Nonmonotonic Reasoning and Self Reference

1. Introduction

During the last decade many major approaches to nonmonotonic reasoning have been developed. The underlying motivation of each of these approaches is a framework which allows more inferences to be drawn than it is deductively possible. Although these approaches look so different in terms of motivation and in their formal machinery, they all seem to have one factor in common: the employment, whether implicit or explicit, in some way or another a term like "inability to infer" or "cannot infer" or "fail to derive" within the inference mechanism of the system. Nonmonotonicity then comes about from the fact that after increasing the amount of information available as premises, some of the conclusions that can be drawn before are no longer derivable. As then what could or could not be inferred initially would have to change in the light of the new information.

In this paper we shall make an attempt at capturing such form of reasoning in a way that seems more natural than any of the other approaches. Although these approaches have based their accounts on some variation of a term like "cannot infer", they have not made the attempt to encode it in the language of the resulting system. Instead, they emphasized the interpretation of the "modal operator" or the "rule" which in some way captures or employs the term, forgetting the main motivation.

On the other hand, encoding a term like "cannot infer" in the language of a logical system (whose inference notion is being encoded) results in an added complexity, namely self-referentiality. Some of the logicians and researchers who are interested in "semantic paradoxes" [via personal communication] tend to believe that the two issues (logics of "truth" and nonmonotonic logics) are orthogonal. However, they do seem to agree that the revision processes employed in some theories of truth bear some similarities to those used in some approaches to nonmonotonic reasoning. We shall, in this paper, make an attempt to bring forward some of these similarities and discuss whether logics of truth and nonmonotonic is a name in L.

It is essential that we should be able to express sentences like "A is not a theorem of S" and to use in the inference mechanism of the underlying system. However, it does not seem necessary, unless the aim is a general theory of revisable inference, that we need to express sentences such as "A equals to the assertion that A is not a theorem of S" which are akin to the liar sentences. We shall discuss the impact of self-referentiality in more details in later sections.

4. Revisable inference through fixed-points

This theory, to which we shall refer as KFG (Kleene-Feferman-Gilmore), is essentially due to Gilmore [4] and Feferman [3] which they have developed for *truth.* The theory although classical in its *external* logic (i.e. the logic of wffs is classical) has a residue of three-valued logic in its *internal* logic of the predicate T which will be interpreted here as expressing theoremhood-of-S. The key idea of the theory is that the extension of "theoremhood-of-S" should be constructed in a predicative layer-by-layer fashion. One starts by assigning "theoremhood-of-S" to those sentences of L which do not contain the predicate "T" according to Kleene semantics for L. At each further stage, the notion of "theoremhood-of-S" for sentences of the previous stage, together with the Tarski scheme, specifies how the notion of "theoremhood-of-S" is to be extended to the next stage.

Kripke [8] has provided a formal account of this analysis. We shall, at this stage, neither present a model theoretic account nor a axiomatization of KFG here, but we shall briefly look at KFG as a modal theory.

4.1. KFG as a modal theory

We now investigate some more of the consequences of KFG and in particular the derivability of certain modal principles.

Theorem 4.1. The following are provable in KFG

(T) $T(A) \to A$

(S4) $T(A) \to T(T(A))$

(IP) $T(A \to B) \to (T(A) \to T(B))$

(BAR) $\forall x T(A) \to T(\forall x A)$

(N) $T(A) \to T(C(A))$

(L) $T(C(A)) \to T(A)$.

The proof is a simple exercise.

(T), (IP), (BAR) and S4 are the characteristic axioms of S4 modal logic. As regards the S5 axiom the following is provable:

$$(C(A) \rightarrow T(C(A))) \rightarrow T(A) \vee T(\neg A)$$

Thus, if S5 is derivable from KFG, then S would decide about every sentence A. I.e. for any sentence A, it is either the case that A is a theorem of S or ¬A is a theorem of S.

Similarly, the addition of the rule of necessitation

(NEC) KFG $\vdash$ A implies KFG $\vdash$ T(A)

renders S decidable.

To see this observe KFG $\vdash$ A V ¬A hence by (NEC) KFG $\vdash$ T(A V ¬A) and so by the axiom for disjunction we have T(A)VT(¬A) i.e. every sentence or its negation is derivable from S.

On the other hand, (N) and (L) are quite interesting. (N) as a modal principle is characterized by the class of models <W,R> in which if w is related to w', there is a world to which both w and w' are related. Withe regard to (L), it is not obvious whether there is a first-order condition on the accessibility relation which corresponds to it. However, from (T) and (L) we may derive McKinsey's axiom:

(Mc) $T(C(A)) \rightarrow C(T(A))$

(Mc) states that all sentences eventually becomes stable [cf. [1]]. (Mc) itself does not correspond to a first order condition on arbitrary R. However, if R is transitive, then it corresponds to the following condition on R:

$$(\forall w)(\exists x)(w \; R \; x \text{ and } (\forall y)(x \; R \; y \text{ implies } x = y))$$

That is, it states the existence of end points.

5. Revisable inference through theory revision

The process of revision presented in the previous section employs a (monotone) non-classical semantics, namely that of Kleene's three valued-logic, at the revision step. Reasons could be given to defend this line of pursuit. However, it may also be argued that the logic of "theoremhood-of-S" is embodied in a classical semantics which is not monotone. This, in fact, is the exact notion we are attempting to capture. In this section, we shall consider a theory GH (Gupta-Herzberger) [5,7] which uses the classical semantic at the revision step.

Before we present the theory, it is important to show that we may construct a classical model of L in which all the Tarski biconditionals for S are true. Such a step would establish that L can consistently meet some essential requirements for a language to allow the expression of sentences which are akin to the liar and related paradoxical sentences. These requirements are that in such a language the ordinary laws of logic hold and the language itself is semantically closed. [For more details on this issue, cf. [5]]. We shall further discuss this issue in the next version.

5.1. The Gupta-Herzberger (GH) semantic theory of revisable inference

GH adopts a theory of revision based completely on classical semantics.

We shall construct a (possibly ordinal) sequence of models, like M above, where the extensions of T and F are continually revised.

Definition 5.1. Let M = <D, I, T, F> be a model of L and let T and F stand for the extensions of positive and negative theoremhood-of-S in the standard model M, i.e. $T = \tau(M)$ and $F = \Phi(M)$. We define M' = <D, I, T', F'> by *Tarski revision* of M as follows:

$T' = \tau(M + T)$ and $F' = \phi(M + F)$.

The important point to observe is that this process of revision is not monotonic since $T(d)=1$ does not imply $T'(d)=1$.

Using this basic step of revision we can define a sequence of positive and negative theoremhood-of-S predicates $T(i)$ and $F(i)$ $(i \geq 0)$ as follows:

(i) $T(0) = T$
$F(0) = F$
(ii) $T(i+1) = T(i)'$
$F(i+1) = F(i)'$

Because the operation of revision is not monotonic, in order to carry the process through to transfinite ordinals, we cannot simply select the union of all the T's at limit ordinals. Here we follow the lead of Herzberger [7].

(i) For limit ordinal k define:
$T(k)(d)=1$ iff $(\exists j)(j < k)(\forall h)(j \leq h < k)(T(h)(d) = 1)$
$F(k)(d)=1$ iff $(\exists j)(j < k)(\forall h)(j \leq h < k)(F(h)(d) = 1)$

It should be pointed out that there are many options concerning the definition of predicate "T" at limit ordinals and no doubt different choices would lead to different theories of "T". The articles by Gupta and Herzberger contain indications of the different choices available, but we shall not pursue this here.

Definition 5.2.

(i) An element d in D is Positively Stable iff $(\exists j)(\forall k \geq j)(T(k)(d) = 1)$.
An element d in D is Negatively Stable iff $(\exists j)(\forall k \geq j)(F(k)(d) = 1)$.
An element d of D is Stable iff d is positively or negatively stable.

We say that d is Positively Stable from an ordinal j (Negatively Stable from j) iff $(\forall k \geq j)(T(k)(d) = 1)$ $((\forall k \geq j)(F(k)(d) = 1))$.

Definition 5.3. An ordinal i is a Stabilization Ordinal iff

(i) For each d in D, d is positively stable iff $T(i)(d)=1$.
For each d in D, d is negatively stable iff $F(i)(d)=1$.

(ii) For each d in D, d is positively(negatively) stable implies that d is positively(negatively) stable from i.

Stabilization ordinals characterize the stable objects exactly. The central result for our purposes is the following.

Theorem 5.1. There exists a stabilization ordinal [7].

We shall be primarily interested in those wff which are valid at such models.

Definition 5.4. A wff is **Sound** iff it is valid at every stabilization ordinal.

In the remaining part of this section we shall turn attention to exploring the *logics* of stable theorems-of-S. In this regard observe that the semantics presented above has a modal flavour to it: for a wff A to be stably a theorem-of-S, A must be true in all models after some point in the revision process. Moreover, at stabilization ordinals T(A) means that A is stably a theorem-of-S and F(A) that ¬A is stably a theorem-of-S.

5.2. GH as a modal theory

We shall begin with the modal logic D which is defined by the following axioms and rules:

(DIS) $T(A) \rightarrow C(A)$

(IP) $T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B))$

(NEC) If D $\vdash$ A then D $\vdash$ T(A)

To establish the soundness of D under the stability interpretation we first establish that each of the axioms (DIS) and (IP) is not just sound (e.g. (DIS) is true at stabilization ordinals) but stably theorems-of-S (e.g. T(DIS) is true at stabilization ordinals). We thus need to establish the soundness of the following axioms:

(SDIS) $T(T(A) \rightarrow C(A))$
(SIP) $T(T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B)))$

Theorem 5.2. (SIP) and (SDIS) are sound.

Theorem 5.3. If D $\vdash$ A then A is stably a theorem-of-S.

The result of this section is the fact that the modal logic D is a consistent logic of revisable inference and moreover all the theorems of D are stably theorems-of-S.

This is hardly striking as D is the only non-trivial logical system which satisfies the meta-theoretic form of Tarski's Biconditionals. That is, $\vdash_D$ A iff $\vdash_D$ T(A).

Theorem 5.4. $\vdash_D$ A iff $\vdash_D$ C(A)

Corollary 5.1. $\vdash_D$ A iff $\vdash_D$ T^n(A) iff $\vdash_D$ C^m(A) for any m,n.

In the next sections we shall look at the axioms of standard modal systems T, S4 and S5.

5.2.1. The modal logic ST

The characteristic axiom for the modal logic T is:

(T) $T(A) \rightarrow A$

Theorem 5.5. T is sound.

Notice that, though (T) implies (DIS), (DIS) is stably a theorem-of-S whereas , (T) is not.

Theorem 5.6. The modal logic T is inconsistent as a logic of revisable inference.

This result is essentially the truth-theoretic version of the result of Montague and Kaplan [11]. Their derivation is based upon the *Hangman* paradox.

In regard to the stable theoremhood-of-S of the T-axiom the following principle is stably a theorem-of-S.

(St) $T(T(A) \rightarrow A) \rightarrow P(A)$

Theorem 5.7. [16] St is stably a theorem-of-S.

The system ST.

(DIS) $T(A) \to C(A)$

(IP) $T(A \to B) \to (T(A) \to T(B))$

(St) $T(T(A) \to A) \to P(A)$

(NEC') if ST $\vdash$ A then ST $\vdash$ T(A)

As a consequence of the previous theorem we have:

Theorem 5.8. [16] If ST $\vdash$ A then A is stably a theorem-of-S.

5.2.2. The modal logic SS4

Next consider the S4 axiom which takes the form:

(S4) $T(A) \to T(T(A))$

Theorem 5.9. S4 is sound.

However, it can easily be shown that S4 cannot be stably a theorem-of-S. Consider the axiom:

(Q) $T(T(A) \to T(T(A))) \to P(A)$

Theorem 5.11. [16] (Q) is stably true.

The system SS4.

(DIS) $T(A) \to C(A)$

(IP) $T(A \to B) \to (T(A) \to T(B))$

(St) $T(T(A) \to A) \to P(A)$

(Q) $T(T(A) \to T(T(A))) \to P(A)$

(NEC') if SS4 $\vdash$ A then SS4 $\vdash$ T(A)

It follows from the previous theorem that SS4 is stable.

Theorem 5.12. [16] If SS4 $\vdash$ A then A is stably true.

5.2.3. The modal logic SS5

The S5 axiom, in the present context, takes the following form

(S5) $C(A) \rightarrow T(C(A))$

S5 is not even safe. In fact, we have:

Theorem 5.13. T + (S5) is inconsistent.

Consider the axiom:

(W) $(C(A) \rightarrow T(C(A))) \rightarrow P(A)$

Theorem 5.14. [16] (W) is stably true.

The system SS5.

(DIS) $T(A) \rightarrow C(A)$
(IP) $T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B))$
(St) $T(T(A) \rightarrow A) \rightarrow P(A)$
(Q) $T(T(A) \rightarrow T(T(A))) \rightarrow P(A)$
(W) $(C(A) \rightarrow T(C(A))) \rightarrow P(A)$
(NEC) if SS5 $\vdash$ A then SS5 $\vdash$ T(A)

It follows from the previous theorem that SS5 is stable.

Theorem 5.15. [16] If SS5 $\vdash$ A then A is stably true.

For more detailed discussion of the logics of stable truth the reader may refer to [16].

6. Revisable inference and nonmonotonic fixed-point theories

From the previous two sections, we may draw the following conclusions:

Even in a fully self-referential language, the theory KFG allows the stipulation (E)

(E) every sentence eventually stabilizes with regard to its positive or negative theoremhood of S.

However, there is an expense for this result, namely that the logic of "T" is three-valued.

The theory GH, on the other hand, where the logic of T is embodied in a classical semantics which is non-monotone, a property which essentially characterize revisable inference, does not allow (E) in a fully self-referential language. As remarked above, depending on S there may not be a need for L to be fully self-referential. In such a case, principles like (T), (S4), (S5) and (E) will cease being problematic for GH.

However, even in a fully self-referential language, the two theories agree on a common basis which is the logical system D. Hence, we may deduce that *D is the minimal logic of revisable inference.* D, as shown above, is characterized by (DIS), (IP) and (NEC) and none of the fixed point approaches such as NMLI, NMLII and AE, to nonmonotonic reasoning seems to have emphasized the principle (DIS). Halpern and Moses (1984) have suggested adding the (DIS) to autoepistemic logic. There, it states that the language L is not an acceptable belief state.

Furthermore, both (T) and (S4) are derivable from KFG and sound in GH. They are only problematic for GH if their stabilization is insisted upon and the problem is due to full self-referentiality of the language. (S5), none-the-less, when added to KFG renders S strongly decidable about the positive and negative theoremhood of every sentence. An issue which can both be argued in favour and against. But, the interesting point to observe is that (T) and (S5) cannot co-exist in a pure classical setting: they lead to inconsistency in GH and McDermott's nonmonotonic S5 collapsed to monotonic S5. This same issue was fully debated by Moore [12] who has suggested dropping (T) in favour of (S5) in contrast to what McDermott has suggested. We have sufficient evidence (not stated here) which supports Moore's suggestion.

References

[1] Benthem J. Van, (1983), The Logic of Time, Reidel Publishing Company.

[2] Clark K. L., (1978), "Negation as Failure", in *Logic and Databases,* H. Gallaire and J. Minker (eds.), Plenum Press, New York, 293-322.

[3] Van Emden M. H. and Kowalski R. A., (1976), "The Semantics of Predicate Logic as a Programming Language", *JACM* 23, 723-742.

[3] Feferman, S., (1984), Towards Useful Type-free Theories 1, *Journal of Symbolic Logic,* Vol. 49, 75-111.

[4] Gilmore, P. C., (1974), The consistency of Partial Set Theory Without Extensionality, *Axiomatic Set Theory,* Proc. Symposia Pure Maths, Vol. XIII, Part II, 147-153, Amer. Math. Soc.

[5] Gupta, A., (1982), Truth and Paradox, *Journal of Philosophical Logic,* Vol. 11, 1-60.

[6] Halpern J. and Moses Y., (1984), Towards a Theory of Knowledge and Ignorance, Preliminary Report, Technical Report RJ 4448 48316, IBM

Research Laboratory, San Jose.

[7] Herzberger, H., (1982), Notes on Naive Semantics, *Journal of Philosophical Logic,* Vol. 11, 61-102.

[8] Kripke, S., (1975), Outline of a Theory of Truth, *Journal of Philosophy,* Vol. 1, xxii, 690-716.

[9] McDermott D. and Doyle J., (1980), Non-Monotonic Logic I, *Artificial Intelligence* 13, 41-72.

[10] McDermott D., (1982), Non-Monotonic Logic II: Non-Monotonic Modal Theories, *JACM* 29 (1), 35-57.

[11] Montague, R. and Kaplan, (1960), A Paradox Regained, *Notre Dame Journal of Formal Logic,* Vol. 1, 79-90.

[12] Moore R., (1985), Semantical Consideration of Nonmonotonic Logic, *Artificial Intelligence* 25, 75-94.

[13] Reiter R., (1978), "On Closed-World Data Bases", in Logic and Data Bases *H. Gallaire and J. Minker (eds.)* Plenum Press, 55-76.

[14] Reiter R., (1980), A Logic for Default Reasoning, *Artificial Intelligence* 13, 81-132.

[15] Smullyan, R. M., (1957), Languages in which Self-Reference Is Possible", *Journal of Symbolic Logic,* 22, 55-67.

[16] Turner, R., (1990), *Truth and Modality for Knowledge Representation,* Pitman, London.