# A Propositional Modal Logic
for the Liar Paradox

Martin Dowd

Keywords: Liar paradox, propositional modal logic

Abstract. A propositional modal language is defined wherein statements may refer to their own truth or falsity. A notion of validity is defined, and a complete proof system given. A system similar to both this and Solovay's modal logic of provability is also studied.

Author's note: The original version of this manuscript is dated 1986.

1. Introduction. The paradox of the liar is the statement "this statement is false". In various forms it has puzzled logicians and philosophers of natural language since the time of the Greeks. Within the last decade, the tools of mathematical logic have been brought to bear on this paradox. It is fair to say that a model which is satisfactory mathematically has been devised. Whether the issues raised by the liar paradox as a statement of natural language have been completely resolved is more open to debate.

Before embarking on an elucidation of the mathematical issues, let us consider the philosophical ones. Certainly there is no question that natural language can refer to statements, in a variety of ways. Statements may be abstract, such as a law of physics, or concrete, such as "your house is on fire", with a continuum of degrees from the abstract to the concrete. One can refer to either an abstract or a concrete statement.

There is also no question that one can attest to the truth of a statement, making a new statement. This is an example of what is called a modality; if S is a statement and $\Box$ is a modal operator then $\Box$S is a new statement. In this case, $\Box$ is "it is true that…"; a classic modality, considered by Aristotle, is "it is necessarily the case that…".

The issue of whether a statement can refer to itself is less clear. Examples of such in natural language are not clear cut, but do exist. For example, a speaker might say "I speak the truth"; presumably the listener may include this statement in the body of statements which the speaker is asserting to be true. One would not expect a man to say "I speak lies", except perhaps in a monologue. In this case the statement itself may be true, and excluded from the body of statements referred to. For another example, a speaker might switch from one language to another, and preface his remarks with an explanation of why he has switched. He might say, "I am speaking in…", which is a statement that refers to itself. More artifical examples are plentiful, such as "this statement consists of six words".

Ordinarily, statements which refer to the truth of other statements refer to statements made in the past. However borderline situations exist, and as Saul Kripke [Kr75] has observed situations which are apparently paradoxical can arise accidentally. For example speaker A might say "B always tells the truth", and speaker B might say "more than half of A's statements are lies". If it should happen that excluding his statement about B, exactly half of A's statements are true, a paradoxical situation arises, if the bodies of statements to which A and B are considered to be referring include these statements themselves.

Paradoxical collections of statements can be uttered by a single individual; for example a speaker might say, "the next statement I make will be true", and then "the last statement I made was false". The truth value of the first statement is undetermined when it is made, since it refers to a situation which is in progress. When the second statement is made the situation becomes paradoxical.

The liar paradox is the simplest form of such paradoxes. There seems no escape from the conclusion that it does not have a truth value. It does not attest to a state of affairs in the world, and so its truth or falsity cannot be determined from facts; indeed it cannot be assigned a truth value. The truth of the statement "this statement is true" likewise cannot be determined from facts, although in this case it can be arbitrarily assigned a truth value.

In the mathematical theory, the sense in which a statement refers to itself is well understood.

Mathematical statements are highly abstract. Indeed, there is a well understood set of formal statements which include any which a working mathematician might make. This set is countable, and can be considered to be a set of integers, as Kurt Godel essentially pointed out in 1933. Since among the statements of mathematics are statements about the integers, one can in this way devise statements which refer to themselves.

One is immediately prompted to ask whether "this statement is false" is a statement about the integers. Within the confines of standard first order logic, the answer is no. In the usual formal language of arithmetic one cannot refer to the truth of statements; that is, the truth modality cannot be defined in terms of primitive notions such as $+$ or $<$. Suitably formalized this fact is known as Tarski's theorem, and is proved by making use of the liar paradox. That is, one notes that if truth were definable then the liar paradox would be a statement of formal arithmetic, which is impossible since statements of formal arithmetic must have a truth value.

One is then prompted to ask whether the truth modality cannot simply be added to the language. The answer to this question is, as Kripke pointed out, yes, provided the basic semantics of formal logic is altered to allow for the possibility that a formal statement might not have a truth value.

Kripke's theory is derived from Godel's; Godel constructed the self-referential statement "this statement is not provable in Peano arithmetic". Provability in Peano arithmetic may be viewed as a modality, and as Godel showed this modality is definable in formal arithmetic. One can then conclude that Godel's statement is not provable in Peano arithmetic, since Peano arithmetic proves only true statements (this can be proved in set theory). Further the statement is therefore true.

Adding the truth modality to the language, one concludes immediately, by an argument similar to the proof of Tarski's theorem, that it cannot be assigned a meaning so that a statement S is true iff $\Box$S is. The modality must sometimes give no answer. In more detail, a unary predicate Tr is added to the language of arithmetic; the semantics of first order arithmetic is altered to allow for predicates to be partially defined on the integers. The meaning of a formula is defined by the usual recursion, except using 3-valued logic for the propositional connectives and quantifiers, so that the meaning of a formula is a partial predicate. An interpretation of Tr is called a fixed point if the meaning of every sentence is the value assigned it by Tr (fixing some well-behaved enumeration of the sentences).

Kripke shows that a least fixed point exists; that it is defined by a transfinite recursion which terminates at ω^1_{CK}, the least nonrecursive ordinal; and that there are fixed points other than the least. As has been observed, sentences fall into six classes, depending on what their truth value may be; namely always true, always false, always undefined, true or undefined, false or undefined, or anything.

This theory clarifies the nature of the liar paradox in formal logic; in systems which possess a truth modality and self-referential statements, some statements cannot be assigned a truth value. Other statements which are not assigned a meaning in the least fixed point may be consistently assigned a truth value; however such an assignment has no mathematical significance.

In mathematics, to have a truth predicate available a higher system is implicitly invoked. No situation is known where a truth predicate for the language itself is of any value; nor does it seem that there is one. Truth is usually understood to be a set, and can be defined in set theory. Of course, truth in set theory cannot be defined; this raises some issues in the foundations of set theory, but none of any significance to other questions. The statement that there is a set which satisfies the same sentences as the universe is an example of a statement which requires expanding the language of set theory. But it is irrelevant whether the truth modality refers to truth in the language of set theory, or in the expanded language, and the former avoids the necessity of dealing with the liar paradox. However, as Kripke shows, and as this paper further illustrates, it is quite easy to deal with.

The concern of this paper is systems simpler than Kripke's which contain both self-reference and a truth modality. These systems are propositional modal systems [Le77]. Smullyan [Sm57] considers such a system, although his is extremely simple. Solovay's [So76] propositional modal logic G is a related system; it does not include a self-reference mechanism, but self reference can be simulated by considering statements such as $p\Leftrightarrow\neg\Box p$.

2. Three-valued Logic. We use $\perp$ to denote the undefined truth value. The semantics of the propositional connectives when $\perp$ is taken to mean "unknown" are given in figure 1a. These are the truth tables which Kripke uses. A possible alternative is given in figure 1b. For statements such as "this statement is false" or "this statement is true", the semantics of $\wedge$ and $\vee$ is irrelevant. Mathematically, either of the two works equally well, and the results of this paper do not depend on which semantics is used. However, the truth tables of figure 1b conform to the principle that statements should be assigned a meaning only when necessary.

p	0	0	0	⊥	⊥	⊥	1	1	1		0	0	0	⊥	⊥	⊥	1	1	1
q	0	⊥	1	0	⊥	1	0	⊥	1		0	⊥	1	0	⊥	1	0	⊥	1
$\neg p$	1	1	1	⊥	⊥	⊥	0	0	0		1	1	1	⊥	⊥	⊥	0	0	0
$p\wedge q$	0	0	0	0	⊥	⊥	0	⊥	1		0	⊥	0	⊥	⊥	⊥	0	⊥	1
$p\vee q$	0	⊥	1	⊥	⊥	1	1	1	1		0	⊥	1	⊥	⊥	⊥	1	⊥	1
$p\Rightarrow q$	1	1	1	⊥	⊥	1	0	⊥	1		1	⊥	1	⊥	⊥	⊥	0	⊥	1
$p\Leftrightarrow q$	1	⊥	0	⊥	⊥	⊥	0	⊥	1		1	⊥	0	⊥	⊥	⊥	0	⊥	1

(a) (b)

Figure 1

We allow 0 and 1 as formulas. An assignment of the values 0,$\perp$, or 1 to the atoms of a formula yields such a value for the formula; if this value is 1 (resp. 0) the assignment is said to satisfy (resp. falsify) the formula. A formula is said to be *valid* if any assignment satisifes it. Validity in this sense, for semantics (a), is known as strong 3-valued logic [Kl52], to be distinguished from other non-classical forms of validity such as intuitionistic logic [RS68].

It is convenient to give the axiom system for validity as a sequent system. Recall that a sequent (see e.g. [Sm68]) is of the form $S\to T$ where S and T are sets of formulas. The left side (i.e. S) of the sequent is considered to be the conjunction of its formulas; and the right side the disjunction. We define the value of a sequent to be 1 if the value of the left side is less than or equal to the value of the right side, and otherwise 0, where the values are ordered $0 \leq\perp\leq 1$. (The conjunction (resp. disjunction) of the empty set is 1 (resp. 0).)

Definition. The system S3G consists of the following axioms and rules:

$$\begin{array}{ll}
\vdash p{\rightarrow}p & \vdash p, \neg p{\rightarrow}q, \neg q \\
\vdash 0{\rightarrow} & \vdash {\rightarrow}\neg 0 \\
\vdash \neg 1{\rightarrow} & \vdash {\rightarrow}1 \\
p{\rightarrow} \vdash \neg\neg p{\rightarrow} & {\rightarrow}p \vdash {\rightarrow}\neg\neg p \\
{\rightarrow}p;\ {\rightarrow}q \vdash {\rightarrow}p{\wedge}q & {\rightarrow}\neg p, \neg q \vdash {\rightarrow}\neg(p{\wedge}q) \\
p, q{\rightarrow} \vdash p{\wedge}q{\rightarrow} & \neg p{\rightarrow};\ \neg q{\rightarrow} \vdash \neg(p{\wedge}q){\rightarrow} \\
{\rightarrow}p, q \vdash {\rightarrow}p{\vee}q & {\rightarrow}\neg p;\ {\rightarrow}\neg q \vdash {\rightarrow}\neg(p{\vee}q) \\
p{\rightarrow};\ q{\rightarrow} \vdash p{\vee}q{\rightarrow} & \neg p, \neg q{\rightarrow} \vdash \neg(p{\vee}q){\rightarrow}
\end{array}$$

As usual in sequent systems, the axioms and rules are schemes. An instance is obtained by substituting arbitrary formulas for atoms, and adjoining arbitrary side formulas on either side of the arrow.

Theorem 1. A sequent is valid iff it is provable in S3G.

Proof. The rules have the property that an assignment satisfies the hypotheses iff it satisfies the conclusion. Hence any derivable sequent is valid (indeed in a derivation from hypotheses any assignment satisfying the hypotheses satisfies the conclusion). Further a valid sequent follows from a simpler valid sequent by a rule, unless it consists entirely of literals, in which case it is an instance of an axiom.

In the modal propositional logic where $\Box$ is to be interpreted as truth, the formula $\Box F$ receives the same value as F. In ordinary logic or indeed in strong 3-valued logic the $\Box$ operator has trivial behavior and can be removed. This is not so in systems considered later; in any case $\Box$ is easily added to S3G. Let MS3G be S3G together with the rules

$$\begin{array}{ll}
{\rightarrow}p \vdash {\rightarrow}\Box p & {\rightarrow}\neg p \vdash {\rightarrow}\neg\Box p \\
p{\rightarrow} \vdash \Box p{\rightarrow} & \neg p{\rightarrow} \vdash \neg\Box p{\rightarrow}
\end{array}$$

The same proof as above shows that this is sound and complete for the sequents with modal formulas.

There is a bothersome point concerning MS3G. Consider the formula $p{\Leftrightarrow}\neg\Box p$; in 2-valued logic this is false, i.e. no statement is equivalent to its falsity. In 3-valued logic this formula has value $\perp$ if p is a statement with value $\perp$ (in either of the above semantics). However if p is the liar paradox, it makes sense for the formula to be true; indeed that the liar paradox is paradoxical seems to require that it be true. One can escape this within the sequent formalism; $p{\rightarrow}\neg\Box p$ and $\neg\Box p{\rightarrow}p$ are what is true.

3. DSR formulas. A simple model including self-reference considers formulas to be directed graphs. A DSR formula is a modal propositional formula, written as a tree, except some leaves may be labelled $\Box$; such a leaf must have a backpointer to an ancestor node (called the target of the backpointer). We consider only the case where all other leaves are labelled 0 or 1.

Given a DSR formula, a labelling of the vertices with $\{0, \perp, 1\}$ is defined to be a fixed point if it obeys the truth tables for the propositional connectives and the modal operator. A fixed point τ is said to extend a fixed point ν if whenever $\nu(v)$ is defined (i.e. 0 or 1) for a vertex v then $\tau(v) = \nu(v)$. It is easy to see that a least fixed point exists. Begin by labelling each $\Box$ leaf with $\perp$, and each 0 or 1 leaf with 0 or 1. Then compute successively the labels of the interior vertices. If a target vertex is assigned 0 or 1, change the label of the leaves pointing to it and start over.

A DSR formula can be converted to a list of equations by introducing an atom for each target vertex. Choose a deepest (i.e. farthest from the root) target vertex v, and a new atom p. Let F be the subformula rooted at v, with backpointers replaced by p. Add $p{=}F$ to the list of equations, delete the descendents of v from the formula, and label v with p. Repeat this step until no target vertices remain. A fixed point corresponds to an assignment to the atoms which satisfies each equation in the list.

The least fixed point may be computed using the list of equations as follows. Determine the value of the right side of the first equation, when all occurrences of atoms are assigned value $\perp$; this determines the value of the atom of the first equation. (This will always be $\perp$ in semantics (b).) The remaining equations may be processed in such an order that the values of the inputs are known. If $\Box$ is understood to be provability, and the underlying logic is 2-valued, then the first equation reduces to one of the forms 0, 1, $p{\Leftrightarrow}\Box p$, or $p{\Leftrightarrow}\neg\Box p$. In the first case, p receives the value 0, and in the last three, the value 1. Continuing, the truth value of the formula may be determined. (It will be shown in the next section that the DSR formulas may be viewed as formulas of arithmetic.)

It is convenient to generalize DSR formulas by starting with a rooted dag rather than a tree. Note that in the fixed point of a tree, identical subformulas (where a subformula is understood to include all backpointers from its leaves) need not have identical labellings. In the least fixed point, however, this is the case. Further each vertex of the tree corresponding to a dag has the same label in the least fixed point of the tree as it does in the least fixed point of the dag.

4. TSR1 Formulas. A TSR1 formula is a propositional formulas whose atoms are of the form $\mathrm{Tr}(t)$ where t is a term. A term is either the variable x; $\langle F\rangle$ where F is a formula; $\mathrm{sr}(x)$; $\mathrm{sr}(\langle F\rangle)$; or $[F]$. A term denotes a formula or a function from formulas to formulas, as follows. The variable x denotes the identity map; $\langle F\rangle$ denotes F; $\mathrm{sr}(x)$ denotes the map $G\mapsto G\frac{\langle G\rangle}{x}$; and $[F]$ the map $G\mapsto F\frac{\langle G\rangle}{x}$. A term t is closed if it denotes a formula; this formula is denoted F_t.

For example "this statement is false" is $\neg\mathrm{Tr}(\mathrm{sr}(\langle\neg\mathrm{Tr}(\mathrm{sr}(x))\rangle))$. The statement " 'this statement is false' is true" is

$$\mathrm{Tr}([\neg\mathrm{Tr}(\mathrm{sr}(\langle\mathrm{Tr}([\neg\mathrm{Tr}(\mathrm{sr}(x))])\rangle))]).$$

Say that a formula is closed if its terms are. If t is $\mathrm{sr}(x)$ or $[F]$ then any formula in the range of the map denoted by t is closed. Thus there is no loss of generality in defining the terms as they are defined. Note that the modal operator $\Box F$ is an abbreviation for $\mathrm{Tr}(\langle F\rangle)$.

Certain closed formulas, such as $\mathrm{Tr}(\langle\mathrm{Tr}(x)\rangle)$, are not meaningful and will be considered malformed. Recursively a formula is malformed if it is open or contains an atom $\mathrm{Tr}(t)$ where F_t is malformed. A formula is a wff if it is not malformed. The definition of a malformed formula is a well founded recursion, but a procedure to determine if a formula is malformed cannot simply call itself recursively, since it might then be invoked for a formula for which it has already been invoked.

It is however decidable whether a TSR1 formula is well-formed. The following procedure takes as input a TSR1 formula, and either produces as output a DSR formula or reports that the input formula is malformed.

```
SV is a substitution value which may be null and is so initially.
BT is a pointer to a node of the output formula and initially is null.
Form($F$)
    Copy $F$ down to atoms.
    For each atom Tr($t$) call Atom($t$).
Atom($t$)
    If $t$ is open and SV is null report malformed.
    Output $\Box$.
    Case $t$ of
        $\langle F\rangle$: Push BT and SV onto a stack; set SV to null;
          call Form($F$); restore BT and SV.
        sr($\langle F\rangle$): Push BT,SV; set BT to current output node;
          set SV to $F$; call form($F$); restore BT,SV.
        $x$: Set $F$ to SV; push BT,SV; set SV to null; call Form($F$);
          restore BT,SV.
        sr($x$): Output backpointer to BT.
        [$F$]: Call Form($F$).
    Endcase
```

Conversely DSR formulas can be translated to TSR1 formulas. Choose a deepest target vertex; a TSR1 formula will be constructed for the subtree rooted at this vertex. This step is then repeated until all target vertices are replaced by TSR1 formulas. To construct the TSR1 formula for a tree with all backpointers to the root, translate $\Box F$ to Tr($\langle F'\rangle$) where F' is the translation of F; the $\Box$'s at the leaves are tranlated to Tr(x). Now use sr to make the formula self-referential.

The semantics of the TSR1 formulas is similar to that of Kripke's model. An assignment τ to the wff's induces an assignment τ' to the well-formed atoms, namely $\tau'(\mathrm{Tr}(t)) = \tau(F_t)$; τ' is then extended to the wff's using the truth tables for the connectives. The partial order $\tau \leq \nu$ on assignments is defined as usual; we say ν extends τ. Define τ to be a fixed point if $\tau' = \tau$. The usual facts now hold, viz. if $\tau \leq \nu$ then $\tau' \leq \nu'$; if $\tau \leq \tau'$ there is a least fixed point τ^* extending τ; and there is an overall least fixed point, namely τ^* where $\tau(F) = \perp$ for each wff F.

Given a TSR1 formula and its DSR translation, for each fixed point of the DSR formula there is a TSR1 fixed point agreeing with it. Represent the TSR1 formula as a dag by identifying common subformulas. Each vertex of the DSR tree corresponds to a vertex of the TSR1 tree; perform the corresponding identification in the DSR tree. Each fixed point of the DSR formula induces an assignment to certain wff's; assign the remaining wff's $\perp$ and take the least fixed point. Since the entire set of fixed points of a DSR formula can be computed, all questions regarding the TSR1 formulas are decidable.

5. TSR2 and TSR3 Formulas. By expanding the terms more general classes of formulas can be obtained. These still correspond to formulas of augmented first order arithmetic, or indeed if the modality is a definable one to formulas of arithmetic. The TSR3 formulas allow arbitrary terms built from an infinite number of variables; the constants $\langle F\rangle$; and for each variable x the binary operation s_x, denoting the map $(G, H) \mapsto G\frac{\langle H\rangle}{x}$. Each closed term t denotes a formula F_t. It is readily seen how to translate TSR1 formulas to TSR3 formulas.

A TSR formula (i.e. formula in a family of this type) can be associated with an infinite tree, obtained by a succession of approximations. The original formula is the first approximation; successive approximations are obtained by replacing closed terms t in atoms Tr(t) by $\langle F_t\rangle$. The infinite tree of a TSR1 formula is periodic, i.e. along any branch the tree repeats. By Konig's lemma these are exactly the infinite trees which can be descibed by trees with backpointers.

The TSR3 infinite trees are not periodic, as the example

$$\mathrm{Tr}(\mathrm{s}_x(\langle \mathrm{Tr}(\mathrm{s}_x(x,x))\rangle, \langle \mathrm{Tr}(\mathrm{s}_x(x,\mathrm{s}_x(\langle \neg\mathrm{Tr}(x)\rangle, x)))\rangle))$$

shows. Nonetheless we conjecture that it is decidable if a TSR3 formula is well-formed, and that the possible truth values in fixed points can be determined. We also conjecture that if $\Box$ is interpreted as provability in Peano arithmetic, then the truth of these statements as statements of arithmetic is decidable.

To conclude this section we give a class TSR2 of formulas intermediate between TSR1 and TSR3 for which the infinite trees are periodic. Fixing an integer k, there are k variables $x_1, \ldots, x_k$; the constants $\langle F\rangle$; and the operator $\mathrm{sub}(t_0, t_1, \ldots, t_k)$ where t_i must be a variable or constant. The operation sub denotes the map $(F_0, F_1, \ldots, F_k) \mapsto F_0\frac{\langle F_1\rangle,\ldots,\langle F_k\rangle}{x_1,\ldots,x_k}$. The following procedure translates TSR2 formulas to DSR formulas. It maintains a stack of records $[F_0, F_1, \ldots, F_k, \mathrm{BT}]$, where the F_i are subformulas of the input and BT is a pointer to a node in the output tree. Also SV is an array of k entries, each a subformula of the input; initially these are null.

```
Form($F$)
    Copy $F$ down to atoms.
    For each atom $\mathrm{Tr}(t)$ call $\mathrm{Atom}(t)$.
Atom($t$)
    If $t$ contains $x_i$ free and $\mathrm{SV}_i$ is null report malformed.
    Output $\Box$.
    Case $t$ of:
        $\langle F\rangle$: Set $G_0$ to $F$ and $G_i$ to null, $1 \le i \le k$.
        $x_i$: Set $G_0$ to $\mathrm{SV}_i$ and $G_i$ to null, $1 \le i \le k$.
        $\mathrm{sub}(t_0, \ldots, t_k)$: Set $G_i$ to $F$ if $t_i$ is $\langle F\rangle$,
            or to $\mathrm{SV}_j$ if $t_i$ is $x_j$, $0 \le i \le k$.
    Endcase
    If a record $[G_0, \ldots, G_k, \mathrm{BT}]$ exists on the stack,
        output a backpointer to BT and return.
    Add a record $[G_0, \ldots, G_k, \text{current output node}]$ to the stack;
        push SV; set $\mathrm{SV}_i$ to $G_i$, $1 \le i \le k$; call $\mathrm{Form}(G_0)$;
        restore SV; remove the top stack record; return.
```

The procedure terminates, since there are only finitely many possible values of the $k+1$-tuple $[G_0, \ldots, G_k]$. Even though the TSR3 formulas are not periodic, they seem to be reasonably simple and can possibly be described in some manner analogous to DSR formulas.

6. TSR Validity. Define a TSR sequent to be valid if it is true in any fixed point. A sequent $\rightarrow F$ or $F\rightarrow$ is valid iff it is true in the least fixed point, but this is not the case in general. Validity is decidable for TSR1 or TSR2 formulas; we conjecture that it is for TSR3 formulas also.

Let TS3G be S3G together with the rules

$$\begin{array}{ll} \rightarrow\langle F_t\rangle \vdash \rightarrow\mathrm{Tr}(t) & \rightarrow\neg\langle F_t\rangle \vdash \rightarrow\neg\mathrm{Tr}(t) \\ \langle F_t\rangle\rightarrow \vdash \mathrm{Tr}(t)\rightarrow & \neg\langle F_t\rangle\rightarrow \vdash \neg\mathrm{Tr}(t)\rightarrow \end{array}$$

In the axioms and rules of S3G, an atom is understood to stand for any TSR formula. The rules have the property that a fixed point satisfies the hypotheses iff it satisfies the conclusion, so the system is sound. Completeness does not follow immediately, since the hypothesis need not be simpler than the conclusion; however it can be shown by an infinitary argument using Hintikka sets.

Lemma 2. The rules of TS3G are reversible, in the sense that if the conclusion is provable then the hypotheses are.

Proof. By induction on the length of the proof of the conclusion. Consider the cases where a formula of one side is $\neg\neg F$, $F \wedge G$, $F \vee G$, or $\mathrm{Tr}(t)$. For each case, if the the last sequent was derived by a rule, either the selected formula was the formula involved, in which case the claim is immediate; or the selected formula was a side formula, in which case by induction the selected formula may be replaced by its precursor(s) in the hypotheses, and the hypotheses may then be derived. Finally if the last sequent was an axiom then one can check that either the hypotheses are axioms, or follow immediately from axioms.

Define an assignment τ to the wff's to be regular if $\tau(0) \leq 0$, $\tau(\neg 1) \leq 0$, $\tau(1) \leq 1$, $\tau(\neg 0) \leq 1$, $\tau(\neg F) \leq \neg\tau(F)$, $\tau(F \wedge G) \leq \tau(F) \wedge \tau(G)$, $\tau(F \vee G) \leq \tau(F) \vee \tau(G)$, and $\tau(\mathrm{Tr}(t)) \leq \tau(F_t)$.

Lemma 3. If τ is regular then $\tau \leq \tau'$.

Proof. By induction on F, $\tau(F) \leq \tau'(F)$.

Define a pair (S, T) of sets of wff's to be deduction-free if $\not\vdash S' \to T'$ for any finite $S' \subseteq S$, $T' \subseteq T$. Define the assignment τ by

$\tau(F) = 1$ if $F \in S$, $\neg F \notin S$, $F \notin T$;

$\tau(F) = 0$ if $F \notin S$, $\neg F \in S$, $\neg F \notin T$;

$\tau(F) = 0$ if $F \in T$, $\neg F \notin T$, $\neg F \notin S$;

$\tau(F) = 1$ if $F \notin T$, $\neg F \in T$, $F \notin S$;

$\tau(F) = \perp$ otherwise.

Define (S, T) to be a Hintikka pair if it is deduction-free and τ is regular. In this case τ^* falsifies every sequent $S' \to T'$, $S' \subseteq S$, $T' \subseteq T$, S', T' finite.

Theorem 4. If (S_1, S_2) is deduction-free there is a Hintikka pair (T_1, T_2) with $S_1 \subseteq T_1$, $S_2 \subseteq T_2$.

Proof. Enumerate S_j as $F_{j1}, F_{j2}, \ldots$. T_j will be enumerated in stages; eventually $T_j = G_{j1}, G_{j2}, \ldots$. If T_j^i is T_j after i stages then $(S_1 \cup T_1^i, S_2 \cup T_2^i)$ will be deduction-free. At stage i, do the following.

```
Append $F_{ji}$ to $T_j$, $j = 1, 2$
Append to $T_1$, case $G_{1i}$ of
  $\mathrm{Tr}(t)$: $F_t$;
  $\neg\mathrm{Tr}(t)$: $\neg F_t$;
  $\neg\neg F$: $F$
  $F \wedge G$: $F$ and $G$;
  $\neg(F \wedge G)$: if $(S_1 \cup T_1 \cup \{\neg F\}, S_2 \cup T_2)$ is deduction-free,$\neg F$;
            else $\neg G$;
  $F \vee G$: if $(S_1 \cup T_1 \cup \{F\}, S_2 \cup T_2)$ is deduction-free, $F$;
            else $G$;
  $\neg(F \vee G)$: $\neg F$ and $\neg G$;
```

```
Endcase.
Complete the appendage, i.e. if $F$ was appended, and if its complement
   was already in $T_1$, append the complement of $G_{1i}$.
Perform the above two steps for $G_{2i}$, with the roles of $\wedge$ and $\vee$
   reversed and other obvious modifications.
Append $T_j$ to itself, $j = 1, 2$.
```

Using the reversability of the rules it is easy to show that $(S_1 \cup T_1^i, S_2 \cup T_2^i)$ is deduction-free after stage i. (Note that it is thus unnecessary to consider the case F in T_{3-j} when completing the appendage.) Finally we claim that if in one of the cases of the definition of regularity the right side is $\perp$, then the left side is also. For example, if either $F, \neg F \in T_1$ or $G, \neg G \in T_1$ then $F{\wedge}G, \neg(F{\wedge}G) \in T_1$; and if, say, $F, \neg F \in T_1$ and $G \notin T_1$ then F was used when $F{\vee}G$ was processed, so $\neg(F{\vee}G) \in T_1$. The last step of stage i ensures that complementary subformulas propogate properly no matter when they are introduced.

Corollary 5. TS3G is complete.

7. TG. Replace Tr by Pr in TSR formulas. Let TG be the system consisting of propositional logic over the connectives $\{\neg, \Rightarrow\}$, with modes ponens the only rule, and the following modal axioms and rules.

$F \vdash \mathrm{Pr}(\langle F\rangle)$

$\vdash \mathrm{Pr}(\langle F{\Rightarrow}G\rangle){\Rightarrow}\mathrm{Pr}(\langle F\rangle){\Rightarrow}(\langle G\rangle)$

$\vdash \mathrm{Pr}(\langle F\rangle){\Rightarrow}\mathrm{Pr}(\langle \mathrm{Pr}(\langle F\rangle)\rangle)$

$\vdash \mathrm{Pr}(t){\Rightarrow}\mathrm{Pr}(\langle F_t\rangle)$

$\vdash \mathrm{Pr}(\langle F_t\rangle){\Rightarrow}\mathrm{Pr}(t)$

As usual $\mathrm{Pr}(\langle F\rangle)$ is abbreviated $\Box F$.

Theorem 6. $\Box F{\Rightarrow}F \vdash F$ and $\vdash \Box(\Box F{\Rightarrow}F){\Rightarrow}\Box F$.

Proof. Let G be $\mathrm{Pr}(\mathrm{sr}(\langle \mathrm{Pr}(\mathrm{sr}(x)){\Rightarrow}F\rangle)){\Rightarrow}F$; then we have $\vdash G{\Rightarrow}(\Box G{\Rightarrow}F)$ and $\vdash (\Box G{\Rightarrow}F){\Rightarrow}G$. The theorem follows by well-known arguments [Bo79].

Define a truth assignment τ to the atoms by letting $\tau(\mathrm{Pr}(t)) = 1$ iff $\vdash F_t$. Define $\models F$ iff τ satisfies F.

Theorem 7. If $\vdash F$ then $\models F$. Proof. Induction on the length of the proof. If F is a propositional axiom clearly τ satisfies F; similarly if F follows by modes ponens. If F is $\Box G$ and follows by necessitation then since $\vdash G$, $\models F$ by definition. If F is $\Box(G{\Rightarrow}H){\Rightarrow}\Box F{\Rightarrow}\Box G$, and if $\models \Box(G{\Rightarrow}H)$ and $\models \Box G$, then $\vdash G{\Rightarrow}H$ and $\vdash G$, so $\vdash H$ and so $\models \Box H$. If F is $\Box G{\Rightarrow}\Box\Box G$ and if $\models \Box G$ then $\vdash G$, so $\vdash \Box G$ and so $\models \Box\Box G$. Finally $\models \mathrm{Pr}(t)$ iff $\vdash F_t$ iff $\models (\langle F_t\rangle)$.

In particular TG is consistent. An alternative consistency proof can be given using canonical structures for propositional modal systems (cf. [Le77]). The truth assignment which assigns every atom false witnesses that the canonical structure for TG exists; TG is therefore consistent. TG is not complete, for the Godel statement $\neg\mathrm{Pr}(\mathrm{sr}(\langle \neg\mathrm{Pr}(\mathrm{sr}(x))\rangle))$ is a wff. Denoting it by F, by obvious arguments if $\vdash F$ then $\vdash \neg\Box F$ and $\vdash \Box F$; hence $\nvdash F$, so $\models \neg\Box F$, so $\models F$; and finally if $\vdash \neg F$ then $\models \Box F$, a contradiction, so $\nvdash \neg F$. It also follows that $\vdash \Box F$ implies $\vdash F$; for if $\vdash \Box F$ then $\models \Box F$ so $\vdash F$.

For TSR1 or TSR2 formulas the theory of TG is decidable. Define $\mathrm{sf}(F)$ to be the least family of wff's containing F; containing G (and H) if it contains $\neg G$, $G{\Rightarrow}H$, or $\Box G$; and containing $\Box F_t$

if it contains $\mathrm{Pr}(t)$. From the algorithms for translating TSR1 or TSR2 formulas to DSR formulas it follows that $\mathrm{sf}(F)$ is finite. Then as in theorem 3.1 of [Le77] if F is consistent with TG there is a finite modal structure containing a world satisfying F. To decide if $\vdash F$, enumerate the theory of TG looking for F, and the finite modal structures looking for one containing a world satisfying $\neg F$. It is also decidable whether $\models F$; by the list of equations computation it suffices to observe that τ satisfies “I am provable” and “I am not provable”. We conjecture that the theory of TG, and whether $\models F$, are decidable for TSR3 formulas.

Let TG$'$ be TG, with the formulas $\Box F \Rightarrow F$ added as axioms; let $\vdash'$ denote provability in TG$'$. If $\models \Box F$ then $\vdash F$ so $\models F$, and so $\models \Box F \Rightarrow F$. It follows that if $\vdash' F$ then $\models F$; in particular TG$'$ is consistent.

Theorem 8. If $\models F$ then $\vdash' F$, for the TSR1 or TSR2 formulas. Proof. If $\vdash F \Leftrightarrow \Box F$ then $\vdash' F$ since Löb's theorem holds in TG. If $\vdash F \Leftrightarrow \neg\Box F$ then

$\vdash \Box F \Rightarrow \Box\Box F$, $\vdash \Box F \Rightarrow \Box\neg\Box F$, $\vdash \Box F \Rightarrow \neg\Box F \Rightarrow 0$, $\vdash \Box\Box F \Rightarrow \Box\neg\Box F \Rightarrow \Box 0$,

and so $\vdash \Box F \Rightarrow \Box 0$. But $\vdash' \Box 0 \Rightarrow 0$, so $\vdash' \neg\Box F$, so $\vdash' F$. The list of equations computation may thus be carried out in TG$'$.

We conjecture that this theorem holds for the TSR3 formulas. **8. Conclusion.** By isolating the essential ingredients of self reference and a truth modality, using the machinery of propositional modal logic, problems related to the liar paradox can be studied in an abstract setting, without the need for arithmetic to provide the self-reference mechanism. Indeed problems of self-reference can be studied for other modalities, such as provability.

Tarski [T39] discusses some general facts in such systems. In particular, consider a Hilbert type system (i.e. using formulas rather than sequents) which satisfies the axioms and rules

$$\begin{array}{ll} F, F \Rightarrow G \vdash G & \vdash \Box(F \Rightarrow G) \Rightarrow \Box F \Rightarrow \Box G \\ F \Rightarrow G, F \Rightarrow \neg G \vdash \neg F & \vdash \Box F \Rightarrow \neg\Box\neg\Box F \\ F \vdash \Box F & \vdash \Box\neg F \Rightarrow \neg\Box\Box F. \end{array}$$

Then if $F \Rightarrow \neg\Box F$ and $\neg F \Rightarrow \Box F$ then $\neg\Box F$ and $\neg\Box\neg F$.

The proof of theorem 7.2 involves the modalities $\vdash$ and $\models$. One example of a possible additional system would axiomatize these modalities in a common system, which applied to itself, and included self-reference. Other examples can be obtained by considering additional modalities or predicates on the formulas, such as ”has no truth value”. The statement ”this statement has no truth value” is presumably false. The predicate ”has no truth value” does not appear to be definable from truth, so has to be added even to Kripke's theory.

Another example is ”there are no self-referential statements”. Here, besides the new predicate quantification over statements must be added to the model. Further it appears that there is more than one kind of self-reference; this statement refers to itself in one way, but the ”self-referential” predicate holds of statements which refer to themselves in another. Any universal statement about statements, such as ”every sentence must contain a verb”, refers to itself in the first way. An interesting question is what kind of ”self-referential” predicates can be defined in Kripke's theory. Finally it might be of some interest to consider how these simple models which deal only with the behavior of predicates applied to statements might be incorporated into more complex models of natural language.

References.

BJ80. G. Boolos and R. Jeffrey, *Computability and Logic*, ambridge University Press.

Bo79. G. Boolos, *The Unprovability of Consistency*, Cambridge University Press.

KL52. S. Kleene, *Introduction to Metamathematics* North-Holland.

Kr75. S. Kripke, "Outline of a theory of truth", Journal of Philosophy LXXII (1975) 690–716.

Le77. E. J. Lemmon, *An Introduction to Modal Logic*, American Philosophical Quarterly Monograph no. 11, 1977.

RS68. H. Rasiowa and R. Sikorski, *The Mathematics of Metamathematics*, Polska Akademia Nauk.

Sm57. R. Smullyan, "Languages in which self-reference is possible", J. Symb. Logic 22 (1957) 55–67.

Sm68. R. Smullyan, *First Order Logic*, Springer-Verlag.

So76. R. Solovay, "Provability interpretations of modal logic", Israel Journal of Mathematics 25 (1976) 287–304.

T39. A. Tarski, "On undecidable statements in enlarged systems of logic and the concept of truth", J. Symbolic Logic 4 (1939) 105–112.