that it contains inconsistent assertions on pp. 91 and 197. They then reason that one of these must be true, and hence that the book is to be destroyed. (The formalization of this is obvious.) The situation might be screwy, but the reasoning is perfectly sound and correct. Yet it cannot be represented in Perlis' approach, just because, as we noted two paragraphs back, $T\underline{\alpha} \vee T\underline{\neg\alpha}$ is not available.

Let me give another example, which concerns the failure of the T-scheme, and which is a slight modification of one of Perlis' own. Suppose we are given that anyone who speaks truly is a human. (Vampires, the other kind of inhabitant of Lower Slobbovia, always lie.) Two speakers, Od and Id, are heard to speak as follows:

Id: Everything I say is not true.
Od: What Id says is not true.

We can show that Od is human as follows. Suppose that what Id says is true. Then everything that Id says is not true. Hence, by *reductio*, what Id says is not true. But Od said just that. Hence he spoke truly. He is therefore a human. I leave a formalization of this to the interested reader. The important point to

[6]Perlis, in effect, admits that his truth predicate just means Kripke-true: "... T[rue] is to be taken to mean Kripke's sense, i.e., grounded and true ...." (p. 312).

note is just that having deduced that what Id says is not true ($\neg Tw$), to then infer that Od spoke the truth ($\neg Tw \supset T\underline{\neg Tw}$) is precisely an instance of (the half of) the T-scheme that does not hold on Perlis' account.

For good measure, we can also infer that Id is human. We have established that what Id said is not true. Thus, Id has said something true; hence he is human. Again, this reasoning cannot be represented in Perlis' construction, just because the principle $\neg T\underline{\alpha} \supset \neg\alpha$ fails.

Notice that there is nothing problematic about either of the above examples due to inconsistency. Both situations are quite consistent. (To see that the second is consistent just suppose that Od and Id are human and that Id has said (at least) one true thing.) There is therefore no paradoxical "funny business". We see that Perlis' construction does not allow for correct and unproblematic cognitive reasoning about truth. Hence, it is not only theoretically incorrect, but also practically inadequate.

4. A More Adequate Solution

I now wish to propose a more adequate solution. The T-scheme, we have seen, must be part of any adequate representation of cognitive reasoning. We have also seen that this gives rise to contradictions. It would appear that this must be accepted. What needs to be rejected is the view that everything may be deduced from a contradiction. After all, the fact that contradictions may arise in self-referential situations is not particularly surprising, or even worrying. What *is* worrying (and also surprising to someone who has not been indoctrinated by a course on Frege/Russel logic) is that once a contradiction has been inferred, everything follows: *ex contradictione quodlibet*. If this rule fails then there is no reason why the contradictions produced by the paradoxes of cognitive reasoning should not be allowed to stand: they need do no harm.[7]

Logics where *ex contradictione* fails are called *paraconsistent logics*, and there are many such, including relevant logics. Some are now familiar to logicians, but to computer scientists they may be less so. (Though some of the more elementary paraconsistent logics have appeared in the AI literature. See, e.g. Levesque [10], Fagin and Halpern [8].) I shall not attempt a review of such logics here. This can be found in [20; 21, Chapter 5].) Instead, I will describe one of the simplest and most natural such logics, LP (see Priest [15; 18, Chapter 5]), and show how it can be applied to the present situation.

LP is obtained by relaxing the classical assumption that sentences cannot be both true and false. Thus, an interpretation assigns to each atomic sentence

[7] In fact, it has been argued quite independently of the paradoxes of cognitive reasoning that inference engines suitable for reasoning from complex data should be paraconsistent. (See, e.g. Belnap [2].) For any but the most simplistic databases and rule systems are liable to be inconsistent. Further, since there is no decision procedure for inconsistency, there is no general and effective way that the inconsistencies can be weeded out. We therefore have to live with them.

one of the truth values $\{1\}$ (true and true only), $\{0\}$ (false and false only), and $\{1, 0\}$ (both). Truth conditions for nonatomic sentences are given in the familiar classical way, except that truth and falsity, now being independent, must each be considered. Thus, let us say that α is true (under an interpretation) iff 1 is in its truth value (under that interpretation); similarly, it is false iff 0 is in its truth value. Then, under an interpretation:

$\neg\alpha$ is true iff α is false,
$\neg\alpha$ is false iff α is true,

$\alpha \wedge \beta$ is true iff α is true and β is true,
$\alpha \wedge \beta$ is false iff α is false or β is false.

Disjunction is treated dually. $\alpha \supset \beta$ is defined as $\neg\alpha \vee \beta$. Quantifiers, as in normal accounts, are just thought of as (possibly infinitary) conjunctions and disjunctions over the domain of interpretation. As can be checked, these truth conditions are sufficient to give all formulas one of the three truth values. Logical consequence is defined in the standard way. An interpretation is a *model* of a formula iff the formula is true in that interpretation; it is a model of a set of formulas iff it is a model of every formula in the set; and

$\Sigma \models \alpha$ iff every model of Σ is a model of α.

It is a simple job to show that this logic is paraconsistent. Take the evaluation that makes p both true and false, and q false only. This makes $p \wedge \neg p$ true (and false) and q not true. (LP might be more familiar to some people as Kleene's strong three-valued logic with middle element designated.)

Let me mention, in passing, one variation on these semantics. This is obtained by allowing any subset of $\{1, 0\}$, including the empty set, to be a truth value. Otherwise details are the same. These semantics are Dunn's semantics for Anderson and Belnap's system of zero-degree entailment. (Discusssed by Belnap [2], used by Levesque [10].) The main difference between these two systems is that the three-valued system validates the law of excluded middle, $\alpha \vee \neg\alpha$, and indeed all classical tautologies, whilst the four-valued system has no logical truths. In the present context, I take this to be a distinct advantage for the three-valued system. For the aim is to capture ordinary reasoning about truth; and the law of excluded middle is an integral part of much of this.

LP has a number of simple proof theories. (For example, a tableau system is given by Lin [11].) I will give a natural deduction system, sound and complete with respect to these semantics. This is obtained by modifying a standard natural deduction system for first-order logic (that of Prawitz' [14]). The modification is simply to replace the ordinary negation rules ($\neg I$ and $\neg E$) by:

$$\begin{array}{c} \bar{\alpha} \\ \vdots \\ \dfrac{\neg\beta \quad \beta}{\neg\alpha} \end{array}\ \text{CON} \qquad \dfrac{}{\alpha \vee \neg\alpha}\ \text{LEM} \qquad \dfrac{\neg\neg\alpha}{\alpha}\ \text{DN}$$

where in CON, α is the only undischarged assumption, and no application of LEM occurs in its subproof. If we delete LEM we obtain a proof theory for zero-degree entailment. If we drop the restriction on CON, we obtain classical logic.

Having got the background logic sorted out, to provide a system to reason about truth, we merely add the two rules:

$$\frac{T\underline{\alpha}}{\alpha}\ TE \qquad \frac{\alpha}{T\underline{\alpha}}\ TI$$

where α is a closed formula. Let us call this system of rules TLP. As it stands, TLP is consistent (that is, no formula of the form $\beta \wedge \neg\beta$ is provable). This can be proved by noting that LP is consistent, and then observing that TLP can be collapsed into LP proofs merely by deleting T's and underlinings.) The consistency is due, however, to the fact that, so far, no self-referential machinery has been provided. As soon as this is provided, inconsistency results. Thus, suppose we can produce a formula, α, such that we can establish $\alpha \equiv \neg T\underline{\alpha}$; it is then a simple matter to deduce $\alpha \wedge \neg\alpha$. I leave this as an elementary exercise.

Although some contradictions are now provable, it would obviously be disastrous if all were (i.e. if the system was trivial). Fortunately, then, it can be shown that this is not the case. It is possible to construct nontrivial TLP models of first-order arithmetic (which certainly contain enough self-referential machinery), which show this. See Dowden [5]. In particular, anything that is Kripke-false is not provable.

5. The Disjunctive Syllogism and Minimal Inconsistency

The inference engine TLP is not subject to the objections I brought against Perlis' account. As may easily be checked, the T-scheme: $T\underline{\alpha} \equiv \alpha$ is provable; and because of the T-rules the T-predicate defines the set of truths in any interpretation. Thus, the account is not subject to Tarski's theorem concerning the indefinability of truth.

There is, however, one important objection. Just because the logic is paraconsistent, some inferences that are classically valid are LP-invalid. *Ex contradictione quodlibet*, of course, fails. However, this is well known to follow from simpler and less intuitively puzzling inferences. One of these must therefore have to fail. In fact, what fails is the disjunctive syllogism:

$$\alpha \quad \neg\alpha \vee \beta \ / \ \beta$$

detachment for material implication (sometimes called modus ponens, though this name is appropriate only if $\supset$ is a genuine implication connective, a claim that most logicians would now doubt—and something the very failure of detachment would show to be false). It might be thought that if the disjunctive syllogism fails then the system of inference is too weak to permit any interesting conclusions. This, however, is quite false as we have already seen. (The illusion may be caused by the fact that in many theorem provers the disjunctive syllogism is the only propositional rule.) In fact, the disjunctive syllogism is the only classically valid inference to fail (in the sense that if this is added to LP classical logic results). Yet it is reasonable to object to my proposal that the failure shows its inadequacy, since this inference is a part of our standard reasoning—about truth or anything else. Indeed, both the examples I gave in the previous section apply detachment to the T-scheme.

There are two ways to meet this objection, both involving extensions of the inferential machinery of LP. The simplest way is as follows. Observe that to obtain an LP counter-example to the disjunctive syllogism (or any other classically valid but LP-invalid inference) we must render the situation inconsistent (by making some formula both true and false). Now, it is both plausible and natural to take consistency as a default assumption. (For a defence of this see [18, Chapter 9].) In that case it makes sense to implement a nonmonotonic logic that implements this default, and which therefore allows the disjunctive syllogism provided that no pertinent inconsistency can be proved.

The simplest way of doing this is as follows. (I outline only the propositional case. Full details are given by Priest [19].) If ν is a propositional LP evaluation, let $\nu!$ be $\{p: p$ is a propositional parameter and $p \wedge \neg p$ is true under $\nu\}$. $\nu!$ is a measure of the inconsistency of an interpretation. Given a set of formulas, Σ, call ν a *minimally inconsistent* (*mi*) model of Σ iff (i) ν is a model of Σ, and (ii) if $\mu!$ is properly contained in $\nu!$ then μ is not a model of Σ. That consistency is a default assumption means that we suppose there to be no more inconsistency than we are forced to suppose; and a natural way of making this idea precise is simply to restrict ourselves to mi models. Thus, define the default consequence relation $\models_m$ as follows:

$$\Sigma \models_m \alpha \text{ iff every mi model of } \Sigma \text{ is a model of } \alpha.$$

This logic, LP_m, is nonmonotonic and paraconsistent. (As may easily be checked $\{\neg p \vee q,\ p\} \models_m q$; but $\{\neg p \vee q,\ p,\ p \wedge \neg p\} \not\models_m q$.) It extends LP, and gives all classical consequences if the premises are consistent. (See Priest [19] for proofs.) Hence, in consistent situations, the disjunctive syllogism and all other classical inferences are valid. In particular, both of the examples of Section 3 (and all other examples where inconsistency does not rear its ugly head) can be represented in terms of LP_m, since these situations are consistent. Moreover, even in inconsistent situations, LP_m still allows us to use the disjunctive syllogism provided only that the inconsistencies do not "get in the way." (Thus, for example, $\{p, \neg p \vee q, r \wedge \neg r\} \models_{\mathrm{m}} q$.) Hence $\mathrm{LP_m}$ validates all classical inferences except where inconsistency would make them naturally doubtful anyway.

6. Relevant Logic

The second way of meeting the objection is to extend the language of LP to include a genuine implication operator, $\rightarrow$, which satisfies (inter alia) detachment (modus ponens)—but not the principle $(\alpha \wedge \neg\alpha) \rightarrow \beta$. The *T*-scheme can now be formulated using this connective, and detachment from it becomes possible. The examples of Section 3, for example, can be represented in this way.

A genuine implication operator can be added to LP in numerous ways. Relevant logicians, in particular, have studied how to give the semantics of such an operator; and LP can be embedded in relevance logics.[8] As I observed in Section 4, the semantics of LP are a fragment of the semantics of zero-degree entailment. One possible approach is therefore to work with the extended semantics. This is unsatisfactory for two reasons. First, one looses all classical tautologies, such as the law of excluded middle (as I observed); second, and in any case, these semantics do not allow for nesting the connective $\rightarrow$, something one would surely want. It is better, therefore, to embed LP semantics in those of a full relevant logic.

This is not the place to go into the semantics of relevant logics in detail. (Details can be found in Dunn [7] or Routley et al. [22].) Let me, however, indicate one embedding. One kind of semantics for relevant logics is based on an algebraic structure of the form $\langle \mathscr{L}, \wedge, \vee, *, \Rightarrow, \mathscr{F} \rangle$, where $\langle \mathscr{L}, \wedge, \vee, * \rangle$ is a De Morgan lattice, $\mathscr{F}$ is a certain filter on the lattice and $\Rightarrow$ is a binary operation satisfying at least the condition: $a \leq b$ iff $a \Rightarrow b \in \mathscr{F}$. An algebraic evaluation is a map from formulas into the lattice such that $\wedge$, $\vee$, $*$, and $\Rightarrow$ are the interpretations of $\wedge$, $\vee$, $\neg$ and $\rightarrow$, respectively. Semantic consequence is defined in terms of membership-of-$\mathscr{F}$ preservation under all evaluations. Given any LP interpretation it is possible to construct such an algebra and embed the interpretation in it. Conversely, any such algebra can be cut down to an LP interpretation. This shows that LP is exactly the extensional (i.e., $\wedge$, $\vee$, $\neg$) fragment of the relevant logic. (Full details can be found in [16, Appendix].)

It is worth observing that many theories based on relevant logics can be shown to be nontrivial even when the *T*-scheme is available. To see this, note that Brady [4] has shown a large class of relevant logics to be nontrivial

[8] A suitable implication operator does not have to be relevant, however. See [18, Chapter 6].

(though inconsistent) when augmented by the abstraction scheme of naive set theory:

$$x \in \{y; \varphi\} \leftrightarrow \varphi(y/x) \qquad \text{(Abs)}$$

where / denotes substitution, and x is free for y in φ. Now, let $\underline{\alpha}$ be $\{x; \alpha\}$, where x is the least variable, in some standard enumeration, not occurring in α; and let Tx be $\emptyset \in x$. Then by (Abs):

$$T\underline{\alpha} \leftrightarrow \emptyset \in \{x; \alpha\} \leftrightarrow \alpha \ .$$

Hence, any such logic can nontrivially model the T-scheme.

7. Final Observations

The last two sections explain different ways of extending LP so that suitable detachments are available. Which of these is preferable on a given occasion may depend on the context. Having a genuine implication connective will not take care of a detachment if the major premise cannot be expressed as a genuine conditional; $\mathrm{LP_m}$ will (consistency permitting). But $\mathrm{LP_m}$ will not allow one to express an indefeasible connection between α and β (i.e., one where one can always get from α to β); having a genuine conditional will. Maybe, on occasions, it will be necessary to use both of these devices, though I have no example of this to offer. At any rate, I take it that, between them, they overcome the objection. Let me finish with four pertinent but miscellaneous comments.

(1) The approach to reasoning about truth that I have advocated accepts the T-scheme and uses a paraconsistent logic to accomodate the consequent inconsistencies. It might be suggested that another possible line is to accept the T-scheme, but accommodate the inconsistencies via some other mechanism, for example, by applying truth maintenance techniques. (See, e.g., Doyle [6].) Thus, for example, starting with an instance of the T-scheme $T\underline{\alpha} \equiv \neg T\underline{\alpha}$ marked IN, the TMS would mark it OUT as soon as it noted that from it and it alone $\alpha \wedge \neg\alpha$ follows.

It would be hubris to claim that no approach like this can be made to work. However, any such approach based on classical logic faces a pretty devastating objection based on Curry paradoxes. (See Priest [16; 18, Chapter 6].) Suppose that the connective $\Rightarrow$ satisfies both detachment and absorption ($\alpha \Rightarrow (\alpha \Rightarrow \beta)$ / $\alpha \Rightarrow \beta$). Suppose that we have suitable self-referential machinery, and thus, for an arbitrary formula, β, can construct a formula $T\underline{\alpha} \Rightarrow \beta$ whose name is $\underline{\alpha}$. The instance of the T-scheme for α is: $T\underline{\alpha} \Leftrightarrow (T\underline{\alpha} \Rightarrow \beta)$. Now, absorption gives $T\underline{\alpha} \Rightarrow \beta$; whence detachment from right to left gives $T\underline{\alpha}$. Putting these together, again by detachment, gives β. Thus an arbitrary formula follows from the T-scheme, even without the help of *ex contradictione*.

Thus, running a TMS on a set of assumptions that includes the T-scheme would be just like running a TMS on the set of assumptions that contains all formulas. The results would be just as arbitrary, and just as meaningless. They would reflect nothing but the order of backtracking.

Just for the record, it is worth noting that LP and certain relevant logics do not fall foul of Curry paradoxes, as the nontriviality results cited above show. This is because LP does not validate detachment for material implication, and suitable relevant logics do not contain absorption (though some relevant logics do). There is as yet no nontriviality proof for LP_m with the T-rules, but the Curry arguments certainly break down. Although $\{T\underline{\alpha} \equiv (T\underline{\alpha} \supset p)\} \models_m p$, $\{T\underline{\alpha} \equiv (T\underline{\alpha} \supset p),\ T\underline{\beta} \equiv (T\underline{\beta} \supset \neg p)\}$ gives neither p nor $\neg p$ in LP_m.[9]

(2) Secondly, while I am on the subject of truth maintenance: it might be suggested that contradictions play an essential role in belief revision (when we find one we revise) and that the use of a paraconsistent logic will stop this. This, however, is false. Using a paraconsistent logic does not prevent the revision of inconsistent beliefs. It just makes revision optional rather than mandatory. The suggestion does raise the question of when contradictions should occasion belief revision, but this is far too big an issue to take further here.[10] It is worth noting, however, that the nonmonotonic logic LP_m indicates one way in which beliefs may be revised in the light of new contradictions.

(3) The third observation concerns other paradoxes of cognitive reasoning. It is not only truth that is known to lead to paradoxes, but plausible conditions on belief, knowledge, proof and other intensional operators similarly lead to contradictions. (See, e.g., Asher and Kamp [1], Thomason [25], Perlis [13] for discussion and references.) It would take me too far afield in this paper to discuss these. But the fact that there is little agreement about how to handle them attests to the fact that all proposed solutions are problematic. Here I note only that these paradoxes in cognitive reasoning can be handled in exactly the same way as those concerning truth: we simply add the appropriate rules of proof for reasoning about knowledge, belief, etc., and allow the contradictions to stand, since they need do no harm.

(4) The final comment concerns the automated implementation of the systems I have described. Though it is simple enough to write algorithms for a number of these (for example, a proof-tree search will do for LP, and a model search will do for propositional LP_m) the problem of efficient algorithms remains largely to be investigated. Only for relevant logics has a start been made on this. Details can be found in Thistlewaite et al. [23, 24] and Bollen [3].

[9] The following are mi-model counter-examples for p and $\neg p$ respectively: p false only, $T\underline{\beta}$ true only, $T\underline{\alpha}$ both; p true only, $T\underline{\alpha}$ true only, $T\underline{\beta}$ both.

[10] The issue is taken up by Priest [18, Chapter 7].

ACKNOWLEDGEMENT

I would like to thank two anonymous referees of *Artificial Intelligence* for their helpful comments on a first draft of this paper. The paper was rewritten while I was a Project Visitor at the Automated Reasoning Project, Australian National University, and a version of it appeared as their Technical Report TR-ARP-2/88. I would also like to express my thanks to them.

REFERENCES

1. Asher, N. and Kamp, J., The knower's paradox and representational theories of attitudes, in: J.Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986).
2. Belnap, N., A useful four-valued logic, in: J.M. Dunn and G. Epstein (Eds.), *Modern Uses of Multiple-Valued Logic* (Reidel, Dordrecht, 1987).
3. Bollen, A., A relevant extension of PROLOG, Research Paper and Implementation, Automated Reasoning Project, Australian National University, Canberra (1988).
4. Brady, R., The non-triviality of dialectical set theory: in: G. Priest, R. Routley and J. Norman (Eds.), *Paraconsistent Logics* (Philosophia Verlag, 1988).
5. Dowden, B., Accepting inconsistencies from the paradoxes, *J. Philos. Logic* **13** (1984) 125–130.
6. Doyle, J., A truth maintenance system, *Artificial Intelligence* **12** (1979) 231–272.
7. Dunn, J.M., Relevance logic and entailment; in: D. Gabbay and F. Gruender (Eds.), *Handbook of Philosophical Logic* **3** (Reidel, Dordrecht, 1985).
8. Fagin, R. and Halpern, J., Belief, awareness and limited reasoning: Preliminary report, in: *Proceedings IJCAI-85*, Los Angeles, CA (1985).
9. Halpern, J.Y., *Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufman, Los Altos, CA, 1986).
10. Levesque, H., A logic of implicit and explicit belief, in: *Proceedings AAAI-84*, Austin, TX (1984).
11. Lin, F., Tableau systems for some paraconsistent logics (to appear).
12. Perlis, D., Languages with self-reference I: Foundations, *Artificial Intelligence* **25** (1985) 301–322.
13. Perlis, D., Languages with self-reference II: Knowledge, belief and modality, *Artificial Intelligence* **34** (1988) 179–212.
14. Prawitz, D., *Proof Theory* (Almqvist and Wiksell, 1965).
15. Priest, G., Logic of paradox, *J. Philos. Logic* **8** (1979) 219–241.
16. Priest, G., Sense, entailment and modus ponens, *J. Philos. Logic* **9** (1980) 415–435.
17. Priest, G., Semantic closure, *Stud. Logica* **43** (1984) 117–129.
18. Priest, G., *In Contradiction* (Nijhoff, Boston, MA, 1987).
19. Priest, G., Consistency by default, Tech. Rept. TR-ARP-3/88, Automated Reasoning Project, Australian National University, Canberra (1988).
20. Priest, G. and Routley, R., Introduction: Paraconsistent logics, *Stud. Logica* **43** (1984) 3–16.
21. Priest, G., Routley, R. and Norman, J. (Eds.), *Paraconsistent Logics* (Philosophia Verlag, 1988).
22. Routley, R. et al., *Relevant Logics and Their Rivals* (Ridgeview, 1982).
23. Thistlewaite, P., McRobbie, M. and Meyer, R., The KRIPKE automated theorem proving system, in: *Proceedings Eighth International Conference on Automated Deduction* (1986).
24. Thistlewaite, P., McRobbie, M. and Meyer, R., *Automated Theorem-Proving in Non-Classical Logics* (Wiley, New York, 1987).
25. Thomason, R., Paradoxes and semantic representation, in: J.Y. Halpern (Ed.), *Theoretical Aspects of Reasoning about Knowledge* (Morgan Kaufmann, Los Altos, CA, 1986).

Received June 1987; revised version received November 1988