All I Know: An Abridged Report[1]

Hector J. Levesque[2]
Dept. of Computer Science
University of Toronto
Toronto, Canada M5S 1A4

Abstract

Current approaches to formalizing non-monotonic reasoning using logics of belief require new metalogical properties over sets of sentences to be defined. This research attempts to show how some of these patterns of reasoning can be captured using only the classical notions of logic (satisfiability, validity, implication). This is done by extending a logic of belief so that it is possible to say that *only* a certain proposition (or finite set of them) is believed. This research also extends previous approaches to handle quantifiers and equality, provides a semantic account of certain types of non-monotonicity, and through a simple proof theory, allows formal derivations to be generated.

I. Introduction

A great deal of attention has been devoted recently to formalisms dealing with various aspects of non-monotonic reasoning [Reiter, to appear, 1988]. Broadly speaking, these can be divided into two camps: those, like the logics of [McDermott and Doyle, 1980] and [Reiter, 1980], which are *consistency-based*, and those, deriving from [McCarthy, 1980] and [McCarthy, 1984], which are based on *minimal models*. In the former case, non-monotonic assumptions are made on the basis of certain hypotheses being consistent with a current theory; in the latter case, non-monotonic assumptions are made on the basis of their being true in all minimal (or otherwise preferred) models of a current theory. For better or for worse, the latter approach seems to be winning, in part, no doubt, because it can be given a compelling model-theoretic account, in addition to its more proof-theoretic formulation.

However, one development that may begin shifting the balance towards consistency-based approaches is the application of logics of knowledge and belief [Halpern and Moses, 1985] and [McArthur, to appear, 1987].[3] Although these have been used in non-monotonic contexts for some time (see [Levesque, 1981], [Konolige, 1982], and [Halpern and Moses, 1984]), only recently have clear and precise connections been established between these logics and the non-monotonic ones [Moore, 1983] and [Konolige, 1987].

What do logics of belief have to do with consistency-based non-monotonicity? The idea, roughly, is this: A "current theory" is no more than a set of beliefs. If these beliefs are closed under logical consequence, then a hypothesis is consistent with a current theory precisely when its negation is not believed. So under this account, non-monotonic assumptions are made based on failing to believe certain other propositions. For example, one might be willing to believe that any bird that is not believed to be flightless can fly. Or perhaps this belief is restricted to certain birds, like those that are currently known. Either way, without claiming that this is the same thing as believing that "Birds generally fly" or anything like that, it does appear that under the right circumstances, the belief leads to the same assumptions as the consistency-based approaches. Moreover, the expectation here is that the model-theoretic accounts of belief deriving from the (reasonably well established) logics of belief can then be used to semantically rationalize these consistency-based systems.

II. Only knowing

This is not to say that logics of belief can be used *as is* to account for non-monotonic reasoning. To see why not, consider how one might explain using belief, why the ever-popular Tweety flies. Assume we take as premises that

1. Tweety is a bird.

2. If a bird can be consistently believed to fly, it flies.

There surely is something missing before we are entitled to write down our favourite non-monotonic conclusion:

3. Tweety flies.

At the very least, we would have to know that our second premise applied to Tweety:

1.5 It is consistent with my beliefs that Tweety flies.

But what justifies this assertion? Clearly not (1) by itself. Rather, it is the fact that, except for (2), (1) *is* by itself.

[1] This research was made possible in part by a grant from the Natural Sciences and Engineering Research Council of Canada. Thanks also to Gerhard Lakemeyer, Ray Reiter, Jim des Rivières, and Bart Selman for proofreading.

[2] Fellow of The Canadian Institute for Advanced Research

[3] For the purposes of this paper, the distinction between knowledge and belief is irrelevant, and the two terms will be used interchangeably.

That is, the understood (non-monotonic) assumption is that there are no other relevant beliefs about Tweety:[4]

1.2 This is *all* I know (about Tweety).

Now (1.5) does seem to follow from (1) and (1.2), so that we are indeed justified in concluding that Tweety flies from (1), (1.2), and (2).

The problem here is that although logics of belief allow us to express and reason with (1), (1.5), (2) and (3), assumption (1.2) cannot be expressed. The approach to this issue taken by Moore and Konolige is to not even try to express it, but instead to characterize (outside the logic itself) sets of beliefs where (1.2) intuitively might be said to hold, and then to examine the properties of such sets, called *stable expansions.*[5] As expected, (1) and (2) have a single stable expansion, and it does indeed contain (3).

However, what is lost by this use of logics of belief is precisely what might have been expected to be gained, namely a precise model-theoretic account of consistency-based non-monotonicity. The concept of a stable expansion (and clearly the key one in the non-monotonic aspect of the inference) is not defined in terms of the semantics of belief, but is a new metalogical property of certain sets of sentences. Because of this, the derivation from only knowing (1) and (2) to knowing (3) must be carried out completely outside the logic, as in McDermott and Doyle's logic or in Reiter's (in their case with appropriate metalogical arguments about fixed points or extensions).

In this paper, we present research that attempts to remedy this situation by augmenting a logic of belief so that propositions similar to (1.2) can be expressed directly within the language. There will be two modal operators, **B** and **O**, where $\mathbf{B}\alpha$ is read (as usual) as "α is believed" and $\mathbf{O}\alpha$ is read as "α is *all* that is believed," or perhaps, "*only* α is believed." It turns out that this latter concept can be given fairly intuitive truth conditions that are remarkably similar to those for belief. We will then establish correspondences to Moore's stable expansions, generalizing them in the process to the quantificational case. The existence of (sometimes multiple) stable expansions will emerge within the logic as valid sentences. Finally, we will exhibit a reasonably standard (though not recursive) proof theory for the logic (that is, with axioms and rules of inference) and show, perhaps for the very first time, a formal derivation of the belief that Tweety flies.

It should be noted that this approach to logic uses it as a specification tool to describe a reasoner rather than as a calculus to be used by one. Thus, there is no notion of an agent "having" a theory in this language, except as stated explicitly using a **B** operator. While the patterns of reasoning to be described may be non-monotonic, the logic itself is perfectly monotonic [Israel, 1980].

[4] Although obviously important, we do not attempt here to deal with relevance, or which beliefs are about what.

[5] To a first approximation, these can be thought of as the fixed-points of McDermott and Doyle's logic, or the extensions of Reiter's.

Because of space limitations, the formal presentation of the logic below will be somewhat terse, and most proofs will be deferred to [Levesque, in preparation].

III. The language and its semantics

The language we consider is called $\mathcal{L}$, and its propositional part is built up in the usual way from propositional letters and the logical connectives $\neg$, $\wedge$ (the others will also be used freely as syntactic abbreviations), and two special unary connectives **B** and **O**. For the quantificational part, we include in addition an infinite stock of predicate symbols of every arity, an infinite collection of (individual) variables, an existential quantifier, and a special two-place equality symbol. For simplicity, we omit function and constant symbols. However, we include a countably infinite set of *standard names* (called parameters in [Levesque, 1984a]), that are considered (like the equality symbol) to be logical symbols. Sentences are formed in the obvious way; in particular, there is no restriction on the relative scope of quantifiers and modal operators. The *objective* sentences are those without any **B** or **O** operators; the *subjective* sentences are those where all non-logical symbols occur within the scope of a **B** or **O**. Sentences without **O** operators are called *basic.* We will use α and β to range over sentences, σ to range over the subjective sentences only, and ϕ and ψ to range over the objective sentences only. Finally, α_n^x is used to name the formula consisting of α with all occurrences of free variable x replaced by standard name n.

Before presenting the semantics of $\mathcal{L}$, a few comments are in order. First, we will be interested in characterizing a system with full logical capabilities and perfect introspection. In other words, beliefs will be closed under logical consequence, anything believed will be known to be believed, and anything not believed will be known not to be believed. This means our notion of belief will satisfy at least the postulates of the modal system weak S5 (see [Halpern and Moses, 1985] or [McArthur, to appear, 1987] for why). However, it will be convenient to give a non-standard semantic account of $\mathcal{L}$ that avoids explicit use of possible worlds. Instead we will take a coarser-grained approach and deal with the truth and falsity of sentences directly. A possible world, then, is modelled by any function w from sentences to $\{0,1\}$ satisfying certain constraints having to do with the interpretation of the logical symbols. We will call such functions *valuations.*

As to the constraints themselves (presented below), there is nothing new about the interpretation of conjunction and negation. The interpretation of equality sentences is based on the convention that standard names are taken to designate distinctly and exhaustively (something one would certainly not want for ordinary *constant* symbols). This exhaustiveness property also means that quantification can be understood substitutionally. This substitutional interpretation imposes no real restrictions on what

sets of sentences will be satisfiable. For example, it will certainly be possible to *believe* $\exists x\alpha$ without believing any of its substitution instances and, as will become clear below, a distinction will remain between $\mathbf{B}\exists x\alpha$ and $\exists x\mathbf{B}\alpha$. Thus, the first four constraints on a function w from the sentences of $\mathcal{L}$ to $\{0,1\}$ are that for every α and β,

1. $w(\alpha \wedge \beta) = \min[w(\alpha), w(\beta)]$,
2. $w(\neg\alpha) = 1 - w(\alpha)$,
3. $w(n_i = n_j) = 1$ iff n_i and n_j are the same standard name,
4. $w(\exists x\alpha) = 1$ iff for some n, $w(\alpha^x_n) = 1$.

We will call any function satisfying these constraints a first-order or *f.o. valuation.* Note that these valuations treat sentences of the form $\mathbf{B}\alpha$ or $\mathbf{O}\alpha$ as atomic sentences.

Turning now to the belief operator, **B**, the by now standard way to give its interpretation is in terms of an *accessibility relation* over worlds: $\mathbf{B}\alpha$ is considered true at some world w iff α is true at every w' that is accessible from w. But what are these accessible worlds? In our case, there are two considerations: (1), an accessible world must make all the beliefs in the original world come out true; and (2), the accessibility relation must be an equivalence relation. So we begin by defining, for any f.o. valuation w, $\Re(w)$ to be the set of all f.o. valuations w' such that for every *basic* α, if $w(\mathbf{B}\alpha) = 1$, then $w'(\alpha) = 1$. To get an equivalence relation, we must also ensure that the same subjective sentences are true in every accessible world. We say that $w \approx w'$ iff for every (subjective) σ, $w(\sigma) = w'(\sigma)$. Intuitively then, the accessible worlds from w are those elements w' of $\Re(w)$ such that $w \approx w'$. Using these definitions, we can now state a constraint on the interpretation of the **B** operator: for every α,[6]

5. $w(\mathbf{B}\alpha) = 1$ iff for every $w' \approx w$, $w' \in \Re(w) \Longrightarrow w'(\alpha) = 1$.

We will call any function w satisfying the first five constraints an autoepistemic or *a.e. valuation.* Not every f.o. valuation is an a.e. valuation (*e.g.*, one that assigns different values to $\mathbf{B}\alpha$ and $\mathbf{B}\neg\neg\alpha$). However, every valuation that is accessible from an a.e. valuation is itself one.

Finally, with regards to the **O** operator, the idea is this: Beliefs are those sentences that are true in all accessible worlds. So to come to believe a new objective sentence means to reduce the set of accessible worlds, keeping only those where the new belief is true. Thus, the more known (in objective terms anyway), the smaller the set of accessible worlds, and vice-versa. Now to say that α is *all* that is known is to say that as little as possible is known compatible with believing α. Thus, the set of accessible worlds is *as large as possible* consistent with believing α, since the larger the set, the less world knowledge represented. Specifically, any valuation that satisfies the same subjective sentences and also satisfies α should be accessible. This leads to our final constraint: for every α,

6. $w(\mathbf{O}\alpha) = 1$ iff for every $w' \approx w$, $w' \in \Re(w) \Longleftrightarrow w'(\alpha) = 1$.[7]

Any function w satisfying all six constraints is called a *logical valuation.* Note once again that not every a.e. valuation is a logical valuation, and that the accessibility relation takes logical valuations to only logical valuations.

For each type of valuation, we say that a set of sentences is *satisfiable* iff some valuation of that type assigns 1 to all its members. A set of sentences *implies* a sentence iff the set together with the negation of the sentence is not satisfiable. Finally, a sentence is *valid* iff it is implied by the empty set. We will usually leave out the "logical" qualifier, except to distinguish a logical valuation (or validity *etc.*) from the other types.

It is easy to see that for objective sentences without equality or standard names, f.o. satisfiability (and thus, f.o. implication and f.o. validity) coincide with their classical definitions. Not sc obvious (by a long shot), is this:

Theorem 1 *A set of basic sentences is a.e. satisfiable iff there is a weak S5 Kripke structure and a world within it where all the sentences are true.*

Thus a.e. satisfiability is the same as weak S5 satisfiability. This theorem justifies our lack of explicit possible worlds and ensures that, for example, standard axiomatizations of weak S5 characterize precisely a.e. validity for basic sentences (and we will present one such below).

IV. Stable sets and expansions

But our primary interest is the notion of only knowing. To justify our interpretation of **O**, we will relate it to the concept of stable expansions. Before doing so, it is useful to consider the properties of the sets of sentences that can be simultaneously believed. We will call a set of basic sentences a *belief set* if there is a logical valuation for which these sentences are precisely the ones believed. In other words, Γ is a belief set iff for some logical valuation w, $\Gamma = \{\beta \mid \beta \text{ is basic and } w(\mathbf{B}\beta) = 1\}$. One important property we can show is that this definition of belief set is the correct quantificational generalization of what Moore calls [Moore, 1983] (following Stalnaker) a *stable theory*:

Theorem 2 *Restricting our attention to basic sentences,*[8] *a set of sentences Γ is a belief set iff Γ is stable, that is, satisfies the following conditions:*

1. *If Γ f.o. implies α, then $\alpha \in \Gamma$.*[9]
2. *If $\alpha \in \Gamma$, then $\mathbf{B}\alpha \in \Gamma$.*

[6] This constraint applies to non-basic sentences, although we have yet to find a need for talking about believing (or only believing) sentences with **O** operators.

[7] Note that this condition differs from the one for belief in exactly one place: an "if" becomes an "iff".

[8] This theorem can be strengthened to handle arbitrary sentences (given a generalized notion of belief set) by extending the first condition below to closure under full logical implication.

[9] Moore required Γ to be closed under tautological consequence, since he only dealt with a propositional language.

3. If $\alpha \notin \Gamma$, then $\neg\mathbf{B}\alpha \in \Gamma$.

For the propositional version of the language, this theorem was proved as Proposition 3 of [Halpern and Moses, 1984] (and apparently independently by R. Moore, M. Fitting, and J. van Benthem). Unfortunately, a new proof was needed because their *proof* fails for a quantificational language, as it depends on the following:

Proposition 1 [Halpern and Moses, 1984] *Stable sets (in the non-quantificational sublanguage) are uniquely determined by their objective subsets.*

With quantifiers, however, the situation is much more complicated:

Theorem 3 *Stable sets (in the quantified language) are not uniquely determined by their objective subsets (and thus neither are belief sets).*

This theorem is proved by showing that there is a difference between believing

$$\{\phi(n_1), \phi(n_3), \phi(n_5), \ldots\},$$

on the one hand, and believing

$$\{\phi(n_1), \phi(n_3), \phi(n_5), \ldots, \exists x(\phi(x) \wedge \neg\mathbf{B}\phi(x))\},$$

on the other, even though both sets involve exactly the same objective sentences. In the latter case, there is the additional information that there is a ϕ apart from the known ones, information that simply cannot be expressed in objective terms.[10] Thus, it is possible to agree on all the objective sentences without yet agreeing on all sentences.

The main result here is the following:

Theorem 4 *Restricting our attention to basic sentences only,[11] for any logical valuation w, $w(\mathbf{O}\alpha) = 1$ iff the belief set of w is a stable expansion of α, that is, the belief set Γ satisfies the fixed-point equation:*

$$\Gamma \text{ is the set of f.o. implications of } \{\alpha\} \cup \{\mathbf{B}\beta \mid \beta \in \Gamma\} \cup \{\neg\mathbf{B}\beta \mid \beta \notin \Gamma\}.$$

So only knowing a sentence means that what is known is a stable expansion of that sentence (or, more intuitively, what is known is derivable from that sentence using logic and introspection alone). This theorem provides for the first time a semantic account (closely related to that of possible worlds) for the notion of a stable expansion which Moore used to rationally reconstruct the non-monotonic logic of [McDermott and Doyle, 1980]. In a subsequent paper [Moore, 1984], Moore provided a possible-world semantics for his autoepistemic logic, but not for the non-monotonic part concerned with stable expansions. In addition, we have generalized the notion of a stable expansion to deal with a quantificational language with equality.

[10] It *could* be expressed if we allowed infinite disjunctions ranging over any set of standard names.

[11] Again this restriction can be removed using logical implication in the definition.

V. Proof theory

The fact that the semantic characterization of $\mathbf{O}\alpha$ uses an "iff" where $\mathbf{B}\alpha$ uses an "if" suggests that it might be worthwhile to look at another operator that uses the "only if" condition alone. The proof theory we are about to present is most conveniently expressed using a new modal operator $\mathbf{N}$ for this only-if condition:

$$w(\mathbf{N}\alpha) = 1 \text{ iff for every } w' \approx w, \quad w'(\alpha) = 0 \implies w' \in \Re(w).$$

$\mathbf{O}\alpha$ can now be *defined* as the conjunction of $\mathbf{B}\alpha$ and $\mathbf{N}\neg\alpha$. Taking $\mathbf{B}\alpha$ as saying "at least α is believed to be true," $\mathbf{N}\alpha$ can be read as "at most α is believed to be false," from which $\mathbf{O}\alpha$ is read as "exactly α is believed."

The remarkable fact about the $\mathbf{N}$ operator is that it behaves exactly like a belief operator, but with respect to the complement of the $\Re$ relation:

$$w(\mathbf{N}\alpha) = 1 \text{ iff for every } w' \approx w, \quad w' \in \overline{\Re}(w) \implies w'(\alpha) = 1.$$

This allows us to produce a proof theory for $\mathcal{L}$ that is very similar to what would be done for two separate believers. The difference is that (1) the two "agents" are mutually introspective (i.e. know about each other's beliefs and non-beliefs), and (2) every world is an element of $\Re$ or $\overline{\Re}$. To handle (1), we include not only the usual introspection axioms like $(\mathbf{N}\alpha \supset \mathbf{NN}\alpha)$, but cross-axioms like $(\mathbf{N}\alpha \supset \mathbf{BN}\alpha)$. To handle (2), we simply stipulate that every falsifiable objective sentence that is true at every member of $\overline{\Re}$ must be false at some member of $\Re$. Overall then, the proof theory is formed by adjoining to any standard objective basis the following axioms:

1. the remaining axioms for weak S5, for both $\mathbf{B}$ and $\mathbf{N}$:
   (a) $\mathbf{L}\phi$, where ϕ is any f.o. valid objective sentence,
   (b) $\mathbf{L}(\alpha \supset \beta) \supset (\mathbf{L}\alpha \supset \mathbf{L}\beta)$,
   (c) $\forall x\mathbf{L}\alpha \supset \mathbf{L}\forall x\alpha$,
   (d) $(\sigma \supset \mathbf{L}\sigma)$, where σ is subjective,
   where $\mathbf{L}$ is either $\mathbf{B}$ or $\mathbf{N}$;
2. $\mathbf{N}\phi \supset \neg\mathbf{B}\phi$, where ϕ is any objective sentence that is falsifiable;[12]
3. $\mathbf{O}\alpha \equiv (\mathbf{B}\alpha \wedge \mathbf{N}\neg\alpha)$, for any α.

The notion of a theorem is defined in the usual way (note that no new rules of inference are introduced).

The first result about this proof system is:

Theorem 5 (Soundness) *Every theorem is valid.*

The proof is by induction on the length of the derivation: the axioms are all clearly valid and the objective rules of inference obviously preserve validity. However, the more substantial result about this simple axiomatization is that for the propositional case anyway, it is also complete:

[12] Note that this set is not r.e. for the full quantificational objective language. Unfortunately, this is the price that must be payed for consistency-based reasoning. In its defense, however, the axiom only requires non-valid *objective* sentences, a relatively well-understood and manageable set.

Theorem 6 (Propositional completeness) *If α is in the propositional subset, then it is a theorem iff it is valid.*[13]

What this shows us is that with a minimum of extra machinery over and above the (modal) axioms necessary for logics of belief, we can account for the semantics of $\mathcal{L}$.

VI. Some applications

What is this logic good for? One application is the formal specification of a Knowledge Representation service: given a certain KB, what are the sentences that are believed? To a first approximation, it's the logical implications of KB. However, if β is some sentence that is not believed, then an introspective system also believes $\neg\mathbf{B}\beta$ (*i.e.*, it realizes that it does not believe β). The problem is that KB does not imply $\neg\mathbf{B}\beta$, nor does $\mathbf{B}$KB imply $\mathbf{B}\neg\mathbf{B}\beta$. So logical implication is not enough. In [Levesque, 1984a], this was handled by moving outside the logic and defining a special ASK operation. But given the **O** operator, we can stay within the logic: $\mathbf{O}$KB *does* imply $\mathbf{B}\neg\mathbf{B}\beta$. In general, the beliefs of an introspective system will be those sentences α such that $(\mathbf{O}\text{KB} \supset \mathbf{B}\alpha)$ is a valid sentence of $\mathcal{L}$.

However, the main application of this logic is to give semantic and/or proof-theoretic arguments involving non-monotonic reasoning. Consider the above example involving Tweety. First, we represent the default as

$$\forall x[\mathsf{Bird}(x) \wedge \neg\mathbf{B}\neg\mathsf{Fly}(x) \supset \mathsf{Fly}(x)].$$[14]

Now believing this and that Tweety is a bird certainly does not imply believing that Tweety flies. But we can show that if this is *all* that is believed, then the belief that Tweety flies does follow:

Theorem 7 *Let* $\beta = \forall x(\mathsf{Bird}(x) \wedge \neg\mathbf{B}\neg\mathsf{Fly}(x) \supset \mathsf{Fly}(x))$. *Then,* $\mathbf{O}[\mathsf{Bird}(\mathsf{tweety}) \wedge \beta] \supset \mathbf{B}\mathsf{Fly}(\mathsf{tweety})$ *is a theorem.*

Proof: We present a formal derivation using natural deduction. The numbers refer to the above proof theory.

a.	$\mathbf{O}[\mathsf{Bird}(\mathsf{tweety}) \wedge \beta]$	Assumption.
b.	$\mathbf{B}[\mathsf{Bird}(\mathsf{tweety}) \wedge \beta]$	From *a* using (3).
c.	$\mathbf{B}\mathsf{Fly}(\mathsf{tweety}) \vee \mathbf{B}\neg\mathsf{Fly}(\mathsf{tweety})$	From *b* using (1).
d.	$\mathbf{N}\neg[\mathsf{Bird}(\mathsf{tweety}) \wedge \beta]$	From *a* using (3).
e.	$\mathbf{N}[\mathsf{Bird}(\mathsf{tweety}) \supset \exists x\neg\mathsf{Fly}(x)]$	From *d* using (1).
f.	$\neg\mathbf{B}[\mathsf{Bird}(\mathsf{tweety}) \supset \exists x\neg\mathsf{Fly}(x)]$	From *e* using (2).
g.	$\neg\mathbf{B}\neg\mathsf{Fly}(\mathsf{tweety})$	From *f* using (1).
h.	$\mathbf{B}\mathsf{Fly}(\mathsf{tweety})$	From *c* and *g*, by classical logic.

Discharging the assumption gives the required result. ∎

A propositional version of this argument can be made in terms of Moore's stable expansions (that is, that there is a single expansion, and it contains the desired conclusion). The significance of this derivation is that the argument only depends on the validity (or in this case, theoremhood) of a certain sentence of $\mathcal{L}$, and so can be carried out completely within the language itself in conventional logical terms. The only unusual step in the derivation is from *e* to *f*, where we infer on the basis of something being all that is believed, that a certain other sentence is *not* believed. This step depends on the fact that $\mathsf{Bird}(\mathsf{tweety}) \supset \exists x\neg\mathsf{Fly}(x)$ is not f.o. valid. Indeed, if not flying was implied by being a bird, this proof would fail (as it should), and $\mathbf{B}\neg\mathsf{Fly}(\mathsf{tweety})$ would be the (correct) conclusion.

This analysis also suggests what happens if we know in addition that Chilly is a bird that does not fly. The problem is that the step from *e* to *f* no longer works since the enlarged knowledge base now implies the existence of a flightless bird. What happens, however, is that since $\mathbf{B}\neg\mathsf{Fly}(\mathsf{chilly})$ is true, so is $\mathbf{NB}\neg\mathsf{Fly}(\mathsf{chilly})$ by (1). The new version of step *e* now uses this to conclude that $\mathbf{N}[\text{KB} \supset \exists x((x \neq \mathsf{chilly}) \wedge \neg\mathsf{Fly}(x))]$, where KB has the facts on Tweety and Chilly. Once again the argument to **N** is not f.o. valid, and so the derivation goes through as before, ending with the belief that Tweety flies. Note that this conclusion depends (quite appropriately) on the fact that Chilly and Tweety are believed to be distinct, a logical property of our standard names.

Although there is no really compelling reason to do so, we can define a non-monotonic *logic* easily enough using **O**. For a finite set of sentences Γ, define $\vdash_{\mathbf{n}}$ by

$$\Gamma \vdash_{\mathbf{n}} \alpha \quad \text{iff} \quad (\mathbf{O}\gamma \supset \mathbf{B}\alpha) \text{ is valid,}$$

where γ is the conjunction of the elements of Γ. Then, in the previous example, we have that

$$\mathsf{Bird}(\mathsf{tweety}),\ \beta \vdash_{\mathbf{n}} \mathsf{Fly}(\mathsf{tweety}),$$

but

$$\mathsf{Bird}(\mathsf{tweety}),\ \beta,\ \neg\mathsf{Fly}(\mathsf{tweety}) \not\vdash_{\mathbf{n}} \mathsf{Fly}(\mathsf{tweety}),$$

so this logic would be truly non-monotonic.

VII. Determinate sentences

The correspondence with stable expansions accounts for many of the properties of this logic. For example, the usual situation with *multiple* expansions also arises here. In our case, this is reflected in the language itself, with interesting consequences. Consider a typical sentence with two expansions, $(\neg\mathbf{B}\phi \supset \psi) \wedge (\neg\mathbf{B}\psi \supset \phi)$. What happens here is that the sentence

$$\mathbf{O}[(\neg\mathbf{B}\phi \supset \psi) \wedge (\neg\mathbf{B}\psi \supset \phi)] \equiv (\mathbf{O}\phi \vee \mathbf{O}\psi),$$

which names the two expansions directly, ends up being valid.[15] Thus, it is possible to only know the sentence in two distinct ways. The logic also specifies what is *common* to both, in that $\mathbf{O}[(\neg\mathbf{B}\phi \supset \psi) \wedge (\neg\mathbf{B}\psi \supset \phi)]$ logically implies, for example, $\mathbf{B}(\phi \vee \psi)$.

---

[13] I believe the axiomatization is also complete for the full language, but I have yet to find a proof. My propositional proof fails for the general case in a subtle and interesting way. See [Levesque, in preparation] for details.

[14] Other versions are possible, such as one where Bird is within the scope of a **B**. Also, in what follows, we will be using tweety and chilly as standard names.

[15] Similarly, the validity of $\neg\mathbf{O}[\neg\mathbf{B}\phi \supset \phi]$ tells us that $(\neg\mathbf{B}\phi \supset \phi)$ has *no* stable expansions.

While the cases of multiple or missing expansions may be interesting in their own right, for those of us interested in Knowledge Representation, sentences with a single stable expansion play a very special role. Call a sentence α *determinate* iff there is a unique (up to $\approx$) w such that $w(\mathbf{O}\alpha) = 1$. Then we have the following:

Theorem 8 *α is determinate iff for every β, one of $(\mathbf{O}\alpha \supset \mathbf{B}\beta)$ or $(\mathbf{O}\alpha \supset \neg\mathbf{B}\beta)$ is valid.*

Thus, determinate sentences tell us exactly what is and what is not known.[16] As such, they can be used as *representations of knowledge*, since they implicitly specify a complete epistemic state. Examples of these include all objective sentences, and all examples outside this section.

One important property of this logic is that it is always possible to represent knowledge in objective terms. Although *believing* does not reduce to believing objective sentences (Theorem 3), *only believing* does:

Theorem 9 *For every determinate α, there is an objective sentence ϕ such that $(\mathbf{O}\alpha \equiv \mathbf{O}\phi)$ is valid.*

Thus, to the extent that an epistemic state can be represented at all, it can be represented in objective terms. In other words, whatever defaults might be used (or whatever other uses of non-objective sentences), if there is a unique end result, it can be described without reference to the modal operators. This theorem offers perhaps some reassurance to those who have been suspicious about these operators all along.

VIII. Conclusion

This research attempts to show that non-trivial non-monotonic behaviour can be formalized using only the classical notions of logic. This is done by extending a logic of belief to include a second modality that can be given a reasonably natural semantic and proof-theoretic account.

As for future research, there are the following topics: formalizing what it means to say that α is all that is known *about* something; developing the concepts for a logically limited notion of belief [Levesque, 1984b]; and the missing quantificational completeness proof. Finally, Konolige's account of default logic [Konolige, 1987] depends on a certain restricted kind of stable expansion, and it remains to be seen how this will fit into the current framework.

References

[Halpern and Moses, 1984] J. Halpern and Y. Moses. Towards a theory of knowledge and ignorance: preliminary report. In *The Non-Monotonic Reasoning Workshop*, pages 125–143, New Paltz, NY, 1984.

[Halpern and Moses, 1985] J. Halpern and Y. Moses. A guide to the modal logics of knowledge and belief: a preliminary draft. In *IJCAI-85*, pages 480–490, Los Angeles, CA, August 1985.

[Israel, 1980] D. Israel. What's wrong with non-monotonic logic? In *AAAI-80*, pages 99–101, Stanford, CA, 1980.

[Konolige, 1982] K. Konolige. Circumscriptive ignorance. In *AAAI-82*, pages 202–204, Pittsburgh, PA, August 1982.

[Konolige, 1987] K. Konolige. *On the Relation Between Default Theories and Autoepistemic Logic.* Technical Report, AI Center, SRI International, Palo Alto, CA, 1987.

[Levesque, 1981] H. Levesque. The interaction with incomplete knowledge bases: a formal treatment. In *IJCAI-81*, pages 240–245, Vancouver, B.C., August 1981.

[Levesque, 1984a] H. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, 1984.

[Levesque, 1984b] H. Levesque. A logic of implicit and explicit belief. In *AAAI-84*, pages 198–202, Austin, TX, 1984.

[Levesque, in preparation] H. Levesque. *All I Know: A Study in Autoepistemic Logic.* Technical Report, Dept. of Computer Science, University of Toronto, Toronto, Canada, in preparation.

[McArthur, to appear, 1987] G. McArthur. Reasoning about knowledge and belief: a review. *Computational Intelligence*, to appear, 1987.

[McCarthy, 1980] J. McCarthy. Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1,2):27–39, 1980.

[McCarthy, 1984] J. McCarthy. Applications of circumscription to formalizing commonsense knowledge. In *The Non-Monotonic Reasoning Workshop*, pages 295–324, New Paltz, NY, 1984.

[McDermott and Doyle, 1980] D. McDermott and J. Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13(1,2):41–72, 1980.

[Moore, 1983] R. Moore. Semantical considerations on nonmonotonic logic. In *IJCAI-83*, pages 272–279, Karlsruhe, West Germany, 1983.

[Moore, 1984] R. Moore. Possible-world semantics for autoepistemic logic. In *The Non-Monotonic Reasoning Workshop*, pages 344–354, New Paltz, NY, 1984.

[Reiter, 1980] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1,2):81–132, 1980.

[Reiter, to appear, 1988] R. Reiter. Nonmonotonic reasoning. *Annual Reviews*, to appear, 1988.

[16] These are also related to the "honest" sentences of [Halpern and Moses, 1984].