Reasoning With Incomplete Information

Investigations of Non-Monotonic Reasoning

By

DAVID WILLIAM ETHERINGTON
B.Sc., The University of Lethbridge, 1977
M.Sc., The University British Columbia, 1982

A THESIS IS SUBMITTED IN PARTIAL FULLFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in
THE FACULTY OF GRADUATE STUDIES
(Department of Computer Science)

We accept this thesis as conforming
to the required standard.

THE UNIVERSITY OF BRITISH COLUMBIA
April 1986

Department of COMPUTER SCIENCE

The University of British Columbia
2075 Wesbrook Place
Vancouver, Canada
V6T 1W5

Date 86-6-19

January 29, 1987

ERRATA

Reasoning with Incomplete Information
Investigations of Nonmonotonic Reasoning

David W. Etherington

The following is a (partial) list of errata.

Page 44: In the definition of the result of a sequence of defaults, the three occurrences of $<\delta_i>$ should be $<\delta_j>$.

Page 50: In (2.ii), delete "and $\gamma_i \notin \{\beta_1, \ldots, \beta_s\}$"

Page 77: In point 3, $x = ux$ should be $x = u$.

Page 96: The last two occurrence of Qa in Example 8.2 should be $\neg Qa$.

Page 116: In (2.ii), delete "and $\gamma_i \notin \{\beta_1, \ldots, \beta_s\}$"

line -2: *LITERALS* (α) should be *LITERALS* $(\alpha \wedge \gamma)$.

Page 131*f*: Every occurrence of $\bigcup\limits_{r=1}^{\infty}$ should be $\bigcup\limits_{r=0}^{\infty}$.

Page 148: The three occurrences of $\left\{ \frac{:\neg Px}{\neg Px} \right]$ should be $\left\{ \frac{:\neg Px}{\neg Px} \right\}$.

Page 149: In the proof of Lemma 8.2.2, after $P\alpha_j \notin CONSEQUENTS\,(GE\,(E, \Delta))$., insert "(The remaining terms can be put into $GD\,(E, \Delta)$ in like manner — again, the existence of M and the domain-closure axiom guarantee that this is possible.)"

Abstract

Intelligent behaviour relies heavily on the ability to reason in the absence of complete information. Until recently, there has been little work done on developing a formal understanding of how such reasoning can be performed. We focus on two aspects of this problem: default or prototypical reasoning, and closed-world or circumscriptive reasoning.

After surveying the work in the field, we concentrate on Reiter's default logic and the various circumscriptive formalisms developed by McCarthy and others. Taking a largely semantic approach, we develop and/or extend model-theoretic semantics for the formalisms in question. These and other tools are then used to chart the capabilities, limitations, and interrelationships of the various approaches.

It is argued that the formal systems considered, while interesting in their own rights, have an important rôle as specification/evaluation tools *vis-à-vis* explicitly computational approaches. An application of these principles is given in the formalization of inheritance networks in the presence of exceptions, using default logic.

Table of Contents

List of Figures

List of Tables

Acknowledgements

My supervisor, Ray Reiter, sparked my interest in AI, and particularly in formal approaches thereto, then fostered it with both his insight and his resistance to inadequately reasoned ideas (as well as with financial and moral support). He, together with Robert Mercer, engaged me in hundreds of fruitful (and not-so-fruitful) discussions which have shaped both the form and content of my research career.

Parts of chapters 4, 5, and 7 appear in [Etherington and Reiter 1983], [Etherington, Mercer, and Reiter 1985], and [Etherington and Mercer 1986]. I gratefully acknowledge Mercer's and Reiter's significant contributions to this thesis.

Many thanks are due to Andrew Adler, Alex Borgida, Paul Gilmore, Akira Kanda, and David Touretzky for their generosity with their time, their ideas, and their encouragement, and to Ron Brachman for all that and more.

Financially, I have benefitted from predoctoral scholarships from the Natural Sciences and Engineering Research Council of Canada and the Isaac Walton Killam trust, as well as a research assistantship sponsored by NSERC grant A7642 to Raymond Reiter. I am also thankful for AT&T Bell Laboratories' support, which I received as a visiting scientist in Murray Hill, during the fall of 1985.

Last and most, thanks to Janine for sharing my victories, commiserating over my defeats, and taking on the task of explaining non-monotonicity to our friends; and to those friends for lots of tea and love.

CHAPTER 1

Incomplete Information

> The perverse maxim that whatever you can get away with is right has its counterpart in the claim that whatever works is clear. I might not *understand* the devices I employ in making useful computations or predictions any more than [one] understands the car [one] drives to bring home the groceries. The utility of a notion testifies not to its clarity, but rather to the philosophical importance of clarifying it.
>
> — Nelson Goodman [1955].

Human common-sense reasoning appears to rely heavily upon the ability to use general rules subject to exceptions; what has been called prototypic or default information. Virtually none of the decisions one makes everyday are made with complete certainty. With little effort, an endless supply of more or less probable scenarios can be constructed which contraindicate any chosen course. Yet people are not paralyzed by indecision; they continue to act and to decide in spite of all this uncertainty.

Science fiction fans will recognize "Insufficient Data" as a favourite cliché: computers are frequently characterized as paralyzed by not having enough information to arrive at a logically sound conclusion. If computers are restricted to sound modes of reasoning based on complete information then Artificial Intelligence is a futile goal. For a variety of reasons, "Intelligence" (whatever it may be) must involve the ability to function without complete information about the world.

In the first place, complete information is hard to come by, even in the most contrived situations. Consider, for example, a simple "blocks-world" situation in which there are three blocks of known dimensions, masses, and locations, and a robot manipulator arm with a known lifting capacity, effective radius, and position. If all of the blocks are of a size and mass within the tolerances of the arm, can the arm be used to stack the blocks? At first glance, the answer seems an obvious "yes". Reflection shows that this might be hasty. Our information about the situation is incomplete. There may be things we know nothing about which may interfere. For example, the arm may be broken. (This argument may not convince those who say, "If so, the actual lifting capacity of the arm (0g) was not really known!".)

Granting this, there may still be a wall between the arm and the blocks – we do not know. We can improve our specification of the problem to avoid such incomplete information by saying that there is nothing between the arm and the blocks (not even air?), but we still cannot unequivocally answer the question. A monkey may be holding back the arm – the perverse mind can generate *possible* reasons for failure indefinitely. Without more information, these cannot be ruled out.

The next step is to add the information that nothing prevents the arm from getting to and lifting the blocks. Now we can safely decide that the arm can lift the blocks. Of course, if *nothing* prevents the goal, we do not need any knowledge of blocks and arms to answer the question: we have given too much away.

Putting a finer point on our knowledge, we might say "nothing prevents the arm from functioning according to specification". We will be charitable, for the moment, and assume that this precludes monkeys. Can the arm lift the blocks now? Well, the blocks may be too slippery, may explode when touched, or any of a number of things "too ridiculous to consider" may happen. It seems that, short of being explicitly told – or actually trying – we can never know enough to decide whether an attempted lift will succeed.

Even in situations where one intuitively should have complete knowledge, incompleteness may result from the impracticality of representing all of the relevant information. For example, an airline database which records flights and the cities which they connect would be overwhelmed if it had to keep track of all of the pairs of cities *not* connected by each flight. If this "negative" information is not explicitly stored, however, how can we decide whether PW819, which connects Vancouver and Guyamas, connects Tokyo and Hong Kong? The traditional approach to this problem has been to invoke the *Closed-World Assumption.* If we assume that we have complete knowledge about all of the positive facts true of the world, we can infer that anything we do not know to be true – such as *CONNECTS(PW819,Tokyo,HongKong)* – is false.

If our knowledge about any aspect of the world may be incomplete, however, this assumption is obviously suspect. Suppose, for example, that we want to start a flight from Vancouver to the New York City area, but do not yet know whether it will actually go to New York or Newark. Perhaps the database also stores information about flights' "home port" for maintenance purposes. We may want to enter Fictictious Airlines 001, with home port Vancouver, so that the maintenance department can gear up for the extra aircraft. The closed-world assumption would then allow us to infer that FAL001 connects Vancouver to neither New York nor Newark (nor anywhere else, for that matter).

To prevent such unwarranted inferences, we must retract our assumption of complete knowledge. Thus we can no longer use the closed-world assumption. As a side-effect, our uncertainty about FAL001 reintroduces uncertainty about whether PW819 connects Tokyo and Hong Kong. In this case, we might decide to manage the uncertainty by explicitly stating for which flights we have complete knowledge. The closed-world assumption can then be used where it is appropriate, and avoided elsewhere.

The closed-world assumption is often made even when its applicability cannot be guaranteed. One can imagine situations – in domains less structured than airline databases – in which it may not be known whether the information at hand is complete. Physicists, for example, periodically believe that they have tracked down the full suite of subatomic particles, and work using this assumption. So far, no-one has been able to say how we will *know* when all such particles have been discovered. In such situations, the best course of action is often to act as though one has complete information until one has reason to suspect otherwise. The question of when to suspect otherwise then becomes quite important. The principles which guide this type of reasoning appear difficult to elucidate. Certainly, knowing nothing is reason to doubt that one knows everything, but where does one draw the line?

Closed-world reasoning takes positive facts as given, and sanctions negative conclusions. Commonsense reasoning often requires a different sort of assumption to be made. Because of the need to act, and the pervasiveness of incomplete information, humans are usually willing to assume – often quite unconsciously – vast numbers of "normalcy" conditions without explicit justification. In planning to get to the airport by going out the front door, getting into one's car, and driving, one assumes that the door will open, the car will start, the airport hasn't moved, and that one's usual route is still passable. These assumptions rarely reach the conscious level, unless circumstances make it likely that they will be violated. For example, at -40 ° C, one might make contingency plans for the car's failing to start.

Should subsequent information or reflection violate any of these "implicit" assumptions, adjustments are made; but the absence of violation need not be *proven* before assumptions are made. The kinds of assumptions which are made to deal with the various forms of incomplete information cannot be sound, in the usual sense of never leading from true premises to false conclusions. This is disappointing to the purist. Unfortunately, if one wants to get anything done, certain assumptions must be made.

If we are willing to forsake soundness, how do we avoid embracing irrationality? The best one can hope for is some form of "justification" for one's assumptions; principles which allow gaps in one's knowledge to be filled and which guarantee that – most of the time – these assumptions will not lead too wildly astray. Deciding what constitutes the "normal" state-of-affairs and when to assume that things are indeed "normal" are important problems. Clearly, one must be very good at detecting abnormal conditions before assuming that everything is normal. Furthermore, once such assumptions have been made, one must be prepared to detect and deal with any conflicting (or apparently conflicting) information which turns up.

1.1. Overview of the thesis

The thesis attempts to pull together a number of threads – aspects of various approaches to reasoning with incomplete information. The results presented fall into two main categories: those which extend our understanding of the capabilities and limitations of particular approaches, and those which explore the interconnections, similarities, and differences between approaches. (Hopefully the latter category is subsumed by the former.)

Chapter 2 presents a detailed survey of a number of important systems for non-monotonic reasoning. We draw together a number of results from the literature and some original observations. The emphasis is on presenting a cohesive picture of the field. The presentation thus attempts to stress the commonalities and essential differences of the various approaches. The reader should be able to come away with an understanding of both the problems and state of the art of the field.

Chapter 3 consists of investigations of the properties of a particular formal system, Reiter's logic[1] for default reasoning. We present a general semantics for default theories, and show how this semantics highlights the essential similarities and dissimilarities between default logic and other non-monotonic systems. We then characterize a broad class of default theories which are well-behaved, in the sense of preserving the coherence of the underlying world-description.

We turn, in chapter 4, to an investigation of inheritance networks with exceptions. We develop a correspondence between such networks and default theories. We then use this correspondence to prove a number of interesting results, including sufficient conditions for the correctness of a network inference algorithm and for the coherence of an inheritance network representation of a body of knowledge. We conclude by showing that Touretzky's [1984a] "inferential distance" algorithm satisfies these criteria.

Chapters 5 through 7 turn from default logic to discuss a quite different approach to incomplete information, the various forms of minimal entailment or circumscription. In chapter 5, we discuss a number of semantically-motivated pessimistic results concerning the capabilities of predicate circumscription. Chapter 6 looks at a generalization of predicate circumscription, called formula circumscription. Model-theories are presented for some variants of this approach, and a number of results (both positive and negative) are proved concerning their power.

The "long-dead" domain circumscription formalism is "resurrected" in chapter 7. We argue that this approach provides an important capability for common-sense and database reasoning systems. We uncover and correct an error in the original presentation, and we show that a niche remains for domain circumscription by refuting subsumption claims made in favour of predicate (and formula) circumscription. We conclude the chapter with some results concerning domain circumscription's capabilities and limitations.

In chapter 8, we return to default logic, this time in the context of our discussion of circumscription. We present a number of results detailing the relationship between these rather disparate formalisms, showing their points of correspondence and their (unfortunately) more-frequent points of divergence.

The thesis concludes with a lengthy discussion of some important open problems and interesting research directions, in chapter 9, and a summary and evaluation of the significance of the work presented, in chapter 10.

Every attempt has been made to make the thesis as self-contained as possible. A familiarity (at times, intimate) with first-order logic is assumed throughout, however. (See [Mendelson 1964] for an introduction.) To preserve continuity, the proofs of the theorems have been relegated to Appendix A, while Appendices B and C contain notational conventions and definitions of logical terms assumed elsewhere in the thesis. The intention has been to keep the degree of logical sophistication required to read the bulk of the thesis to a minimum.

[1] Some have objected to the use of the term "logic" (and even "formal") for the systems we discuss here. Rather than debate this issue, we encourage those who find the terminology objectionable to substitute whatever term(s) they find appropriate.

CHAPTER 2

Approaches to Incomplete Knowledge

Traditional logics suffer from
the 'Monotonicity Problem'

— Drew McDermott

In traditional logical systems, extending a set of axioms can never prevent the derivation of conclusions derivable from the original set. More formally, if S and S' are arbitrary sets of formulae then:

$$S \subseteq S' \rightarrow \{w \mid S \vdash w\} \subseteq \{w \mid S' \vdash w\}.$$ [1]

The addition of formulae to a set *monotonically* increases what can be proved from that set; hence such logics are sometimes called *monotonic*.

Recently, it has been noted [McCarthy 1977, Minsky 1975] that monotonic logics seem inadequate to capture the tentative nature of human reasoning. Since people's knowledge about the world is necessarily incomplete, there will always be times when they will be forced to draw conclusions based on an incomplete specification of pertinent details of the situation. Under such circumstances, *assumptions* are made (implicitly or explicitly) about the state of the unknown factors. Because these assumptions are not irrefutable, they may have to be withdrawn at some later time should new evidence prove them invalid. If this happens, the new evidence will prevent some assumptions from being used; hence all conclusions which can be arrived at only in conjunction with those assumptions will no longer be derivable. This causes any system which attempts to reason consistently using assumptions to exhibit non-monotonic behaviour.

Common-sense conclusions are often based on both supporting evidence and the absence of contradictory evidence. Traditional logics cannot emulate this form of reasoning because they lack any means for considering the absence of knowledge. A number of systems have been developed to address this shortcoming, by augmenting a traditional first-order logic with some mechanism for predicating conclusions on the absence of specific knowledge.

In AI, logic-based attempts to solve the problems presented by incomplete information have fallen into two categories. (For the purposes of this thesis, we ignore "probabilistic" approaches.) The first category includes those which assume that all of the relevant *positive* information (*e.g.*, which individuals exist, which predicates are satisfied by which individuals) is known. From this

[1] $S \vdash w$ means w is provable from premises S.

assumption, it follows that anything which is not "known" to be true must be false. Negative facts[2] can thus be omitted, since they can be inferred from the absence of their positive counterparts. Such assumptions of complete positive knowledge underlie PLANNER's "THNOT" [Hewitt 1972] and related negation operators in AI programming languages, semantic networks, and databases [Reiter 1978a, b], as well as more formal reasoning techniques such as predicate completion [Clark 1978], and circumscription [McCarthy 1980, 1986].

In contrast, many have wanted to represent and use what would generally described as "default" or "prototypic" information. Defaults are used to fill gaps in knowledge. *In the absence of specific evidence*, they allow a system to make (hopefully) enlightened "guesses", instead of reserving judgement or assuming that whatever is unknown is false. Non-monotonic logic [McDermott & Doyle 1980], default logic [Reiter 1980a], truth maintenance systems [Doyle 1979, McAllester 1978, 1980], and various network- and frame-based procedural knowledge representation schemes [Quillian 1968, Minsky 1975] all embody this idea.

The two approaches are not mutually exclusive – each of these reasoning techniques has been used to achieve the other. Comparisons of the power of the two paradigms are most notable for their absence from the literature, however. The discussion in the remainder of this chapter does not provide such a comparison, although some points of correspondence are indicated.

2.1. Closed-World Reasoning

Negative facts – those which state what is *not* true about the world – vastly outnumber positive facts. For example, in a discussion at a sufficiently high level, everything which is at some place is *not* at *every* other place. Similarly, if Tumnus is a cat, he is not a dog, a fish, or a tree (among other things). The amount of negative information about a world increases geometrically with the size of the Herbrand Universe. One would like to avoid having to explicitly represent all such information. The information must somehow be available, however – at some point it may become useful to know that Tumnus is not a dog.

In certain situations, it is reasonable to assume that one knows all of the relevant truths. For example, it is reasonable to assume that a company's inventory database lists all parts supplied by that company, that one's T4 slips list all deductions from one's income, and that one's electricity will not be cut off tomorrow. Such assumptions are justified either by the design and intended function of the instrument in question or, as in the latter example, by the implicit belief that if a fact were important enough – such as the impending cessation of one's electric services – one would presumably have heard about it.

If one assumes "total knowledge about the domain being represented", it is no longer necessary to explicitly represent negative infomation. Negative facts may simply be inferred from the absence of their positive counterparts. Reiter [1978a] calls this assumption the *Closed-World Assumption* (CWA), since it implies a closed domain in which all truths are known. The closed-

[2] A fact is negative *iff* all of the literals in its clausal form are negative.

world assumption on a knowledge-base, *KB*, corresponds roughly to an inference rule of the form:

If $KB \not\vdash P$ *then infer* $\neg P$

applicable to positive facts, *P*. This rule can be paraphrased as "If *P* is not provable from the knowledge-base, assume $\neg P$."

2.1.1. Naive Closure

Reiter provides the following syntactic realization of the CWA, which we will call *naive closure (NC)*:[3] Define $\overline{EKB}$, the negative extension of KB, as follows:

$$\overline{EKB} = \{ \neg P\vec{\alpha} \mid P \text{ is an } n\text{-ary predicate letter, } \alpha \text{ is an } n\text{-tuple of ground terms, and } KB \not\vdash P\vec{\alpha} \}.$$

Then the naive closure of *KB* is defined as those formulae provable from $KB \cup \overline{EKB}$. We write $KB \vdash_{NC}$.

It is important to notice that naive closure extends the knowledge-base by adding a set of ground literals. Universal statements capturing the CWA for particular predicates do not generally follow from the naive closure of the knowledge-base.[4] For example, if $KB = \{ Penguin(Opus) \}$, then

$$\overline{EKB} = \{ \neg Penguin(Tweety), \neg Penguin(Fred), \ldots \}$$

but

$$KB \not\vdash_{NC} \forall x.\ x \neq Opus \supset \neg Penguin(x).$$

To see this, notice that we can construct a model for $KB \cup \overline{EKB}$ with a domain element, say α, which does not correspond to any named by *KB* or $\overline{EKB}$ and set $Penguin(\alpha)$ true.

A semantic characterization of this type of closed-world reasoning can be given in terms of minimal Herbrand models, as outlined below. We introduce the notion of minimal model in greater generality than is immediately required. This will help simplify subsequent discussion and clarify the relationships between the various formalisms we will be discussing.

In general, if we are given an ordering relation, $\leq$ on some class of interpretations, $\mathbf{I}$, we say that $I \in \mathbf{I}$ is *minimal* in $\mathbf{I}$ iff $\forall I' \in \mathbf{I}.\ \neg(I' \leq I)$ or $(I' = I)$. For our present purposes, let $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{Z}$ be disjoint sets of predicate-letters which jointly exhaust the supply of predicate-letters of the language. We can define an ordering $\leq$ on sets of Herbrand interpretations as follows:[5] ,[6]

$$I_1 \leq I_2 \equiv \forall P \in \mathbf{P}.\ |P|_{I_1} \subseteq |P|_{I_2}, \text{ and } \forall Q \in \mathbf{Q}.\ |Q|_{I_1} = |Q|_{I_2}.$$

In other words, the extensions of predicates in $\mathbf{Q}$ are identical, and those in $\mathbf{P}$ are (not necessarily

[3] A confusion prevalent in the literature conflates 'the CWA' with what we are calling 'naive closure'.

[4] To simplify the discussion, we assume here that there is at least one ground term.

[5] Note that this is *not* the standard mathematical notion of substructure or submodel.

[6] We will use $|I|$ to represent the domain of the interpretation, *I*, and $|P|_I$, $|t|_I$ to represent the interpretation in *I* of the predicate, *P*, and term, *t*, respectively.

proper) subsets. Observe that nothing is said about the interpretations of the predicates in **Z**.

Returning to the semantics of the CWA, it can be shown [van Emden and Kowalski 1976] that the naive closure of *KB* corresponds to minimality in the set of Herbrand models of *KB*, in the following sense. Let **P** be the set of all predicate symbols of *L* (hence $\mathbf{Q} = \mathbf{Z} = \{\ \}$). Then, if *KB* is Horn and consistent, there is a unique minimal element, *M*, in the class of Herbrand models of *KB* (in fact, $M = \bigcap \{ M \mid M \text{ is a Herbrand model of } KB \}$). Furthermore, for ground clause, L, $KB \cup \overline{EKB} \vdash L$ iff $M \models L$.

The class of Herbrand models of a theory is interesting for common-sense reasoning because each Herbrand model contains precisely the individuals for which the theory provides names. Intuitively, this is attractive for closed-world reasoning, since one would imagine that a closed world would contain no spurious individuals. Unfortunately, as we have seen, in general both *KB* and $KB \cup \overline{EKB}$ may have models with individuals not corresponding to any name. This accounts for the fact that, while $KB \cup \overline{EKB}$ agrees with *KB*'s minimal Herbrand model, *M*, for ground clauses, there may be facts true in *M* which do not follow from the naive closure of *KB* (specifically, those which entail there being exactly the named individuals). If the knowledge-base entails that there are only finitely-many individuals, it can be shown that *M* is the only model (up to isomorphism) for $KB \cup \overline{EKB}$.

Despite its attractiveness as a means of implicitly representing negative knowledge, closed-world reasoning is not without shortcomings and pitfalls. The most obvious of these is that there is no room for genuinely incomplete knowledge under the CWA – anything which is not known will be assumed false. To see the problems presented by incomplete information, consider a database consisting of only $BLOCK(A) \vee BLOCK(B)$. Since it is possible to derive neither $BLOCK(A)$ nor $BLOCK(B)$, naive closure allows the derivation of $\neg BLOCK(A)$ and $\neg BLOCK(B)$. It is easy to see that such situations lead to inconsistent conclusions.

The fact that some classically consistent databases are not consistent with naive closure leads to the question, "Under what circumstances can naive closure be consistently employed?" There is no complete characterization of suitable databases, and the only known sufficient condition is that the database be Horn and consistent. Purely negative information (clauses without positive literals) plays no part in closed-world query evaluation for such databases. Since negative information can be reconstructed using the CWA, it can be ignored without loss of deductive power [Reiter 1978a].

A more subtle drawback is that the " $\not\vdash$ " relation is not effectively computable, since first-order provability is only semi-decidable. Thus, even where naive closure preserves consistency, it may be impossible to even enumerate all of its consequences. While first-order logic is semi-decidable, and its theorems recursively enumerable, neither of these holds for first-order logic + naive closure.

2.1.2. Negation As Failure To Derive

AI programming languages (*e.g.,* PROLOG [Roussel 1975], PLANNER [Hewitt 1972]) have often addressed the problem of negative knowledge by adopting a weakened form of the CWA. They represent only positive information, assuming that whatever cannot be shown to be true must be false. Such systems embody a weakened form of the CWA because they do not fully implement the " $\not\vdash$ " relation. A derivation of $\neg P$ typically consists of an unsuccessful exhaustive search for a derivation of P. This technique is called ***negation as failure*** (NAF). Because the search-space may not be finite, the search for a derivation of P may never fail, even when P truly does not follow from the knowledge-base. Thus, NAF may not be able to find all of the negative facts implied by the CWA.

In PROLOG, an attempt to prove the literal, $\neg P$, consists of (recursively) attempting to prove P. If this fails, having exhausted the potential proofs for P, then the proof of $\neg P$ succeeds. This is the only inference rule for negation, and it is applicable only when P is a positive ground literal. Clark [1978] justifies this approach to negation by showing that the inference of $\neg P$ from a database, DB, by NAF corresponds to a proof of $\neg P$ from an extended database which is implicitly given by DB. (This extended database is discussed in detail in the next section.) Clark shows that NAF can be viewed as a derived inference rule, a heuristic for deriving negative facts which are (under the CWA) implicit in the database.

Because of the requirement of finite failure, the syntactic form of the database, as well as its logical content, can play a role in what can be derived by NAF. For example, Sheperdson [1984] points out that while the databases:

$$DB_1 = \{ \; Pa \; \}$$

and:

$$DB_2 = \{ \; \neg Pa \supset Pa \; \}$$

are logically equivalent, PROLOG can prove Pa only from DB_1. An attempt to prove Pa from DB_2 leads to an infinite proof tree: The subgoal $\neg Pa$ is set up, leading to a further subgoal of (failure to prove) Pa, ***ad infinitum***. Although the attempt to prove Pa obviously fails, it does not ***finitely*** fail, so the failure proof never returns. Of course, both databases logically entail Pa, and the CWA functions correctly in either.

2.1.3. Database Completion

The CWA allows a system to act on the assumption that "the objects that can be shown to have a certain property by reasoning from certain facts are all the objects that have that property" [McCarthy 1980]. It does not, however, allow the reasoner to derive this assumption. Such systems can never be "conscious" of the underlying principles which they are implicitly assuming. Clark [1978] remedies this shortcoming by making the completeness assumptions explicit in the database. All of the information about a particular relation in the database, DB, is gathered together and a *completion axiom* is added which states that a particular tuple satisfies the relation only in those cases where DB says it must. Applying this process to all of the relations in DB yields the *completed database* (C(DB)). This completion of the database makes explicit the assumptions of total world knowledge.

The database is viewed as a set of clauses, each with at most one distinguished positive literal. A clause is said to be *about* the predicate occurring in its distinguished positive literal. All of the clauses in DB about each n-ary predicate, P, are gathered together and converted to equivalent implications with $P(x_1,...,x_n)$ as their consequents. This implicative form makes clear all of the conditions which DB gives as sufficient for P. Predicate completion asserts that these conditions are also necessary, thus yielding a *definition* for P. If $E_1(\vec{x}),...,E_k(\vec{x})$ are the left-hand sides of all of the implications for $P(\vec{x})$ in DB, then the *completion axiom* for P in DB is:

$$\forall \vec{x}.\ P(\vec{x}) \supset [E_1(\vec{x}) \vee ... \vee E_k(\vec{x})].$$

If there are no axioms about a predicate, the completion axiom says that that predicate is universally false. The *completed database*, C(DB), is the original database, together with the completion axioms for each predicate. For example, the theory:

$$Bird(Tweety) \qquad (1)$$
$$\forall x.\ Penguin(x) \supset Bird(x) \qquad (2)$$
$$\forall x.\ Bird(x) \wedge \neg Penguin(x) \supset Flies(x) \qquad (3)$$

gives rise to the following implications about *Bird*:

$$\forall x.\ x = Tweety \supset Bird(x) \qquad (1')$$
$$\forall x.\ Penguin(x) \supset Bird(x) \qquad (2')$$

(*Bird* does not occur positively in (3)). Thus, the completion axiom for *Bird*, given these axioms is:

$$\forall x.\ Bird(x) \supset x = Tweety \vee Penguin(x)\ . \qquad (4)$$

Similarly, the completion axiom for *Flies* is:

$$\forall x.\ Flies(x) \supset Bird(x) \wedge \neg Penguin(x)\ , \qquad (5)$$

and the completion axiom for *Penguin* is:

$$\forall x.\ \neg Penguin(x)\ . \qquad (6)$$

(We have assumed that (3) is about *Flies*.) Hence, C({(1), (2), (3)}) = {(1′), (2′), (3)-(6)}, which says the only birds are *Tweety* and the penguins, and all non-penguin birds fly. Furthermore, there are no penguins, so all (and only) birds fly.

Besides the original theory and the completion axioms, Clark adds "Unique Names Axioms" [Reiter 1978a]. These are inequality axioms stating that different names denote different objects. Thus, for example, if we add:

$$Penguin(Opus)$$

to the DB (1)-(3), we get the new completed database:

$$C(DB') = \{\ (1'),\ (2'),\ (3)\text{–}(5),\ \forall x.\ Penguin(x) \equiv x = Opus,\ Opus \neq Tweety\ \},$$

which entails $Flies(Tweety)$ and $\neg Flies(Opus)$. Without the unique names axiom, $C(DB')$ would entail neither $Flies(Tweety)$ nor $\neg Flies(Opus)$.

When restricted to Horn databases, which have at most one positive literal, database completion preserves consistency. However, if clauses are allowed to have more than one positive literal problems may result. For example, the clausal form of (3), above, is:

$$\neg Bird(x) \vee Penguin(x) \vee Flies(x)\ .$$

We arbitrarily decided that (3) was about *Flies* (because it illustrated our point), but we could as easily have chosen *Penguin*. It is easy to see that our choice makes the completion of:

$$DB = \{\ (1)\text{–}(3),\ \neg Flies(Tweety)\ \}$$

inconsistent. Because (3) is taken to be about *Flies*, it is not taken into account when calculating the completion of *Penguin*, even though it can be used to infer $Penguin(Tweety)$. Hence, the completion axiom stating that there are no penguins can still be derived, even though it is now inconsistent with *DB*.

Database completion can sometimes be consistently extended to non-Horn theories by treating a clause with positive literals, $L_1,...,L_k$, as k clauses, each *about* a different L_i. This may allow database completion to be applied to databases containing incomplete information without introducing inconsistencies. For example, the database:

$$BLOCK(A) \vee BLOCK(B),$$

which is not Horn and is inconsistent with naive closure, can be rewritten as:

$$\left\{ \begin{array}{l} \forall x.\ [\neg BLOCK(A) \wedge x = B \supset BLOCK(x)]\ , \\ \forall x.\ [\neg BLOCK(B) \wedge x = A \supset BLOCK(x)] \end{array} \right\}$$

These result in the consistent completed database:

$$\{\forall x.\ [BLOCK(x) \equiv (\neg BLOCK(A) \wedge x = B) \vee (\neg BLOCK(B) \wedge x = A)],\ A \neq B\ \}\ ,$$

or equivalently,

$$\{[\forall x.\ BLOCK(x) \equiv x = A] \vee [\forall x.\ BLOCK(x) \equiv x = B],\ A \neq B\ \}\ .$$

Notice that the completed database states that there is exactly one block, and it must be either *A* or *B*. The disjunction in the original database, which did not exclude the possibility of two blocks, has become "exclusive" in the completed database.

This approach has two drawbacks. First, the price paid for preserving consistency is weakened conjectures. For example, if axiom (3) is treated as also being about penguins, the completion axiom for *Penguin* in (1)–(3) becomes:

$$\forall x.\ Penguin(x) \supset Bird(x) \wedge \neg Flies(x)\ ,$$

and the completed database no longer allows us to conclude that *Tweety* does not fly. In fact, it is a simple corollary of results by Reiter [1982] and those in chapter 5 that predicate completion cannot be used to conjecture positive facts (such as $Flies(Tweety)$) without risk of inconsistency.

A more serious drawback, however, is that this weakened form still does not guarantee consistency. Shepherdson [1984] shows that the database:

$$P(a) \vee P(a) \tag{7}$$

has an inconsistent completion, namely:

$$\forall x.\ P(x) \equiv x = a \wedge \neg P(a)\ .$$

This is especially disturbing, since (7) is equivalent to the trivial database, $P(a)$. Perhaps consistency can be guaranteed by restricting databases to some normal form which precludes (7), but excluding all problematic cases would presumably require a sophisticated algorithm, capable of determining when one set of clauses subsumes another. Such an algorithm would lose some of the advantages of simplicity and directness which predicate completion enjoys. Normal forms aside, the precise limits of the consistent applicability of predicate completion are as yet unknown.

This illustrates what is simultaneously a strength and a weakness of database completion. The manipulations involved in completing the database are deterministic syntactic transformations. Any database can thus be effectively completed with relatively little effort. This same fact, however, means that logically equivalent databases may have different completions. Thus, the syntactic forms of formulae take on semantic significance, which is foreign to most logical systems. Besides sometimes leading to inconsistency, this seems to argue against Clark's view that the completion axioms are somehow implicit in the database.

Reiter [1984] explores the effects of adding completion axioms to normal relational databases. He demonstrates applications of these techniques to problems involving some types of incomplete information commonly encountered in the database field, such as null values and disjunctive information.

Database completion is more powerful than a first-order system augmented by NAF. Clark shows that, for PROLOG programs, the structure of a failure proof is isomorphic to that of a first-order proof from the completed database. Conversely, the completion of the database:

$$DB = \{Penguin(Opus)\}$$

is:

$$C(DB) = \{\forall x.\ [Penguin(x) \equiv x = Opus]\} \tag{8}$$

from which $\forall x.\ [x \neq Opus \supset \neg Penguin(x)]$ follows by first-order reasoning. For any particular $x \neq Opus$, NAF applied to DB can show $\neg Penguin(x)$, but the universal summary (8) is beyond its capabilities. (This follows from the fact that NAF is weaker than naive closure and naive closure cannot derive the universal summary.)

Database completion does not avoid all of the problems of NAF simply because all of the deductions are first-order. There will still be propositions which are not decided by the completed database – for example, propositions corresponding to those for which the exhaustive search for a failure proof never terminates. Consider the database:

$$DB = \{\ Penguin(Opus),\ \forall x.\ Penguin(father(x)) \supset Penguin(x)\ \}$$

which says that the property of being a penguin is handed down from father to son. NAF cannot prove $\neg Penguin(Bruce)$ because the search for a derivation of $Penguin(Bruce)$ will search forever for a penguin among *Bruce*'s paternal ancestors. The completed database,

$$C(DB) = \{\ \forall x.\ Penguin(x) \equiv x = Opus \lor Penguin(father(x)),\ Bruce \neq Opus\ \}$$

also fails to entail $\neg Penguin(Bruce)$. Because of the circularity in the definition for *Penguin*, it cannot prove the nonpenguinity of his father.

2.1.4. Generalized Realizations of the CWA

The CWA is the assumption of complete knowledge about which positive facts are true in the world. As we have seen, this assumption is not always appropriate, and can lead to inconsistency if made in situations where knowledge is genuinely incomplete. This has led a number of researchers to develop more sophisticated knowledge-closing operators which are able to handle incompleteness in certain aspects of the KB without completely retreating to the "Open-World Assumption" that what is known is precisely what follows from what is explicitly stated.

To specify the *generalized closed-world assumption (GCWA)*, Minker [1982] also uses minimal models to characterize what follows from the closure of the database. Restricting his attention to clausal databases (hence to universal theories) with a finite set of terms, Minker considers the set of minimal Herbrand models of the database. (For non-Horn theories, there may not be a unique minimal Herbrand model.)

The GCWA augments the database with the negations of all the literals which are false in all of its minimal Herbrand models. It can be shown that the resulting extended database is consistent iff the original database is, and that no new positive clauses are derivable from the augmented database.

To illustrate the idea, consider the theory $\{BLOCK(A) \lor BLOCK(B), BLOCK(C) \lor \neg BLOCK(D)\}$. This database has nine Herbrand models:

$M_1 = \{BLOCK(A), BLOCK(B), BLOCK(C), BLOCK(D)\}$
$M_2 = \{BLOCK(A), BLOCK(B), BLOCK(C), \neg BLOCK(D)\}$
$M_3 = \{BLOCK(A), BLOCK(B), \neg BLOCK(C), \neg BLOCK(D)\}$
$M_4 = \{BLOCK(A), \neg BLOCK(B), BLOCK(C), BLOCK(D)\}$
$M_5 = \{BLOCK(A), \neg BLOCK(B), BLOCK(C), \neg BLOCK(D)\}$
$M_6 = \{BLOCK(A), \neg BLOCK(B), \neg BLOCK(C), \neg BLOCK(D)\}$
$M_7 = \{\neg BLOCK(A), BLOCK(B), BLOCK(C), BLOCK(D)\}$
$M_8 = \{\neg BLOCK(A), BLOCK(B), BLOCK(C), \neg BLOCK(D)\}$
$M_9 = \{\neg BLOCK(A), BLOCK(B), \neg BLOCK(C), \neg BLOCK(D)\}$

of which M_6 and M_9 are minimal. Accordingly, the GCWA sanctions $\neg BLOCK(C)$ and $\neg BLOCK(D)$, since they are both false in all minimal Herbrand models, but yields no conclusions about which of A and B are blocks. Thus, where the database could consistently be construed as closed, the GCWA closes it, but where it is known to be incomplete (*i.e.*, $BLOCK(A) \lor BLOCK(B)$), no conclusion is drawn.

Because Horn theories have unique minimal Herbrand models, it is easily seen that this definition of the GCWA corresponds to naive closure for Horn theories. The GCWA has the advantage that it does not overcommit itself to the principle that all positive information is known. Faced with a situation where some positive information is clearly not known, judgement is reserved, rather than blundering into inconsistency.

Minker also provides a syntactic definition of the GCWA, which he proves corresponds to the semantic characterization given above. The database, *DB*, is extended by adding $\overline{\overline{EDB}}$, the set of negations of ground atomic formulae occurring in minimal positive clauses derivable from *DB*. Specifically:

$$\overline{\overline{EDB}} = \{ \neg P\vec{c} \mid \forall K.\ DB \not\vdash (P\vec{c} \vee K), \textit{ where } K \textit{ is a disjunction of } \\ \textit{0 or more positive literals such that } DB \not\vdash K \}$$

It is easily seen that, for Horn theories, this reduces to:

$$\overline{\overline{EDB}} = \overline{EDB} = \{ \neg P\vec{c} \mid DB \not\vdash P\vec{c} \}$$

– the closure set generated by naive closure – since for Horn *DB* and a positive clause, *K*, $DB \vdash (P\vec{c} \vee K)$ iff $DB \vdash P\vec{c}$ or $DB \vdash K$.

Minker proves the GCWA preserves consistency – $DB \cup \overline{\overline{EDB}}$ is consistent iff *DB* is – and introduced no new positive information – if *K* is a positive clause, then $DB \cup \overline{\overline{EDB}} \vdash K$ iff $DB \vdash K$. These facts, together with the fact that the GCWA subsumes naive closure indicate that the GCWA is an interesting extension. Of course, the GCWA is even less tractable than naive closure (to the extent that *either* can be said to be tractable), since it involves multiple $\not\vdash$ tests for each literal. This suggests that naive closure might be preferred in those cases (Horn theories) where it is applicable.

Gelfond and Przymusinska [1985] have developed an extension of the GCWA and naive closure. Their "careful closure procedure" differs from the GCWA (and naive closure) in that the effects of closing the world can be constrained by indicating precisely which predicates may be affected.

The predicates of the theory are divided into three sets, **P**, **Q**, and **Z**. **P** consists of those aspects of the world which are to be closed; **Q** contains the predicates which are not to be affected by the closure; and the predicates in **Z** may be affected in any way (consistent with the knowledge-base) necessary to achieve maximum "closed-mindedness" about **P**.

This arrangement allows greater flexibility in closing the world. Firstly, by requiring that certain predicates not be affected by the closure (those in **Q**), one can avoid inadvertently making conclusions about, for example, the price of tea in China while one's intention was to conclude that the availability of tea at the local supermarket has not changed. Secondly, allowing the predicates in **Z** to vary weakens the GCWA/naive closure restriction that no new positive facts be derivable from the closure of the database. This means that if one is confident that one has all the positive information about **P**, but knows only certain constraints on the relationship between **P** and **Z**, then **Z** can vary as necessary to establish the minimal extensions for **P**.

The "careful closure" of DB with respect to ($\mathbf{P}$, $\mathbf{Q}$, $\mathbf{Z}$) is defined as $DB^* = DB \cup \overline{\overline{\overline{EDB}}}$, where

$$\overline{\overline{\overline{EDB}}} = \{\neg P\vec{c} \mid \forall \{L_1,\ldots,L_n\} \subseteq (\mathbf{P}^+ \cup \mathbf{Q}^+ \cup \mathbf{Q}^-).\ DB \not\vdash L_1 \vee \ldots \vee L_n,\ P\vec{c} \notin \{L_i\},$$
$$\text{or } \exists k<n.\ DB \vdash L_1 \vee \ldots \vee L_k\}$$ [7]

Intuitively, one can assume $\neg P\vec{c}$ unless this would allow the derivation of new facts about $\mathbf{Q}$ and/or positive $\mathbf{P}$.

The semantic definition of careful closure again involves a variant of the notion of minimal Herbrand model outlined earlier, this time in its full generality ($\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{Z}$ may all be non-empty). Gelfond and Przymusinska show that, for a universal knowledge-base, KB, every minimal Herbrand model of KB satisfies KB^*, and that KB^* is consistent iff KB is.

It is easy to see that if $\mathbf{Q} = \mathbf{Z} = \{\ \}$ then the above semantic characterization is the same as that for the GCWA. Furthermore, if the knowledge-base is also Horn, the same is true for naive closure. Since Gelfond and Przymusinska do not require that the knowledge-base be function free nor have a finite set of constants, this observation shows that these restrictions given in the development of the GCWA were unnecessary.

2.1.5. Circumscription

McCarthy [1977, 1980, 1986] has presented a number of rules of conjecture for closed-world reasoning. These rules are based on syntactic manipulations, rather than consistency. Instead of the undecidability of appeals to non-provability on which some approaches to non-monotonic reasoning are based, these "circumscriptive" formalisms simply add new axioms (conjectures). These conjectures force minimal, "closed-world", interpretations on particular aspects of the underlying incomplete theory.

2.1.5.1. Predicate Circumscription

The most widely studied of these rules of conjecture is "predicate circumscription" [McCarthy 1980]. Predicate circumscription allows explicit completeness assumptions, similar to Clark's completion axioms, to be conjectured as they are required. This provides a means for closing off the world with respect to a particular predicate at a particular time. A schema for a set of first-order sentences is generated. This schema is then instantiated by substituting suitable predicates for the predicate variables it contains. The particular substitution(s) chosen determine which individuals are conjectured to comprise the entire extension of the circumscribed predicate.

The semantic intuition underlying predicate circumscription is the now-familiar notion that closed-world reasoning about one or more predicates of a theory corresponds to truth in all models

[7] If $\mathbf{R}$ is a set of predicates, we use $\mathbf{R}^+$ and $\mathbf{R}^-$, respectively, to indicate the positive and negative *ground* literals over prediates in $\mathbf{R}$.

of the theory which are minimal in those predicates. Specifically, let $T(P_1,...,P_n)$ be a first-order theory, some (but not necessarily all) of whose predicates are those in $\mathbf{P} = \{P_1,...,P_n\}$. A model M of T is a **P**-*submodel* of a model M' of T (written $M \leq {}_{\mathbf{P}}M'$) iff the extension of each P_i in M is a subset of its extension in M', and M and M' are otherwise identical. M is a **P**-*minimal model of* T iff every **P**-submodel of M is identical to M.

For finitely axiomatizable theories, $T(P_1,...,P_n)$, McCarthy [1980] proposes realizing predicate circumscription syntactically by adding the following axiom schema to T:

$$\left[T(\Phi_1,...,\Phi_n) \wedge \bigwedge_{i=1}^{n}[\forall \vec{x}.\ (\Phi_i\vec{x} \supset P_i\vec{x})]\right] \supset \bigwedge_{i=1}^{n}[\forall \vec{x}.\ (P_i\vec{x} \supset \Phi_i\vec{x})].$$

Here $\Phi_1,...,\Phi_n$ are predicate variables, with the same arities as $P_1,...,P_n$, respectively. $T(\Phi_1,...,\Phi_n)$ is the sentence obtained by conjoining the sentences of T, then replacing every occurrence of $P_1,...,P_n$ in T by $\Phi_1,...,\Phi_n$, respectively. The above schema is called *the (joint) circumscription schema of* $P_1,...,P_n$ *in* T. Let $CLOSURE_{\mathbf{P}}(T)$ – *the closure of* T *with respect to* $\mathbf{P} = \{P_1,...,P_n\}$ – denote the theory consisting of T together with the above axiom schema. McCarthy formally identifies reasoning about T under the closed-world assumption with respect to the predicates **P** with first-order deductions from the theory $CLOSURE_{\mathbf{P}}(T)$.

McCarthy [1980] shows that any instance of the schema resulting from circumscribing a single predicate P in a sentence $T(P)$ is true in all $\{P\}$-minimal models of T. This generalizes directly to the joint circumscription of multiple predicates. An argument due to Davis [1980] can be used to show that no general "completeness" result can be obtained identifying the "circumscriptive theorems" with precisely those formulae true in all minimal models of the theory. Minker and Perlis [1983, 1984a] prove a "finitary" completeness result, however. Specifically, if the original theory (or the circumscribed version) entails that the minimized predicates have finite extensions, the minimal models of the original theory are all (and only) the models of the circumscribed theory.

McCarthy considers the blocks-world example, discussed previously, in which all that is known is:

$$BLOCK(A) \vee BLOCK(B)^{8} \tag{9}$$

If the predicate variable, Θ, in the circumscription of (9):

$$[\Theta(A) \vee \Theta(B)] \wedge \forall x.\ [\Theta(x) \supset BLOCK(x)] \supset \forall x.\ [BLOCK(x) \supset \Theta(x)]$$

is replaced successively by the predicates $x = A$ and $x = B$, the conjecture:

$$\forall x.\ [BLOCK(x) \supset x = A] \vee \forall x.\ [BLOCK(x) \supset x = B] \tag{10}$$

can be derived. As did the completed database, (10) says that there is only one block: A or B. Again, the conjecture closes the world and puts the "exclusive" interpretation on the original disjunction.

[8] Recall that this theory is NOT consistent with its naive closure.

The choice of substituends is crucial in determining what can be obtained by circumscription. It is not clear, in general, how these substituends are to be chosen. McCarthy suggests that the desired goal directs the choice of appropriate substitutions. It remains to be seen whether this can be translated into general rules.

The relationships between predicate circumscription and the various forms of closed-world reasoning are only partially understood. Reiter [1982] shows that predicate circumscription can sometimes be used to derive the database completion axioms. McCarthy circumscriptively derives the induction axiom for arithmetic, which shows that predicate circumscription is more powerful than database completion.

Doyle [1984] has observed that circumscription is related to the idea of implicit definability as it occurs in Mathematical Logic. A set of axioms, A, *implicitly defines* a predicate, P, if A forces a "unique" interpretation for P, or, more formally, if

$$A(\Phi) \supset [\forall \vec{x}.\ P\vec{x} \equiv \Phi\vec{x}]$$

is valid for each expression, Φ, of the same arity as P. It is easy to see that this schema implies the circumscription schema.

Beth's Definability Theorem [Beth 1953] guarantees that if A implicitly defines P then A *explicitly defines* P. That is,

$$A \vdash \forall \vec{x}.\ P\vec{x} \equiv \phi\vec{x}$$

where ϕ is some expression using only symbols of A (exclusive of P). This result is much-studied in logic, and the known consequences include methods for finding an appropriate ϕ. In those cases where the circumscription schema actually implicitly defines P, these techniques can be used to reduce the schema to an explicit definition axiom.

Circumscription does not always result in an implicit definition for P. In general, it is not even decidable whether P is implicitly defined. In the $Block(A) \vee Block(B)$ example cited above, for example, all that is obtainable is a *disjunctive definition*,

$$[\forall x.\ Block(x) \equiv x = A] \vee [\forall x.\ Block(x) \equiv x = B]\ .$$

There are techniques for finding disjunctive definitions with k disjuncts, where such definitions exist, but it is undecidable in general whether a disjunctive definition (or a disjunctive definition of size k) exists.

Doyle suggests that there may be profit in searching the Mathematical Logic literature (and enquiring of mathematical logicians) for results which may shed light on such questions as:

1) When does circumscription implicitly define P? Disjunctively? When does it fail? Are there interesting cases which can be characterized? Recognized?

2) What does circumscription do when it fails to define P?

3) When are new axioms irrelevant to prior circumscriptions? That is, when is the addition of new information guaranteed not to invalidate circumscriptively derived explicit definitions?

4) How can the revision of circumscriptive conclusions in the face of new information be mechanized?

We have discovered a number of surprising limitations on the applicability and efficacy of predicate circumscription. These are detailed in chapter 5.

2.1.5.2. Formula Circumscription

Many of the limitations of predicate circumscription stem from the fact that only those predicates being minimized are allowed to vary. McCarthy [1986] has developed a generalized form of circumscription which addresses this problem. This new formalism, formula circumscription, retains many of the attractive features of its predecessor, without some of its limitations. The formula circumscription axiom looks like:

$$\forall \vec{\Phi}.\ T(\vec{\Phi}) \wedge [\forall \vec{x}.\ E(\vec{\Phi},\vec{x}) \supset E(\vec{P},\vec{x})] \supset [\forall \vec{x}.\ E(\vec{P},\vec{x}) \supset E(\vec{\Phi},\vec{x})]$$

where $E(\vec{P},\vec{x})$ is any well-formed expression whose free individual variables are among $\vec{x} = x_1,\ldots,x_k$, and in which some of the predicate variables $\vec{P} = P_1,\ldots,P_n$ occur free; $E(\vec{\Phi},\vec{x})$ is the result of replacing each free occurrence of the predicate variables, P_i , in $E(\vec{P},\vec{x})$ with predicate variables, Φ_i , of the same arity.

There are three main differences between the predicate circumscription schema and the formula circumscription axiom. First, the former is a first-order axiom schema, while the latter is a second-order axiom. McCarthy suggests that this is advantageous because it allows the results of one circumscription to participate in subsequent circumscriptions. However, this feature is not essential; the second-order axiom can be replaced with a first-order schema. Although weaker, the first-order-schema variant appears adequate for many applications [Perlis and Minker 1986]. Investigations into the relative advantages and disadvantages of second-order axioms *vs* first-order schemas for circumscription are still continuing, and the question of the value of adopting a second-order logic remains undecided.

The second new feature of formula circumscription is that arbitrary predicate expressions, rather than simple predicates, may be minimized. McCarthy [1984, personal communication] suggests that this is an inessential change, since the same effect could be indirectly obtained by introducing new predicates, with axioms defining these predicates as equivalent to the required expression. While this is true for formula circumscription, we show in chapter 5 that predicate circumscription cannot deal with such definitions. We also discuss additional mechanisms which are sometimes used to augment predicate circumscription which allow definitions to be circumscribed. These mechanisms do not always preserve consistency, however.

The third, and most significant, innovation is that the predicates allowed to vary are no longer identified with those being minimized. This is reflected in the fact that $\vec{P}$ [alternately, $\vec{\Phi}$] may contain predicate variables not occurring in $E(\vec{P},\vec{x})$ [respectively, $E(\vec{\Phi},\vec{x})$] (and *vice versa*). This separation allows circumscription to operate in richly connected worlds. Provided predicates which would be altered by the minimization of the expression in question are among those identified as "variable", circumscription can have the desired effect.

Chapter 6 describes a model theory we have developed for formula circumscription, along the lines of McCarthy's [1980] semantics for predicate circumscription. For formula circumscription, the appropriate notion of submodel is one in which the extensions of the variable predicates are allowed to expand or contract, provided that the extension of $E(\vec{P},\vec{x})$ contracts. The extensions of the predicate parameters (those predicates which are not among the predicates designated as variable) must be identical in a model and its submodels. A model is minimal if it has no proper submodels. It is shown that formula circumscription is sound with respect to this model theory; anything derivable from the circumscribed theory is true in all minimal models of the original theory.

Perlis and Minker 1986] consider the completeness of the first-order-schema variant of formula circumscription with respect to this model theory. They present results analogous to their finitary completeness results for predicate circumscription [Minker and Perlis 1983, 1984a]. These results partially answer some of Doyle's [1984] questions about the relationship between circumscription and explicit/disjunctive definability, at least inasmuch as they establish explicit and disjunctive definability as sufficient conditions for the completeness of formula circumscription. These results have yet to be extended to the case of second-order formula circumscription.

Lifschitz [1984] has studied second-order formula circumscription and derived certain conditions under which the second-order circumscription axiom can be reduced to an equivalent first-order axiom. Such equivalences improve the usefulness of formula circumscription, in some cases, by eliminating both the need for a second-order logic and the problem of finding the "right" substitutions.

Lifschitz defines a formula to be *separable in P* iff it can be written in the form:

$$\bigvee_i \left[C_i \wedge [\forall \vec{x}.\ E_i(\vec{x}) \supset P(\vec{x})] \wedge [\forall \vec{x}.\ P(\vec{x}) \supset F_i(\vec{x})]\right]$$

where C_i, E_i, and F_i are P-free formulae. Essentially, a formula is separable if it is not recursive in P. Lifschitz proves that the second-order formula resulting from circumscribing P in a separable formula, A, allowing only P to vary is equivalent to a first-order formula with about the same logical complexity as A.

In itself, this result is not very exciting, since second-order circumscription of P with only P variable is subject to the same limitations chapter 5 outlines for predicate circumscription. Lifschitz also shows, however, that the circumscription of P in A with P and Y variable is equivalent to the circumscription of P in $[\exists\, Y.\ A]$ with only P variable. Furthermore, if A is separable in Y, then $[\exists\, Y.\ A]$ is equivalent to a first-order formula with complexity lower than A. While these transformations do not always preserve separability [Reiter, personal communication], it appears that these techniques may be useful for eliminating the second-order quantifiers introduced by formula circumscription – without re-introducing the awkwardness of axiom schemata and "right" substitutions.

Another innovation due to Lifschitz is to minimize according to arbitrary pre-orders (reflexive, transitive binary relations), rather than simple subset relations. Specifically, if $\mathbf{X}$ is an n-tuple of predicate, function, and/or constant letters of T, and $\mathbf{X}'$ is an n-tuple of predicate, function, and/or individual variables of corresponding types and arities, then the generalized circumscription axiom has the form:

$$\mathcal{T}(\mathbf{X}) \wedge \forall \mathbf{X}'.\ \mathcal{T}(\mathbf{X}') \wedge (\mathbf{X}' \leq_R \mathbf{X}) \supset (\mathbf{X} \leq_R \mathbf{X}')$$

where $\leq_R$ is an appropriate pre-order.

The use of pre-orders allows a number of interesting and potentially useful extensions to circumscription. For example, the pre-order $\mathbf{X} \leq_R \mathbf{Y}$ defined by

$$(\forall x.\ X_1 x \supset Y_1 x) \wedge ((\forall x.\ Y_1 x \supset X_1 x) \supset (\forall x.\ X_2 x \supset Y_2 x))$$

allows the joint minimization of the unary predicates X_1 and X_2, with the minimization of X_1 having a "higher priority".

The effect of allowing $\mathbf{X}$ to include constant and function letters is to allow constants and functions to vary during the minimization process. It appears that – for languages with finite sets of constants – it is possible to circumscriptively conjecture new facts about equality, including unique names axioms, by allowing constants to vary. Unfortunately, Lifschitz neither motivates nor discusses the variability of terms in detail. A semantic explanation of the process involved has yet to appear. In chapter 6, we show that allowing circumscriptively variable terms corresponds to weakening the definition of submodel in the semantic characterization of formula circumscription by dropping the requirement that a model and its submodels share identical interpretations of constant and function symbols. We also show that this approach can lead to some unexpected consequences.

2.1.5.3. Domain Circumscription

In database and commonsense reasoning, it is often necessary to assume that the only individuals whose existence is relevant to some task are those required to exist by what is known about the task. In such situations, the *domain-closure assumption* is made [Reiter 1980a]. This is the assumption that the "world" contains only individuals whose existence is required by the available information. Reiter observes that this assumption is implicit in relational database theory, where it is entailed by the manner in which universal queries are treated. Thus, for example, in the education database:

Teacher(Smith)	*Student(Brown)*
Teacher(Jones)	*Student(Black)*
Teacher(Plato)	*Student(Aristotle)*

with an integrity constraint specifying that the sets of teachers and students are disjoint, even the simple query, "Who are all of the teachers?" cannot be answered without implicitly assuming that the domain consists of only the listed individuals.

In cases where there are only finitely many individuals, this assumption can be stated using *domain-closure axioms*. These are axioms of the form:

$$\forall x.\ x = t_1 \vee \ldots \vee x = t_n \tag{11}$$

where the t_i are ground terms. Any model satisfying (11) will have at most n distinct individuals in its domain, those corresponding to the t_i. Reiter [1980a, 1984] shows that domain-closure axioms have an important role in logically formalizing the theory of relational databases.

Even when the domain cannot be enumerated to form a domain closure axiom, useful restrictions can sometimes be put on the size and composition of the domain by conjecturing that it coincides with the extension of some predicate or function whose extension is (partly) known. For example, in the education database discussed above, if it is known is that teachers are employees and students are not, assuming domain closure allows one to conjecture that teachers are the only employees. By conjecturing that the domain consists only of teachers and students (*i.e.*, $\forall x.\ Teacher(x) \vee Student(x)$), it becomes possible to deduce that there are no non-teacher employees (regardless of whether all of the teachers and students are known).

Domain-closure axioms are also important with respect to a variety of closed-world reasoning formalisms. Perlis and Minker [1986], for example, show that the effects of predicate and formula circumscription [McCarthy 1980, 1986] can be more precisely characterized in conjunction with closed-domain theories. Similarly, Clark [1978] requires domain-closure axioms in the development of his predicate completion approach.

Given the importance of domain-closure axioms, the question arises: Why not explicitly add them to theories? Probably the most important reason is that the appropriate domain-closure axiom may not be obvious. The repercussions of choosing too strong or too weak an axiom (inconsistency or loss of useful conjectures, respectively) argues in favour of a more automatic approach. Furthermore, as the state of the world (or the system's knowledge) changes to bring more entities into consideration, the same mechanism could be used to generate new domain-closure axioms. In certain cases, domain circumscription provides such an automatic mechanism.

Actually the first of the circumscriptive formalisms, domain circumscription [McCarthy 1977, 1980; Davis 1980] is intended to be a syntactic realization of the model-theoretic domain-closure assumption. It provides a mechanism for conjecturing domain closure axioms, eliminating the need to explicitly state them.

To circumscribe the domain of a sentence, A, McCarthy proposes adding the schema:

$$Axiom(\Phi) \wedge A^{\Phi} \supset \forall x.\ \Phi(x) \tag{12}$$

to A. $Axiom(\Phi)$ is the conjunction of $\Phi\alpha$ for each constant symbol α and $\forall x_1...x_n.\ [\Phi x_1 \wedge\cdots\wedge \Phi x_n] \supset \Phi f x_1...x_n$ for each n-ary function symbol f. A^{Φ} is the result of rewriting A, replacing each universal or existential quantifier, '$\forall x.$' or '$\exists x.$', in A with '$\forall x.\Phi x \supset$ ' or '$\exists x.\Phi x \wedge$ ', respectively.

This axiom schema represents the conjecture that the domain of discourse is no larger than it must be given the sentence A. For any predicate, Φ, if Φ is true for all individuals whose existence is given by the constant terms, through function application, or by existential quantification, and if all individuals in Φ's extension satisfy all of the universally quantified formulae, then Φ is assumed to contain the entire domain. If the extension of some predicate meeting these requirements is known, then the domain is (assumed to be) completely known.

The semantic intuition underlying domain circumscription is *minimal entailment*: only those models with minimal domains should be considered in determining the consequences of the given information. In this connection, a model, M, of a sentence is said to be a *submodel* of another model, N, if M is the restriction of N to a subset of N's domain. A model is said to be *minimal* if it has no proper submodels. Davis [1980] shows that every instance of (12) is true in all minimal models of the original sentence A. This result is correct for those theories with at least one constant symbol. In chapter 7, however, we show that inconsistency results when circumscribing theories whose prenex normal forms contain no leading existential quantifiers and no constant symbols. We also present a simple, easily motivated solution. This leads to a revised version of domain circumscription which is shown to preserve consistency.

2.1.6. Restricting Closed-World Inferences

One may want to do closed-world reasoning to form conjectures about the underlying principles governing a situation. In this case, one is making universal (inductive) conjectures about the state of the world. This is the type of reasoning which is involved in deducing laws, such as "an unsupported object drops when released". In many cases, however, closed-world reasoning yields stronger conjectures than may be desirable. For example, it is often sufficient to conclude that the situation immediately at hand does not have certain properties. In day-to-day reasoning, one is usually interested in forming particular conjectures in aid of completing a particular deduction. These conjectures should be of as limited scope as possible while still strong enough to allow the desired goal to be achieved. Thus, for example, if we knew that Tweety is a bird and that all birds except penguins fly, we might want to conjecture that Tweety could fly (and hence that Tweety is not a penguin). It is unlikely that we would want to conjecture that there are no penguins *at all*, however.

"Protected Circumscription" [Minker & Perlis 1984b] provides one means for delimiting the effects of closed-world reasoning. To prevent the circumscription of P in a theory, A, from conjecturing that S's are not P's, the predicate, S, is protected by weakening the circumscription schema to:

$$A(\Phi) \wedge [\forall \vec{x}.\ (\Phi\vec{x} \wedge \neg S\vec{x}) \supset P\vec{x}] \supset [\forall \vec{x}.\ (P\vec{x} \wedge \neg S\vec{x}) \supset \Phi\vec{x}]\ .$$

The conclusions of protected circumscription apply only to those individuals that do not satisfy the protected predicate. Thus, for example, McCarthy [1984, personal communication] has suggested that one may wish to conclude only that there are no penguins *present*. Assuming that there is a predicate, $Present(x)$, which says that an individual is in the immediate vicinity, protecting $\neg Present$ while circumscribing $Penguin$ will result in conjectures which say nothing about those penguins which are not present.

Using formula circumscription, the scope of conjectures can be limited by conjoining a protecting predicate with the expression to be minimized, and not allowing the protecting predicate to vary. For example, to minimize present penguins with respect to a theory, A, while protecting possible "absent" penguins, the following circumscription axiom suffices:

$$\begin{aligned}\forall \Phi.\ A(\Phi) \wedge [\forall x.\ \Phi x \wedge Present(x) \supset Penguin(x) \wedge Present(x)] \\ \supset [\forall x.\ Penguin(x) \wedge Present(x) \supset \Phi x \wedge Present(x)]\ ,\end{aligned}$$

which says nothing new about absent penguins.

2.1.7. Semantic Interconnections

Gelfond, Przymusinska, and Przymusinski [1985] have extended the "careful closure" notion of Gelfond and Przymusinska [1985], by allowing the theory to be augmented with the negations of arbitrary formulae meeting admissibility criteria. This is more powerful than adding only negations of ground atomic formulae. Gelfond, Przymusinska, and Przymusinski restrict their attention to *fixed-domain* theories, those with axioms stating that there are finitely many individuals, and that each term of the language denotes a unique individual. Let **P**, **Q**, and **Z** be as in section 2.1. Then a formula, K, not involving literals from **Z**, is *free for negation* iff there is no ground clause, B, made up of literals in $\mathbf{P}^+ \cup \mathbf{Q}^+ \cup \mathbf{Q}^-$ such that $T \vdash K \vee B$ and $T \nvdash B$. Then the *extended CWA for T* is defined as:

$$ECWA(T) = T \cup \{ \neg K | K \text{ is free for negation in } T \} .$$

Using the same partial-order relation on models as Gelfond and Przymusinska [1985] (see section 2.1.4), Gelfond, Przymusinska, and Przymusinski claim that the set of formulae free for negation in T are precisely those whose negations are true in every minimal model of T. (Here we refer to minimality over all, not just Herbrand, models.) Thus, for consistent, function-free, fixed-domain theories, T, $ECWA(T)$ is consistent,[9] and corresponds precisely to the formulae true in all minimal models of T. It follows that the free-for-negation formulae characterize the results of formula circumscription for such theories.

In fact, because of the fixed-domain property, one need only consider those K which are conjunctions of literals from $\mathbf{P}^+ \cup \mathbf{Q}^+ \cup \mathbf{Q}^-$. It can be shown that $ECWA(DB)$ corresponds, syntactically, to the careful closure of DB if DB is a fixed-domain theory. The semantic correspondence follows from the fact that every model of a fixed-domain theory is isomorphic to a Herbrand model, and hence every minimal model to a minimal Herbrand model.

By suitably matching the model-theory to the proof-theory, it is possible to show that, for fixed-domain theories, predicate circumscription corresponds to the GCWA and, for Horn theories, to naive closure. These observations show how central the notion of minimal model is to the various formalisms for closed-world reasoning. The two forms of minimization – of extensions of predicates and of the domain of the model (hence producing a fixed-domain model) – suffice to connect them all.

[9] Every consistent, finite-domain theory has at least one minimal model.

2.2. Default or Prototypical Reasoning

> Never utter these words: 'I do not know this, therefore it is false.' One must study to know, know to understand, understand to judge.
>
> — Apothegm of Neruda

All of the approaches discussed so far provide ways of becoming more "closed-minded". Each functions by restricting the set of models for the given axioms. The goal has been to allow only minimal models, in which only a minimal set of predicate instances or domain elements necessary to satisfy the axioms is allowed.

The complementary approach also involves restricting the set of models considered. Rather than focussing on minimality, the systems discussed in the sequel provide more flexibility in determining which models are considered "interesting".

2.2.1. Default Logic

Reiter [1978a, 1980a] addresses the problem of incomplete information by allowing new inference rules to be added to a standard first-order logic. These rules sanction their conclusions provided that the set of beliefs satisfies the conditions outlined in their premises. Unlike standard logic, these premises are allowed to refer both to what is known and to what is not known. The latter property allows rules to be added that specify inferences that will be made only when specific information is missing. These inferences can be used to tailor the completion of partial knowledge, unlike closed-world reasoning, which involves a uniform completion strategy.

2.2.1.1. Default Theories

A *default* is any expression of the form:[10]

$$\frac{A(\vec{x})\colon B_1(\vec{x}), \ldots, B_m(\vec{x})}{w(\vec{x})}$$

where $A(\vec{x})$, $B_i(\vec{x})$, and $w(\vec{x})$ are all formulae whose free variables are among those in $\vec{x} = x_1, \ldots, x_n$. A, B_i, and w are called the *prerequisite*, *justifications*, and *consequent* of the default, respectively. If none of A, B_i, and w contain free variables, the default is said to be *closed.* If the prerequisite is empty, it may be taken to be any tautology. Two classes of defaults having only a single justification, $B(\vec{x})$, are distinguished. Those with $B(\vec{x}) = w(\vec{x})$, are said to be *normal,* while those with $B(\vec{x}) = w(\vec{x}) \wedge C(\vec{x})$, for some $C(\vec{x})$, are called *semi-normal.* Virtually all of the defaults

[10] This notation differs from Reiter's in the omission of the "M" preceeding each of the B_i's. Since they are implicit in the positional notation, they have been omitted as a notational convenience.

occurring in the literature fall into one of these two categories. (Łukaszewicz [1985] argues that the remaining class of single-justification defaults, where $B(\overline{x}) \not\models w(\overline{x})$ are ill-motivated, and we know of no application for multi-justification defaults.)

Defaults serve as rules of inference or conjecture, augmenting those normally provided by first-order logic. Under certain conditions, they sanction inferences which could not be made within a strictly first-order framework. If their prerequisites are known and their justifications are "consistent" (*i.e.*, their negations are not provable), then their consequents can be inferred. Thus the term "justification" is seen to be somewhat misleading, since justifications need not be known, merely consistent.[11] The consequent's status is akin to that of a belief, subject to revision should the justifications be denied at some future time. It is this characteristic which induces the non-monotonic behavior of defaults.

Default rules can be seen to have a great deal in common with many previously mentioned approaches. For example, the Closed-World Assumption states:

$$If \not\vdash w \; infer \neg w$$

which can be represented in default logic by:

$$\frac{:\neg w}{\neg w} \tag{13}$$

In fact, (13) will later be referred to as the "Closed-World" default. The DEFAULT assignments which can be attached to frame slots in KRL [Bobrow & Winograd 1977] also appear to be related. KRL provides a mechanism for obtaining a value for a slot in the absence of a "better" value. A KRL default value, *d*, for a slot, *s*, in a frame instance, *f*, can be viewed as:

$$\text{If} \not\vdash s(f) \neq d \text{ infer } s(f) = d$$

or, in default logic, as:

$$\frac{:s(f) = d}{s(f) = d}$$

Similar mechanisms are available in many other frame-based knowledge representation schemes [Minsky 1975].

A closely related approach is Sandewall's [1972] "*Unless*" operator. "*Unless*(*P*)" is interpreted as "$\not\vdash P$", and "*Unless*" terms are allowed in the construction of wffs, with results like:

$$A \wedge Unless(B) \supset C$$

which corresponds roughly to:

$$\frac{A : \neg B}{C} .$$

"*Unless*" was originally proposed as a solution to the frame problem [Hayes 1973]. Rather than having to have explicit axioms stating that the properties of objects remained invariant from situation to situation unless explicitly changed, Sandewall suggested that these "frame axioms" be

[11] In a modal logic with the operator K (know) the justifications B_i might appear as $\neg K \neg B_i$.

replaced by a *frame inference rule* like:

$$\frac{\begin{array}{l} IS(object,property,situation)\\ Unless(ENDS(object,property,Successor(situation,act)))\end{array}}{IS(object,property,Successor(situation,act))}$$

which can be interpreted: If an object has a property in a situation, it can be concluded to retain that property in the successor situation resulting from performing 'act', unless it can be shown otherwise.

No formation rules were provided for "*Unless*", however, so questionable formulae such as:

$$A \supset Unless(B)$$

can be constructed. The semantics of such formulae are, at best, difficult to determine. Sandewall also fails to provide any formal understanding of the impact of the "*Unless*" rule on the underlying logic. Default logic has, to some extent, remedied these shortcomings.

2.2.1.2. Closed Default Theories and Their Extensions

A *default theory*, Δ, is an ordered pair, (D, W). D is a set of defaults; W is a set of first-order formulae. Reiter [1980a] describes the extensions of a default theory as "acceptable sets of beliefs that one may hold about an incompletely specified world, W". D is viewed as extending the first-order knowledge of W in order to provide information not derivable from W.

Since defaults allow reference to what is not provable in the determination of what is provable, the "theorems" of a default theory are not so easy to generate as are those of a first-order theory. What is provable both determines and is determined by what is not provable. To avoid this apparent circularity, the theorems of a default theory are defined by a fixed-point construction. An extension, E, for Δ is required to have the following properties:

$$W \subseteq E$$

$$Th_L(E) = E$$

For each default, $\dfrac{A\colon B_1,\ldots,B_m}{w} \in D$, if $A \in E$, and $\neg B_1,\ldots,\neg B_m \notin E$

then $w \in E$.

These properties state that E must contain all the known facts, that E must be closed under the $\vdash$ relation, and that the consequent of any default whose prerequisite is satisfied by E, and whose justifications are consistent with E, must also be in E. Reiter defines an extension for a closed default theory to be a minimal fixed-point of an operator having the above characteristics.

The extensions of a default theory select restricted subsets of the models of the underlying first-order theory, W. Any model for an extension of Δ will also be a model for W, but the converse is generally not true. Default theories need not always have extensions, even when W is consistent. There are, however, certain classes of theories for which the existence of at least one extension is guaranteed. Theories with only normal defaults have been shown always to have extensions [Reiter 1980a]. In chapter 3, we prove the same result for certain classes of theories with semi-normal defaults.

Reiter [1980a] presents an iterative mechanism for deciding whether a set of formulae forms an extension for a theory, Δ. The method is, unfortunately, not suitable for constructing extensions. This is because an oracle is required which can decide whether a particular formula's negation will be in the set. Reiter [1980a] and Etherington [1982] also present constructive mechanisms applicable to normal theories and to arbitrary finite theories, respectively.

Some examples of defaults were presented in the preceeding section. The following example illustrates the extensions induced by the closed-world default on the theory:

$$W = \{BLOCK(A) \vee BLOCK(B)\}.$$

The closed-world default is really a default schema which is applicable to any positive ground literal. In this case, it results in the following set of normal defaults:

$$D = \left\{ \frac{:\neg BLOCK(A)}{\neg BLOCK(A)}, \frac{:\neg BLOCK(B)}{\neg BLOCK(B)} \right\}$$

The theory, (D, W), has two extensions, E_1 and E_2.

$$E_1 = Th(\{\neg BLOCK(A), BLOCK(B)\})$$

$$E_2 = Th(\{BLOCK(A), \neg BLOCK(B)\})$$

Note that $\overline{E} = Th(\{BLOCK(A), BLOCK(B)\})$ is *not* an extension. Like database completion and circumscription, the closed-world default sanctions the exclusive interpretation of disjunctions to which it is applied. Intuitively, this is because the defaults force as many things to be false as possible, resulting in extensions whose models may be minimal models for W. More precisely, $\overline{E}$ is not an extension because it violates the minimality condition of the definition of extensions. (Were W also to contain both $BLOCK(A)$ and $BLOCK(B)$, $\overline{E}$ would be the only extension.)

Notice how the extensions E_1 and E_2 manifest W's inconsistency with the CWA. The inconsistent assignments for $BLOCK(A)$ and $BLOCK(B)$ are still obtainable, but they are separated into orthogonal, self-consistent extensions. In fact, Reiter has shown that the extensions of any default theory will always be self-consistent provided that the first-order theory W is consistent, and that all the extensions of a normal default theory will be (pairwise) mutually inconsistent.

2.2.1.3. General Default Theories

In contrast to closed defaults, an *open* default is one in which at least one of $A(\vec{x})$, $B_i(\vec{x})$, or $w(\vec{x})$ contain free variables in $\vec{x}$. An open default is interpreted as standing for the set of closed defaults obtainable by replacing its free variables by ground terms. If the set of ground terms is infinite this results in a default theory with an infinite set of defaults.

Most interesting default theories are not closed. Consider what, by now, must be the archetypal default theory:

$$W = \left\{ \begin{array}{l} \forall x.\ \mathit{Penguin}(x) \supset \mathit{Bird}(x), \\ \forall x.\ \mathit{Penguin}(x) \supset \neg \mathit{Can\text{-}Fly}(x), \\ \forall x.\ \mathit{Dead\text{-}Bird}(x) \supset \mathit{Bird}(x), \\ \forall x.\ \mathit{Dead\text{-}Bird}(x) \supset \neg \mathit{Can\text{-}Fly}(x), \\ \forall x.\ \mathit{Ostrich}(x) \supset \mathit{Bird}(x), \\ \forall x.\ \mathit{Ostrich}(x) \supset \neg \mathit{Can\text{-}Fly}(x), \\ \mathit{Bird}(\mathit{Tweety}) \end{array} \right\}$$

$$D = \left\{ \frac{\mathit{Bird}(x)\ :\ \mathit{Can\text{-}Fly}(x)}{\mathit{Can\text{-}Fly}(x)} \right\}$$

The default, which is not closed, might be interpreted as "If x is a bird, and it is consistent that x can fly, conclude that it can". This theory allows one to conclude, for an arbitrary bird (*e.g.*, Tweety), that it can fly – unless one is told that it cannot, or that it is a penguin, an ostrich, or dead. The conclusion may later have to be revoked should Tweety turn out to be a penguin, but common sense seems to sanction the same conclusion. This is partly because people tend to assume that they have the relevant information in most situations (*c.f.* linguists' use of Grice's Conversational Implicatures [Grice 1975]: one of these is that all information necessary to interpret an utterance is expected to be contained in the utterance.)

2.2.1.4. Interacting Defaults

Their broad applicability and the guarantee of coherence makes normal defaults attractive for knowledge representation and reasoning. There are, however, some types of knowledge which normal defaults cannot completely characterize. For example, Reiter and Criscuolo [1983] have noticed that defaults sometimes interact with one another, and that normal defaults cannot adequately constrain these interactions. One manifestation of this occurs when two defaults with distinct but not mutually exclusive prerequisites have contradictory consequents. In such circumstances it is not always clear which default should be applied. Commonsense reasoning usually prefers one of the competing defaults by virtue of its prerequisite being more specific, making the default applicable for only a subset of those individuals for which the competing default is applicable. This preference cannot be enforced using only normal defaults. For example, assume we are given:

Typical adults are employed.
Typical high-school dropouts are adults.
Typical high-school dropouts are not employed.

This may be expressed by the following normal defaults:

$$\left\{ \frac{Adult(x) : Employed(x)}{Employed(x)}, \frac{Dropout(x) : Adult(x)}{Adult(x)}, \frac{Dropout(x) : \neg Employed(x)}{\neg Employed(x)} \right\}.$$

For a given a dropout, this theory can be seen to have two extensions which differ on his/her state of employment. Intuition dictates that we assume s/he is unemployed. Careful consideration shows that the conflict arises because typical dropouts are not *typical* adults; this atypicality should block the transitivity from *Dropout* through *Adult* to *Employed.* The first default incorporates no explicit reference to these exceptional circumstances which should block its application. One way to address this problem is to require that the case under consideration not be a known exceptional case. This requirement is then added to the justification. Thus the first default above becomes:

$$\frac{Adult(x) : Employed(x) \bigwedge \neg Dropout(x)}{Employed(x)},$$

which is not applicable to known dropouts.

Semi-normal defaults can be used to resolve the ambiguities resulting from the interactions between defaults. This is done by making interactions explicit, as exceptions to the applicability of defaults. There are three major objections to this approach, however.

First, the complexity of theories with semi-normal defaults is substantially greater than of theories with normal defaults. Application of a default may force conclusions obtained from previously applied defaults to be retracted. This phenomenon, which cannot occur with normal default theories, precludes the type of straightforward proof theory developed by Reiter [1980a] for normal theories.

Secondly, it is possible to so overconstrain the interactions between defaults that the resulting theory has no extension. Chapter 3 explores ways of guaranteeing that this does not happen, but, for complicated theories with many interactions, it may be difficult to detect such overconstraining.

Finally, interactions must be noticed and explicitly dealt with at the time new knowledge is given to the system. In a large, complicated, system this is likely to be an enormous task. The contributors of new knowledge may not be aware of all possible interations between their contributions and the remainder of the knowledge base. In security-conscious environments, contributors may not even be allowed access to some of the information which interacts with their contribution. Touretzky [1984a,b] argues that explicit control of interactions in default theories is inappropriate, for the reasons outlined above and because many of the ambiguities introduced by such interactions can be resolved using more general principles. In semantic network systems, which can be viewed as corresponding to default theories (see chapter 4), the standard such principle is the "shortest-path heuristic", which resolves ambiguities by preferring whichever conclusion can be reached by traversing the smallest number of network arcs. Etherington [1982] shows how to construct networks which defeat the shortest-path heuristic and other simple-minded ambiguity resolution techniques.

Touretzky [1984a] presents a more sophisticated ambiguity resolution device, the *inferential distance topology,* which appears to capture the intention of the shortest-path heuristic without its naive realization. He exploits the subclass/superclass relations, which are one of the *raisons d'être* for semantic networks, to arbitrate between rival conclusions. In the "Dropout" example, above, since *Dropout* is a (default) subclass of *Adult*, the inferential distance ordering perfers conclusions associated with *Dropout* (*i.e.*, unemployed) over those associated with *Adult* (*i.e.*, employed), in accord with our intuitions. In chapter 4, we discuss Touretzky's approach in more detail, and show its relationship to default logic.

In spite of the fact that it is applicable only in subclass/superclass hierarchies, the success of Touretzky's approach in agreeing with the intuitively acceptable conclusions (a vague criterion, to be sure) suggests that it may be possible to elucidate some set of general principles which avoid the necessity of *ad hoc* manipulations of the knowledge base. Finding and evaluating such principles remains an important open problem.

2.2.2. Minimizing Abnormality

Default reasoning can involve conjecturing both positive and negative instances of predicates. This would seem to preclude the use of any of the closed-world or circumscriptive formalisms, discussed earlier, in situations where general default reasoning is required. (In chapter 5, this is shown conclusively in the case of predicate circumscription.) Expanding on an idea first presented (to our knowledge) by Levesque [1982], McCarthy [1986] and Grosof [1984] have explored the possibility of using formula circumscription for default reasoning. Essentially, the idea involved is that if defaults represent the properties of "normal" individuals, then there is "something abnormal" about an individual who does not fit the default patern. By appropriately axiomatizing abnormality, it is possible to do default reasoning by circumscribing abnormality.

An individual may be normal in some respects and abnormal in others; few, if any, are ever totally "typical". Thus, some allowance must be made for these differing aspects of abnormality. McCarthy explicitly introduces these aspects into his ontology, speaking of the (ab)normality of particular aspects of an individual. Grosof, preferring not to proliferate objects unduly, instead has a variety of abnormality predicates, each corresponding to abnormality of a particular aspect in McCarthy's notation.

An example helps to clarify the method. We follow McCarthy's notation:

$$\left.\begin{array}{l}
\forall x.\ \mathit{Thing}(x) \wedge \neg ab(aspect_1(x)) \supset \neg \mathit{Fly}(x) \\
\forall x.\ \mathit{Bird}(x) \supset ab(aspect_1(x)) \\
\forall x.\ \mathit{Bird}(x) \wedge \neg ab(aspect_2(x)) \supset \mathit{Fly}(x) \\
\forall x.\ \mathit{Penguin}(x) \supset ab(aspect_2(x)) \\
\forall x.\ \mathit{Penguin}(x) \supset \mathit{Bird}(x) \\
\forall x.\ \mathit{Penguin}(x) \wedge \neg ab(aspect_3(x)) \supset \neg \mathit{Fly}(x) \\
\forall x.\ \mathit{Penguin\text{-}in\text{-}his\text{-}dreams}(x) \supset ab(aspect_3(x)) \\
\forall x.\ \mathit{Penguin\text{-}in\text{-}his\text{-}dreams}(x) \supset \mathit{Penguin}(x) \\
\forall x.\ \mathit{Penguin\text{-}in\text{-}his\text{-}dreams}(x) \wedge \neg ab(aspect_4(x)) \supset \mathit{Fly}(x) \\
\forall x.\ \mathit{Ostrich}(x) \supset ab(aspect_2(x)) \\
\forall x.\ \mathit{Ostrich}(x) \supset \mathit{Bird}(x) \\
\forall x.\ \mathit{Ostrich}(x) \wedge \neg ab(aspect_5(x)) \supset \neg \mathit{Fly}(x)
\end{array}\right\} \qquad (14)$$

$aspect_1(x)$–$aspect_5(x)$ are the aspects, and ab is the abnormality predicate on aspects of individuals. Given $\mathit{Thing}(\mathit{Theodore})$, circumscribing $ab(x)$ varying ab and Fly allows us to conclude $\neg ab(aspect_1(\mathit{Theodore}))$ and hence $\neg \mathit{Fly}(\mathit{Theodore})$. Given $\mathit{Bird}(\mathit{Tweety})$, circumscription will yield $ab(aspect_1(\mathit{Tweety}))$, $\neg ab(aspect_2(\mathit{Tweety}))$, and $\mathit{Fly}(\mathit{Tweety})$. If Opus is a penguin, the conjectures will be $ab(aspect_1(\mathit{Opus}))$, $ab(aspect_2(\mathit{Opus}))$, $\neg ab(aspect_3(\mathit{Opus}))$, and $\neg \mathit{Fly}(\mathit{Opus})$.

This reformulation of default reasoning as closed-world reasoning about abnormality can deal with many of the problems of interacting defaults that forced the consideration of semi-normal default theories. The direct default representation of the above example looks like:

$$\frac{\mathit{Thing}(x) : \neg \mathit{Fly}(x)}{\neg \mathit{Fly}(x)}, \quad \frac{\mathit{Ostrich}(x) : \neg \mathit{Fly}(x)}{\neg \mathit{Fly}(x)},$$

$$\frac{\mathit{Bird}(x) : \mathit{Fly}(x) \wedge \neg \mathit{Ostrich}(x) \wedge \neg(\mathit{Penguin}(x) \wedge \neg \mathit{Penguin\text{-}in\text{-}his\text{-}dreams}(x))}{\mathit{Fly}(x)},$$

$$\frac{\mathit{Penguin}(x) : \mathit{Fly}(x) \wedge \neg \mathit{Penguin\text{-}in\text{-}his\text{-}dreams}(x))}{\neg \mathit{Fly}(x)}$$

In order to preserve a unique extension, the complicated interactions between the default statements must be explicitly reflected in the rules. Introducing abnormality allows a *normal* default representation consisting of the first-order axioms (14), together with the single closed-world default:

$$\frac{: \neg ab(x)}{\neg ab(x)}.$$

Grosof [1984] has developed a translation scheme, using abnormality predicates, which he claims produces representations of normal default theories in a form suitable for circumscriptive default reasoning.[12] He is currently seeking a way of extending this approach to arbitrary semi-normal default theories. The related problem – whether ab can be used to reduce semi-normal default theories to normal default theories – also remains open.

[12] In fact, the representation correctly translates only prerequisite-free normal defaults.

McCarthy [1984, personal communication] has discovered that it is possible to run into interaction problems with *ab* predicates. Augmenting (14) with:

$$\forall x.\ Canary(x) \wedge \neg ab(aspect_6(x)) \supset Bird(x)$$
$$\forall x.\ Gangster\text{--}Canary(x) \supset Canary(x)$$
$$\forall x.\ Gangster\text{--}Canary(x) \supset ab(aspect_6(x))\ ,$$

to allow the possibility that "canary" may be used in the sense of old gangster-movies, may result in ambiguity. *Dinsdale* the *Canary* must be abnormal with respect to either $aspect_6$ or $aspect_1$. Circumscribing *ab* can only conjecture either that *Dinsdale* flies (because he is a bird and hence abnormal in $aspect_1$), or that he is an abnormal canary (in $aspect_6$). McCarthy [1986] has proposed a variant of formula circumscription, prioritized circumscription, which allows several expressions to be simultaneously minimized according to some particular precedence. This can eliminate undesirable interactions, but at the cost that the precedence must be explicitly worked out before circumscribing. The criticisms applied to semi-normal default representations, that interactions must be known and accommodated when knowledge is represented, apply equally to the prioritized circumscription of abnormality representations. Whether such interactions can be dealt with without destroying the conceptual clarity and naturalness of the *ab* representation scheme is unknown.

2.2.3. Non-Monotonic Logic

McDermott and Doyle [1980, McDermott 1982] propose a formalism complementary to default logic, which they call non-monotonic logic (NML). Unlike default logic, which uses the notion of consistency only at the "meta" level (in the inference rules), NML centres around the introduction of consistency into the object language. The first incarnation of NML [McDermott & Doyle 1980] consists of a standard first-order logic, augmented with an "*M*" operator, roughly equivalent to the familiar " $\not\vdash \neg$ ". The set of theorems is defined as the intersection of all of the fixed-points of an operator, *NM*. Essentially, *NM* produces the logical closure of the original theory together with as many assertions of the form *Mq* as possible. The set of theorems can be contrasted with the extensions of a default theory, each of which is a fixed-point. This indicates that non-monotonic theoremhood is, in some sense, a more conservative or restrictive concept than extension membership. Moore [1983a] suggests that this difference can be understood by viewing fixed-points as sets of beliefs an agent might come to hold given his premises, while the intersection of the fixed-points determines what an outside observer could infer about the agent's beliefs knowing only his premises. In fact, the extensions of default theories and the fixed-points of non-monotonic theories are incomparable in general. The two formalisms often agree, as would intuitively be expected, given that any default:

$$\frac{A\ :\ B_1,\ldots,B_m}{w}$$

can be approximated in NML by:

$$A \wedge MB_1 \wedge \cdots \wedge MB_m \supset w\ .$$

There are, however, default theories which have extensions even though the corresponding non-monotonic theories have no fixed-points, and *vice versa* (see [Reiter 1980a] for examples).

Davis [1980] suggests that it might be impossible to assign a reasonable semantics to the M operator were it included in the object language. McDermott and Doyle point out that MP, intuitively read as "P is consistent", is not necessarily inconsistent with $\neg P$! Moore [1983a] observes that this is caused by the lack of any prohibition, in the fixed-point construction, against $\neg P$ and MP being contained in a single fixed-point. Thus, in Moore's terms, $\neg P$ may be believed without the statement "$\neg P$ is believed" ($\neg MP$ or $L\neg P$) being believed. This allows weaker interpretations to be placed on M than the intended "is consistent". These and other problems led to the recasting of the theory in terms of a more classical modal logic [McDermott 1982].[13] The resulting non-monotonic S5 is unfortunately redundant, since it is no more powerful than S5. Because of this, McDermott suggests falling back to non-monotonic S4 or non-monotonic T. This suggestion is peculiar, since McDermott acknowledges that the characteristic axiom of S5 ($\neg LP \supset L\neg LP$) - if P is not believed, it is believed not to be believed - seems appropriate for any belief system. However, the collapse of non-monotonic S5 was seen to force this retreat.

Moore [1983a, b] argues that this retreat is ill-motivated. He goes on to show that the collapse of non-monotonic S5 is actually due to the axiom $LP \supset P$, which says that whatever is believed is true. While this axiom is appropriate for knowledge, Moore claims that a non-monotonic system is actually dealing with belief, and that an axiom stating the infallibility of an agent should be *expected* to lead to peculiar consequences.

Aside from the question of their appropriateness, McDermott presents no proofs of the consistency of non-monotonic T and S4. Such proofs are a necessary step in the development of non-monotonic T and S4.

In the second paper on NML, McDermott [1982] acknowledges the restrictiveness of believing only those formulae in the intersection of all the fixed-points of a theory. He proposes a "brave robot" which would believe all of the formulae of some particular fixed-point. Such an approach is required in order to provide an intuitively satisfactory semantics for Mp: "p is consistent with what is believed".

The availability of the "M" terms in the language has advantages and disadvantages. For example, it can be shown that sentences of the form:

$$p \supset Mq$$

where p and q are arbitrary formulae, are either redundant or inconsistent [Etherington & Mercer, 1982, unpublished notes]. (This follows because the "theorems" of any NML theory must include all formulae Mp which are not inconsistent.) Such sentences cannot be formed in default logic, but are readily available in NML (as they are in Sandewall's formalism).

[13] A discussion of modal logics is beyond the scope of this proposal. See [Hughes and Cresswel 1972] for an introduction.

On the positive side, the default rules can be manipulated by the theory. For example, in the normal default theory with no axioms and the defaults:

$$\left\{ \frac{A : B}{B}, \ \frac{\neg A : B}{B} \right\}$$

nothing can be inferred about B. The corresponding non-monotonic theory:

$$\{A \wedge MB \supset B, \neg A \wedge MB \supset B\}$$

implies MB and $MB \supset B$, from which B can be inferred. This appears to be more in accord with normal commonsense reasoning.

Finally, $Lp \equiv p$ is a thesis of NML. While most modal logicians would agree that "p is provable" implies "p is true", the converse is usually not accepted. Hughes and Cresswell [1972, p28] conclude that "no intuitively plausible modal system" would have such a thesis. This indicates that there may be fundamental problems with NML.

2.2.4. Autoepistemic Logic

Moore [1983a, b] provides a detailed criticism and reconstruction of NML. He begins by distinguishing between default reasoning and "autoepistemic" reasoning. The latter is defined to be what goes on in an ideally rational agent reasoning about her own beliefs. It is this type of reasoning – not default reasoning – that NML attempts to model, according to Moore.

Moore sees a NML axiom of the form:

$$\forall x.Bird(x) \wedge M(Can\text{-}Fly(x)) \supset Can\text{-}Fly(x) \qquad (15)$$

as saying not "Typical birds can fly", as McDermott and Doyle interpret it, but rather "The only birds which do not fly are those *known* not to fly". Read in this way, axioms such as (15) become statements about the state of an agent's knowledge, not about typical individuals.[14]

Having made this distinction, Moore points out that default and autoepistemic reasoning are nonmonotonic for different reasons. Default reasoning is tentative, and thus defeasible. It provides plausible grounds for holding certain beliefs, but these beliefs may have to be retracted should those grounds prove to have been merely plausible, rather than true. Autoepistemic reasoning makes only valid inferences. Provided that the premises are true, the conclusions follow with all of the force of logic behind them. Non-monotonicity enters because autoepistemic statements are *context-sensitive* or *indexical*. They explicily refer to the entire knowledge context that contains them. Thus, their meaning changes depending on what is known. Obviously, what follows from not knowing α will hold when α is not known, but may not hold if α is learned.

Moore argues that the possible sets of beliefs an ideally rational agent can hold based on a consistent set of premises, A, are those sets, T, such that

[14] Similar arguments can be applied to default logic and other consistency-based non-monotonic formalisms.

$$T = Th(A \cup \{LP|P \in T\} \cup \{\neg LP|P \notin T\}) ,$$

where LP means "P is believed". These sets he calls the *stable expansions of A*. A stable expansion includes the premises, accurately characterizes what is and what is not believed, and includes no beliefs not supported by the premises. Moore shows that stable expansions contain all and only those formulae which are true in every interpretation which satisfies all of the premises and makes LP true for every formula, P, in the expansion. The important thing to notice here is that if the premises (which may contain implications from what is/is not believed) are true, and the set of beliefs corresponds to the beliefs contained in a particular expansion, then all of (and only) the formulae in that expansion can be true. This intuitively corresponds to the idea that the different conclusions one can draw from incompletely specified knowledge will be completely determined by what one chooses to believe.

Unlike NML, autoepistemic logic (AEL) is a propositional modal logic. No provision is made for individual variables or quantifiers. Moore [1984, personal communication] suggests that it should be a relatively easy matter to extend AEL to its first-order counterpart, provided that the M operator is applied only to closed formulae. This means that no M occurs within the scope of a quantifier, so the problems of "quantifying in" to the scope of a modality are avoided. Unfortunately, many of the statements one would like to make using a first-order version of AEL involve quantifying in. For example, to say "All of the α's are known" seems to require an axiom of the form:

$$\forall x.\ \alpha(x) \supset L\alpha(x)$$

(in NML, this would be written as $\forall x.\ M\neg\alpha(x) \supset \neg\alpha(x)$), but the free '$x$' in '$L\alpha(x)$' is captured by the universal quantifier outside the modal context. The problems of quantifying in are an important topic in the study of modal logics (*c.f.* [Linsky 1971]), but the problem has yet to be studied in depth from the perspective of non-monotonicity.

Like many other non-monotonic reasoning systems, AEL was presented non-constructively [Moore 1983a, b]. Neither the semantic basis nor the syntactic realization of that semantics provided a mechanism for enumerating the theorems of a given theory. Moore [1984] has recently developed an alternative semantic characterization based on the familiar Kripke-style possible-worlds structures. In this semantics, it is possible to enumerate all of the interpretations (each model is finitely specifiable and, if the language is finite, there are only finitely many), decide which of these are models, and determine what is true in those. Moore [1984, personal communication] points out that, for theories with few propositional constants, this is an easy task. One might hope for a more direct means of arriving at the theorems of an autoepistemic theory, but this remains an open problem.

2.2.5. KFOPC

Levesque [1982, 1984] approaches the problem of incomplete information differently. Instead of immediately addressing the task of completing the incompleteness, he considers what an incomplete knowledge-base might be expected to know about its own knowledge, and what one might reasonably ask or tell such a knowledge base. This entails questions of what constitutes a reasonable answer to a query under incompleteness, and what an incomplete knowledge base can be expected to know after being told a particular fact. After developing a logical framework in which to talk about (incomplete) knowledge, Levesque discusses ways in which this framework can be applied to completing knowledge bases.

In formalizing knowledge (or "rational belief"), Levesque [1982] makes three basic assumptions about the nature of knowledge bases. These are:

1) Consistency: The knowledge in a knowledge base is self-consistent. There is some possible state-of-affairs that makes everything that is known true.

2) Competence: Every unknown sentence is false in some world compatible with everything that is known. *I.e.*, all logical consequences of what is known are known.

3) Closure: The knowledge base has complete and accurate self-knowledge. Any sentence which deals only with the state of the knowledge base will be known to be true or false according as it is true or false, respectively, for the knowledge base.

These assumptions lead to a first-order modal logic of knowledge (KFOPC), roughly similar to "weak" S5 [Stalnaker 1980]. This logic characterizes the beliefs of an ideally rational agent capable of reflecting on her own beliefs. Any sentence which is not known is known to be unknown (and *vice versa*), knowledge is logically closed, and the agent believes in the veracity of her beliefs. This last statement does not mean that every belief is true, merely that the agent believes all of her beliefs to be true.

One of the most surprising aspects of this system is that, although KFOPC allows one to tell/ask the knowledge base things that cannot generally be phrased in first-order logic, the query and update mechanisms, and the knowledge base itself, are all first-order representable. Levesque presents a function that translates any update or query into an equivalent first-order sentence with respect to a particular knowledge base. Unfortunately, the mapping is not a partial recursive function: In the general case, choosing first-order representability means that effectiveness must be traded for some heuristic component. Levesque does not explore whether there are effective subcases.

Non-monotonicity enters KFOPC in two ways. The first, most obvious, comes from the assumption of closure. That is, if KB does not know P, it knows this (*i.e.*, $KB \vdash K\neg KP$). Clearly, though, $KB \cup \{P\} \vdash KP$, and hence $KB \cup \{P\} \not\vdash K\neg KP$. As Levesque points out, $\neg KP$ means only that P is not currently known, not that it will never be known. Levesque's solution to this problem is to have self-knowledge come only from introspection: from implicit, rather than explicit, statements. Such a KB will never contain statements of the form $\neg KP$:

> "... any assertion will be a statement about the world and not the *KB*. If the assertion talks about what is known ... it is only doing so to help make a statement about the world. Thus, there is no way to tell a *KB* about itself... . This is to be expected, however, since a *KB* has been assumed to have complete and accurate knowledge of itself at any time."

The non-monotonicity of statements about lack of knowledge is not particularly problematic when treated in this way.

A second form of non-monotonicity arises when lack of knowledge is used as a premise in deductions. For example, the "Flying Birds" default can be expressed in KFOPC as:

$$\forall x.\ Bird(x) \wedge \neg K \neg Fly(x) \supset Fly(x) \tag{16}$$

– any bird not known not to fly can fly. Such statements allow conclusions about what is true in the world to be based on what is not known by the current knowledge base. Since the state of the knowledge base can change, non-monotonicity can extend to encompass more than purely introspective statements.

This representation of defaults in KFOPC is subject to the problems of interaction that plague normal default theories [Reiter & Criscuolo 1983]. For example, the fact that Australian birds are non-fliers by default can be written:

$$\forall x.\ Australian\text{-}bird(x) \supset Bird(x)$$
$$\forall x.\ Australian\text{-}bird(x) \wedge \neg KFly(x) \supset \neg Fly(x)$$

but Australian birds can also be conjectured to fly by virtue of being birds, through (16). Again, the logic provides no means of deciding between these alternatives.

A more serious problem arises when the knowledge base knows that some unknown individual is atypical. For example, the knowledge base:

$Bird(B1)$,
$Bird(B2)$,
$\neg Fly(B1) \vee \neg Fly(B2)$

is *inconsistent* with the default (16) because there is a bird ($B1$ or $B2$), not known not to fly, which nonetheless does not fly. In fact, this type of reasoning is not really default reasoning at all, but what Moore [1983a] calls "Autoepistemic Reasoning". What is really involved is a valid form of inference: from the premise that all exceptional cases are known to the conclusion that an individual, not known to be exceptional, is typical. If all of the exceptional cases are *not* known, as in the above example, then the "default" is simply *false*. This means that defaults cannot simply be stated as axioms, since the logic is too rigid to allow for occasional violations by particular individuals.

Levesque addresses these problems by "preprocessing" the defaults. This is done by means of a mapping on defaults that rejects any which are contradicted by the knowledge base or which, given the current state of the knowledge base, conflict with another default. *All* conflicting defaults are rejected.

This leads to a strongly conservative interpretation of defaults. If two defaults cannot be applied together because they are mutually inconsistent, neither will be applied. Not having any grounds for deciding between them, KFOPC chooses to reject both. This is similar to non-monotonic logic [McDermott & Doyle 1980], which sanctions only those beliefs in every fixed-point of its theories. In contrast, default logic [Reiter 1980] and auto-epistemic logic [Moore 1983a] sanction multiple sets of beliefs, one supporting each of the alternatives. Circumscriptive default reasoning conjectures the disjunction of the two alternatives. The only defaults assumed in KFOPC are those which are assumable independently of all other defaults.

Levesque discusses some interesting techniques for circumventing these difficulties in certain cases. To avoid the problem of interacting defaults without requiring defaults to explicitly allow for exceptional cases, Levesque suggests a representation scheme involving a "typical"-predicate-forming operator, $\nabla{:}k$. If P is a predicate letter, $\nabla{:}kP(\vec{x})$ is interpreted as saying that $\vec{x}$ is typical with respect to the k^{th} aspect of Pness. The properties of typical individuals are stated as first-order axioms. For example, the KB:

$$\forall x.\ \nabla{:}1Bird(x) \supset Fly(x)$$
$$\forall x.\ \nabla{:}1Bird(x) \supset \neg Australian\text{-}bird(x)$$
$$\forall x.\ Australian\text{-}bird(x) \supset Bird(x)$$
$$\forall x.\ \nabla{:}1Australian\text{-}bird(x) \supset \neg Fly(x)$$

says that all birds typical in aspect 1 of "Birdness" fly and are not Australian-birds, while $\nabla{:}1$*Australian-birds* are birds which do not fly. Such axioms do not state default properties of classes of individuals. Rather, they state properties that all *typical* individuals of those classes *must* (or must not) have. Default reasoning is performed by conjecturing that individuals *are* typical. For example, the statements:

$$\forall x.\ Bird(x) \wedge \neg K \neg \nabla{:}1Bird(x) \supset \nabla{:}1Bird(x) \tag{17}$$
$$\forall x.\ Australian\text{-}bird(x) \wedge \neg K \neg \nabla{:}1Australian\text{-}bird(x) \supset \nabla{:}1Australian\text{-}bird(x) \tag{18}$$

say that – unless known otherwise – birds and Australian birds are typical in the specified aspects. If *Tweety* is a bird not known not to fly (nor be otherwise atypical in aspect 1), (17) says that she is typical in aspect 1, and hence flies. An Australian bird, *Oscar*, on the other hand, is known to be an atypical bird in aspect 1, so (17) is inapplicable. The default (18) is applicable, however, so $\nabla{:}1Australian\text{-}bird(Oscar)$ can be conjectured, and hence $\neg Fly(Oscar)$.

To prevent defaults which are contradicted for particular individuals (or classes) from being rejected outright, axioms can be added which explicitly state that those individuals (or members of those classes) are atypical in the relevant aspects. For example, given the knowledge base:

$$\begin{array}{ll} Quaker(john), & \forall x.\ \nabla Quaker(x) \supset Pacifist(x), \\ Republican(george), & \forall x.\ \nabla Republican(x) \supset \neg Pacifist(x), \\ Quaker(nixon) \wedge Republican(nixon), & \end{array}$$

$$\forall x.\ Quaker(x) \wedge \neg K \neg \nabla Quaker(x) \supset \nabla Quaker(x) \tag{19}$$
$$\forall x.\ Republican(x) \wedge \neg K \neg \nabla Republican(x) \supset \nabla Republican(x) \tag{20}$$

the defaults, (19) and (20), stating that Republicans and Quakers are typical Republicans and Quakers, respectively, are not applicable for any individual because they are mutually contradictory for *nixon*. *nixon* violates the defaults: he is not known to be an atypical Quaker, so (19) sanctions $\nabla Quaker(nixon)$; similarly, (20) sanctions $\nabla Republican(nixon)$; but these conclusions are mutually inconsistent. Under the conservative interpretation of defaults, KFOPC rejects both (19) and (20) for this knowledge-base. Hence, *john* and *george* cannot be concluded to be typical, and so $Pacifist(john)$ and $\neg Pacifist(george)$ cannot be concluded. To remedy this, axioms stating that the typical Republican is not a Quaker (and *vice versa*) can be added:

$$\forall x.\ \nabla Quaker(x) \supset \neg Republican(x)$$
$$\forall x.\ \nabla Republican(x) \supset \neg Quaker(x)\ .$$

With these additional facts, *nixon* no longer constitutes a violation of the defaults, since

$$K\neg\nabla Republican(nixon) \wedge K\neg\nabla Quaker(nixon)$$

follows from the knowledge base.

Typical-predicates can also be used to specify a precedence-hierarchy among multiple, potentially-conflicting, defaults. For example, the axioms:

$$\forall x.\ \nabla{:}1student(x) \supset undergrad(x)$$
$$\forall x.\ \nabla{:}2student(x) \supset undergrad(x) \vee MSc(x)$$
$$\forall x.\ \nabla{:}3student(x) \supset undergrad(x) \vee MSc(x) \vee PhD(x)$$

together with the defaults:

$$\forall x.\ student(x) \wedge \neg K\neg\nabla{:}kstudent(x) \supset \nabla{:}kstudent(x) \qquad \text{for } k = 1,2,3. \tag{21}$$

will result in a theory that will assume *student*'s are *undergrad*'s if possible, otherwise *MSc*'s if possible, and otherwise *PhD*'s if possible.

Levesque gives numerous examples showing that these strategies can be combined to obtain remarkably subtle control of the interactions between defaults, without modifying the structure of the defaults themselves. All that is required is the addition of new axioms. He cites three advantages of this formalization of default reasoning:

1) A default need not be discarded and replaced when a subclass that typically fails to satisfy the default is discovered. Additional axioms can be added, stating the inapplicability of the default for members of that subclass.

2) The knowledge given to the knowledge base is more structured. Instead of arbitrary defaults, properties of typical individuals are listed.

3) Only a single type of default (*e.g.*, (21)) need be considered, and only one of these for each typical-predicate, $\nabla{:}kP$.

Levesque's use of typical-predicates as a representation scheme for defaults corresponds directly to McCarthy's subsequent use of abnormality-predicates. There is a straightforward mapping between Levesque's axiomatizations using ∇-predicates and McCarthy's using *ab*-predicates. The defaults, $\forall x.\ Px \wedge \neg K\neg\nabla{:}kPx \supset \nabla{:}kPx$, then correspond to minimizations of the corresponding abnormality-predicates. It is not yet known whether this mapping constitutes a translation, whether the two approaches lead to the same conjectures for corresponding default theories. There are striking similarities, however. For example, tangled hierarchies, wherein members of one class may typically – but not always – be members of another, are problematic in both paradigms. McCarthy's "Gangster and Canaries" example, discussed earlier, requires additional KFOPC rules, beyond those required by the straighforward abnormality/typicality representation, to express priorities or preferences for particular kinds of atypicality. Capturing these priorities in KFOPC appears to involve a loss of clarity and naturalness of representation similar to that incurred by McCarthy's introduction of priorities into circumscriptive abnormality theories.

In spite of Levesque's insights into representing default knowledge, default reasoning in KFOPC remains largely unexplored. Similarly, the application of Levesque's ideas on typical- (abnormal-) predicates to default reasoning based on other formalisms has only begun. Both of these areas promise to provide important insights into reasoning about incompletely specified worlds, and deserve further exploration.

2.2.6. Objections to Non-Monotonic Formalisms

Kramosil [1975] claims to have shown that any formalized theory which allows unprovability as a premise in deductions must either be "meaningless", or no more powerful than the corresponding first-order theory without rules involving such premises. He presents two "proofs" to support his claim. Careful examination shows that the first result follows from a definition of "formalized theory" which expressly excludes any theory which exhibits the types of behavior common to non-monotonic theories. The second result is based on an incorrect definition of "proof" and hence of "theoremhood" and is itself meaningless. As the paper stands, it shows only that non-monotonic theories must behave differently than monotonic theories in those cases where the former can derive results unobtainable using the latter.

Kramosil was not the only one to be uncomfortable with opening the "Pandora's Box" of non-monotonicity. Sandewall [1972] notes that the "Unless" operator has "some dirty logical properties". Considering the example:

A
$A \wedge Unless(B) \supset C$
$A \wedge Unless(C) \supset B$

he observes that either B and C can be theorems, but, in general, not both simultaneously. Reiter [1978b] makes a similar observation in an early paper, stating that:

> Such behavior [is] clearly unacceptable. At the very least, we must demand of a default theory that it satisfy a kind of 'Church-Rosser' property: No matter what the order in which the theorems of a theory are derived, the resulting set of theorems will be unique.

It appears that the Church-Rosser property is a necessary casualty if non-monotonicity is accepted.

A further problem which must be faced by those embracing consistency- or unprovability-based approaches to non-monotonicity is that the non-theorems of a first-order theory are not recursively enumerable. This means that the rules of inference in theories involving the $\not\vdash$ operator cannot be effective in general. It follows that the theorems are not recursively enumerable. By contrast, in monotonic logics, the rules of inference MUST be effective and the theorems MUST be recursively enumerable.

Finally, the very non-monotonicity which makes such theories interesting means that "theorems" may have to be retracted if the assumptions on which they are based are refuted (either by new knowledge or changes in the state of the world). To be useful, a non-monotonic reasoning system must be able to remember which assumptions underly each theorem and be able to unwind the potentially complex chain of deductions founded on retracted justifications.

CHAPTER 3

Default Logic

If the wheel is fixed,
I would still take a chance.
If we're treading on thin ice,
Then we might as well dance.

— Jesse Winchester

In this chapter, we explore default logic in some detail. We present a model-theoretic semantics for arbitrary default theories, thus rectifying a major deficiency. The remaining sections investigate the causes of incoherence in certain default theories. This leads to a strong sufficient (although not necessary) syntactic condition for the existence of extensions for particular theories.

3.1. The Semantics of Default Theories

In his development of default logic, Reiter provided a fixed-point characterization of the extensions of a default theory, but no model-theoretic semantics for the logic. Etherington [1982, 1983] observes that the semantics can be viewed in terms of restrictions of the set of models of the underlying theory. Łukaszewicz [1985] formalizes this intuition for normal default theories. Because of the well-behaved nature of these theories, this is relatively straightforward. The resulting semantic characterization amounts to considering the Tarskian semantics of each of the partial extensions constructed by proceding monotonically toward an extension by satisfying, at each step, the next applicable normal default (according to some arbitrary ordering of the defaults) by making its consequent true. If, after each default in the sequence has been considered, no more defaults from D are applicable, the resulting set, together with the first-order theory, W, yields an extension. Since each step affirms a formula consistent with those affirmed previously, the set of models contracts monotonically. The intersection of the sets of models from each stage is precisely the set of models of the extension.

This semantics can perhaps best be envisaged as a transition network, whose nodes are subsets of $\mathbf{M}$, the set of all models of W, with arcs labelled by defaults, as follows: From the node corresponding to a set of models N, for every $\delta = \frac{\alpha : \beta}{\beta} \in D$, an arc labelled δ leads (i) back to N if no model in N satisfies β or some satisfy $\neg\alpha$, or (ii) to the node corresponding to the set:

$\{ N \mid N \in \mathbf{N}$ and $N \models \beta \}$, otherwise. Each leaf – a node all of whose outbound links loop back – reachable from $\mathbf{M}$ corresponds to the set of models of some extension of Δ. Furthermore, the set of models of each extension of Δ corresponds to such a leaf node. The set of arc-labels for every path from root to leaf gives the generating defaults for the extension corresponding to the node.

This approach does not apply directly to non-normal defaults, since the property of semi-monotonicity which guarantees its success holds only for normal defaults [Reiter 1980a, theorem 3.2]. Łukaszewicz partially addresses this problem by presenting a translation scheme from non-normal defaults to normal defaults. He argues that, of single-justification defaults, only normal and semi-normal defaults have reasonable interpretations. Non-semi-normal defaults are therefore translated to semi-normal defaults by conjoining the consequent to the justification:

$$\frac{\alpha : \beta}{\gamma} \rightarrow \frac{\alpha : \beta \wedge \gamma}{\gamma}$$

The translation from semi-normal to normal, which is somewhat more controversial, involves replacing the consequent with the justification:

$$\frac{\alpha : \beta \wedge \gamma}{\gamma} \rightarrow \frac{\alpha : \beta \wedge \gamma}{\beta \wedge \gamma} .$$

This makes sense, Łukaszewicz argues, so long as α's which are also γ's are typically β's. That is, so long as one could reasonably augment the theory with

$$\frac{\alpha \wedge \gamma : \beta}{\beta} .$$

One can imagine situations where this is not appropriate. For example, a system for legal reasoning might want to have a rule suggesting that those with motives who *might* be guilty should be suspects:

$$\frac{\mathit{has\text{-}motive}(x) \ : \ \mathit{guilty}(x)}{\mathit{suspect}(x)} .$$

It is clearly reasonable to translate this to:

$$\frac{\mathit{has\text{-}motive}(x) \ : \ \mathit{suspect}(x) \wedge \mathit{guilty}(x)}{\mathit{suspect}(x)} ,$$

allowing that there may be reasons not to include someone on the list of suspects even without knowing their innocence. It is *not* reasonable to follow through by asserting the guilt of all suspects:

$$\frac{\mathit{has\text{-}motive}(x) \ : \ \mathit{suspect}(x) \wedge \mathit{guilty}(x)}{\mathit{suspect}(x) \wedge \mathit{guilty}(x)} !$$

Thus, while the semantics Łukaszewicz outlines covers many cases, there is reason to want a semantics which covers more than normal defaults. To this we now turn.

Because of the failure of semi-monotonicity for non-normal theories, simply applying one default after another will not, in general, lead to extensions. It is necessary to ensure that the application of each default does not violate the justifications of already-applied defaults. If we augment Łukaszewicz's semantics by encoding some information in each state about the set of defaults which led to a particular state, we can determine whether a node is on a viable path toward an extension. The precise details are these:

Definition: Satisfiability, admissibility, and applicability

Let X be a set of models; Γ a set of formulae; α, β, and ω formulae, and $\delta = \frac{\alpha : \beta}{\omega}$ a default. Then

i) α is *X-satisfiable* (*X-valid*) iff $\exists x \in X.\ x \models \alpha \qquad (\forall x \in X.\ x \models \alpha)$

ii) Γ is *X-admissible* (*X permits* Γ) iff $\forall \gamma \in \Gamma.\, \exists x \in X.\ x \models \gamma$

iii) δ is *X-applicable* iff α is X valid and β is X-satisfiable. ■

Definition: Result of a default

Let X, Γ, and δ be as above. Then the *result* of δ in (X, Γ) is:

$$\delta(X, \Gamma) = \begin{cases} (\mathrm{X}, \Gamma) \text{ if } \delta \text{ is not } X\text{-applicable and } \Gamma \text{ is } X\text{-admissible,} \\ ((X - \{N \mid N \models \neg\omega\}),\ (\Gamma \cup \{\beta\})) \text{ if } \delta \text{ is } X\text{-applicable and } \Gamma \text{ is } X\text{-admissible, and} \\ \perp \text{ otherwise.} \end{cases}$$ ■

Definition: Result of a sequence of defaults

Let X and Γ be as above, and let $<\delta_i>$ be a sequence of defaults. Then the *result* of $<\delta_i>$ is:

$$<\delta_i>(X, \Gamma) = (\cap X_i,\ \cup \Gamma_i) \text{ where } \begin{cases} X_0 = X; \qquad \Gamma_0 = \Gamma; \qquad \text{and} \\ (X_{i+1}, \Gamma_{i+1}) = \delta_i(X_i, \Gamma_i), \quad i \geq 0\ . \end{cases}$$ ■

Definition: Stability

Let Y be a non-empty set of models, Γ a set of formulae, and $\Delta = (D, W)$ a default theory. Then (Y, Γ) is *stable for* Δ iff

(1) $(Y, \Gamma) = <\delta_i>(X, \{\ \})$ for $X = \{M \mid M \models W\}$, and some $\{\delta_i\} \subseteq D$,

(2) $\forall \delta \in D.\ \delta(Y, \Gamma) = (Y, \Gamma)$, and

(3) Γ is Y-admissible. ■

In other words, a set of models and a set of constraints is stable for a default theory, (D, W), if they are the result of some sequence of defaults in D applied to the set of models of W and no constraints, if no default in D produces any change in this result, and the constraints are satisfied by the set of models. Note that condition (2), together with the definition of "result" means that condition (3) is redundant. We include it for conceptual clarity. The soundness and completeness results for this semantics are given by Theorems 3.1 and 3.2, respectively.

Theorem 3.1 – Soundness

If E is an extension for Δ, then there is some set Γ such that
$(\{M | M \models E\}, \Gamma)$ is stable for Δ. ■

Theorem 3.2 – Completeness

If (X, Γ) is stable for Δ then X is the set of models for some extension of Δ.
(*I.e.*, $Th(X)$ is an extension for Δ.) ■

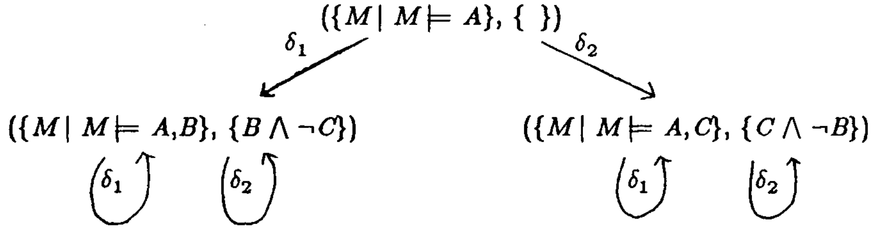
Returning to the transition network analogy, the nodes are now pairs consisting of a subset of **M** and a subset of the justifications of the defaults in *D*. Now Δ's extensions correspond to those leaf nodes, (X, Γ), where X permits Γ. We say that such nodes are *viable*. If all leaf nodes are $\perp$, the theory has no extensions. Again, the generating defaults for the extension $Th(X)$ are those defaults labelling arcs on any path from $(\mathbf{M}, \{\ \})$ to (X, Γ'), for any Γ'.

Example 3.1

Consider the default theory:

$$\Delta = \left[D = \left\{ \delta_1 = \frac{A : B \wedge \neg C}{B}, \delta_2 = \frac{A : C \wedge \neg B}{C} \right\}, \quad W = \{A\} \right]$$

This produces the following transition network.



$(\{M | M \models A\}, \{\ \})$
δ_1
δ_2
$(\{M | M \models A, B\}, \{B \wedge \neg C\})$
$(\{M | M \models A, C\}, \{C \wedge \neg B\})$
δ_1
δ_2
δ_1
δ_2

Both leaves are viable, so the theory has two extensions, $Th(\{A, B\})$ and $Th(\{A, C\})$. ■

Example 3.2

The incoherent theory:

$$\Delta = \left[\left\{ \delta = \frac{:\neg A}{A} \right\}, \{\ \} \right]$$

gives rise to:

$$\begin{tikzcd}
(\{M \mid M \models P \vee \neg P\}, \{\ \}) \arrow[d, bend right, "\delta"'] \\
(\{M \mid M \models A\}, \{\neg A\}) \arrow[d, bend right, "\delta"'] \\
\perp
\end{tikzcd}$$

in which the leaf is not viable. Hence this theory has no extension. ■

It is instructive to compare this model-set restriction semantics with the minimal-model semantics of closed-world reasoning presented in chapter 2. There, the semantics of the closure of a theory was defined in terms of a restriction of the set of models of the underlying theory, according to the principle of minimization. The model-set restriction semantics for default logic similarly provides a principle for determining subsets of the models of a first-order theory which characterize acceptable belief-sets, on the basis of maximal satisfaction of the set of defaults. There are several significant differences, however. Firstly, rather than an ordering on individual models, this semantics imposes an ordering on sets of models. Secondly, the ordering is defined in terms of accessibility via a sequence of defaults, rather than strictly in terms of intrinsic features of the models themselves. Finally, each extension is determined by a single extremum of the ordering, rather than by the set of all extrema.

The first of these differences results because the extensions of a default theory – unlike the models of a first-order theory – are not complete. They do not decide every formula. Because they incompletely specify the world, sets of models – rather than single models – are required to allow for undecided formulae. Using situations [Barwise and Perry 1983] – incomplete model-descriptions – instead of sets of models might lead to a closer correspondence. Intuitively, certainly, one can simply view the model-sets as partial model-descriptions without ill effect.

The second deviation results from the fact that defaults are general inference rules. Consequently, the submodel(-set) relation is potentially more complex for default logic. Lifschitz' [1984] recent work allowing arbitrary pre-orders as well as simple subset orderings may void this difference, but the question remains open.

The fact that individual extrema determine extensions is the result of the "brave" (in McDermott's [1982] terminology) character of default logic. Reiter's presentation of default logic defined each extension as an acceptable set of beliefs, with the intention that a reasoner would somehow "choose" a single extension within which to reason about the world. Other non-monotonic formalisms (see chapter 2) are based on "cautious" approaches which accept a default conclusion only if it occurs in *all* acceptable sets of beliefs. One can easily construct a variant of default logic which pursues a "cautious" course. (The converse is not obviously true for all "cautious" systems, as we see in chapter 8.) Such a system would define the theorems of a default theory to be those formulae true in all extensions, with the obvious change to the semantics: the theorems would then be defined as those formulae true in all models of all viable leaves.

3.2. Coherence of Default Theories

Extensions play a fundamental role in default logic. An extension is a set of beliefs which are in some sense "justified" or "reasonable" in light of what is known about a world. Formally, extensions are attractive because they are both grounded and complete: A formula enters an extension, E, only if it is in W, if it is provable from other formulae in E, or if it is the consequent of a default whose prerequisites are in E and whose justifications are not denied by E; furthermore, every formula which meets these requirements is in E. The first of these restrictions prevents extensions from containing spurious, unsupported beliefs. The second ensures that justified beliefs are not ignored. These restrictions are analogous to those which define the theorems of a first-order theory.

Since the individual extensions of a default theory *are* both grounded and complete, it is quite natural to require any default inference system to restrict its conclusions to a single common extension. If no extension of a theory contains a formula, then it is not in any acceptable set of beliefs associated with that theory. If conclusions are drawn from different extensions, they may be incompatible. Consider the blocks-world example from the previous chapter. In that example, both $\neg Block(A)$ and $\neg Block(B)$ are reasonable assumptions. They are drawn from different extensions, however, and concluding both leads to inconsistency.

Since reasonable conclusions must reside in an extension of the default theory under consideration, it is clearly important to know whether every theory has extensions. Simply put, the answer is "No". For example, the theory:

$$W = \{ \}$$

$$D = \left\{ \frac{:A}{\neg A} \right\}$$

has no extension. Such theories are *incoherent;* they support no reasonable set of beliefs about the world. Beyond pointing out the existence of incoherent theories, the most useful answer would include a syntactic characterization of which theories have or do not have extensions. While no such characterization is known, there are sufficient conditions which guarantee extensions. We present three such conditions below, in order of increasing utility.

A theory, $(\{ \}, W)$, with no defaults has a unique extension, $Th(W)$, the logical closure of the underlying first-order theory. Of course, this is a trivial default theory. We mention it only to emphasize that, since default logic is a superset of first-order logic, the required results obtain for the area of overlap.

The distinctions between commonly encountered types of defaults lead to more enlightening results. Any default of the form:

$$\frac{\alpha : \beta}{\beta}$$

is said to be *normal*. Normal defaults are sufficient for knowledge representation and reasoning in many naturally occurring contexts. In fact, they can express any rule whose application is subject only to first-order prerequisites and the consistency of its conclusion with the rest of what is believed. Rules like:

"Assume a bird can fly unless you know otherwise.", or

"Assume a thing is not a block unless it is required to be."

translate easily into normal defaults:

$$\frac{Bird(x) : Can\text{-}fly(x)}{Can\text{-}fly(x)} \quad \text{and} \quad \frac{: \neg Block(x)}{\neg Block(x)} .$$

The consequent of a normal default is equivalent to its justification. Intuitively, this makes the default inapplicable where the consequent has been denied. Such defaults cannot introduce inconsistencies, they cannot refute the justifications of other, already applied, normal defaults, nor can they refute their own justifications. This gives rise to well-behaved theories. Any theory involving only normal defaults (a *normal theory*) must have at least one extension [Reiter 1980a].

Any default of the form:

$$\frac{\alpha : \beta \wedge \gamma}{\beta}$$

is said to be *semi-normal*. Semi-normal defaults differ from normal defaults by having justifications which entail but are not entailed by their consequents. The assurances of well-behavedness associated with normal theories do not carry over to theories with semi-normal defaults. For example, the theory:

$$W = \{ \}$$

$$D = \left\{ \frac{:A \wedge \neg B}{A}, \frac{:B \wedge \neg C}{B}, \frac{:C \wedge \neg A}{C} \right\} \tag{1}$$

has no extension. This appears to be a somewhat artificial example, inasmuch as we have been unable to find a natural situation which fits this pattern. Which semi-normal theories, then, are assured of extensions? Do all "natural" theories have extensions? Perhaps pathological examples are merely formal curiosities? We do not purport to answer these questions — partly because of the difficulty of delimiting the class of "natural" theories. There is, however, a large class of semi-normal theories which are coherent. We characterize this class, which appears to be sufficient for many common applications, in the next section.

3.3. Ordered Default Theories

There appears to be a unifying characteristic among default theories without extensions. Consider again the theory:

$$W = \{\ \}$$

$$D = \left\{ \frac{:A}{\neg A} \right\}$$

which has no extension. The only reasonable candidates are $\overline{E_1} = Th(\{\ \})$ or $\overline{E_2} = Th(\{\neg A\})$. A is consistent with $\overline{E_1}$, so to be an extension $\overline{E_1}$ must contain $\neg A$, which it does not. Similarly, A is inconsistent with $\overline{E_2}$, so $\overline{E_2}$ cannot contain $\neg A$. The problem is that the default's justification is denied by its consequent; not applying the default forces its application, and vice versa. Returning to the semi-normal theory (1), we see that applying any one default leaves one other applicable. Applying any two, however, results in the denial of the non-normal part of the justifications of at least one of them. Any set small enough to be an extension is too small; any set large enough is too large. This behaviour is characteristic of theories with no extension; the requirement that extensions be closed under the default rules forces the application of defaults whose consequents lead to the denial of justifications of other applied defaults.

The exact source of the problem can be further isolated by recalling that all normal theories have extensions. Since the justification and consequent of normal defaults are identical, no applicable default can refute the justifications of an already applied default: applied normal defaults have already asserted their justifications. This means that any normal default capable of refuting those justifications is inapplicable, since its justifications have already been refuted. It follows that that part of the justification which distinguishes non-normal defaults from normal defaults is integrally involved in making a theory incoherent. Restricting our attention to semi-normal default theories, we see that once a default has been applied, only those conjuncts of its justification not entailed by its consequent are susceptible to refutation by other defaults. These conjuncts play a key role in the discussion below.

The conflict between closure under defaults and consistency of justifications can occur only if some formula depends on the absence of another and at the same time may serve to support the inference of that formula. In the theory (1) above, for example, A depends on the absence of B, B on that of C, and C on that of A. Hence inferring A would block the inference of C, allowing the inference B, which would invalidate the inference of A, and similarly for B and C.

The examples presented so far have involved defaults in their simplest form:

$$\frac{\alpha : \beta_1 \wedge \dots \wedge \beta_n}{\omega}$$

where α, ω and β_i are all literals (*i.e.*, atomic formulae or negations of atomic formulae). The problem of determining dependencies is more complicated when α, ω and β_i are allowed to be arbitrary first-order formulae. For example, the consequent of a default may be an implication; applying that default would introduce new dependencies. The essential idea remains the same, however: determine whether the dependencies involve potentially unresolvable circularities. The following definitions outline a syntactic method for determining whether such circularities exist within a semi-normal theory.

Definition: $\ll$ and $\underline{\ll}$

Let $\Delta = (D, W)$ be a closed,[1] semi-normal default theory. Without loss of generality, assume all formulae are in clausal form. The partial relations, $\underline{\ll}$ and $\ll$, on *Literals* × *Literals*, are defined as follows:

(1) If $\alpha \in W$ then $\alpha = (\alpha_1 \vee ... \vee \alpha_n)$, for some $n \geq 1$.
For all α_i, $\alpha_j \in \{\alpha_1,...,\alpha_n\}$, if $\alpha_i \neq \alpha_j$, let $\neg\alpha_i \underline{\ll} \alpha_j$.

(2) If $\delta \in D$ then $\delta = \dfrac{\alpha : \beta \wedge \gamma}{\beta}$. Let α_1, ... α_r , β_1, ... β_s , and γ_1, ... γ_t be the literals of the clausal forms of α, β, and γ, respectively. Then

(i) If $\alpha_i \in \{\alpha_1,...,\alpha_r\}$ and $\beta_j \in \{\beta_1,...,\beta_s\}$ let $\alpha_i \underline{\ll} \beta_j$.

(ii) If $\gamma_i \in \{\gamma_1,...,\gamma_t\}$, $\beta_j \in \{\beta_1,...,\beta_s\}$ and $\gamma_i \notin \{\beta_1,...,\beta_s\}$ let $\neg\gamma_i \ll \beta_j$.

(iii) Also, $\beta = \beta_1 \wedge ... \wedge \beta_m$, for some $m \geq 1$.
For each $i \leq m$, $\beta_i = (\beta_{i,1} \vee ... \vee \beta_{i,m_i})$, where $m_i \geq 1$.
Thus if $\beta_{i,j}$, $\beta_{i,k} \in \{\beta_{1,1},...,\beta_{m,m_m}\}$ and $\beta_{i,j} \neq \beta_{i,k}$ let $\neg\beta_{i,j} \underline{\ll} \beta_{i,k}$.

(3) The expected transitivity relationships hold for $\ll$ and $\underline{\ll}$. *I.e.*,

(i) If $\alpha \underline{\ll} \beta$ and $\beta \underline{\ll} \gamma$ then $\alpha \underline{\ll} \gamma$.

(ii) If $\alpha \ll \beta$ and $\beta \ll \gamma$ then $\alpha \ll \gamma$.

(iii) If $\alpha \ll \beta$ and $\beta \underline{\ll} \gamma$ or $\alpha \underline{\ll} \beta$ and $\beta \ll \gamma$ then $\alpha \ll \gamma$. ■

The definition is complex, but the intention is that $\alpha \underline{\ll} \beta$ or $\alpha \ll \beta$ if there is any way that α could figure in an inference of β in the theory as it stands. The intuition behind parts (1) and (2.iii) is that any disjunction of n literals can be interpreted as an implication of any one of those literals. *E.g.*, $(\alpha_1 \vee ... \vee \alpha_n) \equiv [(\neg\alpha_1 \wedge ... \wedge \neg\alpha_{j-1} \wedge \neg\alpha_{j+1} \wedge ... \wedge \neg\alpha_n) \supset \alpha_j]$. The special prominence we have alluded to for the conjuncts in a justification not entailed by the consequent is reflected in part (2.ii) by the use of the distinguished " $\ll$ " relation. The negation, $\neg\gamma_i$, occurs in part (2.ii) since it is not knowing $\neg\gamma_i$ which makes γ_i consistent.

[1] The definition is readily extensible to open theories using a technique given in [Reiter 1980a].

Definition: Orderedness

A semi-normal default theory is said to be *ordered* if and only if there is no literal, α, such that $\alpha \ll \alpha$. ■

An ordered theory has no potentially unresolvable circular dependencies. The theory in example (1) is not ordered, since $B \ll A$, $C \ll B$, and $A \ll C$; hence $A \ll A$. The theory:

$$W = \{\ \}$$
$$D = \left\{ \frac{: A \wedge \neg B}{A},\ \frac{: B \wedge \neg D}{B},\ \frac{: (C \supset D) \wedge \neg A}{(C \supset D)} \right\} \tag{2}$$

is also not ordered. The defaults give rise to the following relationships:

$$\{B \ll A\},\quad \{D \ll B\},\quad \text{and}\quad \{C \underline{\ll} D,\quad \neg D \underline{\ll} \neg C,\quad A \ll \neg C,\quad A \ll D\},$$

respectively. Hence $A \ll D \ll B \ll A$.

The significance of orderedness for semi-normal default theories is shown by Theorem 3.3.

Theorem 3.3 – Coherence

If a semi-normal default theory is ordered, then it has at least one extension. ■

Normal theories are clearly ordered, since only non-normal defaults give rise to " $\ll$ " relationships. Thus the coherence of all normal theories is a corollary of Theorem 3.3. This is encouraging inasmuch as it suggests that orderedness is not merely a special purpose gimmick but, rather, it subsumes an existing, widely applicable characterization.

It is important to notice that orderedness is only a sufficient condition for existence of extensions. Non-ordered theories have potentially unresolvable circularities but, for one reason or another, these circularities do not always interfere. The theory (2) is not ordered, but it does have an extension: $Th(\{B, (C \supset D)\})$. The circularity would cause problems, however, if C were added to W: the resulting theory has no extensions. In other cases, two or more potential circularities may cancel each other out. At present, we do not know whether the given condition can be strengthened to one which is both necessary and sufficient for the coherence of semi-normal theories and yet is still decidable.

3.4. Constructing Extensions

Having delineated a large class of theories which have extensions, we turn to the problem of generating extensions. Reiter [1980a] shows that extensions need not be recursively enumerable, and that it is not generally semi-decidable whether a formula is in any extension of a theory. Faced with such pessimism, further exploration might seem pointless. Still, there are tractable subcases.

Etherington [1982] presents a procedure which can generate all the extensions of an arbitrary finite default theory.[2] The procedure centres on a relaxation style constraint propagation technique. Extensions are constructed by a series of successive approximations. Each approximation, H_j, is built up from the first-order components in W by applying defaults, one at a time. At each step, the default to be applied is chosen from those, not yet applied, whose prerequisites are "known" and whose justifications are consistent with both the previous approximation and the current state of the current approximation. When no more defaults are applicable, the procedure continues with the next approximation. If two successive approximations are the same, the procedure is said to *converge.*

The choice of which default to apply at each step of the inner loop may introduce a degree of non-determinism. Generality requires this non-determinism, however, since theories do not necessarily have unique extensions. Deterministic procedures can be constructed for theories which have unique extensions, or if full generality is not required.

In the presentation of the procedure, below, $CONSEQUENT(\frac{\alpha : \beta}{\gamma})$ is defined to be γ.

```
$H_0 \leftarrow W$; $j \leftarrow 0$;
repeat
  $j \leftarrow j + 1$; $h_0 \leftarrow W$; $GD_0 \leftarrow \{\ \}$; $i \leftarrow 0$;
  repeat
    $D_i \leftarrow \left\{ \frac{\alpha : \beta}{\gamma} \in D \mid (h_i \vdash \alpha), (h_i \nvdash \neg\beta), (H_{j-1} \nvdash \neg\beta) \right\}$;
    if $\neg null(D_i - GD_i)$ then
      choose $\delta$ from $(D_i - GD_i)$;
      $GD_{i+1} \leftarrow GD_i \cup \{\delta\}$;
      $h_{i+1} \leftarrow h_i \cup \{CONSEQUENT(\delta)\}$; endif;
    $i \leftarrow i + 1$;
  until $null(D_{i-1} - GD_{i-1})$;
  $H_j = h_{i-1}$
until $H_j = H_{j-1}$
```

To see how this procedure works, consider the theory:

$W = \{A\}$

[2] A finite theory is one with only finitely many variables, constant symbols, predicate letters, and defaults. No function symbols are allowed, except the 0-ary function symbols, the constants. These restrictions make the universe of discourse (or Herbrand Universe) finite, ensuring only a finite number of closed instances of open defaults.

$$D = \left\{ \frac{A:B}{B}, \frac{A:C}{C}, \frac{B:D}{D}, \frac{B:\neg D \wedge \neg C}{\neg D} \right\},$$

which has the unique extension, $Th(\{A,B,C,D\})$. The procedure can generate any of the following sequences of approximations:

$H_0=\{A\}$		
$H_1=\{A,B,\neg D,C\}$	$H_0=\{A\}$	$H_0=\{A\}$
$H_2=\{A,B,C\}$	$H_1=\{A,C,B,D\}$	$H_1=\{A,B,C,D\}$
$H_3=\{A,B,D,C\}$	$H_2=H_1$	$H_2=H_1$
$H_4=H_3$		

(The formulae in each approximation are listed in the order in which they are derived.) In the first sequence of approximations, $\neg D$ occurs in H_1 because it can be inferred in h_2 *before* C is inferred in h_3.

Etherington [1982] proves:

There is a converging computation such that $H_n = H_{n-1}$ and $Th(H_n) = E$ if and only if E is an extension for the default theory (D, W).

In other words, the procedure can return every extension, and only extensions are returned. This result falls short in two respects: First, while the procedure can converge on every extension, there are appeals to *non-provability*. In general, such tests are not computable, since arbitrary first-order formulae are involved. There are computable subcases, however. If the set:

$$W \cup \left\{\alpha \mid \frac{\alpha:\beta}{\gamma} \in D\right\} \cup \left\{\beta \mid \frac{\alpha:\beta}{\gamma} \in D\right\}$$

belongs to a decision class for first-order provability, extensions are computable. Propositional theories and function-free, monadic theories fall into this class, as do finite theories, provided W is also finite.

The second shortcoming is that some finite theories admit non-converging computations. The procedure may never terminate even though the theory has an extension and each step is computable. In such cases, the procedure cycles forever between two or more distinct H_j's. Fortunately this cyclic behaviour seems to be caused by features similar to those which make theories incoherent. We have characterized certain classes of ordered theories for which the procedure is more well-behaved. Theorem 3.4 shows that one such class is the class of ordered, network theories.

Definition: Network Default Theory

A default theory, $\Delta = (D, W)$, is a *network theory* iff it satisfies the following conditions:

(1) W contains only:

a) Literals (*i.e.*, Atomic formulae or their negations), and

b) Disjuncts of the form $(\alpha \vee \beta)$ where α and β are literals.

(2) D contains only normal and semi-normal defaults of the form:

$$\frac{\alpha:\beta}{\beta} \quad \text{or} \quad \frac{\alpha:\beta \wedge \gamma_1 \wedge \ldots \wedge \gamma_n}{\beta}$$

where α, β, and γ_i are literals. ■

Theorem 3.4 – Convervence

For finite, ordered, network theories, the procedure given above always converges on an extension. ■

We will have more to say about network theories in the next chapter.

We conjecture that Theorem 3.4 can be generalized to apply to arbitrary ordered semi-normal theories, but we have no proof. The proof may require a more restrictive definition of D_i in the procedure, *viz*:

$$D_i = \left\{ \frac{\alpha : \beta}{\gamma} \in D \mid \alpha \in h_i \, , \, (h_i \cup H_{j-1}) \nvdash \neg\beta \right\}$$

instead of:

$$D_i = \left\{ \frac{\alpha : \beta}{\gamma} \in D \mid \alpha \in h_i \, , \, h_i \nvdash \neg\beta \, , H_{j-1} \nvdash \neg\beta \right\}$$

but it can be shown that all the results of [Etherington 1982] and those of this chapter still hold for the stronger version, so this should present no problem.

For normal theories, an even stronger result can be proved:

Theorem 3.5 – Strong Convergence

For finite normal theories, the procedure given above always converges on an extension immediately — *i.e.*, $Th(H_1)$ is always an extension. ■

CHAPTER 4

Inheritance Networks with Exceptions

A centipede was happy, quite,
Until a frog, in fun,
Said, "Pray, which leg comes after which?"
This raised his mind to such a pitch
He lay distracted in a ditch,
Considering how to run.

One of the problems with the non-monotonic formalisms we have discussed to this point is their intractability. Default logic, in the general case, is not even semi-decidable. Because of the need to build systems which have good computational properties, many researchers have sacrificed formal precision. While this has sometimes led to very fast "inference" mechanisms, there has often been little more than vague intuitions about exactly *what* these mechanisms infer.

As the field matures and systems capable of assuming responsibility for such things as nuclear reactors and medical diagnosis are touted as "on the horizon", it becomes increasingly important that it be understood what such systems "consider" justifiable inferences.

The argument has long been made that, because of the general intractability of formal systems, it is unreasonable to consider them for practical applications. This is taken as support for the use of systems such as semantic networks which, although not completely understood, can compute quickly. This argument falls down on two points. The first is that most of these fast inference algorithms are applicable to a limited class of problems. It could well be that – for these problems – formal systems such as default logic are just as tractable, and fast implementations may be possible. Secondly, even if formal systems are not implemented directly in an inference system, they may be useful as specification tools. In this way, an implementation could either be shown always to reach justified conclusions or, at the very least, to deviate in well-understood ways from justified conclusions. In the former case, the fast algorithm could actually be viewed as an *implementation* of an appropriately-restricted version of the general formal system; in the latter case, at least would-be purchasers of such systems could make enlightened decisions about the risks involved.

In this chapter, we employ default logic to outline a specification for "inheritance" reasoning in the presence of exceptions. Semantic networks have been widely adopted as a representational mechanism for AI. In such networks, "inference" is equated with inheritance of properties by nodes from their superiors. Recent work has considered the effects of allowing exceptions to inheritance within networks [Brachman 1982; Etherington and Reiter 1983; Fahlman 1979; Fahlman *et al* 1981; Touretzky 1982, 1984a; Winograd 1980]. Such exceptions represent either explicit or implicit cancellation of the normal property inheritance which networks enjoy.

In the absence of exceptions, an inheritance network is a taxonomy organized by the usual IS-A relation, as in Figure 4.1. Schubert [1976] and Hayes [1977] have argued that such networks correspond quite naturally to certain theories of first-order logic. *E.g.*,

$NAUTILUS(Fred)$
$\forall x.\ NAUTILUS(x) \supset CEPHALOPOD(x)$
$\forall x.\ CEPHALOPOD(x) \supset MOLLUSC(x)$
$\forall x.\ MOLLUSC(x) \supset INVERTEBRATE(x)$
$\ldots$

Such a correspondence can be viewed as providing the semantics which "semantic" networks had previously lacked [Woods 1975].

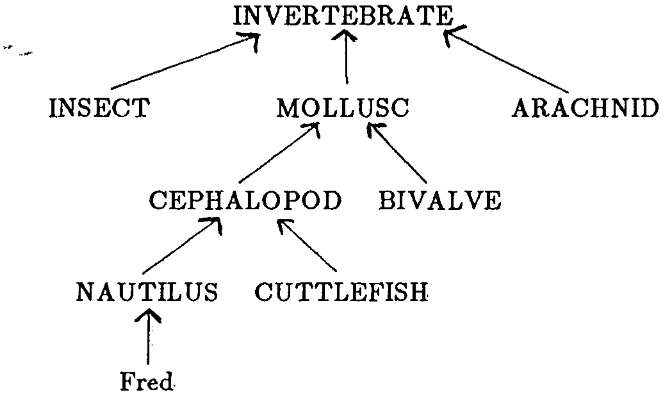


Figure 4.1 — Fragment of a taxonomy.

The significant features of this semantics are these:

(1) Inheritance is a logical property of the representation. Given that $NAUTILUS(Fred)$, $MOLLUSC(Fred)$ is provable from the given formulae. Inheritance is the repeated application of *modus ponens.*

(2) The node labels of such a network are unary predicates: *e.g.*, $NAUTILUS(*)$, $INVERTEBRATE(*)$.

(3) No exceptions to inheritance are possible. If Fred is a nautilus, he must be an invertebrate, regardless of any other properties he enjoys.

Unfortunately, this correspondence no longer applies when exceptions to inheritance are allowed. The logical properties of networks change drastically when exceptions are permitted. For example, consider the following facts about elephants:

(1) Elephants are gray, except for albino elephants.
(2) All albino elephants are elephants.

Common-sense reasoning about "elephants" allows one, given an individual elephant not known to be an albino, to infer that she is gray. Subsequent discovery — perhaps by observation — that she is an albino elephant forces the retraction of the conclusion about her grayness. Thus, common-sense reasoning about exceptions is non-monotonic; new information can invalidate previously derived facts. This non-monotonicity precludes the use of first-order representations, like those used for taxonomies, for formalizing networks with exceptions.

We establish a correspondence between networks with exceptions and network default theories. This correspondence provides a formal semantics and a notion of correct inference for such networks. As was the case for taxonomies, inheritance emerges as a logical feature of the representation. Those properties $P_1,\ldots,P_n$ which an individual, b, inherits prove to be precisely those for which $P_1(b),\ldots,P_n(b)$ all belong to a common extension of the default theory. Should the theory have multiple extensions — an undesirable feature, as we shall see — then b may inherit different sets of properties depending on which extension is chosen. We consider two radically different remedies for this problem.

To see how defaults might be used to represent networks with exceptions, consider the elephant example, which can be represented by the default theory:

$$W = \left\{ \forall x.\ \mathit{Albino\text{–}Elephant}(x) \supset \mathit{Elephant}(x) \right\}$$

$$D = \left\{ \frac{\mathit{Elephant}(x)\ :\ \mathit{Gray}(x) \wedge \neg\mathit{Albino\text{–}Elephant}(x)}{\mathit{Gray}(x)} \right\}.$$

It is easy to see that if we are told only *Elephant*(*Fred*) then, so far as we know, *Gray*(*Fred*) $\wedge$ ¬*Albino–Elephant*(*Fred*) is consistent; hence *Gray*(*Fred*) may be inferred. Given only *Albino–Elephant*(*Sue*) one can conclude *Elephant*(*Sue*) using first-order knowledge, but *Albino–Elephant*(*Sue*) "blocks" the application of the default, preventing the derivation of *Gray*(*Sue*), as required.

We adopt a network representation with seven link types. Other approaches to inheritance may omit one or more of these, but our formalism subsumes these. The seven link types,[1] with their translations to default logic, are:

(1) Strict IS-A: *A*.———▶.*B*: *A*'s are always *B*'s. Since this is universally true, we identify it with the first-order formula: $\forall x.\ A(x) \supset B(x)$.

(2) Membership: *a*∘———▶.*A*: The individual *a* belongs to the class *A*. We represent this with the first-order fact $A(a)$.

(3) Strict ISN'T-A: *A*.-|||▶.*B*: *A*'s are never *B*'s.
Again, this is a universal statement, identified with: $\forall x.\ A(x) \supset \neg B(x)$.

(4) Non-membership: *a*∘-|||▶.*A*: The individual *a* does not belong to the class *A*. We represent this with the first-order fact $\neg A(a)$.

(5) Default IS-A: *A*.———>.*B*: Normally *A*'s are *B*'s, but there may be exceptions.
To provide for exceptions, we identify this with a default:

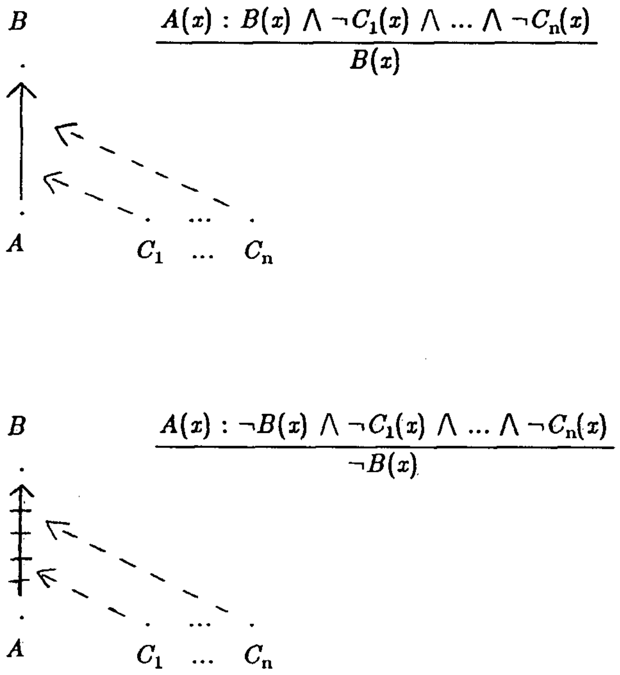
[1] Note that strict and default links are distinguished graphically by solid and open arrowheads, respectively.

$$\frac{A(x) : B(x)}{B(x)}$$

(6) Default ISN'T-A: A.┼┼┼>.B: Normally A's are not B's, but exceptions are allowed. Identified with:

$$\frac{A(x) : \neg B(x)}{\neg B(x)}$$

(7) Exception: A.------>
The exception link has no independent semantics; it serves only to make explicit the exceptions, if any, to the above default links. There must always be a default link at the head of an exception link; the exception then alters the semantics of that default link. There are two types of default links with exceptions; their graphical structures and translations are shown in Figure 4.2.



$$\frac{A(x) : B(x) \wedge \neg C_1(x) \wedge ... \wedge \neg C_n(x)}{B(x)}$$

$$\frac{A(x) : \neg B(x) \wedge \neg C_1(x) \wedge ... \wedge \neg C_n(x)}{\neg B(x)}$$

Figure 4.2 — Links with exceptions.

We illustrate with an example from [Fahlman *et al* 1981].

Molluscs are normally shell-bearers.
Cephalopods must be Molluscs but normally are *not* shell-bearers.
Nautili must be Cephalopods and must be shell-bearers.

Our network representation of these facts is given in Figure 4.3.

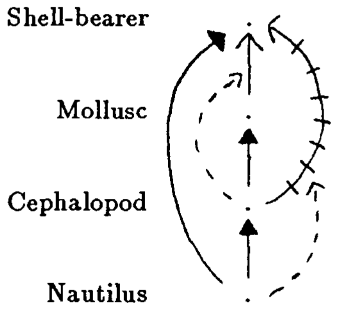


Figure 4.3 — Network representation of our knowledge about Molluscs.

The corresponding default logic representation is:

$$D = \left\{ \frac{M(x) : Sb(x) \wedge \neg C(x)}{Sb(x)}, \frac{C(x) : \neg Sb(x) \wedge \neg N(x)}{\neg Sb(x)} \right\},$$

$$W = \left\{ (x).\ C(x) \supset M(x),\ (x).\ N(x) \supset C(x),\ (x).\ N(x) \supset Sb(x) \right\}$$

Given a particular Nautilus, this theory has a unique extension in which it is also a Cephalopod, a Mollusc, and a Shell-bearer. A Cephalopod not known to be a Nautilus will turn out to be a Mollusc with no shell.

It is instructive to compare our network representations with those of NETL [Fahlman *et al* 1981]. A basic difference is that in NETL there are no strict links; all IS-A and ISN'T-A links are potentially cancellable and hence are defaults. Moreover, Fahlman *et al* allow explicit exception (*UNCANCEL) links only for ISN'T-A (*CANCEL) links. If we restrict the graph of Figure 4.3 to NETL-like links, we get Figure 4.4(a), which is essentially the graph given by Fahlman.

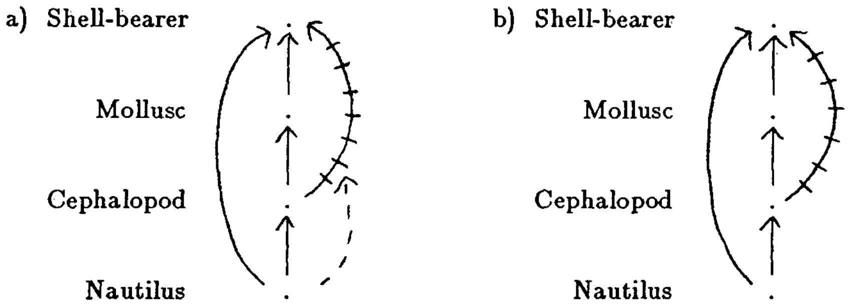


Figure 4.4 — NETL-like network representations of our knowledge about Molluscs.

The network in Figure 4.4(a) corresponds to the defaults:

$$\left\{\begin{array}{l}\dfrac{M(x) : Sb(x)}{Sb(x)}, \dfrac{C(x) : M(x)}{M(x)}, \dfrac{N(x) : C(x)}{C(x)}, \\ \dfrac{C(x) : \neg Sb(x) \wedge \neg N(x)}{\neg Sb(x)}, \dfrac{N(x) : Sb(x)}{Sb(x)}\end{array}\right\}.$$

As before, a given Nautilus will also be a Cephalopod, a Mollusc, and a Shell-bearer. A Cephalopod not known to be a Nautilus, however, gives rise to *two* extensions, corresponding to an ambivalence about whether or not it has a shell. While counter-intuitive, this merely indicates that an exception to shell-bearing, namely being a Cephalopod, has not been explicitly represented in the network. The ambiguity can be resolved by making the exception explicit, as in Figure 4.3. On the other hand, representations which do not permit exception links to point to IS-A links cannot make this exception explicit in the graphical representation.

Other versions of NETL (and many other inheritance reasoners) do not allow explicit exception links at all. If only default IS-A and ISN'T-A links are allowed, the representation of the Nautilus example becomes that of Figure 4.4(b), which corresponds to the defaults:

$$\left\{\begin{array}{l}\dfrac{M(x) : Sb(x)}{Sb(x)}, \dfrac{C(x) : M(x)}{M(x)}, \dfrac{N(x) : C(x)}{C(x)}, \\ \dfrac{C(x) : \neg Sb(x)}{\neg Sb(x)}, \dfrac{N(x) : Sb(x)}{Sb(x)}\end{array}\right\}.$$

In such theories, there is a further ambiguity about whether a Nautilus is a Shell-bearer.

How then do such systems conjecture that a Cephalopod is not a Shell-bearer, without also conjecturing that it is a Shell-bearer? Such ambiguities are typically resolved by means of an inference procedure which prefers shortest paths. Interpreted in terms of default logic, this "shortest path heuristic" is intended to favour one extension of the default theory. Thus, in the networks of Figure 4.4, the paths from Cephalopod to ¬Shell-bearer are shorter than those to Shell-bearer so that the former win. Unfortunately, this heuristic is not sufficient to replace the excluded exception type in all cases. Reiter and Criscuolo [1983] and Etherington [1982] show that it can lead to conclusions which are unintuitive or even invalid — *i.e.*, not in any extension. Fahlman *et al* [1981] and Touretzky [1981, personal communication; 1982] have also observed that shortest path algorithms can lead to anomalous conclusions. They describe attempts to restrict the form of networks to exclude structures which admit such problems. One effect of these restrictions appears to be to permit only networks whose corresponding default theories have unique extensions.

An inference algorithm for network structures is correct only if it can be shown to derive conclusions all of which lie within a single extension of the underlying default theory. This criterion rules out shortest path inference for unrestricted networks. This is unfortunate, since shortest path inference has been popular for its relative efficiency and ease of implementation.

On a more positive note, any network constructed using the seven link-types given above corresponds to a network default theory. By insisting that any network constructed must correspond to an ordered theory, the coherence of a network knowledge representation system can be assured. For such systems, the procedure given in chapter 3 is a correct and always converging inference algorithm.

It turns out that orderedness can be assured without reference to the full complexity of the mechanism described in chapter 3. It is easy to see that any acyclic network gives rise to an ordered theory. The same is true if only the subgraph consisting of all IS-A links and explicit exceptions thereto has no cycles involving at least one exception link, or if there are no explicit exceptions to IS-A links.

Theorem 4.1

Any network in which the subgraph of IS-A links and explicit exceptions thereto is acyclic corresponds to an ordered theory. ■

Corollary 4.2

Any acyclic network corresponds to an ordered theory. ■

Corollary 4.3

Any network with no explicit exceptions to IS-A links corresponds to an ordered theory. ■

Corollary 4.4

The networks mentioned in theorem 4.1 and corollaries 4.2 and 4.3 are coherent. ■

In addition to pointing out the inadequacies of shortest path inferencing and to providing sufficient conditions for coherence and a correct inference mechanism, the formal reconstruction of inheritance we have presented clarifies some of the outstanding problems in network inference. One of these, how to perform inferences in parallel, is considered in the next section.

4.1. Parallel Network Inference Algorithms

The computational complexity of inheritance problems, combined with some encouraging examples, has sparked interest in the possibility of determining inheritance in parallel. Fahlman [1979] has proposed a massively parallel machine architecture, NETL. This architecture assigns one processor to each predicate in the knowledge base. "Inferencing" is performed by nodes passing "markers" to adjacent nodes in response to their own state and that of their immediate neighbours. Fahlman suggests that such architectures could achieve logarithmic speed improvements over traditional serial machines.

The formalization of inheritance networks as default theories suggests, however, that there might be severe limitations to this approach. For example, correct inference requires that all conclusions share a common extension. For networks with more than one extension, inter-extension interference effects must be prevented. This seems impossible for a one pass parallel algorithm with purely local communication, especially in view of the inadequacies of the shortest path heuristic.

Even in knowledge bases with unique extensions, structures requiring an arbitrarily large radius of communication can be created. For example [Etherington 1982], the default theories corresponding to the networks in Figure 4.5 each have unique extensions. A network inference algorithm must reach F before propagating through B in the first network and conversely in the second. The salient distinctions between the two networks are not local; hence they cannot be utilized to guide a purely local inference mechanism to the correct choices. Similar networks can be constructed which defeat marker-passing algorithms with any fixed radius.

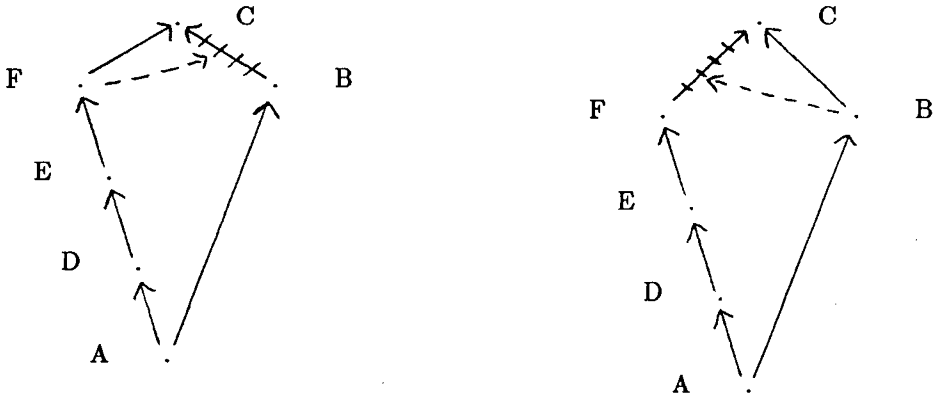


Figure 4.5 — Problems for local inheritance algorithms.

This has prompted Touretzky [1981, personal communication; 1984a] to characterize a restricted class of network structures which admit parallel inferencing algorithms. In part, his restrictions appear to exclude networks whose corresponding default theory has more than one extension. Unfortunately, it is unclear how these restrictions affect the expressive power of the resulting networks. Moreover, Touretzky [1982, personal communication; 1983] has shown that it is not possible to determine on a parallel marker-passing machine whether a network satisfies these restrictions.

Provided the network in question corresponds to an ordered theory, a form of limited parallelism can be achieved without sacrificing correctness. The key to this result lies in partitioning the network into subnetworks which are suitable for parallel processing. Essentially, each node in the network is numbered according to the number of exception links apon which it depends. This assigns each node to the lowest "level" possible while preserving the ordering amongst the nodes induced by the " $\ll$ " and " $\lesssim$ " relations. Since the network is ordered, this can be done in parallel, in finite time proportional to the longest chain in the network. Processing then proceeds in k parallel steps, where k is the number of the highest level to which nodes were assigned. At step n, all links having exceptions which were asserted at step n–1 are disabled. The resulting sub-network, consisting of all remaining links impinging on nodes at levels less than or equal to n, is processed in parallel, ignoring exception links, with markers propagating from nodes asserted at step n–1. The "nodes asserted at level 0" are those in $Th(W)$. These correspond to the nodes for which the network is "activated". The result after step k is an extension.[2]

There are two caveats associated with this procedure: If both positive and negative markers reach a node in the same step, one must be chosen. Either choice will lead to an extension; we do not consider other ramifications of such choices here. Second, the algorithm assumes that all strict links propagate instantaneously. If this is not the case, each step in the algorithm must be followed by propagation along strict links, resolving conflicts as above. Note that conflicts are always resolved by changing assignments at the current level.

Provided that the inviolability of strict links is maintained, that default links are active only if their prerequisites are asserted and their justifications have not been denied, and that no node and its negation are asserted together (conflict resolution), *any* reasonable propagation algorithm (parallel or otherwise) may be used at each step.[3]

To illustrate the construction, we apply it to the moderately complex network of Figure 4.6. Rather than restrict ourselves to a particular parallel propagation algorithm at each step, we present a table showing all possibilities.

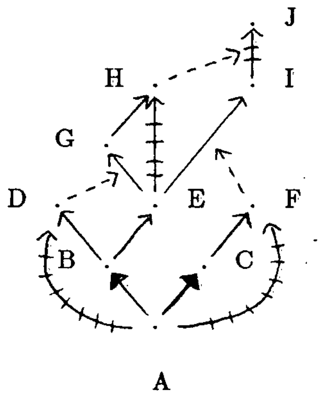


Figure 4.6— A multi-level inheritance graph.

The corresponding default theory, simplified to the propositional case and "activated" for A, is:

$$W = \left\{ A,\ (A \supset B),\ (A \supset C) \right\}$$

$$D = \left\{ \frac{A : \neg D}{\neg D},\ \frac{A : \neg F}{\neg F},\ \frac{B : D}{D},\ \frac{C : F}{F},\ \frac{B : E}{E},\ \frac{E : \neg H}{\neg H}, \right.$$

[2] This construction is that used in the proof of Theorem 3.3, where it is shown to yield an extension.

[3] To see this, it is necessary only to note that each step is, effectively, dealing with a normal theory. Arguments similar to those used in the proof of Theorem 3.5 can be used to show that the order of propagation is immaterial.

$$\frac{E: G \wedge \neg D}{G}, \quad \frac{G: H}{H}, \quad \frac{E: I \wedge \neg F}{I},$$

$$\left.\frac{I: \neg J \wedge \neg H}{\neg J}\right\}$$

The defaults above have been grouped according to the level to which their consequents are assigned (see Table 4.1). Table 4.2 shows the possibilities at each step; alternatives are shown in separate columns, with major rows corresponding to steps in the algorithm.

Level	Literals
1	A, B, C, D, ¬D, E, F, ¬F, ¬H
2	G, H, I
3	¬J

Table 4.1 — Levels of literals.

Step 1	A, B, C, E, ¬H			
	D, F	D, ¬F	¬D, F	¬D, ¬F
Step 2		I	G	I, G
Step 3		¬J		¬J

Table 4.2 — Possible outcomes using different propagation schemes.

Thus the algorithm can, depending on the nature of the parallel marker propagation procedure, find:

$$E_0 = Th(W \cup \{A, B, C, E, \neg H, D, F\})$$
$$E_1 = Th(W \cup \{A, B, C, E, \neg H, D, \neg F, I, \neg J\})$$
$$E_2 = Th(W \cup \{A, B, C, E, \neg H, \neg D, F, G\})$$
$$E_3 = Th(W \cup \{A, B, C, E, \neg H, \neg D, \neg F, I, G, \neg J\})$$

all of which are extensions. Significantly, no choice of parallel marker-passing procedure will enable the algorithm to find the theory's other two extensions:

$$E_4 = Th(W \cup \{A, B, C, E, H, \neg D, F, G\})$$
$$E_5 = Th(W \cup \{A, B, C, E, H, \neg D, \neg F, G, I\})$$

because $\neg H$ is at level 1 and so can (and must) be inferred at step 1. H, being at level 2, is thus precluded before it can be inferred. We have not yet characterized the biases which this inability to find all extensions would induce in a reasoner.

Another potential problem with this approach stems from the fact that many network inference systems "prefer" one link-type over another (*e.g.*, negation may override assertion). By breaking the network into sub-networks which are processed in turn, the ability to globally assert these preferences may be lost. We have three responses to this. Firstly, if network structure is restricted, in the manner suggested by Touretzky [1981, personal communication], so that resulting theories have unique extensions, the above algorithm produces the same results as any correct procedure. Secondly, many of these preferences are not well-defined, and break down when pressed (*c.f.* race conditions in [Fahlman *et al* 1981]). The inability to exhibit incorrect behaviour can hardly be called a liability. Finally, given a well-defined preference scheme, it must preserve correctness: all inferences must lie in a single extension. If such a scheme exists which cannot be implemented within the confines outlined above, some other inference procedure will be required. Given the problems already observed with parallelism, we doubt that a parallel or quasi-parallel, single-pass, marker-passing algorithm can be found which takes global considerations into account (at least in unrestricted networks).[4]

Touretzky [1984a] has recently developed a well-defined notion of preference, which we discuss in the next section. The above algorithm does not necessarily produce the conclusions this scheme sanctions, but Touretzky observes that there appears to be no parallel marker-passing algorithms which respect this preference-order for all networks.

4.2. Theory Preference

The formalisation of inheritance, above, uses semi-normal links to represent default links with explicit exceptions. We argue that such explicit exceptions are generally necessary to ensure that the resulting theory has a unique extension. This is important for systems whose inference mechanism is incapable of guaranteeing that all the conclusions it draws from the network representation lie within a single extension of the corresponding default theory. Otherwise, the correctness of the system's "beliefs" must be questionable.

Touretzky [1984a, 1984b] argues that our reformulation of inheritance in terms of semi-normal defaults is inappropriate for two reasons. Firstly, adding new information to the knowledge-base requires modification of the defaults already in the knowledge-base. These become increasingly complex as the knowledge-base grows. Secondly, the translation of a link depends on other links in the network. The translation, Touretzky claims, ignores the essentially "hierarchical" nature of inheritance networks, which he views as their chief asset – both in terms of representational conciseness and computational efficiency.

These criticisms suggest that a (common) misapprehension about default logic has occurred. It is commonly believed that a default logic based reasoning system must be able to find any of the extensions of a default theory, and must view them all as equally acceptable sets of beliefs. In fact, while extensions are all acceptable, the logic says nothing about preference of one to another. It has always been assumed that an agent would "choose" one extension or another, but nothing has been said about how such choices should be made. There is no reason not to (and, perhaps, good reason to) exploit extra-logical properties of the knowledge-base (*e.g.*, hierarchical structure) to establish preferred extensions.

[4] Cottrell [1985] has experimented with a multi-pass, "connectionist", parallel architecture which shows some promise here, although no correctness proofs have been forthcoming. Connectionist architectures are beyond the scope of this thesis, however.

To our knowledge, the first algorithm capable of correctly reasoning with an inheritance network in parallel was that presented in the preceding section (see also [Etherington 1983]). Because of the partitioning of the network, the algorithm is incapable of finding some extensions of some default theories; it is not complete. This algorithm is correct; all of its conclusions lie within a single extension. However, it does not necessarily produce the preferred extension, based on the intuitive semantics for inheritance networks.

Touretzky [1984a] has developed a more sophisticated algorithm, based on the "inferential distance" topology. This inferential distance algorithm is applicable to networks without explicit exception links, and is correct, in the sense that all of its conclusions lie within a single extension. Furthermore, the inferential distance concept is based on the principle that ambiguous inheritance should be, when possible, resolved by appealing to the subclass/superclass relation which forms the basis of inheritance.

Inferential distance is somewhere between the "brave" and "cautious" ends of the spectrum of non-monotonic reasoning systems. Essentially, if an individual could inherit property P by virtue of the fact that she IS-A B, and property $\neg P$ because she IS-A C, then the ambiguity is resolved as follows: If C IS-A B and not *vice versa*, inherit $\neg P$; otherwise, if B IS-A C and not *vice versa*, inherit P; otherwise, inherit neither. Conceptually, the inferential distance algorithm eliminates those extensions which do not satisfy the hierarchical nature of the representation, then draws those conclusions which hold in all of the remaining extensions.

This approach captures the semantic intuition (properties associated with subclasses should override those associated with superclasses) which is the fundamental *raison d'ètre* for inheritance representations. It also avoids the pitfalls of incorrect behaviour which curse shortest-path inference algorithms, as evidenced by Theorem 4.5.

Theorem 4.5

> In the absence of "no-conclusion" links, all of the ground facts returned by Touretzky's inferential distance algorithm lie within a single extension of the default theory which corresponds to the inheritance graph in question. ■

To illustrate the inferential distance algorithm, consider the network from Figure 4.4(b). Because *Nautilus* is a subclass of *Cephalopod*, which is a subclass of *Mollusc*, inferential distance gives the desired results: *Nautili* are *Shell-Bearers*, while *Cephalopods* not known to be *Nautili* are not. In the network of Figure 4.7, neither *Republican* nor *Quaker* is a subclass of the other. Thus inferential distance sanctions no conclusions about whether *Nixon* is a *Pacifist*.

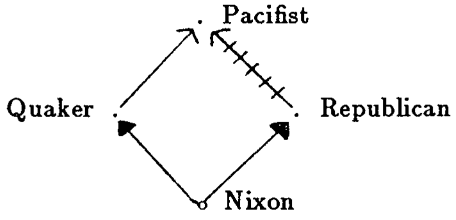


Figure 4.7— A genuinely ambiguous inheritance graph.

Theorem 4.5 only begins to explore the connections between Touretzky's work and that reported in chapters 3 and 4 of this thesis (and in [Reiter 1980a]). We have shown that ground facts returned by inferential distance – *e.g.*, "Clyde is an elephant", or "Clyde is not grey" – belong to a common extension of the corresponding default theory. Inferential distance also sanctions normative conclusions, such as "Albino-elephants are [typically] herbivores". We have not explored the relationship such statements inferred under inferential distance bear to the underlying default theory.

Touretzky also allows what he calls "no-conclusion" links. These links allow inheritance to be blocked without explicit cancellation. Default logic has no analogue for the no-conclusion link, and we have excluded them from consideration here. It appears that it would be straightforward to add a similar capacity to default logic, assuming that such links actually prove useful. The proof of theorem 4.5 suggests that its generalization to networks with no-conclusion links *vis-à-vis* such an extended default logic would present no problems.

Touretzky [1984a] provides a detailed exploration of the properties of inferential distance inheritance reasoning, including a constructive mechanism for determining the 'grounded expansions' (analogous to extensions) of a network. Many of his results bear a superficial similarity in form and proof to those in [Reiter 1980a] and in chapter 3 of this thesis. His proofs rely on partial acyclicity conditions which seem similar to the orderedness conditions we describe. We have speculated (as has Touretzky) that the results in [Touretzky 1984a] and those contained herein may prove to be closely related.

Finally, Touretzky [1984a, 1985] explores the applications of inferential distance to "inheritable relations", citing examples such as

Citizens dislike crooks.
Elected crooks are crooks.
Gullible citizens don't dislike elected crooks.

In this example, citizens generally dislike elected crooks, but Fred, the gullible citizen, doesn't dislike Dick, the elected crook. A complete treatment of the relation between Touretzky's work and default logic should try to extend the correspondence presented here to include Touretzky's inferential distance treatment of inheritable relations.

Touretzky shows that, in general, parallel marker-passing algorithms cannot derive the conclusions sanctioned by the inferential distance algorithm. He also shows that an arbitrary network can be "conditioned", by adding logically-redundant links, in such a way that a parallel marker-passing algorithm *can* return correct results. Unfortunately, this conditioning, which must be performed each time the network is modified, is expensive (Touretzky [1984a] gives a polynomial-time algorithm which adds $O(N^2)$ links in the worst case) and is apparently not amenable to parallel marker-passing implementation [Touretzky 1982, personal communication; 1983].

We conclude that, for conditioned networks, there are correct (in the sense we have described) parallel, marker-passing algorithms for determining inheritance in the presence of exceptions. Such algorithms can be viewed as fast inference algorithms for reasoning with the tractable class of default theories which correspond to conditioned networks.

CHAPTER 5

Predicate Circumscription

In this chapter we focus on predicate circumscription, as presented in [McCarthy 1980]. Our objective is to establish various results concerning the consistency of this formalism, and to describe some limitations of its ability to conjecture new information. One such limitation is that predicate circumscription cannot account for the standard kinds of default reasoning. Another limitation relates to equality; predicate circumscription yields no new ground facts about the equality predicate for a large class of first-order theories. This has important consequences for the so-called "unique names" and "domain closure" assumptions.

5.1. Formal Preliminaries

Predicate circumscription was discussed in detail in chapter 2. We repeat some of the technical details here for convenience. The semantic intuition underlying predicate circumscription is that closed-world reasoning about one or more predicates of a theory corresponds to truth in all models of the theory which are minimal in those predicates. Specifically, let $T(P_1,...,P_n)$ be a first-order theory, some (but not necessarily all) of whose predicates are $\mathbf{P} = \{P_1,...,P_n\}$. A model M of T is a **P**-*submodel* of a model M' of T iff the extension of each P_i in M is a subset of its extension in M', and M and M' are otherwise identical. M is a **P**-*minimal model of* T iff every **P**-submodel of M is identical to M.

For finite theories, $T(P_1,...,P_n)$, McCarthy [1980] proposes realizing predicate circumscription syntactically by adding the following axiom schema to T:

$$\left[T(\Phi_1,...,\Phi_n) \wedge \bigwedge_{i=1}^{n} [\forall \vec{x}.\ (\Phi_i\vec{x} \supset P_i\vec{x})]\right] \supset \bigwedge_{i=1}^{n} [\forall \vec{x}.\ (P_i\vec{x} \supset \Phi_i\vec{x})].$$

Here $\Phi_1,...,\Phi_n$ are predicate variables with the same arities as $P_1,...,P_n$, respectively. $T(\Phi_1,...,\Phi_n)$ is the sentence obtained by conjoining the sentences of T, then replacing every occurrence of $P_1,...,P_n$ in T by $\Phi_1,...,\Phi_n$, respectively. The above schema is called *the (joint) circumscription schema of* $P_1,...,P_n$ *in* T. Let $CLOSURE_{\mathbf{P}}(T)$ – *the closure of* T *with respect to* $\mathbf{P} = \{P_1,...,P_n\}$ – denote the theory consisting of T together with the above axiom schema. McCarthy formally identifies reasoning about T under the closed-world assumption with respect to the predicates **P** with first-order deductions from the theory $CLOSURE_{\mathbf{P}}(T)$.

McCarthy [1980] shows that any instance of the schema resulting from circumscribing a single predicate P in a sentence $T(P)$ is true in all $\{P\}$-minimal models of T. This generalizes directly to the joint circumscription of multiple predicates; we omit the proof of this. We use this generalization extensively in the proofs of the results of this chapter. Because predicate circumscription is applicable only to finitely axiomatizable theories, we will restrict our attention to such theories.

5.2. On the Consistency of Predicate Circumscription

The minimal model semantics of predicate circumscription guarantees that $CLOSURE_{\mathbf{P}}(T)$ is consistent whenever T has **P**-minimal models. This suggests that certain consistent first-order theories lacking minimal models may have inconsistent closures. Indeed, this can happen, as we now show.

Example 5.1 – An inconsistent circumscription

Consider the following consistent theory:

$$T = \left\{ \begin{array}{l} \exists x.\ \mathbf{N}x \land \forall y.\ [\mathbf{N}y \supset x \neq succ(y)] \\ \forall x.\ \mathbf{N}x \supset \mathbf{N}succ(x) \\ \forall xy.\ succ(x) = succ(y) \supset x = y \end{array} \right\}$$

In any model of T, the extension of **N** contains a sequence of elements isomorphic to the natural numbers. An $\{\mathbf{N}\}$-submodel can always be constructed by deleting a finite initial segment of this sequence. Hence every model of T has a proper $\{\mathbf{N}\}$-submodel, so T has no $\{\mathbf{N}\}$-minimal models.

Circumscribing **N** in this theory, and letting Φx be $[\mathbf{N}x \land \exists y.\ x = succ(y) \land \mathbf{N}y]$ yields $\forall x.\ \mathbf{N}x \supset \exists y.\ [\mathbf{N}y \land x = succ(y)]$ which contradicts the first axiom. ∎

In view of this example, it is natural to seek classes of first-order theories for which predicate circumscription does not introduce inconsistencies. The "well-founded" theories form such a class. We say that a first-order theory is *well-founded* iff each of its models has a **P**-minimal submodel for every finite set of predicates **P**. Any consistent well-founded theory obviously has at least one **P**-minimal model. Since every instance of the circumscription schema of **P** in a theory T is true in all **P**-minimal models of T, we have:

Theorem 5.1

If T is a consistent well-founded theory, then $CLOSURE_{\mathbf{P}}(T)$ is consistent for any set **P** of predicates of T. *I.e.*, predicate circumscription preserves consistency for well-founded theories. ∎

Which theories are well-founded? We know of no complete syntactic characterization, but a partial answer comes from a generalization of a result on universal theories due to Bossu and Siegel [1985]. A first-order theory is *universal* iff the prenex normal form of each of its formulae contains no existential quantifiers.

Theorem 5.2

Universal theories are well-founded. ∎

In view of Theorem 5.1, we know that predicate circumscription preserves consistency for universal theories:

Corollary 5.3

If T is a consistent universal theory, then $CLOSURE_{\mathbf{P}}(T)$ is consistent for any set $\mathbf{P}$ of predicates of T. ∎

Notice that the class of universal theories includes the Horn theories, which have attracted considerable attention from the PROLOG, AI, and Database communities.

Lifschitz [1985b] has generalized theorem 5.2 to include the class of "almost universal" theories. A theory is *almost universal in* **P** iff it has the form $\forall \vec{x}.\ A$, where A does not contain positive occurrences of $P \in \mathbf{P}$ within the scope of quantifiers. Almost universal theories include universal theories as well as the "separable" theories of Lifschitz [1985a] (see § 2.1.5.2).

5.3. Well-Founded Theories and Predicate Circumscription

The property of well-foundedness, taken together with the "soundness" of predicate circumscription with respect to the set of minimal models allows us to characterize the power of predicate circumscription. This leads to some rather surprising results. In this section we describe some limitations of predicate circumscription with respect to well-founded theories. The first such result is that predicate circumscription yields no new positive ground instances of any of the predicates being circumscribed.

Theorem 5.4

Suppose that T is a well-founded theory, $P \in \mathbf{P}$ is an n-ary predicate, and $\vec{\alpha}_1, \ldots, \vec{\alpha}_k$ are n-tuples of ground terms. Then

$$CLOSURE_{\mathbf{P}}(T) \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k \iff T \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k\,. \qquad ∎$$

On reflection, this is not too surprising, since circumscription is intended to minimize the extensions of those predicates being circumscribed. New positive instances of such predicates should not arise from this minimization.

A more interesting – even startling – result is that no new ground instances, positive or negative, of uncircumscribed predicates can be derived by predicate circumscription.

Theorem 5.5

Suppose that T is a well-founded theory, $P \notin \mathbf{P}$ is an n-ary predicate, and $\vec{\alpha}_1,...,\vec{\alpha}_k$ are n-tuples of ground terms. Then

(i) $CLOSURE_{\mathbf{P}}(T) \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k \iff T \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k$, and

(ii) $CLOSURE_{\mathbf{P}}(T) \vdash \neg P\vec{\alpha}_1 \vee ... \vee \neg P\vec{\alpha}_k \iff T \vdash \neg P\vec{\alpha}_1 \vee ... \vee \neg P\vec{\alpha}_k$. ∎

In summary, Theorems 5.4 and 5.5 tell us that the only new ground literals that can be conjectured by predicate circumscription of well-founded theories are negative instances of one of the predicates being circumscribed. An unfortunate consequence of this result is that the usual kinds of default reasoning cannot be realized by predicate circumscription. To see why, consider the standard AI example concerning whether birds fly, given that "by default" birds fly. The relevant facts may be represented in various ways, two of which follow:

1) In this representation, all of the exceptions to flight are listed explicitly in the axiom sanctioning the conclusion that birds can fly.

   $\forall x.\ Bird(x) \wedge \neg Penguin(x) \wedge \neg Ostrich(x) \wedge \neg Dead(x) \wedge ... \supset Can\text{-}Fly(x)$

   In addition, there are various IS-A axioms, as well as mutual exclusion axioms:

   $\forall x.\ Canary(x) \supset Bird(x)$
   $\forall x.\ Penguin(x) \supset Bird(x)$
   $\forall x.\ \neg(Canary(x) \wedge Penguin(x))$
   $\forall x.\ \neg(Penguin(x) \wedge Ostrich(x))$

2) In this representation, due to McCarthy [1986], a new predicate, *ab*, standing for "abnormal", is introduced. One then states that "normal" birds can fly:

   $\forall x.\ Bird(x) \wedge \neg ab(x) \supset Can\text{-}Fly(x)$

   The abnormal birds are listed:

   $\forall x.\ Penguin(x) \supset ab(x)$
   $\forall x.\ Ostrich(x) \supset ab(x)$

   Finally, one includes the IS-A and mutual exclusion axioms as in (1) above.

Both representations (1) and (2) are universal, and hence well-founded, theories. Therefore, if *Bird*(*Tweety*) is given, Theorems 5.4 and 5.5(i) tell us that the default assumption *Can–Fly*(*Tweety*) cannot be conjectured by predicate circumscription.

Careful readers of [McCarthy 1980] might find Theorems 5.4 and 5.5 inconsistent with the results in Section 7 of that paper. In the blocks-world example presented there to illustrate predicate circumscription, the ground instance $on(A,C,result(move(A,C),s_0))$ can be derived by circumscribing a different predicate, $\lambda z.prevents(z,move(A,C),s_0)$. This appears to violate Theorem 5.5(i). This discrepancy stems from the fact that in formulating the circumscription schema for this example, McCarthy uses specializations of some of the original axioms (*i.e.*, the axioms which specify what can *prevent* a *move* from succeeding), and omits one of the axioms (*i.e.*, the axiom which states that if nothing *prevents* a *move* from succeeding, the *move* will be successful). Thus, only part of the theory enters into the circumscription for his example, whereas Theorems 5.4 and 5.5 suppose that the entire theory is used in proposing a circumscription schema.

5.4. Equality

We now consider some limitations of predicate circumscription with respect to the treatment of equality. These limitations will be seen to have consequences for two special cases of closed-world reasoning, namely deriving the "unique names assumption" and the "domain closure assumption".

5.4.1. The Unique-Names Assumption

When told that Tom, Dick and Harry are friends, one naturally assumes that 'Tom', 'Dick' and 'Harry' denote distinct individuals: Tom $\neq$ Dick, Tom $\neq$ Harry, Dick $\neq$ Harry. For a more general example, consider a setting in which one is told that Tom's telephone number is the same as Sue's, and that Bill's number is 555-1234, which is different from Mary's number. Thus, we have:

tel-no(Tom) = tel-no(Sue)
tel-no(Bill) = 555-1234
tel-no(Mary) $\neq$ 555-1234

One would naturally assume from this information that tel-no(Tom) $\neq$ 555-1234, and that tel-no(Tom) $\neq$ tel-no(Mary).

In general, the unique-names assumption is invoked whenever one can assume that all of the relevant information about the equality of individuals has been specified. All pairs of individuals not specified as identical are assumed to be different. This assumption arises in a number of settings, for example in the theory of databases [Reiter 1980b], and in connection with the semantics of negation in PROLOG [Clark 1978]. Virtually every AI reasoning system, with the exception of those based on theorem-provers, implicitly makes this assumption. Because of Clark's results, we know that this is also the case for PROLOG based AI systems.

Unique-names axioms are also important for closed-world reasoning using predicate circumscription. For example, if all we know is that *Opus* is a *Penguin*, we can circumscriptively conjecture $\forall x.\ Penguin(x) \equiv x = Opus$. We cannot use this to deduce $\neg Penguin(Tweety)$, however, unless we know $Opus \neq Tweety$.

How then can we formalize reasoning under the unique-names assumption? The natural first attempt is to circumscribe the equality predicate in the theory under consideration. To that end, we shall assume that the theory T contains the following axioms which define the equality predicate, =, for the theory:

$\forall x.\ x = x$
$\forall xy.\ x = y \supset y = x$
$\forall xyz.\ x = y \wedge y = z \supset x = z$
$\forall x_1,...,x_n, y_1,...,y_n.\ x_1 = y_1 \wedge ... \wedge x_n = y_n \wedge P(x_1,...,x_n)$
$\qquad \supset P(y_1,...,y_n)$, for each n-ary predicate symbol P of T.
$\forall x_1,...,x_n, y_1,...,y_n.\ x_1 = y_1 \wedge ... \wedge x_n = y_n$
$\qquad \supset f(x_1,...,x_n) = f(y_1,...,y_n)$, for each n-ary function symbol f of T.

When T is finite, it is therefore possible to circumscribe the equality predicate, since the resulting schema is finite. The next result informs us that doing so yields nothing new.

Theorem 5.6 (Reiter)

Let T be a first-order theory containing axioms which define the equality predicate, =. Then $T \vdash CLOSURE_{\{=\}}(T)$. ■

In view of this result, one might attempt to capture the unique names assumption by jointly circumscribing several predicates of the theory, not just the equality predicate. We do not know whether there are any theories for which this might work, but it cannot succeed for well-founded theories. No new ground equalities or inequalities can be derived by circumscribing a well-founded theory, regardless of the predicates circumscribed.

Theorem 5.7

Suppose that T is a well-founded theory containing axioms which define the equality predicate; $\alpha_1,...,\alpha_k$; $\beta_1,...,\beta_k$ are ground terms, and $\mathbf{P}$ is a set of some of the predicates of T. Then

(i) $CLOSURE_{\mathbf{P}}(T) \vdash (\bigvee_{i=1}^{k} \alpha_i = \beta_i) \iff T \vdash (\bigvee_{i=1}^{k} \alpha_i = \beta_i)$, and

(ii) $CLOSURE_{\mathbf{P}}(T) \vdash (\bigvee_{i=1}^{k} \alpha_i \neq \beta_i) \iff T \vdash (\bigvee_{i=1}^{k} \alpha_i \neq \beta_i)$. ■

Corollary 5.8

Suppose that T is a well-founded theory containing axioms which define the equality predicate; P is an n-ary predicate; and $\alpha_1,...,\alpha_k$; $\beta_1,...,\beta_k$ are ground terms. Then

$$CLOSURE_{\mathbf{P}}(T) \vdash \neg P(\alpha_1,...,\alpha_k) \Rightarrow T \vdash (\bigvee_{i=1}^{k} \alpha_i \neq \beta_i) \text{ or } T \not\vdash \neg P(\beta_1,...,\beta_k) .$$ ■

Returning to the "Penguin" example above, we see that predicate circumscription cannot conjecture $\neg Penguin(Tweety)$ unless it is known that $Opus \neq Tweety$; otherwise we could derive $Opus \neq Tweety$ from $CLOSURE(\{Penguin(Opus)\})$, contradicting Theorem 5.7.

This last restriction is somewhat puzzling. The model-theory fixes the domain and the interpretations of constants and function symbols when determining minimal models. Given the soundness of predicate circumscription with respect to this model-theory, it is easy to see why identity is not influenced by the set of minimal models. If the equality predicate is interpreted as a congruence relation, rather than as identity (*i.e.,* if non-normal models are allowed, where pairs of distinct domain elements are permitted to be in the extension of '='), the situation is less clear. Essentially, it can be shown that, for any pair of terms for which one might hope to circumscriptively conjecture (in)equality, there are minimal models which support either side of the issue. So much for the semantic explanation. There remain two questions. What feature of the circumscription schema gives rise to this anomaly? What does this tell us about circumscription? A partial answer to the first question is that Leibniz' principle of substitutivity – equals are everywhere intersubstitutible preserving truth – makes a stronger statement about equality than the circumscription schema. The second question remains unanswered.

In a recent paper, McCarthy [1986] proposes a circumscriptive approach to the unique-names assumption by introducing two equality predicates. One of these is the standard equality predicate, but restricted to arguments which are names of objects. The other equality predicate, $e(x,y)$, means that the names x and y denote the same object. e is axiomatized as an equivalence relation which does not, however, satisfy the full principle of substitution, in contrast to "normal" equality. This failure of full substitutivity for the predicate e prevents our Theorems 5.6 and 5.7 from applying to e. Benjamin Grosof (personal communication) has independently proposed a similar approach to the unique-names assumption. He has also observed that our Theorem 5.6 applies to McCarthy's [1986] more general notion of circumscription.

5.4.2. The Domain Closure Assumption

The domain-closure assumption is the assumption that, in a given first-order theory T, the universe of discourse is restricted to the smallest set which contains those individuals mentioned in T, and which is closed under the application of those functions mentioned in T. Domain circumscription [McCarthy 1977, 1980] is a proposed formalization of this assumption. McCarthy [1980] suggests that domain circumscription might be reduced to predicate circumscription. This is in fact false, as shown by Theorems 5.9 and 5.10.

The simplest setting in which the domain-closure assumption can arise is for a theory with a finite Herbrand Universe $\{c_1,...,c_n\}$. In this case we might want to conjecture the *domain-closure axiom* for this theory: $\forall x.\ x = c_1 \vee ... \vee x = c_n$. Such an axiom is important for the theory of first-order databases [Reiter 1980b]. No such axiom can arise from predicate circumscription for well-founded theories.

Theorem 5.9

Suppose that T is a well-founded theory; $t_1,\ldots,t_n$ are ground terms; and $\mathbf{P}$ is a set of some of the predicate symbols of T. Then

$$\mathit{CLOSURE}_{\mathbf{P}}(T) \vdash \forall x.\ x = t_1 \vee \ldots \vee x = t_n \Longleftrightarrow$$
$$T \vdash \forall x.\ x = t_1 \vee \ldots \vee x = t_n\,. \qquad \blacksquare$$

Theorem 5.10

If T is a well-founded theory and T has a model with some domain, D, then so does $\mathit{CLOSURE}_{\mathbf{P}}(T)$. $\blacksquare$

5.4.3. Some Misconceptions

There are a number of common misconceptions about the use of predicate circumscription, which we discuss briefly, below.

It has been proposed that arbitrary formulae could be circumscribed using predicate circumscription by including a new predicate letter and a definition declaring it to be equivalent to the expression to be circumscribed. This will not work, in general.

Theorem 5.11

If $T \vdash \forall \vec{x}.\ P\vec{x} \equiv \Phi\vec{x}$ for some expression $\Phi\vec{x}$, not involving predicate letters from $\mathbf{P}$, then $T \vdash \mathit{CLOSURE}_{\mathbf{P}}(T)$. $\blacksquare$

This result seems to be related to Doyle's [1984] comments on implicit definability. Since the theory already contains a *definition* for P, circumscription cannot further constrain P. As it is generally undecidable whether $T \vdash \forall \vec{x}.\ P\vec{x} \equiv \Phi\vec{x}$ for a particular Φ (let alone all Φ), it follows that one cannot decide which predicates to circumscribe.

Corollary 5.12

It is generally undecidable whether $\mathit{CLOSURE}_{\mathbf{P}}(T)$ is stronger than T. $\blacksquare$

It is widely (and correctly) believed that $\mathit{CLOSURE}_{\mathbf{P}}(\{\exists x.\ Px\}) \models \exists ! x.\ Px$ (*i.e.*, there is a unique P). There appears to be some misunderstanding about how this is achieved, however. After some experimentation, the idea of skolemization comes to mind and, indeed, $\mathit{CLOSURE}_{\mathbf{P}}(P\alpha) \models \exists ! x.\ Px$ (actually $\forall x.\ Px \equiv x = \alpha$). Skolemization, however, can change the set of minimal models of a theory (and hence the results of circumscription). To see this, notice that in Example 5.1, T has no minimal models, but the skolemized form of T is universal and hence well-founded. There has been a tendency to believe that the skolemized form of a theory T is equivalent to T, which is false. In fact, skolemization preserves *satisfiability*, not derivability; the *existence* of models, not the set of models. The actual circumscriptive derivation of $\exists !x.\ Px$ from $\exists x.\ Px$ involves the substitution of a *binary* predicate for Φx, *viz* $x = u$, where u is a variable distinct from x.

The skeleton of a correct circumscriptive derivation in a natural deduction system of $\exists !x.\ Px$ from $\exists x.\ Px$ follows:

$$\begin{array}{lll}
1 & [(\exists x.\ \Phi x) \wedge (\forall x.\ \Phi x \supset Px)] \supset (\forall x.\ Px \supset \Phi x)] & CLOSURE_{\{P\}}(\exists x. Px) \\
2 & [(\exists x.\ x = u) \wedge (\forall x.\ x = u \supset Px)] & \\
 & \qquad \supset (\forall x.\ Px \supset x = u)] & 1,\ [x = u/\Phi x] \\
3 & \forall u.\ [(\exists x.\ x = u) \wedge (\forall x.\ x = ux \supset Px)] & \\
 & \qquad \supset (\forall x.\ Px \supset x = u)] & 2,\ \textit{universal generalization} \\
4 & \exists x.\ Px & \textit{given} \\
5 & P\alpha & \textit{hypothesis} \\
6 & [(\exists x.\ x = \alpha) \wedge (\forall x.\ x = \alpha \supset Px)] & \\
 & \qquad \supset (\forall x.\ Px \supset x = \alpha)] & 3,\ \textit{universal instantiation} \\
7 & P\alpha \supset (\forall x.\ Px \supset x = \alpha) & 6,\ \textit{tautology} \\
8 & \forall x.\ Px \supset x = \alpha & 5,7,\ \textit{tautology} \\
9 & P\alpha \wedge (\forall x.\ Px \supset x = \alpha) & 5,8,\ \textit{tautology} \\
10 & \exists y.\ Py \wedge (\forall x.\ Px \supset x = y) & 4,5,9,\ \textit{existential generalization} \\
11 & \exists !x.\ Px & 10,\ \textit{definition}
\end{array}$$

There have been implicit [McCarthy 1980] and explicit [Genesereth and Nilsson 1987] suggestions that the way around some of the limitations of predicate circumscription might be to circumscribe only "relevant" portions of the theory. The idea is that, by weakening $T(\Phi)$ – hence eliminating some of the conditions that Φ must satisfy – perhaps some more useful results will obtain. Obviously, one must be careful; circumscribing P in Pa, leaving out Pa will produce $\forall x.\ \neg Px$. This, being inconsistent with Pa, is perhaps *too* useful. One idea is to distinguish, amongst the positive literals in each clause of the theory, one which the clause is said to be "about". Then only those clauses "about" P are taken into account in forming $T(\Phi)$. This may indeed allow positive facts to be derived. For example, consider

$$T = \begin{cases} \forall x.\ Bird(x) \wedge \neg Penguin(x) \supset Flies(x) \\ Bird(Tweety),\ Penguin(Opus),\ Opus \neq Tweety \\ \forall x.\ Penguin(x) \supset Bird(x) \end{cases}$$

If the first axiom is taken to be about *Flies*, then we get

$$\Phi Opus \wedge [\forall x.\ \Phi x \supset Px] \supset [\forall x.\ Px \supset \Phi x]$$

when we circumscribe (in this fashion) *Penguin* in T. From this we can derive $\forall x.\ Penguin(x) \equiv x = Opus$ and $Flies(Tweety)$! There are two drawbacks with this approach, however. The first is that its semantics are unknown. They are *not* those of predicate circumscription, and there is no known model-theory or "soundness" result corresponding to that for predicate circumscription. Thus it is not clear what this approach computes. More seriously, consistency is not necessarily preserved; the first axiom of T is also "about" *Penguins*, in the sense that $T^* = (T \cup \{\neg Flies(Tweety)\}) \vdash Penguin(Tweety)$. Taking the above approach to circumscribing T^* will result in an inconsistency.

5.5. What to Circumscribe?

One obvious problem with using circumscription in a given setting is knowing just what to circumscribe. Some of our results provide clues in this direction. (Corollary 5.12 shows that clues are the best that can be hoped for, in general.) Theorem 5.5 tells us that if we wish to use predicate circumscription to conjecture $\neg P(\vec{\alpha})$ in some well-founded theory then we must include P among the predicates being circumscribed. Theorems 5.4 and 5.5 tell us that predicate circumscription will not do at all if we wish to conjecture $P(\vec{\alpha})$, as is the case for most forms of default reasoning, so that we must appeal to some other mechanism, such as McCarthy's more general form of circumscription, discussed in the next chapter.

CHAPTER 6

Generalizations of Circumscription

6.1. Formula Circumscription

McCarthy [1986] has recently formulated a generalization of predicate circumscription, called formula circumscription. This generalization provides for the minimization of arbitrary first-order expressions rather than simple predicates. It also provides for the treatment of designated predicates as variables of the minimization. In this version of circumscription some of the limitations of Theorem 5.5 no longer apply. Thus, as some of McCarthy's examples show, it is possible to circumscribe a predicate P, treating another predicate Q as variable, and derive new positive and negative ground instances of Q. In particular, McCarthy's new formalism appears adequate for the treatment of some forms of default reasoning, as his examples show.

Many of the limitations of predicate circumscription stem from the fact that only those predicates being minimized are allowed to vary. Formula circumscription retains many of the attractive features of its predecessor, without some of its limitations. McCarthy's definition of the formula circumscription of $E(\mathbf{P},\vec{x})$ in the theory $T(\mathbf{P})$ takes the form of the second-order axiom, (22).

$$T(\mathbf{P}) \wedge \forall\Phi.\ T(\Phi) \wedge [\forall\vec{x}.\ E(\Phi,\vec{x}) \supset E(\mathbf{P},\vec{x})] \supset [\forall\vec{x}.\ E(\mathbf{P},\vec{x}) \supset E(\Phi,\vec{x})] \tag{22}$$

where $E(\mathbf{P},\vec{x})$ is any well-formed expression whose free individual variables are among $\vec{x} = x_1,\ldots,x_k$, and in which some of the predicate variables $\mathbf{P} = \{P_1,\ldots,P_n\}$ occur free; $E(\Phi,\vec{x})$ is the result of replacing each free occurrence of the predicate letters, P_i, in $E(\mathbf{P},\vec{x})$ with predicate variables, Φ_i, of the same arity.

Not everyone is convinced of the need for second-order logic for circumscription [Perlis and Minker 1986]. A first-order schema version of formula circumscription, (23), is obtained by deleting the second-order quantifier, $\forall\Phi$.

$$T(\mathbf{P}) \wedge T(\Phi) \wedge [\forall\vec{x}.\ E(\Phi,\vec{x}) \supset E(\mathbf{P},\vec{x})] \supset [\forall\vec{x}.\ E(\mathbf{P},\vec{x}) \supset E(\Phi,\vec{x})] \tag{23}$$

We will sometimes write $CLOSURE(T;\ \mathbf{P};\ E(\mathbf{P},\vec{x}))$ for either axiom (22) or schema (23), indicating minimization of the expression $E(\mathbf{P},\vec{x})$, with the predicates $\mathbf{P}$ treated as variable, in the theory, T.

McCarthy presented only a syntactic characterization of formula circumscription. Motivated by a belief in the importance of semantic characterizations for reasoning systems, and by the striking consequences of exploring the semantics of predicate circumscription, we explored the possibility that an appropriate generalization of the minimal-model semantics of predicate circumscription would characterize formula circumscription.[1] This led us to a form of the generalized minimal-model semantics which has since been used in the explication of a variety of closed-world reasoning formalisms (see §2.1). The precise details are given below.

Definition: $M \leq_{E(\mathbf{P},\vec{x})} M'$

Let $T(\mathbf{P})$ be a finitely axiomatized (first- or second-order) theory, some (but not necessarily all) of whose predicates are those in $\mathbf{P}$; let $E(\mathbf{P},\vec{x})$ be a formula whose free variables are among $\vec{x} = x_1,\ldots,x_n$, and in which some of the predicate variables $\mathbf{P} = \{P_1,\ldots,P_n\}$ occur free; and let M, M' be models of T. We say M is an $E(\mathbf{P},\vec{x})$-*submodel* of M' (written $M \leq_{E(\mathbf{P},\vec{x})} M'$) iff

(i) $|M| = |M'|$,

(ii) If t is a term, then $|t|_M = |t|_{M'}$,

(iii) If $Q \notin \mathbf{P}$ is a predicate letter of T, then $|Q|_M = |Q|_{M'}$, and

(iv) $|E(\mathbf{P},\vec{x})|_M \subseteq |E(\mathbf{P},\vec{x})|_{M'}$. ■

Definition: $E(\mathbf{P},\vec{x})$-Minimal Model

A model, M, of T is $E(\mathbf{P},\vec{x})$-*minimal* iff T has no model, M', such that $M' \leq_{E(\mathbf{P},\vec{x})} M$ and $\neg(M \leq_{E(\mathbf{P},\vec{x})} M')$. ■

That this is the correct semantics is suggested by Theorems 6.1 and 6.2. Theorem 6.2 is applicable only to the first-order-schema version of formula circumscription; Theorem 6.1 applies both to that and to second-order formula circumscription.

Theorem 6.1 – Soundness

CLOSURE(T; $\mathbf{P}$; $E(\mathbf{P},\vec{x})$) is satisfied by every $E(\mathbf{P},\vec{x})$-minimal model of T. ■

Theorem 6.2 – Finitary Completeness (Perlis and Minker)

If all models of T have finite extensions for each $P \in \mathbf{P}$ (modulo equality), then M satisfies every instance of *CLOSURE*(T; $\mathbf{P}$; $E(\mathbf{P},\vec{x})$) only if M is an $E(\mathbf{P},\vec{x})$-minimal model of T. ■

[1] Lifschitz [1985, personal communication] argues that the model-theory for second-order logic provides sufficient semantics for the generalized forms of circumscription. While this may be true, the explicit notion of minimality leads to useful insights, as is indicated in the sequel.

Perlis and Minker [1986] actually prove a slightly stronger result, applicable if all models for $CLOSURE(T; \mathbf{P}; E(\mathbf{P},\vec{x}))$ have finite extensions for each $P \in \mathbf{P}$. Of course, no general completeness result could be forthcoming. There is a unique (up to isomorphism) minimal model for the standard axiomatization of the natural numbers, but there is no recursive first-order axiomatization which uniquely characterizes this model. If circumscription were complete, it could be used to conjecture such a first-order axiomatization.

It is worthwhile determining which of the differences between predicate circumscription and formula circumscription are really necessary. As McCarthy has suggested, the minimization of arbitrary expressions is not.

Theorem 6.3

> The ability to minimize arbitrary expressions, $E(\mathbf{P},\vec{x})$, instead of simple sets of predicates, is an inessential extension, provided predicates other than those being minimized are allowed to vary. ■

Theorem 6.3 tells us that it suffices to circumscribe predicates. To see this, observe that one can simply extend the language with a new predicate symbol, ψ and add the axiom:

$$\forall \vec{x}.\ \psi \vec{x} \equiv E(\mathbf{P},\vec{x})$$

to the theory. Circumscribing $\psi\vec{x}$ in the extended theory with $\mathbf{P} \cup \{\psi\}$ variable results in a *conservative extension* (no new theorems over the original language are derivable) of the circumscription of $E(\mathbf{P},\vec{x})$ in the original theory.

6.2. Generalized Circumscription

McCarthy's formula circumscription has lately been generalized by Lifschitz [1984], exploiting pre-orders, as discussed in §2.1.5.2. Lifschitz' generalized form is:

$$T(\mathbf{X}) \wedge \forall \mathbf{X}'.\ T(\mathbf{X}') \wedge (\mathbf{X}' \leq_R \mathbf{X}) \supset (\mathbf{X} \leq_R \mathbf{X}') \qquad (24)$$

where $\leq_R$ denotes the pre-order on tuples of (predicate, function, and individual) variables induced by a reflexive, transitive relation, R. We call this *generalized circumscription*, and write $CLOSURE(T; \mathbf{X}; R)$ for (24) or the corresponding first-order schema. This formulation allows for arbitrary ordering relations to drive the minimization, and provides for the denotations of terms (constant and function letters) to be affected by the minimization process.

The extended minimal-model semantics outlined above is amenable to this further generalization. The most significant change from the forms we have seen to this point is that the denotations of some constant and function terms may change between a model and its submodels. The appropriate definitions are:

Definition: $M \leq_{(\mathbf{X},R)} M'$

Let $T(\mathbf{P})$ be a finitely axiomatized (first- or second-order) theory, whose predicate, function and constant letters include (but need not be limited to) those in $\mathbf{X}$; let R be a binary relation on tuples of type $\mathbf{X}$; let $\leq_R$ be the pre-order induced by R; and let M, M' be models of T. Then M is an $(\mathbf{X},R)$-*submodel* of M' (written $M \leq_{(\mathbf{X},R)} M'$) iff

(i) $|M| = |M'|$,

(ii) If t is a term and $t \notin \mathbf{X}$, then $|t|_M = |t|_{M'}$,

(iii) If $Q \notin \mathbf{X}$ is a predicate letter of T, then $|Q|_M = |Q|_{M'}$, and

(iv) $<|\mathbf{X}|_M, |\mathbf{X}|_{M'}> \in R$. ■

Definition: $(\mathbf{X},R)$-Minimal Model

A model, M, of T is $(\mathbf{X},R)$-*minimal* iff T has no model, M', such that $M' \leq_{(\mathbf{X},R)} M$ and $\neg(M \leq_{(\mathbf{X},R)} M')$. ■

We have shown that generalized circumscription is sound *vis-à-vis* the set of minimal models specified by this model theory.

Theorem 6.4 – Soundness

CLOSURE(T; $\mathbf{X}$; R) is satisfied by every $(\mathbf{X},R)$-minimal model of T. ■

We do not know whether there is an analogue of Theorem 6.2 (finitary completeness) for generalized circumscription.

The provision for variable terms leads to some surprising results. These include new positive equality statements, and the provability of new positive or negative ground facts in predicates not included among those specified as variable.

Proposition 6.5

If terms are allowed to vary, then new ground equality statements
may result from generalized circumscription. ■

Proposition 6.6

If terms are allowed to vary, then new ground facts involving predicates
$Q \notin \mathbf{X}$ may result from *CLOSURE*(T; $\mathbf{X}$; R). ■

Example 6.1

Consider the theory $T = \{Pa,\ Pb,\ Qb\}$. $CLOSURE(T;\ \{P,\ a\};\ \{P\})$ is

$$Pa \wedge Pb \wedge Qb \wedge \forall \Phi.\ \forall u.\ [\Phi u \wedge \Phi b \wedge (\forall x.\ \Phi x \supset Px)] \supset (\forall x.\ Px \supset \Phi x)$$

Instantiation with $[x = b/\Phi x]$ and $[b/u]$ gives $\forall x.\ Px \supset x = b$, from which we can infer $a = b$ and hence Qa. ■

6.3. Well-Founded Theories

As with all of the forms of minimal-model semantics we have discussed in this thesis, that for generalized circumscription provides for certain elements to differ between a model and its submodels while others remain fixed. Despite Theorems 6.2-6.4, it is not necessarily clear that the syntactic manipulations of generalized (or formula) circumscription respect the intent expressed by this semantic characterization. It is conceivable that all models reflecting a particular configuration of supposedly fixed attributes might have no minimal submodels. The semantics then fails to guarantee that circumscription will not affect these supposedly "inviolable" facets. It is natural to question whether there is any property analogous to the well-foundedness property we discussed for predicate circumscription, which would address this concern. In fact, as we shall see, there is such a notion. Let us redefine the term "well-founded" as follows:

Definition – Well-Foundedness

The theory, T, is *well-founded with respect to* $(\mathbf{X}, R)$ iff every model of T has an $(\mathbf{X}, R)$-minimal submodel. ■

This definition is slightly weaker than that given in chapter 5, where we required that every model of T have a $\mathbf{P}$-minimal submodel for every finite tuple of predicates, $\mathbf{P}$. This weaker definition, relativized to $(\mathbf{X}, R)$, is sufficient for deciding whether a particular circumscription is well-behaved. The more direct generalization of the definition of chapter 5 is so strong that it excludes all theories.

Proposition 6.7 (Lifschitz)

Universal theories are not necessarily well-founded if constants are allowed to vary. ■

Example 6.2 (Lifschitz)

The natural-number example of Example 5.1, with the existentially specified individual replaced by the constant '0':

$$\mathrm{N}0 \wedge \forall x.\ \mathrm{N}x \supset succ(x) \neq 0$$
$$\forall x.\ \mathrm{N}x \supset \mathrm{N}succ(x)$$
$$\forall xy.\ succ(x) = succ(y) \supset x = y$$

is not well-founded with respect to minimization of N with {N, 0} variable. Since the denotation of 0 is allowed to change from model to submodel, the infinite chains of models presented in Example 5.1 serve to show that this theory has no minimal models. ■

Proposition 6.8

No class of theories is well-founded with respect to all pre-orders. ■

Example 6.3

Consider the theory with no proper axioms, and minimize the expression $E(P,x) = Px \wedge [\forall x.\ \neg Px] \wedge [\exists x.\ Px \wedge \neg Psx]$. Consider a model in which P is interpreted by the natural numbers, and s by the successor function. Clearly any non-empty initial subset of the natural numbers produces a proper submodel, but the model with the empty interpretation for P makes E true everywhere. ■

Proposition 6.8 and Example 6.3 can best be understood in terms of Theorem 6.3. Minimization of $E(P,x)$ in T is equivalent to minimization of ψx, with $\{\psi,\ P\}$ variable, in

$$T' = \left\{ \forall x.\ \psi x \equiv \left[Px \wedge [\forall x.\ \neg Px] \wedge [\exists x.\ Px \wedge \neg Psx] \right] \right\}$$

which does not belong to any of the known classes of well-founded theories (because P occurs positively within the scope of existential quantifiers). In some sense, allowing arbitrary pre-orders enables one to "import" arbitrary axioms into the theory.

With these examples in mind, we will restrict our attention in the sequel to the case of simple minimization of some of the predicates of **X**. In other words, we will consider a generalization of joint predicate circumscription, in which other predicates and terms may be allowed to vary. We will write $\leq_{(\mathbf{X},\mathbf{P})}$ for the pre-order determined by the joint minimization of each of the predicates in **P**, allowing the predicates and terms of **X** to vary. (**X** is assumed to contain all of the predicate symbols of **P**.)

The question remains, "Are there any theories which are well-founded?" Fortunately, the answer is "Yes". (This result has been proved independently (using rather different techniques) by Lifschitz [1985].)

Theorem 6.9

> If T is a universal theory, and **X**, **P** are finite tuples of predicate letters, then T is well-founded with respect to $\leq_{(\mathbf{X},\mathbf{P})}$. ■

The existence of well-founded theories proved most distressing in the context of predicate circumscription. What are the repercussions of Theorem 6.9 for generalized circumscription? Certainly, they are less pessimistic. Generalized circumscription affords much greater control over which aspects of models must remain fixed when constructing submodels. This means that generalized circumscription is not driven, willy-nilly, to avoid conclusions which lead to the derivation of new positive information. Thus, for well-founded theories, generalized circumscription allows useful conclusions to be drawn without sacrificing a clear semantic intuition of exactly what is open to conjecture. Also on the positive front, we have Corollary 6.10:

Corollary 6.10

> If T is consistent and well-founded with respect to (**X**, **P**), then $CLOSURE(T;\ \mathbf{X};\ \mathbf{P})$ is consistent. ■

It is natural to question the extent to which the negative results of chapter 5 apply to generalized circumscription. It is clear that, in the case where only the minimized predicates are allowed to vary, that all the results in chapter 5 continue to hold, since in this case generalized circumscription reduces to predicate circumscription. Furthermore, Theorem 5.4 and an appropriate version of Theorem 5.5 continue to hold, even with variable predicates.

Theorem 6.11

> If T is well-founded with respect to (**X**, **P**); $P \in \mathbf{P}$ is an n-ary predicate; **X** a set of predicate letters; and $\vec{\alpha}_1,\dots,\vec{\alpha}_k$ are n-tuples of ground terms; then
>
> $$CLOSURE(T;\ \mathbf{X};\ \mathbf{P}) \vdash P\vec{\alpha}_1 \vee \dots \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee \dots \vee P\vec{\alpha}_k\ .$$ ■

Theorem 6.12

> If T is well-founded with respect to (**X**, **P**); **X** is a set of predicate letters; $P \notin \mathbf{P} \cup \mathbf{X}$ is an n-ary predicate; and $\vec{\alpha}_1,\dots,\vec{\alpha}_k$ are n-tuples of ground terms; then
>
> (i) $CLOSURE(T;\ \mathbf{X};\ \mathbf{P}) \vdash P\vec{\alpha}_1 \vee \dots \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee \dots \vee P\vec{\alpha}_k$, and
>
> (ii) $CLOSURE(T;\ \mathbf{X};\ \mathbf{P}) \vdash \neg P\vec{\alpha}_1 \vee \dots \vee \neg P\vec{\alpha}_k \Longleftrightarrow T \vdash \neg P\vec{\alpha}_1 \vee \dots \vee \neg P\vec{\alpha}_k$. ■

The fact that the model-theory outlined in §6.2 for generalized circumscription (even with variable terms) restricts the submodel relationship to models with identical domains suggests that generalized circumscription (and *a fortiori* formula circumscription) cannot be used to conjecture domain closure axioms. For well-founded theories, this is the case.

Theorem 6.13

> If T is well-founded for $(\mathbf{P},R)$ and T has a model with domain, D, then so does $CLOSURE(T(\mathbf{P});\mathbf{P};R)$. ■

Thus neither generalized circumscription without variable terms nor formula circumscription subsumes domain circumscription.

Equality appears to remain problematic if only predicates are variable, but we have not proven an analogue of Theorem 5.7. Theorem 5.6 continues to apply even if terms are allowed to vary.

Theorem 6.14

> If T is a first-order theory containing axioms which define the equality predicate, $=$, then $T \vdash CLOSURE(T,\mathbf{X},\{=\})$. ■

It appears that unique names axioms are derivable (for theories with finite domains) given variable terms, however [Lifschitz 1984]. Unfortunately, we have seen that variable terms can be problematic. The general formulation of closed-world reasoning about equality using generalized circumscription with variable terms remains an open question.

Also open are the questions of analogues of Theorems 5.4 and 5.5 *vis à vis* arbitrary preorders and/or variable terms. Because of the failure of well-foundedness for these forms of circumscription, the tools we have used in this chapter and in chapter 5 do not apply to these more general problems. Proposition 6.6 suggests that such analogues may not be forthcoming.

CHAPTER 7

Domain Circumscription

In chapter 2, we discussed the motivation for and one realization of domain circumscription. In this chapter, we investigate the formalism more thoroughly.

Domain circumscription [McCarthy 1977, 1980; Davis 1980] is intended to be a syntactic realization of the model-theoretic domain-closure assumption. It provides a mechanism for conjecturing domain-closure axioms, eliminating the need to explicitly state them. To circumscribe the domain of a sentence, A, the schema:

$$Axiom(\Phi) \wedge A^{\Phi} \supset \forall x.\ \Phi(x) \tag{25}$$

is added to A. $Axiom(\Phi)$ is the conjunction of $\Phi\alpha$ for each constant symbol α and $\forall x_1 ... x_n.\ [\Phi x_1 \wedge ... \wedge \Phi x_n] \supset \Phi f x_1 ... x_n$ for each n-ary function symbol f. A^{Φ} is the result of rewriting A, replacing each universal or existential quantifier, '$\forall x.$' or '$\exists x.$', in A with '$\forall x.\Phi x \supset$ ' or '$\exists x.\Phi x \wedge$ ', respectively.

7.1. A Revised Domain Circumscription Axiom Schema

As was noted in §2.1.5.3, the appropriate model-theoretic characterization for domain-closure involves restriction of models to progressively smaller domains, preserving agreement over common terms. This notion of submodel corresponds roughly to the standard notion of "substructure". It is slightly stronger, however, in the sense that substructures are not required to be models of the theory in question.

Davis [1980] shows that every instance of (25) is true in all minimal models of the original sentence A. This result is correct for most theories. However, inconsistency results when circumscribing universal theories (theories whose prenex normal forms contain no leading existential quantifiers) with no constant symbols. For example, consider the relational theory:

$$A = \{\ \forall x.\ Px\ \}.$$

Because there are no constant or function symbols, $Axiom(\Phi)$ is empty, so the domain circumscription schema for A is:

$$\left[\forall x.\ \Phi x \supset Px\right] \supset \forall x.\ \Phi x\ .$$

Mercer [1984, personal communication] has noted that substituting $\neg Px$ for Φx gives:

$$\left[\forall x.\ Px\right] \supset \forall x.\ \neg Px$$

which is clearly inconsistent with A.

The root of this problem is that, for such theories, Φ can be chosen to be universally false. Models of first-order theories must have at least one domain element, so the conjecture that everything is a Φ (and hence there is nothing) is inconsistent. Having isolated the problem, we have developed a simple, easily motivated solution. Since models must have non-empty domains, those Φ's which are identically false must be excluded. To achieve this, the conjunct $\exists x.\ \Phi x$ is added to the left-hand-side of the circumscription schema (25), giving:

$$\exists x.\ \Phi x \wedge Axiom(\Phi) \wedge A^{\Phi} \supset \forall x.\ \Phi(x) \tag{26}$$

Davis' proof is easily corrected and amended to apply to this revised schema. Schemas (25) and (26) are equivalent in all but the problematic cases outlined above. If A contains a constant symbol, α, then $\Phi\alpha$ occurs on the left of (25), and this entails $\exists x.\ \Phi x$. Similarly, if A has any leading existential quantifiers, then $\exists x.\ \Phi x$ already occurs in (25). In those cases where $\exists x.\ \Phi x$ is not entailed by the left-hand-side of (25), (25) results in inconsistency. The revised schema may still take a consistent theory with no minimal models to an inconsistent circumscription (for an example, see [Davis [1980]), but so long as A has a minimal model, (26) preserves consistency.

Theorem 7.1 – Soundness

Every instance of schema (26) is true in every minimal model of the original theory. ∎

7.2. Some Properties of Domain Circumscription

In this section we consider some properties of domain circumscription. We examine their consequences with respect to using domain circumscription to formalize the domain-closure assumption. To better illustrate the properties of domain circumscription, we refer to the following example.

Example 7.1

Let $T = \{Pa, Pc, Qb, Qc\}$. T has the following minimal models. (We use the corresponding boldface letter for the interpretations of constant terms, and α, β, and γ represent the equivalence classes $\{\mathbf{a}, \mathbf{c}\}$, $\{\mathbf{b}, \mathbf{c}\}$, and $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, respectively.)

M_1: $|M_1| = \{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$

$|P|_{M_1} = \{\mathbf{a}, \mathbf{c}\}$

$|Q|_{M_1} = \{\mathbf{b}, \mathbf{c}\}$

$|{=}|_{M_1} = \{(\mathrm{a,a}), (\mathrm{b,b}), (\mathrm{c,c})\}$

M_2: $|M_2| = \{\alpha, \mathrm{b}\}$
$|P|_{M_2} = \{\alpha\}$
$|Q|_{M_2} = \{\mathrm{b}, \alpha\}$
$|{=}|_{M_2} = \{(\mathrm{a,a}), (\mathrm{b,b}), (\mathrm{c,c}), (\mathrm{a,c}), (\mathrm{c,a})\}$

M_3: $|M_3| = \{\mathrm{a}, \beta\}$
$|P|_{M_3} = \{\mathrm{a}, \beta\}$
$|Q|_{M_3} = \{\beta\}$
$|{=}|_{M_3} = \{(\mathrm{a,a}), (\mathrm{b,b}), (\mathrm{c,c}), (\mathrm{b,c}), (\mathrm{c,b})\}$

M_4: $|M_4| = \{\gamma\}$
$|P|_{M_4} = \{\gamma\}$
$|Q|_{M_4} = \{\gamma\}$
$|{=}|_{M_4} = \{(\mathrm{a,a}), (\mathrm{b,b}), (\mathrm{c,c}), (\mathrm{a,b}), (\mathrm{b,a}), (\mathrm{a,c}), (\mathrm{c,a}), (\mathrm{b,c}), (\mathrm{c,b})\}$ ∎

Several important features are evident in the above example. First, every model of T has one of M_1 - M_4 as a minimal submodel. As with other forms of circumscription and their corresponding notions of minimality, it is interesting to know whether there is a class of theories each of whose models has a minimal submodel (*i.e.*, well-founded theories). It is for such theories that domain circumscription corresponds most closely with one's intuitions. In the case of domain circumscription, the mathematical logic literature provides a sufficient condition (*c.f.* [Barwise 1977, p 62]).

Proposition 7.2 (Łoś-Tarski Theorem)

> Universal theories (possibly with function symbols) are well-founded for domain circumscription. ∎

It is also clear that theories with only finite models are well-founded.

Second, because the domain circumscription schema is satisfied by every minimal model, domain circumscription does not produce any new ground-term equalities or inequalities, for well-founded theories. (The same limitation also applies to predicate and formula circumscription.)

Theorem 7.3

If T is a well-founded theory which contains axioms which define the equality predicate, $=$, and $\alpha_1,\ldots,\alpha_n$, $\beta_1,\ldots,\beta_n$ are ground terms, then

(i) $T \vdash (\bigvee_{i=1}^{n} \alpha_i = \beta_i) \Longleftrightarrow DC(T) \vdash (\bigvee_{i=1}^{n} \alpha_i = \beta_i)$

(ii) $T \vdash (\bigvee_{i=1}^{n} \alpha_i \neq \beta_i) \Longleftrightarrow DC(T) \vdash (\bigvee_{i=1}^{n} \alpha_i \neq \beta_i)$ ■

The automatic generation of all possible ground term inequalities to capture the unique-names assumption [Reiter 1980b] remains a thorny issue in knowledge representation.

Third, the ambiguity of the usual statement of the domain-closure assumption is revealed. Only M_4 has the minimum *number* of individuals necessary to satisfy T (*i.e.*, 1), yet each of M_1 — M_4 has only individuals named (and hence required to exist) by T. Domain circumscription captures a weak sense of the domain-closure assumption which does not decide between these interpretations. Based on common applications of the domain-closure assumption (typically in conjunction with some form of unique-names assumption), this weak sense appears to be the preferred sense.

While new ground equality statements are not generally forthcoming, the results of domain circumscription do interact with the equality theory in interesting ways. The circumscription of T in Example 7.1 entails $a = b \wedge b = c \supset \exists x \forall y.\ x = y$, for example. The circumscription of $\{\exists x.\ Px,\ \ \exists x.\ Qx\}$ entails $\exists x.\ Px \wedge Qx \supset \exists x \forall y.\ x = y$. Such formulae seem to precisely capture the difference between the various minimal models of the original theory. In fact, a completeness result for domain circumscription can be obtained. This result guarantees that, for theories with only finite models (among others), the set of minimal models of the original theory constitutes exactly the set of models of the circumscribed theory. Such a precise characterization is very encouraging. The proof of this result is analogous to Perlis and Minker's [1986] finitary completeness proof for predicate and formula circumscription.

Theorem 7.4 – Finitary Completeness

If T is a finitely axiomatizable theory, and every model of T is finite, then only the minimal models of T satisfy every instance of schema (26) for T. ■

In the statement of Theorem 7.4, the requirement that all of T's models be finite is stronger than necessary. The theorem holds even if only the models which satisfy schema (26) are finite.

Corollary 7.5

If T is a finitely axiomatizable theory, and every model of $T \cup$ schema (26) is finite, then only the minimal models of T satisfy every instance of schema (26) for T. ■

7.3. Related Formalisms

McCarthy [1980] claims that domain circumscription is a special case of predicate circumscription, in that the domain circumscription schema for a theory, A, can be derived by predicate circumscription of a theory, A', which is a conservative extension of A. In view of this, it might appear that interest in domain circumscription is pointless. Apart from the fact that domain circumscription is a more direct and somewhat simpler approach to domain-closure, and that the model theory of domain circumscription perhaps better captures our intuitions about the conjectures involved, there is another reason to reject this argument for abandonment. McCarthy's demonstration of this subsumption actually rests on a strengthened form of predicate circumscription which allows axioms of the original theory to be ignored during the circumscription process. As we noted in chapter 5, this form of circumscription does not always preserve consistency, even for theories with minimal models. Ordinary predicate circumscription cannot, in general, yield the domain circumscription schema. In fact, this is fortunate, since the form of domain circumscription McCarthy was trying to emulate introduced inconsistencies into some theories with minimal models.

Our revised form of domain circumscription, which preserves consistency for minimally modelable theories, is still not obtainable using predicate circumscription. In chapter 5, we showed that predicate circumscription is too weak to conjecture domain-closure axioms. Since domain circumscription can conjecture such axioms, it follows that it is not subsumed by its predicate cousin. In chapter 6, we showed that neither formula circumscription nor generalized circumscription without variable terms subsumes domain circumscription, in general. Our semantic characterization suggests that it is unlikely that any form of generalized circumscription can conjecture domain-closure axioms. It appears, therefore, that domain circumscription continues to fill a niche among the various mechanisms for closed-world reasoning.

CHAPTER 8

Connections Between Default Logic and Circumscription

In chapter 3, we observed that the model-set semantics for default logic bears a superficial resemblance to the minimal-model semantics of the various forms of circumscription. Chapters 5 and 6 considered the feasibility of doing default reasoning using circumscription. We now consider the relationships between default logic and circumscription in more detail.

The natural question is whether either form subsumes the other. Is there a direct correspondence between default theories and circumscription, or *vice versa*?

Proposition 8.1

> Default logic can reach conclusions which cannot be obtained by generalized circumscription without variable terms. ■

Example 8.1

> The default theory $\left[\left\{\frac{: a \neq b}{a \neq b}\right\}, \{\ \}\right]$ has a unique extension, containing $a \neq b$. In chapter 6, we showed that generalized circumscription without variable terms cannot conjecture new inequalities. ■

The converse of proposition 8.1 is apparently false. Assuming that *CLOSURE*(T; X; R) is consistent, the theory

$$\left[\left\{\frac{: I}{I} \mid I \text{ is an instance of } CLOSURE(T;\ \mathrm{X};\ R)\right\},\ T\right]$$

obviously produces the required results. Perhaps this is not what one has in mind when one asks if default logic can capture circumscription, however! We will return to this question in later sections.

8.1. "Translation" from Default Logic to Circumscription

In view of proposition 8.1, the title of this section might seem paradoxical. There has been some work on partial translations, however. Grosof [1984] presents two equivalent translation schemes for normal default theories, one involving '*ab*' predicates (discussed in §2.2.2), the other involving minimizing arbitrary expressions. We discuss the former.

The translation scheme carries the first-order axioms, *W*, over unchanged. For each closed normal default, $\frac{\alpha_i : \beta_i}{\beta_i}$, the axiom $\alpha_i \wedge \neg\beta_i \supset ab(i)$ is added. Then *ab* is circumscribed in the resulting theory, varying *ab* and each predicate which occurs in any of the β_i's. Grosof observes that this "translation" actually differs from default logic in a number of respects. First, the equality predicate is not affected by the circumscriptive theory. Grosof proposes to exclude defaults about equality to remedy this, but this is insufficient. Any default which affects equality will not behave "correctly" in the circumscriptive theory. A further difference is that the circumscriptive theory inherits circumscription's "cautious" nature. The multiplicity of extensions of a default theory are reflected in disjunctive statements in the translated theory. Finally, Grosof's translation of the normal default $\frac{\alpha : \beta}{\beta}$ actually more closely corresponds to the default $\frac{: \alpha \supset \beta}{\alpha \supset \beta}$, since the translation allows the conjecture of $\neg\alpha$ from $\neg\beta$, something Grosof appears not to have noticed. Even allowing for these discrepancies, Grosof presents no more than intuitive arguments and examples in support of the correctness of the translation scheme.

Imielinski [1985] takes the complementary tack of defining a translation scheme to be adequate if the theory and its translation produce precisely the same conclusions, and furthermore the translation scheme is "modular". Modularity requires that the translation of the defaults and first-order facts must be independent.

Imielinski views the translation of a set of defaults to consist of a collection of first-order facts and a pre-order relation. Both of these must be determined from the defaults alone, without reference to the specific facts at hand. This is a desirable property, since one does not wish to have to recompute one's representation of knowledge (in addition to the necessary adjustments to the set of one's conjectures) every time a new fact is learned.

Given these strictures, Imielinski is able to prove that even normal defaults are not modularly translatable to generalized circumscription. There are *some* defaults which do have modular translations, however. These are the semi-normal defaults without prerequisites (*e.g.*, $\frac{: \alpha \wedge \beta}{\beta}$).

These results highlight the necessity of the fundamental distinction between the model-set-restriction semantics of default logic (see chapter 3) and the minimal-model semantics of circumscription. The prerequisites of the defaults are required to be provable. This is a global characteristic of the set of models. The submodel relation, however, is only able to consider pairs of models. Prerequisite-free defaults fit nicely into circumscription precisely because they are prerequisite-free. There are no (global) provability requirements, only consistency requirements. Consistency can be determined by the existence of a single model, so can be locally determined.

There remains the question of whether the requirement of identical sets of theorems is too strong. Imielinski's theorem, proving that normal default theories are not modularly translatable, rests on the fact that any modular translation of the default $\frac{A : B}{B}$, where the sets $\{A, B\}$ and $\{A, \neg B\}$ are both consistent, will necessarily yield $A \supset B$ as a theorem (assuming $W \not\vdash \neg B$). While this may be true, if an extension contains A or B, it will also contain $A \supset B$. It appears that the offending implication is offensive only in those cases where it cannot be used to deduce anything "useful". More convincingly, we have noted that default logic is a "brave" reasoner while circumscription is "cautious". It seems reasonable to expect that a circumscriptive translation of default logic would reflect this cautious nature, perhaps returning those facts true in *all* extensions. Finally, circumscriptive conjectures apply to all individuals, whereas those resulting from open defaults apply only to individuals with names in the language. It might be reasonable to expect that circumscriptive versions of default theories with open defaults would therefore prove stronger conjectures (at least for theories without domain closure axioms).

These considerations suggest that Imielinski's results might be taken as a "worst case" scenario, leaving open the possibility of acceptable translation schemes for defaults with prerequisites, given a weaker notion of "acceptable". We do not further consider this possibility here.

8.2. Translations from Circumscription to Default Logic

The other side of the coin we have been examining is whether default logic can be used to perform circumscription (in any but the trivial sense mentioned at the beginning of this chapter). The previous section outlined a number of the very different capabilities of the two formalisms: brave *vs* cautious, effects on equality, global (provability) *vs* local (consistency) comparisons in the model-theory (proof-theory), and statements about "unnamed" individuals. In all but the last of these categories, default logic came out on the stronger end. This suggests that the search for a direct implementation of circumscription in default logic might be more successful that the converse attempt. The answer to this is, "Yes, and no.". There is one facet of generalized circumscription which is completely absent from default logic. That is the ability to specify which predicates are to be allowed to vary during the circumscription process. In default logic, there is no way to restrict the repercussions of the defaults to some particular set of predicates (and/or individuals). Thus we have Theorem 8.2.

Theorem 8.2

If $T \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n$ and $T \vdash \alpha_i \neq \alpha_j$, for $i \neq j$ for ground terms $\alpha_1, ..., \alpha_n$; and $\mathbf{X}$ includes all of the predicates of L; then those formulae true in every extension of $\Delta = \left[\left\{ \frac{: \neg Px}{\neg Px} \right\}, T \right]$ are precisely those entailed by *CLOSURE*$(T; \mathbf{X}; \{P\})$. ■

Corollary 8.3

If E is an extension of Δ, then every model of E is an $(\mathrm{X},\{P\})$-minimal model of T. ■

Corollary 8.4

If M is an $(\mathrm{X},\{P\})$-minimal model of T, then M is a model for some extension of Δ. ■

Corollary 8.5

Δ captures the brave circumscription of P in T with every predicate variable. ■

Notice that Theorem 8.2 requires that T have unique-name axioms as well as domain-closure axioms. If we drop the requirement for unique-name axioms, then the default theory becomes stronger than the circumscriptive theory, in the sense that Corollary 8.3 continues to hold but Theorem 8.2 and Corollary 8.4 do not. We have not yet determined whether these results generalize to the joint minimization of several predicates. Because of the limitation of open defaults to named individuals, none of the results generalize to theories without domain-closure axioms.

Proposition 8.6

If T does not entail a domain-closure axiom, and $T \not\vdash \forall x.\ \neg Px$, then every extension for Δ has models which are not $(\mathrm{X},\ \{P\})$-minimal. ■

Even more pessimistic is the result that fixed predicates preclude such a straightforward translation of circumscription to default logic, even for closed-domain, unique-name theories.

Theorem 8.7

There are theories, T, such that $T \vdash \forall x.\ x = \alpha_1 \vee \ldots \vee x = \alpha_n$ and $T \vdash \alpha_i \neq \alpha_j$, for $i \neq j$ and yet no combination of the extensions of $\Delta = \left[\left\{\frac{:\neg Px}{\neg Px}\right\},\ T\right]$ precisely characterizes the $(\mathrm{X},\ \{P\})$-minimal models of T. ■

We experimented with an extended version of default logic which allowed for the specification of "fixed" predicates. Although we were able to show that the results in [Reiter 1980a, chapters 2 and 3] hold for this logic, and – for finite theories – the obvious generalization of the model-set restriction semantics of chapter 3 applies, we abandoned this approach when it proved incapable of yielding an analogue for Theorem 8.2 in the presence of fixed predicates. (The best that could be guaranteed was that those ground literals in P contained in all extensions were true in all minimal models. This is significantly weaker – sufficiently so that we doubt that the (abundant) extra machinery required is worthwhile.

Example 8.2

Let T be $\{\forall x.\ x = a \vee x = b,\ a \neq b,\ \neg Pa \wedge \neg Pb \supset Qa\}$ and let Q be fixed.
The P-minimal models of T are (loosely represented):

$\{\ Pa,\ \neg Pb,\ \neg Qa\}$
$\{\neg Pa,\ \ Pb,\ \neg Qa\}$
$\{\neg Pa,\ \neg Pb,\ \ Qa\}$

There are no ground literals in P true in every P-minimal model. However,

$$CLOSURE(T;\ \{P\};\ \{P\}) \vdash (\exists x.\ Px \equiv Qa) \wedge (\neg Pa \vee \neg Pb)\ .$$

In other words, one can circumscriptively conjecture that there is exactly one P if Qa, and none otherwise. ■

Gelfond and Przymusinska [1985] prove the weak result alluded to above for their version of Minker's generalized closed-world assumption, which allows fixed predicates. Gelfond, Przymusinska, and Przymusinski [1985] prove a much stronger result for their extended closed-world assumption.

Proposition 8.8 (Gelfond, Przymusinska, and Przymusinski)

A structure, M, is a model for $ECWA(T)$ iff it is a minimal model for T. ■

At first glance this might suggest that there should be some analogous result for *some* default theory. It appears that the ECWA actually achieves this power by the subterfuge discussed near the beginning of this chapter, by adding every instance of the circumscription schema. This is certainly the case in the absence of variable predicates.

Proposition 8.9

If there are no variable predicates ($Z = \{\ \}$), then $ECWA(T)$ adds to T every instance of the circumscription schema. ■

It seems that any generalized translation from circumscription to default logic (for finite theories) – if such a thing exists, short of adding defaults for each instance of the circumscription schema – requires more power than the closed-world default provides. The existence of an appropriate translation remains open.

CHAPTER 9

Open Problems

Don't confront me with my failings ...
I have not forgotten them.

– Jackson Browne

Throughout the thesis, a catalogue of open problems has been compiled. Rather than recapitulate this list of specific problems, this chapter addresses a broader, philosophical perspective. We consider a general research programme, instead of a litany of isolated potholes in need of filling.

Although there has been considerable activity in the area of non-monotonic reasoning, along with some remarkable successes, very little attention has been focussed on the dynamics of non-monotonicity. As this promises to be a particularly fruitful avenue of investigation, this chapter addresses two aspects of this problem: how new information is assimilated into a theory involving assumptions, and how non-monotonic inference rules are acquired and employed.

These two areas are intimately related. A major goal for future research should be to develop a unifying framework which makes their interrelationships more apparent. This point of view may be expected to provide new insights into both non-monotonic reasoning and updates. Furthermore, much of the work that has been done treating these problems in isolation can, hopefully, be reinterpreted to advantage from this more general standpoint.

9.1. Principles of Non-Monotonic Reasoning

The important issue of non-monotonicity which remains unaddressed is not primarily how conclusions are obtained given some facts and some non-monotonic inference rules. Rather, the question is how non-monotonic rules are formulated, determined to be applicable, and applied. This question can be illustrated by considering the circumscriptive examples of §2.1.5.2. Given a representation of the facts about the "world", certain predicates must be circumscribed, other predicates specified as variable, appropriate substitutions discovered, and then the required conjectures are obtained. As much of the problem lies in these "ancillary" tasks of deciding what and how to circumscribe as in the closed-world reasoning achieved by actually performing the circumscription. To date, most of the work in non-monotonic reasoning (including this thesis) has focussed more on developing mechanisms for performing certain specialized reasoning tasks than on underlying principles or even an understanding of when and how to employ the mechanisms once they are developed.

The central question is: can we discover ways to make non-monotonic reasoning automatic and/or goal-directed? *I.e.*, are there features of particular problems which can guide the completion of an incomplete knowledge-base, without external intervention, to solve those problems? A first approximation to a theory of non-monotonic theory construction was outlined by Reiter [1978a]. He explained non-monotonic reasoning in terms of the closed-world assumption. Reiter's idea was that reasoners might assume their knowledge about relevant aspects of the situation to be complete. Closed-world reasoning sanctions exactly those conclusions true in a world *completely characterized* by what is known.

Such a clear, simple, uniform characterization of non-monotonic reasoning appeals to introspective intuitions about the simplicity and naturalness of commonsense reasoning. Unfortunately, it proved simplistic as well as simple. Not every knowledge state uniquely characterizes a state of the world. Assuming the real world is that world characterized by what is known is a dubious step when no world is so characterized! Research since 1978 has focussed on mechanisms which avoid the shortcomings of the naive interpretations of the CWA. Little effort has been directed to finding a corresponding intuitive explication of the underlying principles.

The minimal-model semantics which we have discussed in one form or another throughout this thesis does not qualify as the intuitive explication we seek, for two reasons. The first – and perhaps less compelling – is that not all theories have minimal models, and it is undecidable whether a particular theory has a minimal model. Certain theories – quite unexpectedly – turn out not to have minimal models. For example, we have shown that the theory:

$$\exists x.\ Nx \land \forall y.\ Ny \supset x \neq succ(y)$$
$$\forall x.\ Nx \supset Nsucc(x)$$
$$\forall xy.\ succ(x) = succ(y) \supset x = y\ ,$$

has no minimal models. This is because any model has a chain of N's isomorphic to the natural numbers, N. But this chain has a subchain, also isomorphic to N, which satisfies the axioms. Hence every model has a proper submodel, and there are no minimal models. But, since every model contains a segment isomorphic to N, and since there are models exactly isomorphic to N, surely commonsense dictates that N is an acceptable minimal model? Minimum-model semantics force the minimization process to go beyond the bounds of commonsense in this case.

More tellingly, minimal-model semantics enter the picture after much of the non-monotonic reasoning process is complete. Only after it has been decided what expression is to be mimimized, and the connections between the minimized expression and the rest of the world have been determined so that variable predicates can be chosen, can the semantic characterization tell us what world(s) the non-monotonic theory characterizes. The semantics sheds no light on these other dimensions of the commonsense reasoning process. Hence, it is not the characterization we seek.

What evidence is there that there is ***any*** underlying principle? Might not the difficulty in finding such a principle stem, in part, from its non-existence? Of course, the only guarantee that the principle we seek exists will be its demonstration. There is evidence which *suggests* that some sort of uniform rules might underlie commonsense reasoning. One indication is the existence of approximations which fill the role of the sought-after rule in limited cases. The CWA is one such rule. Others include minimal-model semantics (for theories with minimal models), the model-set-restriction semantics for default logic, and the inferential distance concept in semantic network reasoning systems. A final example is "Occam's Razor", a hypothesis-ranking rule which suggests that the simplest explanation for any phenomenon is the best.

Of course, there may be no uniform underlying principles. So be it. That is, in itself, interesting. Besides, if humans use no uniform procedures at all, we can still hope to uncover heuristics which can help guide the task of commonsense reasoning. For example, even a way to automatically determine, for some class of theories, which expressions to circumscribe and/or which predicates to vary based on the goal at hand and the current knowledge state would be a significant contribution.

9.2. Update

The problems of updating theories with information inconsistent with their current state are obviously problems of non-monotonic inference: such new facts must force the retraction of previously accepted facts if consistency is to be preserved. A second major open problem is to develop a view of updates which integrates them with other forms of non-monotonic reasoning. Instead of blind addition and deletion – which obviously will not work – or the proliferation of alternate theories – which increases uncertainty – it seems appropriate to view updates as new information which leads to the *reasoned* assertion or retraction of facts.

The exact form that this research might take is unclear. The final result will likely be heavily influenced by work in five areas:

1) Relevance Logic [Anderson and Belnap 1975]: in Relevance Logic, contradictions do not automatically lead to chaos. The repercussions of the various facts in an inconsistent theory can be explored without introducing "artifacts" of the inconsistency. This seems like an ideal environment for investigating the effects of contrary updates.

2) Counterfactuals and Hypotheticals [Rescher 1964, 1976; Lewis 1973]: These branches of philosophy deal with what would be true in a world which differs from the real world in that (at least) certain specified facts hold. The update problem can easily be construed in these terms. One might therefore expect this work to shed light on updating.

3) Change-recording, correcting, and knowledge-adding updates: Wilkins [1983] and Keller [& Wilkins 1984a, b] distinguish different kinds of updates depending on whether the update expresses a change in the state of the world, an error in the database, or simply new knowledge. In a database with incomplete information, an update can be expected to have different semantics depending on to which of these categories it belongs.

4) Non-monotonic reasoning systems appear to provide useful theoretical tools for examining the repercussions of updates. Updates contrary to what was inferred by default can be made to automatically exclude these offending defaults after the update. Reiter [1980a] has considered updates to default theories in limited circumstances. He shows that certain classes of updates are knowledge-conserving; they do not force the rejection of any conclusions.

5) Belief Revision Systems: The assumption-based approach to belief revision [Martins 1983, de Kleer 1984] provides an attractive book-keeping system for dealing with straightforward repercussions of changing sets of assumptions. Reiter and Grosof [1985, personal communications]

have each worked on formalizing these systems in default logic.

Non-monotonic reasoning and update are intimately connected: non-monotonic reasoning is non-monotonic precisely because of its behaviour when confronted by updates. In fact, it is possible to view what we have been calling non-monotonic reasoning as a monotonic, valid, form of inference. Any update which forces assumptions to be retracted can be construed as contrary to the original knowledge-base (*i.e.*, assumptions are viewed as entailed by the knowledge-base under a modified entailment relation [Israel 1980, Nutter 1983].) Under this view, non-monotonicity becomes strictly a problem of dealing with contrary updates.

The problem of updates is also important within the context of non-monotonic reasoning. Given a system for drawing non-monotonic inferences, one is faced with the problem of adapting to new information. Even updates which do not represent a change in the state of the world are problematic when non-monotonicity is involved. The obvious problem is that contrary information may have been previously inferred by default. In such cases, the conflict can perhaps be detected. The default inference can then simply be revoked (if the system remembers its default genesis) or various consistency restoration techniques can be applied to reject some set of "offending" beliefs.

The update problem in non-monotonic theories is compounded by the fact that inferences may have been based on the absence of what is now being asserted. In such circumstances, there may be no inconsistencies to signal the necessity of belief revision. Unless the assumptions underlying facts in the knowledge-base can be examined for compatability with updates in the same way that the facts themselves are, nothing can prevent the knowledge-base from being "catapulted" into self-supporting - but otherwise unjustifed - belief sets. For example, the default theory:

$$D = \left\{ \frac{: P \wedge \neg Q}{P} \right\}, \qquad W = \left\{ P \equiv R \right\}$$

leads to the beliefs P and R. Unless care is taken, belief in R may support belief in P after Q is asserted, even though R was originally inferred because of a lack of belief in Q. Work on truth-maintenance systems [Doyle 1979; Doyle and London 1980] has shed some light on these problems.

In a related vein, there are issues of how knowledge representation languages should be designed to address these issues. Work on both database theory and non-monotonicity has tended to deal with tenseless languages, viewing the knowledge-base as a snap-shot of some state-of-affairs. Update is seen as an atomic process of transforming from one snap-shot to the next, with the state of the knowledge-base defined only before and after - not during - the update. Other work in AI has embraced time - either reservedly, by adopting "situations" or "states" and "fluents" which transform the world from one state to another [McCarthy & Hayes 1969; Moore 1979], or wholeheartedly, by adopting a full-blown temporal logic [McDermott 1981; Allen 1984], or somewhere in between. Perhaps the best way to deal with non-monotonicity is monotonically, by representing the state of an agent's beliefs at a particular time.

CHAPTER 10

Conclusions

> I don't understand it. I don't even understand
> the *people* who understand it.
>
> — Queen Juliana of The Netherlands

10.1. Default Logic and Inheritance

We presented a correspondence between default theories and inheritance networks with exceptions, analogous to that outlined by Hayes [1977] between first-order theories and exception-free inheritance networks. This correspondence allowed us to specify minimum correctness criteria for any inheritance-determining algorithm, identifying the notion of correct inference with that of derivability within a single extension of the corresponding default theory. These criteria show that proposed parallel marker-passing implementations of inheritance networks with exceptions are not feasible for general theories. Correct behaviour would require that severe (and difficult to define) constraints be placed on the structure of the inheritance networks they could represent and reason with.

Given a notion of correct inference, it became possible to question whether inheritance networks with exceptions are always coherent, in the sense of always representing a reasonable set of beliefs. Inheritance graphs are typically acyclic. We showed that acyclic networks are coherent and, in fact, that weaker criteria are sufficient to ensure coherence. This led to a generalization of the notion of acyclicity which can be applied to default theories, called "orderedness". The ordered theories constitute a natural class of theories all of which have at least one extension. We provided an inference algorithm for ordered inheritance networks with exceptions which is provably correct with respect to this concept of derivability.

Our formulation suggests that it may not be possible to correctly realize massively parallel marker-passing hardware of the kind envisaged by NETL which is applicable to arbitrary inheritance graphs. It appears that the best that can be achieved for such networks is a restricted, quasi-parallel inference algorithm. We have sketched such an algorithm, but have shown that not every set of conclusions justified by the network is accessible to it. It remains to be seen whether the limitations imposed by the algorithm are acceptable. Fortunately, these pessimistic observations do not preclude parallel architectures for suitably restricted networks. We have shown that Touretzky's inferential distance algorithm produces correct conclusions. Touretzky shows how to restrict a network so that parallel marker-passing produces the same conclusions as the inferential distance algorithm. We conclude that, for such restricted networks, parallel marker-passing is correct.

We have shown default logic to be a useful tool for formalizing the reasoning processes involved in AI systems. Such a specification provides a method for evaluating correctness and a metric by which various approaches can be measured and compared. A default logic specification of a system can provide both a more complete visualization of how the system performs and a guarantee that that performance is coherent. To facilitate such applications, we have presented a number of results on default logic. These include a semantics for arbitrary single-justification default theories, a characterization of a large class of theories for which coherent reasoning is always possible (*i.e.*, theories which always have at least one extension), and a totally correct inference algorithm for a subclass of these theories.

It might be — and has been — argued that a declarative formalism such as default logic is inadequate for the tasks of knowledge representation and reasoning. While we clearly disagree with this position, we expect default logic to be useful even to "proceduralists". Even if some system were fundamentally more than the sum of its declarative content, default logic could be used to formalize that declarative content. The non-declarative "control" information could then be treated as an inference algorithm for the resulting default theory. The correctness of the system would be determined by whether this inference algorithm was correct with respect to the proof theory of default logic.

Defaults, in one form or another, are extremely common in AI. Reiter [1978b, 1980a] discusses a wide variety of common situations to which they can be applied, including several AI knowledge representation schemes. Many of these may be amenable to analysis using an approach similar to that which we have used for inheritance networks. If some are not, two possibilities arise: the features not so amenable may prove incorrect or inessential, or they may point out shortcomings of default logic. Either result would raise interesting questions.

10.2. Predicate Circumscription

Although a model-theory for predicate circumscription has been available since 1980, together with an attendant soundness result, very little was known about the strengths and weaknesses of predicate circumscription until recently. We explored the constraints imposed by circumscription's model-theory and were surprised to find them very rigid indeed. Previous expectations for predicate circumscription had been very high; examples in the literature had pushed the technique beyond the safety of its semantic justifications, and this fact had gone unnoticed.

Predicate circumscription (and formula circumscription) can lead to inconsistent conjectures when applied to theories without minimal models. In retrospect, this is not surprising, but it does not appear to have occurred to anyone until we discovered an example. This is perhaps attributable to the schematic nature of predicate circumscription. Not every substitution produces inconsistency, so unless an inconsistent substitution is discovered, circumscription of theories without minimal models may appear simply ineffectual. The existence of theories with inconsistent circumscriptions suggests that one must be careful to circumscribe only those theories with minimal models. Alas, it is undecidable which theories have minimal models.

We have characterized a class of theories, which we call *well-founded*, which always have minimal models. We then explored the properties of predicate circumscription *vis-à-vis* these well-founded theories. We discovered that the semantic characterization of predicate circumscription – so intuitively appealing on the surface – rigidly constrained the effectiveness of circumscription in conjecturing new ground facts. The only ground facts which predicate circumscription can conjecture are negative instances of one of the predicates being circumscribed – and then only insofar as such conjectures provide no new information about the extensions of non-circumscribed predicates. Furthermore, the equality predicate is somehow resistant to predicate (and formula) circumscription.

10.3. Generalizations of Circumscription

The success of our model-theoretic investigations into predicate circumscription (pessimistic though the results were) suggested that a similar exploration of the various generalized forms of circumscription might also prove worthwhile. McCarthy [1986] did not provide a model-theory for formula circumscription, however. The first task for this investigation, thus, was to develop a model-theory. The model-theory presented is a generalization of that of predicate circumscription, with appropriate changes to accomodate the introduction of variable predicates. The minimization of expressions, rather than predicates, also forces modifications to the definitions of submodel and minimal model. The soundness (and, for certain classes of theories, completeness) of formula circumscription with respect to this model-theory has been proven.

Universal theories always have minimal models regardless of the predicates varied or minimized. For these theories, the consistency of generalized circumscription is assured. In fact, the proof shows that every model of a universal theory has at least one minimal submodel. As a corollary of this, generalized circumscription of universal theories does not affect the extensions of any predicates not designated as variable. For such theories, the repercussions of circumscription do not extend beyond those predicates explicitly indicated as liable to change.

Lifschitz [1984, 1985a,b] has developed extensions to circumscription allowing constants and functions to be treated as variables during the minimization process, and allowing arbitrary pre-orders to be specified; minimization proceeds according to this pre-order. Suitable modifications to the generalized circumscription model-theory, which accommodate these extensions, were presented. Lifschitz' innovations were shown to be sound with respect to this model theory. We examined the effects of some of these formulations on the existence of minimal models, on consistency, and on the types of conjectures which can be obtained.

10.4. Domain Circumscription

McCarthy [1980] claims that domain circumscription is a special case of predicate circumscription. We showed that the demonstration actually rests on a strengthened form of predicate circumscription which does not always preserve consistency, even for theories with minimal models. We showed that none of predicate circumscription, formula circumscription, or generalized circumscription without variable terms supercedes domain circumscription, in general. We conjectured that even variable terms are unlikely to suffice to make generalized circumscription subsume domain circumscription.

In fact, the domain circumscription schema presented by McCarthy [1980] and Davis [1980] is also too strong. Certain theories with minimal models turn out to have inconsistent domain circumscriptions. After isolating the problem, we outlined a straightforward correction which preserves the appealing semantic characterization presented by Davis [1980], and proved its correctness.

We have also noted the ambiguity of the domain-closure assumption, as it is usually stated. We argue that the most common disambiguation agrees with the results obtained from domain circumscription. Also, we conjectured that the completeness of domain circumscription for certain classes of theories might be provable.

10.5. Relations Between Circumscription and Default Logic

We have considered the relationship between default logic and circumscription. We showed that, in some cases, the closed-world default coincides with circumscription; that, in a particularly useless way, default logic subsumes circumscription; and that default logic is capable of affecting the equality theory while predicate, formula, and domain circumscription are not.

We showed that the introduction of fixed predicates and applications to open domains each provide circumscription with capabilities not available using simple closed-world default theories.

Finally, we used semantic comparisons to highlight a number of the essential differences between the two approaches. This allowed us to suggest that some of the work on translations between the two formalisms may not have noticed the essential characteristics which should be carefully considered in determining adequacy conditions for translations.

References

Allen, J.F. [1984], "Towards a General Theory of Action and Time", *Artificial Intelligence 23 (2)*, North-Holland, July, 1984, pp 123-154.

Anderson, A.R., and Belnap, N.D. [1975], *Entailment: The Logic of Relevance and Necessity*, Princeton University Press, Princeton, NJ, 1975.

Barwise, Jon (ed) [1977], *Handbook of Mathematical Logic*, North-Holland, New York, 1977.

Barwise, Jon, and Perry, John [1983], *Situations and Attitudes*, MIT Press, Cambridge, MA, 1983.

Beth, E.W. [1953], "On Padoa's method in the theory of definitions", *Indag. Math 15*, 1953, pp 330-339.

Bobrow, D.G. and Winograd, T. [1977], "An Overview of KRL-0, a Knowledge Representation Language", *Cognitive Science 1(1)*.

Bossu, G. and Siegel, P. [1985], "Saturation, Non-Monotonic Reasoning, and the Closed-World Assumption", *Artificial Intelligence 25(1)*, North-Holland, pp 13-63, 1985.

Brachman, R. [1982], "What 'IS-A' Is and Isn't", *Proc. Canadian Soc. for Computational Studies of Intelligence-82*, Saskatoon, Sask., May 17-19, pp 212-220.

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

Cottrell, G.W. [1985]. "Parallelism in Inheritance Hierarchies with Exceptions", *Proc. Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, Aug. 18-23, 1985, pp 194-202.

Davis, M. [1980], "The Mathematics of Non-Monotonic Reasoning", *Artificial Intelligence 13*, North-Holland, pp 73-80.

de Kleer, J. [1984], "Choices Without Backtracking", *Proc. Amer. Assoc. for Artificial Intelligence-84*, Austin, TX, pp 79-85.

Doyle, J. [1979], *A Truth Maintenance System*, AI Memo 521, MIT Artificial Intelligence Laboratory, Cambridge, Mass.

Doyle, J. [1982a], "Some theories of reasoned assumptions: An essay in rational psychology", Technical Report, Carnegie-Mellon University.

Doyle, J. [1982b], *The Ins and Outs of Reason Maintenance*, Technical Report, Carnegie-Mellon University, 1982.

Doyle, J. [1984], "Circumscription and Implicit Definability", Preprint, *Amer. Assoc. for Artificial Intelligence Workshop on Non-Monotonic Reasoning*, New Paltz, NY, October 1984.

Doyle, J., and London, P. [1980], *A Selected Descriptor-Indexed Bibliography to the Literature on Belief Revision,* AI Memo 568, MIT, Cambridge, Mass.

Etherington, D.W. [1982], *Finite Default Theories*, M.Sc. Thesis, Dept. Computer Science, University of British Columbia.

Etherington, D.W. [1983], *Formalizing Non-Monotonic Reasoning Systems*, Technical Report 83-1, Department of Computer Science, University of British Columbia, 1983.
(To appear in *Artificial Intelligence*).

Etherington, D.W. and Mercer, R.E. [1986], "Domain Circumscription Revisited", *Proc. Canadian Society for Computational Studies of Intelligence-86*, Montreal, May 1986.

Etherington, D.W., Mercer, R.E., and Reiter, R. [1985], "On the adequacy of predicate circumscription for closed-world reasoning", *Computational Intelligence 1(1)*, February 1985.

Etherington, D.W., and Reiter, R. [1983], "On Inheritance Hierarchies With Exceptions", *Proc. American Assoc. for Artificial Intelligence-83*, Washington, D.C., August 24-26, pp 104-108.

Fagin, R., Ullman, J.D., and Vardi, M.Y. [1983], "On the Semantics of Updates in Databases", *Proc. of the Second ACM SIGACT-SIGMOD Symp. on Principles of Database Systems,* Atlanta, GA, pp 352-365.

Fahlman, S.E. [1979], *NETL: A System for Representing and Using Real-World Knowledge*, MIT Press, Cambridge, Mass.

Fahlman, S.E. [1982], "Three Flavors of Parallelism", *Proc. Canadian Soc. for Computational Studies of Intelligence-82,* Saskatoon, Sask., May 17-19, pp 230-235.

Fahlman, S.E., Touretzky, D.S., and van Roggen, W. [1981], "Cancellation in a Parallel Semantic Network", *Proc. Seventh International Joint Conference on Artificial Intelligence*, Vancouver, B.C., Aug. 24-28, pp 257-263.

Gallaire, H. and Minker, J. (eds) [1978], *Logic and Data Bases*, Plenum Press, 1978.

Gelfond, M., and Przymusinska, H. [1985], *Negation as Failure: Careful Closure Procedure*, University of Texas at El Paso, unpublished draft, 1985.

Genesereth, M. and Nilsson, N. [1987], draft of chapter entitled "Nonmonotonic Reasoning", To appear in *Fundamentals of Artificial Intelligence*, Morgan Kaufmann Publishers, Los Altos, CA, forthcoming 1987.

Gelfond, M., Przymusinska, H., and Przymusinski, T. [1985], *The Extended Closed-World Assumption and its Relation to Parallel Circumscription*, University of Texas at El Paso, unpublished draft, 1985.

Goodman, Nelson [1955], *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, MA, 1955.

Grice, H.P. [1975], "Logic and Conversation", in *Syntax and Semantics, Vol 3: Speech Acts*, P. Cole and J.L. Morgan (eds.), Academic Press.

Grosof, B. [1984], "Default Reasoning as Circumscription", Technical Report, Stanford University, Stanford, CA.

Hayes, P.J. [1973], "The Frame Problem and Related Problems in Artificial Intelligence", in *Artificial and Human Thinking*, A. Elithorn and D. Jones (eds.), Jossey-Bass Inc., San Francisco.

Hayes, P.J. [1977], "In Defense of Logic", *Proc. Fifth International Joint Conference on Artificial Intelligence.* Cambridge, Mass., pp 559-565.

Hewitt, C. [1972], *Description and Theoretical Analysis (Using Schemata) of PLANNER: A Language for Proving Theorems and Manipulating Models in a Robot,* AI Memo 251, MIT Project MAC, Cambridge, Mass.

Hughes, G.E. and Cresswel, M.J. [1972], *An Introduction to Modal Logic*, Methuen and Co. Ltd., London.

Imielinski, T. [1985], "Results on Translating Defaults to Circumscription", *Proc. Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, Aug. 18-23, 1985, pp 114-120.

Israel, D.J. [1980], "What's wrong with Non-Monotonic Logic", *Proc. Amer. Assoc. for Artificial Intelligence-80*, 1980.

Kaplan, S.J. and Davidson, J. [1981], "Interpreting natural language database updates", *Proc. 19th Ann. Meeting of the Assoc. for Computational Linguistics*, Stanford, 1981, pp 139-142.

Keller, A.M. and Wilkins, M.W. [1984a], "Approaches for Updating Databases with Incomplete Information and Nulls", *Proc. IEEE Computer Data Engineering Conference*, April 1984, Los Angeles.

Keller, A.M. and Wilkins, M.W. [1984b], "On the Use of an Extended Relational Model to Handle Changing Incomplete Information", *IEEE Transactions on Software Engineering*, 1984.

Kramosil, I. [1975], "A Note on Deduction Rules with Negative Premises", *Proc. Fourth International Joint Conference on Artificial Intelligence.*

Levesque, H.J. [1982], *A Formal Treatment of Incomplete Knowledge*, Technical Report 3, Fairchild Laboratory for Artificial Intelligence Research, Menlo Park, CA.

Levesque, H.J. [1984], "Foundations of a Functional Approach to Knowledge Representation", *Artificial Intelligence 23(2)*, July, 1984, North-Holland, pp 155-212.

Lewis, D. [1973], *Counterfactuals*, Harvard University Press, 1973.

Lifschitz, V. [1984], "Some Results on Circumscription", Technical Report, Stanford University, Stanford, CA.

Lifschitz, V. [1985a], "Prioritized Circumscription and Separable Formulas", Technical Report, Stanford University, Stanford, CA.

Lifschitz, V. [1985b], "On the Satisfiability of Circumscription", Technical Report, Stanford University, Stanford, CA.

Linsky, L. (ed) [1971], *Reference and Modality*, Oxford University Press, London, 1971.

Lipsky, W. Jr. [1979], "On Semantic Issues Connected with Incomplete Information Databases", *ACM Transactions on Database Systems 4 (3)*, Sept. 1979, pp 262-296.

Łukaszewicz, W. [1984], "Non-monotonic logic for default theories", *Proc. Sixth European Conference on Artificial Intelligence*, 1985, pp 165-193.

Łukaszewicz, W. [1985], "Two results on default logic", *Proc. Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, CA, Aug. 18-23, 1985, pp 459-461.

Martins, J. [1983], *Reasoning in Multiple Belief Spaces*, PhD thesis, SUNY at Buffalo, Computer Science Technical Report 203.

McAllester, D.A. [1978], *A Three-Valued Truth Maintenance System*, AI Memo 473, MIT, Cambridge, Mass.

McAllester, D.A. [1980], *An Outlook On Truth Maintenance*, AI Memo 551, MIT, Cambridge, Mass.

McCarthy, J. [1977], "Epistemological Problems of Artificial Intelligence", *Proc. Fifth International Joint Conference on Artificial Intelligence.* 1977, pp 1038-1044.

McCarthy, J. [1980], "Circumscription – a Form of Non-Monotonic Reasoning", *Artificial Intelligence 13(1,2)*, North-Holland, 1980, pp 27-39.

McCarthy, J. [1986], "Applications of Circumscription to Formalizing Commonsense Knowledge", *Artificial Intelligence 28*, 1986, pp 89-116.

McCarthy, J. and Hayes, P.J. [1969], "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in *Machine Intelligence 4*, B. Meltzer and D. Michie (eds.), Edinburgh University Press, Edinburgh.

McDermott, D. [1981], "A Temporal Logic for Reasoning About Processes and Plans", Research Report 196, Computer Science Department, Yale University, New Haven, CT, 1981, (also in *Cognitive Science 6(2)*, 1982).

McDermott, D. [1982], "Non-Monotonic Logic II", *J.ACM 29(1)*.

McDermott, D., and Doyle, J. [1980], "Non-Monotonic Logic I", *Artificial Intelligence 13(1,2)*, (April 1980), North-Holland, pp 41-72.

Mendelson, E. [1964], *Introduction to Mathematical Logic*, Van Nostrand Reinhold, New York.

Minker, J. [1982], "On Indefinite Databases and the Closed-World Assumption", *Proc. Sixth Conf. on Automated Deduction*, New York, 7-9 June, 1982, Springer-Verlag, NY.

Minker, J. and Perlis, D. [1983], *On the Semantics of Circumscription*, Technical Report, University of Maryland.

Minker, J. and Perlis, D. [1984a], *Circumscription: Finitary Completeness Results*, Unpublished Draft, University of Maryland.

Minker, J. and Perlis, D. [1984b], "Applications of Protected Circumscription", *Lecture Notes in Computer Science 170: Proc. of the Seventh Conference on Automated Deduction*, pp 414-425, Springer.

Minsky, M. [1975], "A Framework for Representing Knowledge", in *The Psychology of Computer Vision*, P. Winston (ed.), McGraw-Hill, New York.

Moore, R. [1983a], "Semantical considerations on non-monotonic logic", *Proc. Eighth International Joint Conference on Artificial Intelligence*. Karlsruhe, West Germany, pp 272-279.

Moore, R. [1983b], *Semantical considerations on non-monotonic logic*, SRI Technical Note 284, Menlo Park, CA.

Moore, R.C. [1984], "Possible-World Semantics for Autoepistemic Logic", Technical Report, SRI, Menlo Park, CA.

Nutter, J.T. [1983], *Default reasoning in AI systems*, MSc Thesis, SUNY at Buffalo, Computer Science Technical Report 204.

Perlis, D. and Minker, J. [1986], "Completeness Results for Circumscription", *Artificial Intelligence 28*, 1986, pp 29-42.

Przymusinski, T. [1985], *Minimal-Model Resolution and Query Answering in Circumscriptive and Closed-World Theories*, University of Texas at El Paso, unpublished draft, 1985.

Quillian, M.R. [1968], "Semantic Memory", in M. Minsky (ed), *Semantic Information Processing*, MIT Press, Cambridge, Mass.

Reiter, R. [1978a], "On Closed-World Data Bases", in [Gallaire and Minker 78], 55-76.

Reiter, R. [1978b], "On Reasoning by Default", *Proc. Second Symposium on Theoretical Issues in Natural Language Processing*, Urbana, Illinois, 1978, 210-218.

Reiter, R. [1980a], "A Logic for Default Reasoning", *Artificial Intelligence 13*, (April 1980), North-Holland, pp 81-132.

Reiter, R. [1980b], "Equality and Domain Closure in First-Order Data Bases", *JACM 27(2)*, 1980, 235-249.

Reiter, R. [1982], "Circumscription Implies Predicate Completion (Sometimes)", *Proc. Amer. Assoc. for Artificial Intelligence-82,* 1982, 418-420.

Reiter, R. [1984], "Towards a Logical Reconstruction of Relational Database Theory", in M.L. Brodie, J. Mylopoulous, and J.W. Schmidt (eds): *On Conceptual Modelling,* Springer-Verlag, New York, pp 191-233.

Reiter, R., and Criscuolo, G. [1983], "Some Representational Issues in Default Reasoning", *Int. J. Computers and Mathematics 9 (1),* (Special Issue on Computational Linguistics), 1983, pp 1-13.

Rescher, N. [1964], *Hypothetical Reasoning,* North-Holland, Amsterdam, 1964.

Rescher, N. [1976], *Plausible Inference,* Van Gorcum, Assen, The Netherlands, 1976.

Roussel, P. [1975], *PROLOG, Manuel de Reference et d'Utilisation,* Group d'Intelligence Artificielle, U.E.R. de Marseille, France.

Sandewall, E. [1972], "An Approach to the Frame Problem and Its Implementation", in *Machine Intelligence 7,* B. Meltzer and D. Michie (eds.), Edinburgh University Press, Edinburgh.

Schubert, L.K. [1976], "Extending the Expressive Power of Semantic Networks", *Artificial Intelligence 7(2),* North-Holland, pp 163-198.

Shepherdson, J.C. [1984], "Negation as Failure: A Comparison of Clark's Completed Data Base and Reiter's Closed-World Assumption", *Journal of Logic Programming 1(1),* June 1984, 51-79.

Stalnaker, R. [1980], "A note on non-monotonic logic", Unpublished note, Dept of Philosophy, Cornell University.

Touretzky, D.S. [1982], "Exceptions in an Inheritance Hierarchy", Unpublished Manuscript, Department of Computer Science, Carnegie-Mellon University.

Touretzky, D.S. [1983], "Multiple Inheritance and Exceptions", Unpublished Manuscript, Department of Computer Science, Carnegie-Mellon University.

Touretzky, D.S. [1984a], *The Mathematics of Inheritance Systems,* PhD Thesis, Dept of Computer Science, Carnegie-Mellon University.

Touretzky, D. [1984b], "Implicit ordering of defaults in inheritance systems", *Proc. Amer. Assoc. for Artificial Intelligence-84.*

Touretzky, D. [1985], "Inheritable relations: a logical extension to inheritance hierarchies", *Proc. Theoretical Approaches to Natural Language Understanding,* Halifax, 28-30 May 1985, pp 55-60.

Wilkins, M.W. [1983], "Partial Information and Alternative Sets", Technical Report, Stanford University, Stanford, CA.

Winograd, T. [1980], "Extended Inference Modes in Reasoning", *Artificial Intelligence 13(1,2),* (April 1980), North-Holland, pp 5-26.

Woods, W.A. [1975], "What's In A Link?", *Representation and Understanding,* Academic Press, pp 35-82.

APPENDIX A

Proofs of Theorems

Background Information

There are a few definitions and results due to Reiter [1980a] on which we draw freely in the following proofs. We reproduce them here for the reader's convenience.

1) Theorem 0.1 [Reiter 1980a, Theorem 2.1]

E is an extension for $\Delta = (D, W)$ if and only if $E = \bigcup_{i=0}^{\infty} E_i$, where

$E_0 = W$, and for $i>0$

$$E_{i+1} = Th(E_i) \cup \{\omega \mid \frac{\alpha : \beta}{\omega} \in D, \alpha \in E_i \text{, and } \neg\beta \notin E\}^1$$

2) The *Generating Defaults* for E with respect to Δ are defined as:

$$GD(E,\Delta) = \{\frac{\alpha : \beta}{\omega} \in D \mid \alpha \in E, \neg\beta \notin E\}$$

3) If D is a set of defaults, then *CONSEQUENTS* (D) is defined, as one would expect, as:

$$\mathit{CONSEQUENTS}(D) = \{\omega \mid \frac{\alpha : \beta}{\omega} \in D\}$$

4) Theorem 0.2 [Reiter 1980a, Theorem 2.5]

If E is an extension for $\Delta = (D, W)$, then

$$E = Th(W \cup \mathit{CONSEQUENTS}(GD(E,\Delta))).$$

5) Theorem 0.3 [Reiter 1980a, Corollary 2.2]

If E is an extension for $\Delta = (D, W)$, then E is consistent if and only if W is.

In the proofs of results from chapters 3 and 4, we will usually assume that formulae are in clausal form: *i.e.*, expressed as a conjunction of disjunctions of literals. We define the functions *CLAUSES* $(\cdot)$ and *LITERALS* $(\cdot)$ as follows:

If $\beta = (\beta_{1,1} \vee ... \vee \beta_{1,m_1}) \wedge ... \wedge (\beta_{m,1} \vee ... \vee \beta_{m,m_m})$ then

$\mathit{CLAUSES}(\beta) = \{(\beta_{i,1} \vee ... \vee \beta_{i,m_i}) \mid 1 \leq i \leq m\}$

$\mathit{LITERALS}(\beta) = \{\beta_{i,j} \mid 1 \leq i \leq m, 1 \leq j \leq m_i\}$

Abusing the notation somewhat we sometimes use *CLAUSES* (Γ), where Γ is a set of formulae, to

[1] Note the explicit reference to E in the definition of E_{i+1}.

refer to $\bigcup_{\gamma \in \Gamma} \mathit{CLAUSES}(\gamma)$.

We will define other notation as it is required.

Definition: Satisfiability, admissibility, and applicability

Let X be a set of models; Γ a set of formulae; α, β, and ω formulae, and $\delta = \frac{\alpha : \beta}{\omega}$ a default. Then

i) α is *X-satisfiable* (*X-valid*) iff $\exists x \in X.\ x \models \alpha \qquad (\forall x \in X.\ x \models \alpha)$

ii) Γ is *X-admissible* (*X permits* Γ) iff $\forall \gamma \in \Gamma.\ \exists x \in X.\ x \models \gamma$

iii) δ is *X-applicable* iff α is X valid and β is X-satisfiable. ■

Definition: Result of a default

Let X, Γ, and δ be as above. Then the *result* of δ in (X, Γ) is:

$$\delta(X, \Gamma) = \begin{cases} (X, \Gamma) \text{ if } \delta \text{ is not } X\text{-applicable and } \Gamma \text{ is } X\text{-admissible,} \\ ((X - \{N \mid N \models \neg\omega\}),\ (\Gamma \cup \{\beta\})) \text{ if } \delta \text{ is } X\text{-applicable and } \Gamma \text{ is } X\text{-admissible, and} \\ \bot \text{ otherwise.} \end{cases}$$ ■

Definition: Result of a sequence of defaults

Let X and Γ be as above, and let $<\delta_i>$ be a sequence of defaults. Then

$$<\delta_i>(X, \Gamma) = (\cap X_i,\ \cup \Gamma_i) \text{ where } \begin{cases} X_0 = X; \quad \Gamma_0 = \Gamma; \quad \text{and} \\ (X_{i+1}, \Gamma_{i+1}) = \delta_i(X_i, \Gamma_i), \quad i \geq 0\,. \end{cases}$$ ■

Definition: Stability

Let Y be a non-empty set of models, Γ a set of formulae, and $\Delta = (D, W)$ a default theory. Then (Y, Γ) is *stable for* Δ iff

(1) $(Y, \Gamma) = <\delta_i>(X, \{\ \})$ for $X = \{M \mid M \models W\}$, and some $\{\delta_i\} \subseteq D$,

(2) $\forall \delta \in D.\ \delta(Y, \Gamma) = (Y, \Gamma)$, and

(3) Γ is Y-admissible. ∎

Theorem 3.1 – Soundness

If E is an extension for Δ, then there is some set Γ such that ($\{M \mid M \models E\}$, Γ) is stable for Δ.

Proof

Define $\mathrm{GD} = \left\{ \delta = \frac{\alpha : \beta}{\omega} \in D \mid \alpha \in E, \neg\beta \notin E \right\}$. From theorem 0.2, we have $E = \mathrm{Th}(W \cup \mathrm{GD})$. There are 2 cases:

$\mathrm{GD} = \{\ \}$:

Then $E = \mathrm{Th}(W)$. Clearly $<>(X,\{\ \}) = (Y,\{\ \})$. Consider $\delta = \frac{\alpha : \beta}{\omega} \in D$. If α is Y-valid and β is Y-satisfiable, then $E \vdash \alpha$, $E \not\vdash \neg\beta$, so $\delta \in \mathrm{GD}$, which is a contradiction. Hence $\delta(X,\{\ \}) = (Y,\{\ \})$. Clearly $\{\ \}$ is Y-admissible. Hence $(Y,\{\ \})$ is stable with respect to Δ.

$\mathrm{GD} \neq \{\ \}$:

Let $\{\delta_1,\ldots\}$ be any ordering of GD. Define δ'_i by $\delta'_i = \delta_j$, where j is the smallest integer such that δ_j is $<\delta'_0\ldots\delta'_{i-1}>(X,\{\ \})$-applicable, and $\delta_j \notin \{\delta'_0,\ldots,\delta'_{i-1}\}$, where $0 \leq i \leq n$.

It can easily be seen that this is well-defined, and uses all of $<\delta_i>$. Obviously, if $\Gamma = \mathrm{JUSTIFICATIONS}(\{\delta'_i\})$, then $\delta \in D$ implies $\delta(Y,\Gamma) = (Y,\Gamma)$. It remains to show that $<\delta'_i>(X,\{\ \}) = (Y,\Gamma)$. It is easily proved that $<\delta'_0\ldots\delta'_i>(X,\{\ \}) = (X_i,\Gamma_i)$, where X_i is the set of all models for $\mathrm{Th}(W \cup \{\omega'_0,\ldots,\omega'_i\})$ – where the ω'_i's are the consequents of the respective δ'_i's – and $\Gamma_i = \mathrm{JUSTIFICATIONS}(\{\delta'_0,\ldots,\delta'_i\})$.

$$\begin{aligned}\text{Hence } <\delta'_i>(X,\{\ \}) &= (\{M \mid M \models (\mathrm{Th}(W \cup \mathrm{GD}))\}, \mathrm{JUSTIFICATIONS}(\mathrm{GD})) \\ &= (\{M \mid M \models E\}, \mathrm{JUSTIFICATIONS}(\mathrm{GD})) \\ &= (Y, \mathrm{JUSTIFICATIONS}(\mathrm{GD}))\ .\end{aligned}$$

Clearly JUSTIFICATIONS(GD)) is Y-admissible. Hence (Y,Γ) is stable for Δ.

QED Theorem 3.1

Theorem 3.2 – Completeness

If (Y, Γ) is stable for Δ then Y is the set of models for some extension of Δ. (*I.e.*, $\{\omega \mid \forall y \in Y.\ y \models \omega\}$ is an extension for Δ.)

Proof

Since (Y,Γ) is stable, $(Y,\Gamma) = <\delta_i>(X,\{\ \})$ where $X = \{M \mid M \models W\}$ and $\{\delta_i\} \subseteq D$. Without loss of generality, let $<\delta_i>$ be infinite. (If finite, replicate δ_n). Define (X_i,Γ_i) as follows: $(X_0,\Gamma_0) = (X,\{\ \})$, and for $i \geq 0$, $(X_{i+1},\Gamma_{i+1}) = \delta_i(X_i,\Gamma_i)$. Then $Y = \cap X_i$, and $\Gamma = \cup \Gamma_i$.

† Since (Y,Γ) is stable, for any default, $\delta = \dfrac{\alpha : \beta}{\omega} \in D$, either δ is not Y-applicable, or ω is Y-valid and $\beta \in \Gamma$. In either event, Γ is Y-admissible.

Assume $\delta_i = \dfrac{\alpha_i : \beta_i}{\omega_i}$. Let F_i be the set of X_i-valid formulae. We show that $F_0 = Th(W)$ and that if $\alpha_i \in F_i$, and $\neg\beta_i \notin F_i$, then $F_{i+1} = Th(F_i \cup \{\omega_i\})$. Otherwise $F_{i+1} = F_i$.

This is trivial for F_0. Assume it is true for F_i, and consider F_{i+1}. Since Γ is Y-admissible, each Γ_i is X_i-admissible. If $\alpha_i \in F_i$, then α_i is X_i-valid. If $\neg\beta_i \notin F_i$, then β_i is X_i-satisfiable. Hence $X_{i+1} = X_i - \{N \mid N \models \neg\omega_i\}$, and $F_{i+1} = Th(F_i \cup \{\omega_i\})$. Otherwise $X_{i+1} = X_i$, so $F_{i+1} = F_i$.

Let $E = \bigcup_{i=1}^{\infty} F_i$. Clearly $Y = \{M \mid M \models E\}$. It remains to show that E is an extension for Δ. Define $E_0 = W$, and $E_{i+1} = Th(E_i) \cup \{\omega \mid \dfrac{\alpha : \beta}{\omega} \in D, \alpha \in E_i, \neg\beta \notin E\}$. We show $E = \bigcup_{i=0}^{\infty} E_i$.

$\bigcup_{i=0}^{\infty} E_i \subseteq E = \bigcup_{i=0}^{\infty} F_i$:

Clearly $E_0 \subseteq F_0 \subseteq E$. Assume $E_i \subseteq E$, and consider $\omega \in E_{i+1}$. Trivially, if $\omega \in Th(E_i)$, $\omega \in E$. Otherwise, there is a default, $\delta = \dfrac{\alpha : \beta}{\omega} \in D$, such that $\alpha \in E_i$ and $\neg\beta \notin E$. Since $\alpha \in E_i$, and $E_i \subseteq E$, α is Y-valid. Similarly, β is Y-satisfiable. By (†), ω is Y-valid, so $\omega \in E$.

$E = \bigcup_{i=0}^{\infty} F_i \subseteq \bigcup_{i=0}^{\infty} E_i$:

Clearly $F_0 = Th(W) \subseteq E_1 \subseteq \bigcup_{i=0}^{\infty} E_i$. Assume $F_i \subseteq \bigcup_{i=0}^{\infty} E_i$ and consider F_{i+1}. Since $\bigcup_{i=0}^{\infty} E_i$ is closed and $F_i \subseteq \bigcup_{i=0}^{\infty} E_i$, it suffices to show that $\alpha_i \in F_i$, and $\neg\beta_i \notin F_i$, whence $\omega_i \in \bigcup_{i=0}^{\infty} E_i$.

If $\alpha_i \in F_i$, and $\neg\beta \notin F_i$ then δ_i is X_i-applicable. Since (Y,Γ) is stable, Γ is Y-admissible. But $\{\beta_i\} \in \Gamma_{i+1} \subseteq \Gamma$, so $E \not\vdash \neg\beta_i$, so $\neg\beta_i \notin E$. $\alpha_i \in F_i \subseteq \bigcup_{i=0}^{\infty} E_i$, so $\alpha_i \in E_j$, for some j.

Thus E is an extension for Δ, by Theorem 0.1.

QED Theorem 3.2

Lemma 3.3.1

If E^i ($i \geq 0$) is an extension for the default theory $\Delta_i = (D_i, E^{i-1})$ and $E^{-1} = W$, then the following are equivalent:

(1) $\alpha \in E^i$

(2) $E^i \vdash \alpha$

(3) $(W \cup \bigcup_{r=0}^{i} \mathit{CONSEQUENTS}(GD(E^r, \Delta_r))) \vdash \alpha$

Proof

(1) $\alpha \in E^i \iff E^i \vdash \alpha$

This follows from the fact that E^i is an extension and thus logically closed.

(2) $E^i \vdash \alpha \iff (W \cup \bigcup_{r=0}^{i} \mathit{CONSEQUENTS}(GD(E^r, \Delta_r))) \vdash \alpha$

If E is an extension for Δ, then by Theorem 0.2 we know that

$$E = Th(W \cup \mathit{CONSEQUENTS}(GD(E, \Delta))).$$

Hence

$$\begin{aligned} E^i &= Th(E^{i-1} \cup \mathit{CONSEQUENTS}(GD(E^i, \Delta_i))) \\ &= Th(Th(E^{i-2} \cup \mathit{CONSEQUENTS}(GD(E^{i-1}, \Delta_{i-1}))) \\ &\qquad \cup \mathit{CONSEQUENTS}(GD(E^i, \Delta_i))) \\ &= Th(Th...(W \cup \mathit{CONSEQUENTS}(GD(E^0, \Delta_0))) \\ &\qquad \cup ... \cup \mathit{CONSEQUENTS}(GD(E^i, \Delta_i))) \end{aligned}$$

Since $Th(Th(A) \cup B) = Th(A \cup B)$,

$$E^i = Th(W \cup \bigcup_{r=0}^{i} \mathit{CONSEQUENTS}(GD(E^r, \Delta_r))) .$$

From this, the result follows by the definition of Th.

QED Lemma 3.3.1

Definition 3.3.2: $\ll$ and $\underline{\ll}$

Let $\Delta = (D,W)$ be a closed, semi-normal default theory. Without loss of generality, assume all formulae are in clausal form. The partial relations, $\underline{\ll}$ and $\ll$, on *Literals* × *Literals*, are defined as follows:

(1) If $\alpha \in W$ then $\alpha = (\alpha_1 \vee ... \vee \alpha_n)$, for some $n \geq 1$.
For all $\alpha_i, \alpha_j \in \{\alpha_1,...,\alpha_n\}$, if $\alpha_i \neq \alpha_j$, let $\neg\alpha_i \underline{\ll} \alpha_j$.
(Since: $(\alpha_1 \vee ... \vee \alpha_n) \equiv [(\neg\alpha_1 \wedge ... \wedge \neg\alpha_{j-1} \wedge \neg\alpha_{j+1} \wedge ... \wedge \neg\alpha_n) \supset \alpha_j])$

(2) If $\delta \in D$ then $\delta = \frac{\alpha : \beta \wedge \gamma}{\beta}$. Let $\alpha_1, ... \alpha_r$, $\beta_1, ... \beta_s$, and $\gamma_1, ... \gamma_t$ be the literals of the clausal forms of α, β, and γ, respectively. Then
(i) If $\alpha_i \in \{\alpha_1,...,\alpha_r\}$ and $\beta_j \in \{\beta_1,...,\beta_s\}$ let $\alpha_i \underline{\ll} \beta_j$.
(ii) If $\gamma_i \in \{\gamma_1,...,\gamma_t\}$, $\beta_j \in \{\beta_1,...,\beta_s\}$ and $\gamma_i \notin \{\beta_1,...,\beta_s\}$ let $\neg\gamma_i \ll \beta_j$.
(iii) Also, $\beta = \beta_1 \wedge ... \wedge \beta_m$, for some $m \geq 1$.
For each $i \leq m$, $\beta_i = (\beta_{i,1} \vee ... \vee \beta_{i,m_i})$, where $m_i \geq 1$.
Thus if $\beta_{i,j}$, $\beta_{i,k} \in \{\beta_{1,1},...,\beta_{m,m_m}\}$ and $\beta_{i,j} \neq \beta_{i,k}$ let $\neg\beta_{i,j} \underline{\ll} \beta_{i,k}$.

(3) The expected transitivity relationships hold for $\ll$ and $\underline{\ll}$. *I.e.*,
(i) If $\alpha \underline{\ll} \beta$ and $\beta \underline{\ll} \gamma$ then $\alpha \underline{\ll} \gamma$.
(ii) If $\alpha \ll \beta$ and $\beta \ll \gamma$ then $\alpha \ll \gamma$.
(iii) If $\alpha \ll \beta$ and $\beta \underline{\ll} \gamma$ or $\alpha \underline{\ll} \beta$ and $\beta \ll \gamma$ then $\alpha \ll \gamma$. ■

Definition 3.3.3: Orderedness

A semi-normal default theory is said to be *ordered* iff there is no literal, α, such that $\alpha \ll \alpha$. ■

Definition 3.3.4: Universe of Δ

For a closed, semi-normal default theory, $\Delta = (D, W)$, define the *Universe of* Δ, U(Δ), as follows:

$$U(\Delta) = \{\alpha \mid \alpha \in \textit{Literals} \text{ and } [\,\exists\xi.\ [(\alpha \vee \xi) \in \textit{CLAUSES}(W \cup \textit{CONSEQUENTS}(D))] \text{ or } [(\neg\alpha \vee \xi) \in \textit{CLAUSES}(W \cup \textit{CONSEQUENTS}(D))]]\}$$

$$\cup \{\alpha_i \mid \exists\alpha,\beta,\gamma.\ \frac{\alpha : \beta}{\gamma} \in D \text{ and } \alpha_i \in \textit{LITERALS}(\alpha)\ \}$$

$$\cup \{\neg\gamma_i \mid \exists\alpha,\beta,\gamma.\ \frac{\alpha : \beta \wedge \gamma}{\beta} \in D \text{ and } \gamma_i \in \textit{LITERALS}(\gamma)\ \}$$

Observe that ξ may be the null clause. ■

Definition 3.3.5: $l : U(\Delta) \mapsto \mathbf{N}$

For a closed, ordered, semi-normal default theory, $\Delta = (D, W)$, we define the function $l : U(\Delta) \mapsto \mathbf{N}$, as follows:

If $\alpha, \beta \in U(\Delta)$ and $\alpha \mathrel{\underline{\ll}} \beta$ then $l(\alpha) \leq l(\beta)$. If $\alpha \ll \beta$ then $l(\beta) \geq l(\alpha)+1$.

If $\beta \in U(\Delta)$ and for no $\alpha \in U(\Delta)$ is $(\alpha \ll \beta)$ or $(\alpha \mathrel{\underline{\ll}} \beta)$ then $l(\beta) = 0$.

If $n \in \mathbf{N}$, $\beta \in U(\Delta)$, and $l(\beta) > n$ then $\exists \alpha \in U(\Delta)$. $(\alpha \ll \beta)$ and $l(\alpha) = n$.

Since Δ is ordered, l is well defined. Observe that l is a total function on $U(\Delta)$ which assigns a natural number to each literal in $U(\Delta)$. $l(\alpha)$ may be thought of as the length of the longest chain of semi-normal defaults which could figure in an inference of α. ■

Definition 3.3.6: l_{MAX}, l_{MIN}

If β is a closed formula, and the clausal form of β is

$$(\beta_{1,1} \vee \ldots \vee \beta_{1,m_1}) \wedge \ldots \wedge (\beta_{m,1} \vee \ldots \vee \beta_{m,m_m}),$$

then define $l_{MAX}(\beta) \equiv MAX(l(\beta_{i,j}))$

$l_{MIN}(\beta) \equiv MIN(l(\beta_{i,j}))$. ■

Lemma 3.3.7

If $\Delta = (D, W)$ is an ordered, closed, semi-normal default theory, then there is a partition, $\{D_i\}$, for D induced by:

$$\forall \delta \in D.\ \delta = \frac{\alpha : \beta \wedge \gamma}{\beta} \text{ and } l_{MIN}(\beta) = i \text{ iff } \delta \in D_i .$$

Proof

Clearly *LITERALS* (*CONSEQUENTS* ($\{\delta \in D\}$)) $\subseteq U(\Delta)$, and l is total on $U(\Delta)$.

Therefore: 1) $\forall\delta \in D.\ \forall i.\ \forall j.\ (\delta \in D_i \wedge \delta \in D_j)$ implies $i = j$.

2) $\forall\delta \in D.\ \exists i.\ (\delta \in D_i)$.

QED Lemma 3.3.7

Corollary 3.3.8

If $\delta \in D_0$, then δ is a normal default.

Proof

If $\delta = \dfrac{\alpha : \beta \wedge \gamma}{\beta} \in D_0$ then $l_{MIN}(\beta) > l_{MAX}(\neg\gamma) \geq 0$.

QED Corollary 3.3.8

Corollary 3.3.9

If $i > 0$ and $D_i \neq \{\ \}$, there is at least one non-normal (*i.e.*, semi-normal) default in D_i.

Proof

If D_i contains only normal defaults, then the minimality of l guarantees that $l_{MIN}(\mathit{CONSEQUENTS}(D_i)) < i$, which is a contradiction.

QED Corollary 3.3.9

Lemma 3.3.10

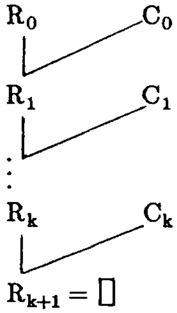
If Γ is consistent, if $l_{MAX}(\neg\beta) < j$, and if $l\,(\gamma)$ is defined for all $\gamma \in \mathit{LITERALS}(\Gamma)$, then there is a linear resolution refutation of β from Γ if and only if there is a linear resolution refutation of β from Ψ, where $\Psi \subseteq \Gamma$ and $\psi \in \Psi$ iff $l_{MIN}(\psi) < j$.

Proof

(→)

The proof is by construction of such a refutation.

Since Γ is consistent, if there is a refutation of β from Γ, there is a refutation with top clause in CLAUSES (β). *I.e.*,



and $R_0 \in$ *CLAUSES* (β), $C_0 \in \Gamma$.

We proceed by induction on the steps in the refutation.

base

Assume β is in clausal form, *i.e.*,

$$\beta = \beta_1 \wedge \dots \wedge \beta_n \quad \text{and} \quad \beta_i = \beta_{i,1} \vee \dots \vee \beta_{i,n_i} \text{ , for i = 1,...,n .}$$

By hypothesis, $l\,(\neg\beta_{i,r}) < j$. Without loss of generality, assume that $R_0 = \beta_1 = \beta_{1,1} \vee \dots \vee \beta_{1,n_1}$, that $C_0 = C_{0,1} \vee \dots \vee C_{0,m_0}$, and that $C_{0,1}$ resolves on $\beta_{1,1}$ to produce R_1. Thus $C_{0,1} = \neg\beta_{1,1}$ so $l\,(C_{0,1}) < j$ and $l_{MIN}(C_0) < j$. It follows that $C_0 \in \Psi$.

Since for i>1, $\neg C_{0,i} \underline{\ll} C_{0,1}$, $l\,(\neg C_{0,i}) \leq l\,(C_{0,1}) < j$. Thus, if $R_1 = R_{1,1} \vee \dots \vee R_{1,t}$ then $\forall s.\ l\,(\neg R_{1,s}) < j$.

step

Assume that $R_i = R_{i,1} \vee \dots \vee R_{i,m}$, that $\forall s.\ l\,(\neg R_{i,s}) < j$, and that $\forall r{<}i.\ C_r \in \Psi$ or $C_r \in \{R_0,\dots,R_{i-1}\}$. Consider the resolution of R_i with C_i . $C_i = C_{i,1} \vee \dots \vee C_{i,m_i}$. Without loss of generality, assume $C_{i,1} = \neg R_{i,1}$. Hence $l\,(C_{i,1}) = l\,(\neg R_{i,1}) < j$ and so $l_{MIN}(C_i) < j$. So $C_i \in \Psi$ or $C_i \in \{R_0,\dots,R_i\}$. For r>1, $l\,(\neg C_{i,r}) \leq l\,(C_{i,1}) < j$. Thus $\forall s.\ l\,(\neg R_{i+1,s}) < j$.

By induction, for every clause, C_i, in the refutation of β, $C_i \in \Psi$ or C_i is a descendent of $\Psi \cup \{\beta\}$. Thus, there is a linear resolution refutation of β from Ψ.

(←)

Trivial: Since $\Psi \subseteq \Gamma$, the refutation from Ψ serves as a refutation from Γ.

QED lemma 3.3.10

Theorem 3.3 – Coherence

If $\Delta = (D, W)$ is an ordered, semi-normal default theory, then Δ has an extension.

Proof

If W is inconsistent, then Δ has the trivial extension, L. Hence assume W is consistent.

We proceed by constructing an extension, E for Δ. First, let $\{D_i\}$ be a partition of D induced by l, as described in Lemma 3.3.7. Recall that by Corollary 3.3.8, if $\delta \in D_0$ then δ is a normal default, and that by Corollary 3.3.9, for $i > 0$, D_i must contain at least one semi-normal default, say

$$\delta = \frac{\alpha : \beta \wedge \gamma}{\beta},$$

and $l_{MAX}(\neg\gamma) < l_{MIN}(\beta)$.

We now construct an extension for Δ.

Let $\Delta_0 = (D_0, W)$. Since Δ_0 is a normal default theory and W is consistent, Δ_0 has a consistent extension, say E^0.

For $i > 0$, construct Δ_i as follows:

$$D_i' = \{\frac{\alpha : \beta}{\beta} \mid \frac{\alpha : \beta}{\beta} \in D_i \vee \frac{\alpha : \beta \wedge \gamma}{\beta} \in D_i\,,\ \neg\gamma \notin E^{i-1}\}$$

$$\Delta_i = (D_i', E^{i-1})$$

Where E^{i-1} is an extension for Δ_{i-1}. Since each Δ_i is a normal default theory, each Δ_i has at least one extension, E^i. Let $E = \bigcup_{i=0}^{\infty} E^i$. Since W is consistent, so is E^0, by Theorem 0.3. Since E^i is an extension for (D_i', E^{i-1}), E^i is consistent if E^{i-1} is, and $E^{i-1} \subseteq E^i$. By induction E is consistent. We now show that E is an extension for Δ. By Theorem 0.1, it is sufficient to show that $E = \bigcup_{i=0}^{\infty} F_i$, where

$F_0 = W$, and for i>0

$F_{i+1} = Th(F_i) \cup \{\omega \mid \frac{\alpha : \beta}{\omega} \in D, \alpha \in F_i$, and $\neg\beta \notin E\}$.

(1) We first show that $\bigcup_{i=0}^{\infty} F_i \subseteq E$.

a) $F_0 = W \subseteq E^0 \subseteq E$.

b) Assume $F_i \subseteq E$. We show that $F_{i+1} \subseteq E$.

$F_{i+1} = Th(F_i) \cup \{\beta \mid \frac{\alpha : \beta \wedge \gamma}{\beta} \in D, \alpha \in F_i, (\neg\beta \vee \neg\gamma) \notin E\}$

i) Since $F_i \subseteq E$ and E is logically closed, $Th(F_i) \subseteq E$.

ii) Consider $\beta \in \{\beta \mid \frac{\alpha : \beta \wedge \gamma}{\beta} \in D, \alpha \in F_i, (\neg\beta \vee \neg\gamma) \notin E\}$.

Since $\alpha \in F_i$, $\alpha \in E$, and hence $\alpha \in E^j$ for some j.

Since $(\neg\beta \vee \neg\gamma) \notin E$, $\neg\gamma \notin E^{j-1}$, so $\frac{\alpha : \beta}{\beta} \in D_j{}'$.

But $\neg\beta \notin E$, so $\neg\beta \notin E^j$.

Therefore, since E^j is an extension for $\Delta_j = (D_j{}', E^{j-1}))$ and $\alpha \in E^j$, $\beta \in E^j$.

Therefore $\beta \in E$.

By induction, $\bigcup_{i=0}^{\infty} F_i \subseteq E$.

(2) Finally, we show that $E \subseteq \bigcup_{i=0}^{\infty} F_i$.

A) Consider $\omega \in E^0$. E^0 is an extension for Δ_0, so by Theorem 0.1 $E^0 = \bigcup_{i=0}^{\infty} G_i$, where

$G_0 = W$, and for i>0

$G_{i+1} = Th(G_i) \cup \{\omega \mid \frac{\alpha : \omega}{\omega} \in D_0, \alpha \in G_i$, and $\neg\omega \notin E^0\}$.

It therefore suffices to show that $\bigcup_{i=0}^{\infty} G_i \subseteq \bigcup_{i=0}^{\infty} F_i$.

a) $G_0 = W = F_0 \subseteq \bigcup_{i=0}^{\infty} F_i$.

b) Assume $G_i \subseteq \bigcup_{i=0}^{\infty} F_i$, and consider $\omega \in G_{i+1}$.

$G_{i+1} = Th(G_i) \cup \{\omega \mid \frac{\alpha : \omega}{\omega} \in D_0, \alpha \in G_i, \neg\omega \notin E^0\}$

i) If $\omega \in Th(G_i)$ then $\omega \in \bigcup_{i=0}^{\infty} F_i$ by hypothesis since $\bigcup_{i=0}^{\infty} F_i$ is logically closed.

ii) Otherwise $\omega \in \{\omega \mid \frac{\alpha : \omega}{\omega} \in D_0, \alpha \in G_i, \neg\omega \notin E^0\}$.

But: 1) If $\omega \in G_{i+1}$ and $E^0 = \bigcup_{i=0}^{\infty} G_i$ then $\omega \in E^0 \subseteq E$.

Since E is consistent, $\neg\omega \notin E$.

2) If $\alpha \in G_i$ then $\alpha \in \bigcup_{i=0}^{\infty} F_i$ by hypothesis, so $\alpha \in F_k$ for some k.

3) $D_0 \subseteq D$

Thus $\omega \in F_{k+1} \subseteq \bigcup_{i=0}^{\infty} F_i$.

By induction, $\bigcup_{i=0}^{\infty} G_i \subseteq \bigcup_{i=0}^{\infty} F_i$.

B) Assume $E^{j-1} \subseteq \bigcup_{i=0}^{\infty} F_i$, and show $E^j \subseteq \bigcup_{i=0}^{\infty} F_i$.

Consider $\omega \in E^j$. E^j is an extension for $\Delta_j = (D_j', E^{j-1})$, so $E^j = \bigcup_{i=0}^{\infty} G_i$, where

$G_0 = E^{j-1}$, and for i>0

$G_{i+1} = Th(G_i) \cup \{\omega \mid \frac{\alpha : \omega}{\omega} \in D_j', \alpha \in G_i \text{, and } \neg\omega \notin E^j\}$.

a) By hypothesis, $G_0 = E^{j-1} \subseteq \bigcup_{i=0}^{\infty} F_i$.

b) Assume $G_i \subseteq \bigcup_{i=0}^{\infty} F_i$ and consider $\omega \in G_{i+1}$.

i) If $\omega \in Th(G_i)$ then $\omega \in \bigcup_{i=0}^{\infty} F_i$ by hypothesis since $\bigcup_{i=0}^{\infty} F_i$ is logically closed.

ii) Otherwise $\omega \in \{\omega \mid \frac{\alpha : \omega}{\omega} \in D_j', \alpha \in G_i \text{, and } \neg\omega \notin E^j\}$.

Since $\alpha \in G_i$, we know that $\alpha \in E^j$ and $\alpha \in \bigcup_{i=0}^{\infty} F_i$. Also, if $\omega \in G_{i+1}$ then $\omega \in E^j$ so $\omega \in E$. Therefore $\neg\omega \notin E$, since E is consistent.

If $\delta = \frac{\alpha : \omega}{\omega} \in D_j'$, then either $\frac{\alpha : \omega}{\omega} \in D$ or $\exists\gamma. \frac{\alpha : \omega \wedge \gamma}{\omega} \in D$. Thus there are two cases:

a) Either $\frac{\alpha : \omega}{\omega} \in D$, $\alpha \in \bigcup_{i=0}^{\infty} F_i$, and $\neg\omega \notin E$ and hence $\omega \in \bigcup_{i=0}^{\infty} F_i$,

b) Or $\frac{\alpha : \omega \wedge \gamma}{\omega} \in D$, $\alpha \in \bigcup_{i=0}^{\infty} F_i$, and $\neg\omega \notin E$.

Clearly, if $(\neg\gamma \vee \neg\omega) \notin E$ then $\omega \in \bigcup_{i=0}^{\infty} F_i$.

Since $\omega \in E$, it can be shown that $(\neg\gamma \vee \neg\omega) \in E$ iff $\neg\gamma \in E$.

We show that $\neg\gamma \notin E$.

Clearly $l_{MAX}(\neg\gamma) < l_{MIN}(\omega) = j$. Assume $\neg\gamma \in E$. Then $\exists r \geq j. (\neg\gamma \in E^r)$.

By Lemma 3.3.1, $(W \cup \bigcup_{i=0}^{r} \textit{CONSEQUENTS}(GD(E^i, \Delta_i))) \vdash \neg\gamma$.

Thus there is a linear resolution refutation of γ from

$$\Gamma = (W \cup \bigcup_{i=0}^{r} \textit{CONSEQUENTS}(GD(E^i, \Delta_i))).$$

Observe that if $\delta \in GD(E^i, \Delta_i)$ then $\delta \in D_i'$ and so $l_{MIN}(\textit{CONSEQUENTS}(\delta)) = i$. By Lemma 3.3.10, the existence of a refutation of γ from Γ, given $l_{MAX}(\neg\gamma) < j$, implies that there is a refutation from

$\Psi \subseteq \Gamma$ such that $\psi \in \Psi \leftrightarrow l_{MIN}(\lambda) < j$. Thus there is a refutation from

$$\Psi = (W \cup \bigcup_{i=0}^{j-1} \mathit{CONSEQUENTS}(GD(E^i, \Delta_i))).$$

Hence $\Psi \vdash \neg\gamma$ and, by Lemma 3.3.1, $\Psi \vdash \neg\gamma$ iff $E^{j-1} \vdash \neg\gamma$. But if $\delta \in D_j'$ then $\neg\gamma \notin E^{j-1}$ and so $E^{j-1} \nvdash \neg\gamma$ since E^{j-1} is logically closed. Hence we obtain a contradiction by assuming that $\neg\gamma \in E$, so $\neg\gamma \notin E$.

Thus $(\neg\gamma \vee \neg\omega) \notin E$ and so $\omega \in \bigcup_{i=0}^{\infty} F_i$.

We see that $G_{i+1} \subseteq \bigcup_{i=0}^{\infty} F_i$, and by induction $\bigcup_{i=0}^{\infty} G_i \subseteq \bigcup_{i=0}^{\infty} F_i$.

Therefore $E^j \subseteq \bigcup_{i=0}^{\infty} F_i$.

By induction, $E \subseteq \bigcup_{i=0}^{\infty} F_i$.

Together, (1) and (2) show that $E = \bigcup_{i=0}^{\infty} F_i$, so E is an extension for Δ.

QED Theorem 3.3

Before presenting the proof of Theorem 3.4, we repeat the definition of the procedure to generate extensions given earlier. Superscripts have been added which serve only as reference points in the proofs. They do not effect the computation.

```
$H_0 \leftarrow W$; $j \leftarrow 0$;
repeat
  $j \leftarrow j + 1$; $h^j_0 \leftarrow W$; $GD^j_0 \leftarrow \{\ \}$; $i \leftarrow 0$;
  repeat
    $D^j_i \leftarrow \{ \frac{\alpha : \beta}{\gamma} \in D \mid (h^j_i \vdash \alpha), (h^j_i \not\vdash \neg\beta), (H_{j-1} \not\vdash \neg\beta) \}$;
    if $\neg$null($D^j_i$ - $GD^j_i$) then
      choose $\delta$ from ($D^j_i$ - $GD^j_i$);
      $GD^j_{i+1} \leftarrow GD^j_i \cup \{\delta\}$;
      $h^j_{i+1} \leftarrow h^j_i \cup$ {CONSEQUENT($\delta$)}; endif;
    $i \leftarrow i + 1$;
  until null($D^j_{i-1}$ - $GD^j_{i-1}$);
  $H_j = h^j_{i-1}$
until $H_j = H_{j-1}$
```

Lemma 3.4.1

If Δ is a finite default theory, then the algorithm can fail to converge only if one of the approximations is repeated. *I.e.*, for some j and some k > j+1, $H_j = H_k$.

Proof

If Δ is finite, there are only a finite number of different combinations possible. Thus there are only a finite number of distinct H_i's which can be constructed. If $H_j = H_{j+1}$, the algorithm converges.

QED Lemma 3.4.1

Lemma 3.4.2

If Δ is a finite, semi-normal default theory, and W is consistent, then

$$H_i \vdash \beta \rightarrow H_i \not\vdash \neg\beta.$$

Proof

Assume $H_i \vdash \beta, \neg\beta$. Let r, s be the smallest integers such that $h^i_r \vdash \beta$, $h^i_s \vdash \neg\beta$. Assume $r \leq s$, so $h^i_{s-1} \nvdash \neg\beta$. By hypothesis, $h^i_s \vdash \beta, \neg\beta$. Now $h^i_s = h^i_{s-1} \cup \{\omega\}$, where

$$\frac{\alpha : \omega \wedge \gamma}{\omega} \in D,\ \alpha \in h^j_{s-1},\ H_{i-1} \nvdash (\neg\omega \vee \neg\gamma), \text{ and } h^i_{s-1} \nvdash (\neg\omega \vee \neg\gamma)\ .$$

But if $h^i_s \vdash \beta, \neg\beta$, then $(h^i_{s-1} \cup \{\omega\}) \vdash \beta, \neg\beta$ so $h^i_{s-1} \vdash \neg\omega$ and hence $h^i_{s-1} \vdash (\neg\omega \vee \neg\gamma)$, which is a contradiction. The proof is similar if $s < r$.

QED Lemma 3.4.2

Definition 3.4.3: Network Default Theory

A default theory, $\Delta = (D, W)$, is *a network theory* if it satisfies the following conditions:

(1) W contains only:

a) Literals (*i.e.*, Atomic formulae or their negations), or

b) Disjuncts of the form $(\alpha \vee \beta)$ where α and β are literals.

(2) D contains only normal and semi-normal defaults of the form:

$$\frac{\alpha : \beta}{\beta} \quad \text{or} \quad \frac{\alpha : \beta \wedge \gamma_1 \wedge \ldots \wedge \gamma_n}{\beta}$$

where α, β, and γ_i are literals. ■

Lemma 3.4.4

If Δ is a finite, ordered, network default theory, if W is consistent, and if β is a literal, then $H_{i-1} \vdash \beta \rightarrow H_i \nvdash \neg\beta$.

Proof

Assume $H_{j-1} \vdash \beta$, and consider $H_j = \bigcup_{i=0}^{\infty} h^j_i$. Assume $H_j \vdash \neg\beta$. The proof proceeds by induction.

base

$h^j_0 = W$. Since $H_{j-1} \nvdash \neg\beta$, clearly $W \nvdash \neg\beta$. Therefore $h^j_0 \nvdash \neg\beta$.

step

Assume $h^j_i \nvdash \neg\beta$ and $h^j_{i+1} \vdash \neg\beta$. $h^j_{i+1} = h^j_i \cup \{\omega\}$, where

$$\frac{\alpha : \gamma \wedge \omega}{\omega} \in D,\ h^j_i \vdash \alpha,\ h^j_i \nvdash (\neg\gamma \vee \neg\omega), \text{ and } H_{j-1} \nvdash (\neg\gamma \vee \neg\omega).$$

Clearly, $\omega \neq \neg\beta$ or else $H_{j-1} \vdash \neg\omega$.

Note that:

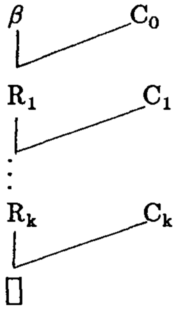
i) H_j contains only disjunctions of two literals.

ii) $h_i^j = W \cup$ *CONSEQUENTS*(GD_i^j)

iii) $GD_i^j \subseteq D$

iv) *CONSEQUENTS*$(GD_i^j) \subseteq$ *Literals*.

Consider a linear resolution refutation of β (*i.e.*, a proof of $\neg\beta$) from h_{i+1}^j, with top clause β. We continue by induction on the structure of this refutation.



base

$\omega \in$ *Literals* and $\omega \neq \neg\beta$ so $C_0 \neq \omega$. Clearly, $C_0 \neq \beta$. Thus $C_0 \in h_i^j$. If $C_0 \in h_i^j$ - W, then $C_0 \in$ *Literals*. But then $C_0 = \neg\beta$ which leads to the contradiction that $h_i^j \vdash \neg\beta$. Thus $C_0 \in W$. Clearly $C_0 \notin$ *Literals*, as above. Hence $C_0 = (\neg\beta \vee \xi)$, with $\xi \in$ *Literals*. Thus $R_1 = \xi \neq \Box$.

step

Assume: i) $\omega \notin \{C_0, \dots, C_{n-1}\}$

ii) $\{C_0, \dots, C_{n-1}\} \subseteq W$

iii) $\{R_1, \dots, R_n\} \subseteq$ *Literals*.

Let $R_n = \eta \in$ *Literals*. If $C_n = \omega$ then $\omega = \neg\eta$ so $W \cup \{\omega\} \vdash \neg\beta$ but $W \subseteq H_{j-1}$ and $H_{j-1} \vdash \beta$, so $H_{j-1} \vdash \beta, \neg\beta$, which contradicts Lemma 3.4.2. Clearly $\eta \neq \neg\beta$, so $C_n \neq \beta$, or else $W \vdash \neg\beta$ which is false. Thus $C_n \in W$. Clearly $C_n \notin$ *Literals*, as above, hence $C_n = (\neg\eta \vee \lambda)$ with $\lambda \in$ *Literals*. Therefore $R_{n+1} = \lambda \neq \Box$.

So: i) $\omega \notin \{C_0, \dots, C_n\}$

ii) $\{C_0, \dots, C_n\} \subseteq W$

iii) $\{R_1, \dots, R_{n+1}\} \subseteq$ *Literals*.

By induction, there is no such resolution refutation and the required result is proved.

QED Lemma 3.4.4

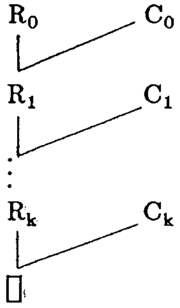
Lemma 3.4.5

If Δ is a finite, ordered, network default theory, and $\{\alpha_1,...,\alpha_n\} \subseteq$ *Literals*, then $H_i \vdash (\alpha_1 \vee ... \vee \alpha_n)$ if and only if $W \vdash (\alpha_1 \vee ... \vee \alpha_n)$ or $H_i \vdash \alpha_j$, for some j.

Proof

($\leftarrow$) Trivial.

($\rightarrow$) Assume false, and consider a linear resolution proof of $(\alpha_1 \vee ... \vee \alpha_n)$ (*i.e.*, a refutation of $(\neg\alpha_1 \wedge ... \wedge \neg\alpha_n))$ from H_i, with top clause $R_0 \in \{\neg\alpha_1,...,\neg\alpha_n\}$.



We know that $C_0 \in H_i \cup \{\neg\alpha_1,...,\neg\alpha_n\}$, and that, for i>0, $C_i \in H_i$ or $C_i \in \{R_j \mid j \leq i\}$ or $C_i \in \{\neg\alpha_1,...,\neg\alpha_n\}$. We proceed by induction.

base

Without loss of generality, assume $R_0 = \neg\alpha_1$. Clearly $\alpha_1 \notin \{\neg\alpha_1,...,\neg\alpha_n\}$, or else $W \vdash (\alpha_1 \vee ... \vee \alpha_n)$, so $C_0 \notin \{\neg\alpha_1,...,\neg\alpha_n\}$. Clearly $C_0 \neq \alpha_1$ or else $H_i \vdash \alpha_1$ which contradicts our assumption. Hence $C_0 = (\alpha_1 \vee \gamma) \in W$, for some $\gamma \in$ *Literals*, and so $R_1 = \gamma \neq \Box$.

step

Assume a) $\{R_0 ,..., R_n\} \subseteq$ *Literals*

b) $\{C_0 ,..., C_{n-1}\} \subseteq W$.

Let $R_n = \eta \in$ *Literals*. If $C_n = \neg\eta \in \{\neg\alpha_1,...,\neg\alpha_n\}$ then $W \vdash (\alpha_1 \vee ... \vee \alpha_n)$ which contradicts our hypothesis. If $C_n = \neg\eta \in H_i \cup \{R_0,...,R_n\}$ then $H_i \vdash \alpha_1$ which also contradicts the hypothesis. Hence $C_n = (\neg\eta \vee \xi) \in W$, with $\xi \in$ *Literals* and $R_{n+1} = \xi \neq \Box$.

By induction, there is no such resolution refutation, and the lemma is proved.

QED Lemma 3.4.5

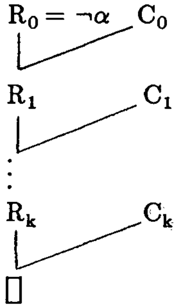
Lemma 3.4.6

If Δ is a finite, ordered, network default theory, and $\alpha \in$ *Literals*, then $H_i \vdash \alpha$ if and only if $W \vdash \alpha$ or $\exists \beta \in$ Literals. $l(\beta) \leq l(\alpha)$, $\beta \in H_i$, and $W \vdash (\beta \supset \alpha)$.

Proof

($\leftarrow$) Trivial.

($\rightarrow$) Assume false and consider a linear resolution proof of α (*i.e.*, a refutation of $\neg\alpha$) from H_i, with top clause $\neg\alpha$. We proceed by induction.



base

Clearly $C_0 \neq \alpha$ or else $\alpha \in H_i$ and $l(\alpha) \leq l(\alpha)$ and $W \vdash (\alpha \supset \alpha)$ which contradicts the hypothesis. Hence $C_0 = (\alpha \vee \gamma) \in W$, for $\gamma \in$ *Literals*. By definition, $l(\neg\gamma) \leq l(\alpha)$. $R_1 = \gamma \neq \Box$. Clearly $W \vdash (\neg\gamma \supset \alpha)$.

step

Assume: a) $\{C_0,\ldots,C_{n-1}\} \subseteq W$

b) $\{R_0,\ldots,R_n\} \subseteq$ *Literals*

c) $l(\neg R_n) \leq l(\alpha)$

d) $W \vdash (\neg R_n \supset \alpha)$

Let $R_n = \eta$. If $C_n = \neg\eta \in H_i$ then $H_i \vdash \alpha$, $\neg\eta \in H_i$, $W \vdash (\neg\eta \supset \alpha)$, and $l(\neg\eta) = l(\neg R_n) \leq l(\alpha)$ which contradicts our assumption. If $C_n = \neg\eta = \neg\alpha$ then $W \vdash \alpha$ which is also a contradiction. Hence $C_n = (\neg\eta \vee \xi) \in W$, with $\xi \in$ *Literals*, $R_{n+1} = \xi \neq \Box$, and $l(\neg R_{n+1}) = l(\neg\xi) \leq l(\neg\eta) = l(\neg R_n) \leq l(\alpha)$. By modus ponens, $W \vdash (\neg\xi \supset \alpha)$.

Thus there is no such refutation, and the result is proved.

QED Lemma 3.4.6

Lemma 3.4.7

If Δ is a finite, ordered, network default theory, and $\alpha \in$ Literals, $\alpha \notin H_i$, $\alpha \notin H_j$, and $\alpha \in H_k$ for i<k<j, then

$$\exists \beta \in \text{Literals.}\ (l\,(\beta) < l\,(\alpha)) \text{ and } \beta \in \bigcup_{i<r\leq j} H_i \Delta H_r .$$

Proof

Let j be the least j>k such that $\alpha \notin H_j$.

Define $D_\alpha = \{\delta \in D \mid \delta = \dfrac{\gamma : \alpha \wedge \omega_1 \wedge \dots \wedge \omega_n}{\alpha}\}$

$$GD_\alpha^i = \bigcup_{r=0}^{\infty} GD_r^i \cap D_\alpha$$

Clearly $GD_\alpha^{j-1} \neq \{\ \}$ and $GD_\alpha^j = \{\ \}$. Consider $\delta \in GD_\alpha^{j-1}$. Since $\delta \notin GD_\alpha^j$ three cases are possible:

1) $H_{j-1} \vdash (\neg\omega_1 \vee \dots \vee \neg\omega_n)$. By Lemma 3.4.5, there is an ω_r , say ω, such that $H_{j-1} \vdash \neg\omega$. By Lemma 3.4.6, there is a $\beta \in H_{j-1}$ such that $l\,(\beta) \leq l\,(\omega)$ and $W \vdash (\beta \supset \omega)$. But then $l\,(\beta) < l\,(\alpha)$. Clearly $\beta \notin H_{j-2}$, so β is the required literal.
2) $H_j \vdash (\neg\omega_1 \vee \dots \vee \neg\omega_n)$. The argument for case 1 applies.
3) $H_j \not\vdash \gamma$. By recursively applying the foregoing arguments to γ, we can construct a set of γ_r's which were in H_{j-1} and are not in H_j. The first of these to go into H_{j-1} must also go into H_j, unless $H_{j-1} \cup H_j$ contains a $\beta \ll \gamma_r \lll \alpha$ which was not in H_i .

QED Lemma 3.4.7

Lemma 3.4.8

If Δ is a finite, ordered, network default theory, and $\alpha \in$ Literals, $\alpha \in H_i$, $\alpha \in H_j$, and $\alpha \notin H_k$ for i<k<j, then either

1) $\exists \beta \in \text{Literals.}\ (l\,(\beta) < l\,(\alpha)) \text{ and } \beta \in \bigcup_{i<r\leq j} H_i \Delta H_r$, or
2) $\exists \beta \in \text{Literals.}\ (l\,(\beta) \leq l\,(\alpha)) \text{ and } \beta \in H_j \text{ and } \beta \notin H_i$.

Proof

Let k be the least k>i such that $\alpha \notin H_k$. Let j be the least j>k such that $\alpha \in H_j$.

Consider $\delta = \frac{\gamma : \alpha \wedge \beta}{\alpha} \in GD_{\alpha}^{j}$. Clearly $GD_{\alpha}^{j} \neq \{\ \}$, and $\delta \notin GD_{\alpha}^{k}$.

Cases: 1) $H_k \vdash \neg\beta$, $H_j \nvdash \neg\beta$. This gives the first of the required conditions, by Lemmas 3.4.5 and 3.4.6.

2) $H_{k-1} \vdash \neg\beta$, $H_j \nvdash \neg\beta$. The argument for case 1 applies.

3) $H_k \nvdash \gamma$, $H_j \vdash \gamma$. By Lemma 3.4.6, $\exists \gamma_1 \lessapprox \alpha.\ \gamma_1 \in H_j$, $\gamma_1 \notin H_k$.

Cases: a) $\gamma_1 \notin H_i$. This is the second of the required conditions.

b) $\gamma_1 \in H_i$. Repeating the above arguments for γ_1 yields a (possibly cyclic) chain of γ_r's such that $\gamma_r \in H_{k-1}$, $\gamma_r \notin H_k$. Consider the first γ_r to go into H_{k-1} . It must also go into H_k , which is a contradiction.

QED Lemma 3.4.8

Theorem 3.4 – Convergence

The procedure presented above always converges when applied to
a finite, ordered, network default theory.

Proof

By Lemma 3.4.1, non-convergence implies there is a cycle. *I.e.*, for some i and some j>i, $H_i = H_j$ and $H_i \neq H_{i+1}$.

Choose $\alpha \in \bigcup_{i<k\leq j} (H_i \Delta H_k)$ such that $\alpha \in$ *Literals* and for every $\beta \in \bigcup_{i<k\leq j} (H_i \Delta H_k)$, $\neg(l\,(\beta) < l\,(\alpha))$.
Thus α is the "least" literal to change state between H_i and H_j. There are two cases:

(1) If $\alpha \notin H_i$ and $\alpha \in H_k$ then, by Lemma 3.4.7, $\exists \beta \in \bigcup_{i<k\leq j} (H_i \Delta H_k).\ l\,(\beta) < l\,(\alpha)$, so α is not the least such α, which is a contradiction.

(2) If $\alpha \in H_i$ and $\alpha \notin H_k$ then, by Lemma 3.4.8, either

a) $\exists \beta \in \bigcup_{i<k\leq j} (H_i \Delta H_k).\ l\,(\beta) < l\,(\alpha)$

so α is not the least such α, which is a contradiction, or

b) $\exists \beta.\ \beta \in H_j$ and $\beta \notin H_i$

which implies that $H_i \neq H_j$ which is also a contradiction.

Therefore, there is no cycle, and so the procedure converges.

QED Theorem 3.4

Theorem 3.5 – Strong Convergence

The procedure given above always converges immediately when applied to a finite, normal default theory $\Delta = (D, W)$ – *i.e.*, $Th(H_1)$ is an extension.

Proof

Etherington [1982] shows that $H_1 = H_2$ if and only if $Th(H_1)$ is an extension for Δ. If W is inconsistent, then $Th(H_1) = L$ which is an extension for Δ. Hence assume W is consistent. To show that $Th(H_1)$ is an extension for Δ, we invoke Theorem 0.1 and show that $Th(H_1) = \bigcup_{i=0}^{\infty} E_i$, where

$$E_0 = W$$

$$E_{i+1} = Th(E_i) \cup \{\omega \mid \frac{\alpha : \omega}{\omega} \in D,\ \alpha \in E_i,\ \neg\omega \notin Th(H_1)\}\ .$$

a) We first show that $\bigcup_{i=0}^{\infty} E_i \subseteq Th(H_1)$. Recall that $H_1 = \bigcup_{i=0}^{\infty} h_i^1$.

base

Clearly $E_0 = W = h_0^1 \subseteq Th(H_1)$.

step

Assume $E_i \subseteq Th(H_1)$ and consider $\omega \in E_{i+1}$.

i) If $\omega \in Th(E_i)$ then $\omega \in Th(H_1)$, by hypothesis and closure.

ii) Otherwise $\omega \in \{\omega \mid \frac{\alpha : \omega}{\omega} \in D,\ \alpha \in E_i,\ \neg\omega \notin Th(H_1)\}$. Therefore $H_1 \not\vdash \neg\omega$. Hence $H_0 \not\vdash \neg\omega$ since $H_0 = W \subseteq H_1$. Also, $\alpha \in E_i$, so $\alpha \in Th(H_1)$, by hypothesis. It follows by [Etherington 1982, Lemma 3.3] that $H_1 \vdash \omega$.

Hence $E_{i+1} \subseteq Th(H_1)$.

b) Finally, we show that $Th(H_1) \subseteq \bigcup_{r=1}^{\infty} E_r$.

Since $\bigcup_{r=1}^{\infty} E_r$ is logically closed, it suffices to show that $H_1 \subseteq \bigcup_{r=1}^{\infty} E_r$.

base

Clearly $h_0^1 = W = E_0 \subseteq \bigcup_{r=1}^{\infty} E_r$.

step

Assume that $h_i^1 \subseteq \bigcup_{r=1}^{\infty} E_r$, and consider h_{i+1}^1 .

$h_{i+1}^1 = h_i^1 \cup \{\omega\}$, for some $\omega \in \mathit{CONSEQUENTS}(D_i^1)$.

Since $h_i^1 \subseteq \bigcup_{r=1}^{\infty} E_r$ by hypothesis, we need only show that $\omega \in \bigcup_{r=1}^{\infty} E_r$.

Since $\omega \in \mathit{CONSEQUENTS}(D_i^1)$, for some $\delta = \frac{\alpha : \omega}{\omega} \in D$, $\alpha \in h_i^1$,

$H_0 \nvdash \neg\omega$, and $h_i^1 \nvdash \neg\omega$.

By hypothesis, since $\alpha \in h_i^1$, $\alpha \in \bigcup_{r=1}^{\infty} E_r$, so $\alpha \in E_j$ for some j.

Since $\omega \in h_{i+1}^1 \subseteq H_1$, it follows by Lemma 3.4.2 that $H_1 \nvdash \neg\omega$.

But then by definition of E_{j+1}, $\omega \in E_{j+1} \subseteq \bigcup_{r=1}^{\infty} E_r$.

Combining (a) and (b), we have the desired result.

QED Theorem 3.5

Theorem 4.1

> Any network in which the subgraph of IS-A links and exceptions thereto is acyclic corresponds to an ordered theory.

Proof

The links corresponding to $\alpha \supset \neg\beta$, $\dfrac{\alpha : \neg\beta}{\neg\beta}$, and $\dfrac{\alpha : \neg\beta \wedge \neg\gamma_1 \wedge \cdots \wedge \neg\gamma_n}{\neg\beta}$ give rise to $\alpha \underline{\ll} \neg\beta$ and $\gamma_i \ll \neg\beta$. There are no links which make a transition from negative to positive or negative to negative, so such links cannot participate in any cycle leading to $\omega \ll \omega$ for any ω. What remains are IS-A links and exceptions thereto.

QED Theorem 4.1

Theorem 4.5

> In the absence of no-conclusion links, all ground facts returned by Touretzky's inferential distance algorithm lie within a single extension of the default theory corresponding to the inheritance network in question.

Proof

We prove that all the ground facts in any "grounded expansion" of the network lie within a single extension. From this the result follows. As a notational shortcut, we will use $\pm P$ to stand for $+P$ or $-P$ (or, occasionally, for P or $\neg P$). The intended meaning should be clear from context.

Let Γ be a network in Touretzky's sense. Let Φ be a grounded expansion for Γ. Define $facts(\Phi) = \{<+\alpha,\pm P> \in C(\Phi) \mid \alpha$ is an individual token$\}$, and
$facts'(\Phi) = \{P\alpha \mid <+\alpha,+P> \in facts(\Phi)\} \cup \{\neg P\alpha \mid <+\alpha,-P> \in facts(\Phi)\}$.

If $<+\alpha,\pm P> \in facts(\Phi)$ then for some $P_1,\ldots,P_n$, we have $<+\alpha,+P_1,\ldots,+P_n,\pm P> \in \Phi$, by definition. Hence, by [Touretzky 1984a, theorem 2.3], $<+\alpha,+P_1>$,..., $<+\alpha,+P_n>$, $<+\alpha,\pm P> \in facts(\Phi)$. Thus $P_1\alpha$,..., $P_n\alpha$, $\pm P\alpha \in facts'(\Phi)$. Furthermore, $<+P_i,+P_{i+1}> \in \Phi$ for $i = 1,\ldots,n{-}1$, and $<+P_n,\pm P> \in \Phi$, by [Touretzky 1984a, theorem 2.3]. Hence they are all in Γ by [Touretzky 1984a, theorem 2.2]. Hence $\dfrac{P_ix : P_{i+1}x}{P_{i+1}x}$ and $\dfrac{P_nx : \pm Px}{\pm Px} \in D$.

We claim that $facts'(\Phi)$ is inconsistent iff W is. By definition, $W = \{\pm R\alpha \mid <+\alpha,\pm R> \in \Gamma$, where α is an individual token$\}$. Therefore, W is inconsistent iff $<+\alpha,+R>$, $<+\alpha,-R> \in \Gamma$, for some α and Γ.

The right-to-left direction of the claim is trivial. For the left-to-right direction, assume that $facts'(\Phi)$ is inconsistent. Then $R\alpha$, $\neg R\alpha \in facts'(\Phi)$ so $<+\alpha,+R>$, $<+\alpha,-R> \in facts(\Phi)$, so $\sigma_1 = <+\alpha,y_1,...,y_j,+R>$ and $\sigma_2 = <+\alpha,x_1,...,x_k,-R> \in \Phi$. So Φ contradicts σ_1 and σ_2, and Φ is inconsistent. Hence Γ is inconsistent, by [Touretzky 1984a, theorem 2.8]. Furthermore, neither σ_1 nor σ_2 is inheritable in Φ, so both are in Γ, since Φ is a grounded expansion of Γ. But then $j = k = 0$, so $<+\alpha,+R>$ and $<+\alpha,-R> \in \Gamma$. Hence, $R\alpha$, $\neg R\alpha \in W$, so W inconsistent. Now if $facts'(\Phi)$ inconsistent, W is inconsistent, so Δ has a unique extension, $Th(L) \supseteq facts'(\Phi)$. In the sequel, we assume $facts'(\Phi)$ consistent.

We show that $E' = Th(facts'(\Phi))$ is an extension for $\Delta' = (D', W)$, where $D' = \left\{ \dfrac{P_i\alpha : \pm P_{i+1}\alpha}{\pm P_{i+1}\alpha} \mid <+\alpha,+P_1,...,\pm P_k> \in \Phi,\ 1 \le i<k \right\}$. Then, by the semi-monotonicity of normal default theories, there will be an extension, $E \supseteq E'$ for Δ, since $D' \subseteq CLOSED\text{-}DEFAULTS(\Delta)$ [Reiter 1980a, theorem 3.2].

As usual, we show that $E' = \bigcup\limits_{i=0}^{\infty} E_i$.

$E' \supseteq \bigcup\limits_{i=0}^{\infty} E_i$: Consider $w = \pm R\alpha \in E_0 = W = \{\pm R\alpha \mid <+\alpha,\pm R> \in \Gamma\}$. Then $<+\alpha,\pm R> \in \Gamma \subseteq \Phi$, so $<+\alpha,\pm R> \in facts(\Phi)$, so $\pm R\alpha \in facts'(\Phi)$. For the inductive step, assume $E_i \subseteq E'$, and consider $w \in E_{i+1}$. If $w \in Th(E_i)$, then $w \in E'$. Otherwise, $w \in \{P_{i+1}\alpha \mid \delta = \dfrac{P_i\alpha : P_{i+1}\alpha}{P_{i+1}\alpha} \in D',\ P_i\alpha \in E_i, \text{ and } \neg P_{i+1}\alpha \notin E'\}$. Since $\delta \in D'$, we have $<+\alpha,+P_1,...,\pm P_k> \in \Phi$, for some $k \ge i+1$. Hence $<+\alpha,+P_i>$, $<+\alpha,\pm P_{i+1}> \in \Phi$, since Φ is a grounded expansion. So $P_i\alpha$, $\pm P_{i+1}\alpha \in facts'(\Phi)$.

$E' \subseteq \bigcup\limits_{i=0}^{\infty} E_i$: Consider $\pm R\alpha \in facts'(\Phi)$. Then $<+\alpha,+R_1,...,+R_j,\pm R> \in \Phi$. By [Touretzky 1984a, theorem 2.3], $<+\alpha,+R_1>$,..., $<+\alpha,+R_j>$, $<+\alpha,\pm R> \in facts(\Phi)$, so $R_1\alpha$,..., $R_j\alpha$, $\pm R\alpha \in facts'(\Phi)$. If $<+\alpha,+R_1> \in \Phi$, then $<+\alpha,+R_1> \in \Gamma$, by [Touretzky 1984a, theorems 2.3, 2.2], so $R_1\alpha \in W \subseteq \bigcup\limits_{i=0}^{\infty} E_i$. For the inductive step, assume $R_1\alpha$,..., $R_k\alpha \in \bigcup\limits_{i=0}^{\infty} E_i$, for $k<j$. We show that $R_{k+1}\alpha \in \bigcup\limits_{i=0}^{\infty} E_i$. Now $\delta = \dfrac{R_k\alpha : R_{k+1}\alpha}{R_{k+1}\alpha} \in D'$. Since $R_k\alpha \in \bigcup\limits_{i=0}^{\infty} E_i$, $R_k\alpha \in E_i$, for some i. Since $<+\alpha,+R_1,...,+R_{k+1}> \in \Phi$, $<+\alpha,R_{k+1}> \in C(\Phi)$ so $<+\alpha,R_{k+1}> \in facts(\Phi)$, so $R_{k+1}\alpha \in facts'(\Phi)$. By the consistency of Γ, $E' \not\vdash \neg R_{k+1}\alpha$, so $R_{k+1}\alpha \in E_{i+1}$. So $R_k\alpha \in \bigcup\limits_{i=0}^{\infty} E_i$ for $1 \le k \le n$, by induction. Similarly for $\pm R\alpha$.

Thus $E' = \bigcup\limits_{i=0}^{\infty} E_i$. So E' is an extension for Δ', by Theorem 0.1.

QED Theorem 4.5

The proof of Theorem 5.1 follows immediately from McCarthy's proof of the soundness of predicate circumscription and the definition of well-foundedness.

Theorem 5.2

Universal theories are well-founded.

Proof

The proof is identical to that of Property 1.3.2 in [Bossu and Seigel 1985]. The definition of submodel used there is less restrictive than that used here, but this does not alter the form of the proof.

QED Theorem 5.2

Theorem 5.4

If T is a well-founded theory, $\vec{\alpha}_1,...,\vec{\alpha}_k$ are n-tuples of ground terms, and $P \in \mathbf{P}$, is an n-ary predicate, then

$$CLOSURE_{\mathbf{P}}(T) \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k .$$

Proof

The right-to-left direction is immediate. We prove the contrapositive of the left-to-right direction. Assume that $CLOSURE_{\mathbf{P}}(T) \vdash \bigvee_{i=1}^{k} P\vec{\alpha}_i$ and $T \not\vdash \bigvee_{i=1}^{k} P\vec{\alpha}_i$. Then T has a model, M, in which $P\vec{\alpha}_i$ is false, for all $i = 1,...,k$. Since T is well-founded, there is a **P**-minimal submodel, M', of M. Furthermore, since the circumscription is true in all **P**-minimal submodels, $P\vec{\alpha}_i$ is true in M', for some $1 \leq i \leq k$. But then M' is not a **P**-submodel of M, and this contradicts the fact that M' is a **P**-minimal submodel of M. Therefore $CLOSURE_{\mathbf{P}}(T) \not\vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k$.

QED Theorem 5.4

Theorem 5.5

If T is a well-founded theory, $\vec{\alpha}_1,...,\vec{\alpha}_k$ are n-tuples of ground terms, and $P \notin \mathbf{P}$ is an n-ary predicate, then

(i) $CLOSURE_{\mathbf{P}}(T) \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee ... \vee P\vec{\alpha}_k$, and

(ii) $$CLOSURE_{\mathbf{P}}(T) \vdash \neg P\vec{\alpha}_1 \vee \ldots \vee \neg P\vec{\alpha}_k \Longleftrightarrow T \vdash \neg P\vec{\alpha}_1 \vee \ldots \vee \neg P\vec{\alpha}_k .$$

Proof

(i) The right-to-left direction is immediate. We prove the contrapositive of the left-to-right direction. Assume $T \not\vdash \bigvee_{i=1}^{k} P\vec{\alpha}_i$. Then there is a model, M, for T in which $P\vec{\alpha}_i$ is false, for all i=1,...,k. Since T is well-founded, there is a **P**-minimal submodel, M', of M. By the definition of submodel, the interpretation of P remains the same in M and M', since $P \notin \mathbf{P}$. Hence $P\vec{\alpha}_i$ is false in M', for all i=1,...,k . Since the circumscription schema is satisfied by all minimal models, $CLOSURE_{\mathbf{P}}(T) \not\vdash \bigvee_{i=1}^{k} P\vec{\alpha}_i$. The proof for (ii) is similar.

QED Theorem 5.5

In the proofs of Theorems 5.6 and 5.7 we use the following notational conventions:

1. $SCHEMA(T,\mathbf{P})$ is the circumscription schema resulting from circumscribing the predicates of **P** in T.
2. $CLOSURE_{\{\,\}}(T) = T$. (The closure of T with respect to the empty set of predicates is defined to be T itself.)
3. If M is a model, $\bigwedge_{i=1}^{0} Q_i$ is true in M. (The empty conjunction is vacuously true in all models.)

Theorem 5.6 (Reiter)

> If T is an arbitrary, finitely-axiomatized theory containing axioms which define the equality predicate, =, then $T \vdash CLOSURE_{\{\text{'='}\}}(T)$.

Proof

Consider the schema resulting from circumscribing '=' in T:

$$SCHEMA(T,\{\text{'='}\}) = [T(\Phi) \wedge \forall xy.\ \Phi xy \supset x = y] \supset \forall xy.\ x = y \supset \Phi xy$$

First, observe that $\vdash (\forall x.\ \Psi xx) \supset (\forall xy.\ x = y \supset \Psi xy)$ for any predicate letter, Ψ. Furthermore, $\forall x.\Phi xx$ is one of the conjuncts of $T(\Phi)$ in $SCHEMA(T,\{\text{'='}\})$ since $\forall x.\ x = x$ must be an axiom of any theory with equality. Thus if any instance of $T(\Phi)$ is true in a model of T, so is the corresponding instance of $\forall xy.\ x = y \supset \Phi xy$. Hence, every instance of $SCHEMA(T,\{\text{'='}\})$ is true in every model of T, so $T \vdash CLOSURE_{\{\text{'='}\}}(T)$.

QED Theorem 5.6

Theorem 5.7

If T is a well-founded theory containing axioms which define the equality predicate; and $\vec{\alpha}$, $\vec{\beta}$ are tuples of ground terms; then

(i) $CLOSURE_{\mathbf{P}}(T) \vdash \vec{\alpha} = \vec{\beta} \Longleftrightarrow T \vdash \vec{\alpha} = \vec{\beta}$, and
(ii) $CLOSURE_{\mathbf{P}}(T) \vdash \vec{\alpha} \neq \vec{\beta} \Longleftrightarrow T \vdash \vec{\alpha} \neq \vec{\beta}$.

Proof

(i) This is a corollary of Theorems 5.4 and 5.5(i).
(ii) The right-to-left direction is immediate. To prove the left-to-right direction, we consider the composition of $\mathbf{P}$. If '=' does not occur in $\mathbf{P}$, the result follows directly from Theorem 5.5(ii). If $\mathbf{P} = \{=\}$, the result follows from Theorem 5.6. Finally, consider $\mathbf{P} = \mathbf{P}' \cup \{\text{'='}\}$, for an arbitrary set of predicates $\mathbf{P}' = \{P_1,...,P_n\}$ not including equality. By Theorem 5.5(ii),

$$CLOSURE_{\mathbf{P}'}(T) \vdash \vec{\alpha} \neq \vec{\beta} \Longleftrightarrow T \vdash \vec{\alpha} \neq \vec{\beta}$$

We show that

$$CLOSURE_{\mathbf{P}}(T) \vdash \vec{\alpha} \neq \vec{\beta} \Longleftrightarrow T \vdash \vec{\alpha} \neq \vec{\beta}\ .$$

We have

$$SCHEMA(T,\mathbf{P}') = \left[T(\Phi_1,...,\Phi_n) \wedge \left(\bigwedge_{i=1}^{n} (\forall \vec{x}.\ \Phi_i\vec{x} \supset P_i\vec{x})\right)\right] \supset \bigwedge_{i=1}^{n} (\forall \vec{x}.\ P_i\vec{x} \supset \Phi_i\vec{x})$$

$$SCHEMA(T,\mathbf{P}) = \left[T(\Phi_1,...,\Phi_n,\Psi) \wedge \left(\bigwedge_{i=1}^{n} (\forall \vec{x}.\ \Phi_i\vec{x} \supset P_i\vec{x})\right) \wedge (\forall xy.\ \Psi xy \supset x = y)\right] \supset \left[\bigwedge_{i=1}^{n} (\forall \vec{x}.\ P_i\vec{x} \supset \Phi_i\vec{x}) \wedge \forall xy.\ x = y \supset \Psi xy\right].$$

Assume $CLOSURE_{\mathbf{P}}(T) \vdash \vec{\alpha} \neq \vec{\beta}$, and $T \not\vdash \vec{\alpha} \neq \vec{\beta}$. It follows that $CLOSURE_{\mathbf{P}'}(T) \not\vdash \vec{\alpha} \neq \vec{\beta}$. Any model of T in which every instance of $SCHEMA(T,\mathbf{P})$ is true is also a model for $CLOSURE_{\mathbf{P}}(T)$. Hence $\vec{\alpha} \neq \vec{\beta}$ is true in that model. Furthermore, there is some model of T in which every instance of $SCHEMA(T,\mathbf{P}')$ is true and $\vec{\alpha} \neq \vec{\beta}$ is false. We show that in every model of T in which every instance of $SCHEMA(T,\mathbf{P}')$ is true, every instance of $SCHEMA(T,\mathbf{P})$ is also true. First observe that $\vdash (\forall x.\ \Psi xx) \supset (\forall xy.\ x = y \supset \Psi xy)$ for any predicate letter, Ψ. Furthermore, $\forall x.\ \Psi xx$ is one of the conjuncts of $T(\Phi_1,...,\Phi_n,\Psi)$ in $SCHEMA(T,\mathbf{P})$. Thus if any instance of $T(\Phi_1,...,\Phi_n,\Psi)$ is true in a model of T, so is the corresponding instance of $\forall xy.\ x = y \supset \Psi xy$. Let M be a model of T where every instance of $SCHEMA(T,\mathbf{P}')$ is true. Consider an instance, I, of $SCHEMA(T,\mathbf{P})$, with the predicates Φ_i' and Ψ' substituted for Φ_i and Ψ, respectively. There are two cases:

1) $\bigwedge_{i=1}^{n} (\forall \vec{x}.\ P_i\vec{x} \supset \Phi_i'\vec{x})$ is true in M. By the observation above, either $T(\Phi_1',...,\Phi_n', \Psi')$ is false in M or $\forall xy.\ x = y \supset \Psi'xy$ is true. In either case, I is true in M.

2) $\bigwedge_{i=1}^{n} (\forall \vec{x}.\ P_i\vec{x} \supset \Phi_i'\vec{x})$ is false in M. But then $T(\Phi_1',...,\Phi_n')$ is false or

$\bigwedge\limits_{i=1}^{n} (\forall \vec{x}.\ \Phi_i'\vec{x} \supset P_i\vec{x})$ is false, since every instance of *SCHEMA*(T,$\mathbf{P}'$) is true in M. In the latter case I is also true in M. In the former case, if $[T(\Phi_1',...,\Phi_n'\ ,\ \Psi') \wedge \forall xy.\ \Psi'xy \supset x = y]$ is false in M, then I is true. Otherwise, by the observation above, $\forall xy.\ x = y \supset \Psi'xy$ is true and, hence, so is $\forall xy.\ x = y \equiv \Psi'xy$. But $T(\Phi_1',...,\Phi_n',\Psi')$ is the result of substituting Ψ' for some of the occurrences of '=' in $T(\Phi_1',...,\Phi_n')$, so $T(\Phi_1',...,\Phi_n',\Psi')$ is false, because $T(\Phi_1',...,\Phi_n')$ is, and this is a contradiction.

Thus, for every model of T, if *SCHEMA*(T,$\mathbf{P}'$) is true, so is *SCHEMA*(T,$\mathbf{P}$). But then $\vec{\alpha} \neq \vec{\beta}$ is true in every model of $CLOSURE_{\mathbf{P}'}(T)$. Hence $CLOSURE_{\mathbf{P}'}(T) \vdash \vec{\alpha} \neq \vec{\beta}$, which is a contradiction, since $T \not\vdash \vec{\alpha} \neq \vec{\beta}$. We conclude that $CLOSURE_{\mathbf{P}}(T) \not\vdash \vec{\alpha} \neq \vec{\beta}$.

QED Theorem 5.7

Corollary 5.8

If T is a well-founded theory containing axioms which define the equality predicate, P is an n-ary predicate, and $\vec{\alpha}$ is an n-tuple of ground terms, then $CLOSURE_{\mathbf{P}}(T) \vdash \neg P\vec{\alpha}$ implies $T \vdash \vec{\alpha} \neq \vec{\beta}$ for all ground n-tuples $\vec{\beta}$ such that $T \vdash P\vec{\beta}$.

Proof

Otherwise $CLOSURE_{\mathbf{P}}(T) \vdash \vec{\alpha} \neq \vec{\beta}$ and $T \not\vdash \vec{\alpha} \neq \vec{\beta}$ which contradicts Theorem 5.7.

QED Corollary 5.8

Theorem 5.9

If T is a well-founded theory; $\alpha_1,...,\alpha_n$ are ground terms; and $\mathbf{P}$ is a set of some of the predicate symbols of T; then

$$CLOSURE_{\mathbf{P}}(T) \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n \Longleftrightarrow T \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n\ .$$

Proof

The right-to-left direction is immediate. For the left-to-right direction, assume that $T \not\vdash \forall x. x = \alpha_1 \vee ... \vee x = \alpha_n$. Then T has a model which falsifies $\forall x. x = \alpha_1 \vee ... \vee x = \alpha_n$. Since T is well-founded, this model has a $\mathbf{P}$-minimal submodel. But $\forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n$ is false in this submodel, because the extension of the equality predicate in this submodel must be a subset of its extension in the original model. Since the circumscription is true in all minimal models,

$CLOSURE_{\mathbf{P}}(T) \not\vdash \forall x.\ x = \alpha_1 \vee \ldots \vee x = \alpha_n$.

QED Theorem 5.9

Theorem 5.10

> If T is a well-founded theory, and T has a model with some domain, D, then so does $CLOSURE_{\mathbf{P}}(T)$.

Proof

McCarthy [1980] shows that $CLOSURE_{\mathbf{P}}(T)$ is true in all minimal models. Since T is well-founded, every model has a minimal submodel. By the definition of submodel, the domain of a minimal submodel of M is the same as that of M.

QED Theorem 5.10

Theorem 5.11

> If $T \vdash \forall \vec{x}.\ P\vec{x} \equiv \Phi\vec{x}$ for some expression $\Phi\vec{x}$, not involving predicate letters from $\mathbf{P}$, then $T \vdash CLOSURE_{\mathbf{P}}(T)$.

Proof

$T(\Psi)$, on the left-hand side of the circumscription schema, includes $\forall \vec{x}.\ \Psi\vec{x} \equiv \Phi\vec{x}$. But any choice of model, M, and predicate, Ψ, which satisfies the LHS clearly already satisfies the RHS, $\forall \vec{x}.\ P\vec{x} \supset \Psi\vec{x}$, since every model of T satisfies $\forall \vec{x}.\ \Phi\vec{x} \equiv P\vec{x}$.

QED Theorem 5.11

Definition: Formula Circumscription

The *circumscription of the formula* $E(\mathbf{P},\vec{x})$ in the theory T, with the predicates $\mathbf{P}$ treated as variable, is given by:

$$T(\mathbf{P}) \wedge \forall \Phi.\ T(\Phi) \wedge [\forall \vec{x}.\ E(\Phi,\vec{x}) \supset E(\mathbf{P},\vec{x})] \supset [\forall \vec{x}.\ E(\mathbf{P},\vec{x}) \supset E(\Phi,\vec{x})]$$

Definition: $M \leq_{E(\mathbf{P},\vec{x})} M'$

Let $T(\mathbf{P})$ be a finitely-axiomatized (first- or second-order) theory, some (but not necessarily all) of whose predicates are those in $\mathbf{P}$; let $E(\mathbf{P},\vec{x})$ be a formula whose free variables are among $\vec{x} = x_1,\ldots,x_n$, and in which some of the predicate variables $\mathbf{P} = \{P_1,\ldots,P_n\}$ occur free; and let M, M' be models of T. We say M is an $E(\mathbf{P},\vec{x})$-*submodel* of M' (written $M \leq_{E(\mathbf{P},\vec{x})} M'$) iff

(i) $|M| = |M'|$,

(ii) If t is a term, then $|t|_M = |t|_{M'}$,

(iii) If $Q \notin \mathbf{P}$ is a predicate letter of T, then $|Q|_M = |Q|_{M'}$, and

(iv) $|E(\mathbf{P},\vec{x})|_M \subseteq |E(\mathbf{P},\vec{x})|_{M'}$. ■

Definition: $E(\mathbf{P},\vec{x})$-Minimal Model

A model, M, of T is $E(\mathbf{P},\vec{x})$-*minimal* iff T has no model, M', such that $M' \leq_{E(\mathbf{P},\vec{x})} M$ and $\neg(M \leq_{E(\mathbf{P},\vec{x})} M')$. ■

Theorem 6.1 – Soundness

CLOSURE(T; $\mathbf{P}$; $E(\mathbf{P},\vec{x})$) is satisfied by every $E(\mathbf{P},\vec{x})$-minimal model of T.

Proof

The proof follows McCarthy's [1980] proof of the soundness of predicate circumscription. Consider a minimal model, M, and an instantiation, with some predicate, Φ, of the schema (or second-order axiom) which makes the left-hand side true and the RHS false. Then by the second conjunct of the LHS, $|E(\mathbf{P},\vec{x})|_M \subseteq |E(\Phi,\vec{x})|_{M'}$ But then a proper submodel, M', could be constructed by letting $\mathbf{P}$ agree with Φ. But this contradicts the fact that M is minimal.

QED Theorem 6.1

Theorem 6.3

The ability to minimize arbitrary expressions, $E(\mathbf{P},\vec{x})$, instead of simple sets of predicates, is an inessential extension, provided predicates other than those being minimized are allowed to vary.

Proof

We show that the theory, T, can be extended by adding a new predicate symbol, Ψ, and the definition $\forall\vec{x}.\ \Psi\vec{x} \equiv E(\mathbf{P},\vec{x})$, and that circumscribing Ψ in the extended theory, T', with $\mathbf{P}$ variable is equivalent to circumscribing $E(\mathbf{P},\vec{x})$ in the original theory. *I.e.*, that

$$T \wedge \left[T(\Phi) \wedge [\forall\vec{x}.\ E(\Phi,\vec{x}) \supset E(\mathbf{P},\vec{x})] \right] \supset [\forall\vec{x}.\ E(\mathbf{P},\vec{x}) \supset E(\Phi,\vec{x})] \tag{27}$$

and

$$T' \wedge \left[T(\Phi,\psi) \wedge [\forall\vec{x}.\ \psi\vec{x} \equiv E(\Phi,\vec{x})] \wedge [\forall\vec{x}.\ \psi\vec{x} \supset \Psi\vec{x}] \right] \supset [\forall\vec{x}.\ \Psi\vec{x} \supset \psi\vec{x}] \tag{28}$$

are equivalent over the language of T.

To see that (27) entails (28), let M be a model which satisfies (27). Since (27) does not mention Ψ, we can interpret Ψ as we choose. Therefore, let $|\Psi|_M = |E(\mathbf{P})|_M$. Clearly, $M \models$ (28). Conversely, let M satisfy (28), and let Φ,ψ be a tuple of predicate variables satisfying the LHS of (28). Clearly, $T' \vdash T$, and $T'(\Phi) \vdash T(\Phi)$. By substitution of equivalents, we get the rest of (27), so $M \models$ (27).

QED Theorem 6.3

Definition: Generalized Circumscription

Let $\mathbf{X}$ be a tuple of predicate, function, and/or constant symbols, and let R be a binary relation on tuples of type $\mathbf{X}$. The *generalized circumscription* of $\mathbf{X}$ in the theory, T, according to the pre-order, $\leq_R$, induced by R is given by:

$$T(\mathbf{X}) \wedge \forall\mathbf{X}'.\ T(\mathbf{X}') \wedge (\mathbf{X}' \leq_R \mathbf{X}) \supset (\mathbf{X} \leq_R \mathbf{X}')$$

Definition: $M \leq_{(\mathbf{X},R)} M'$

Let $T(\mathbf{P})$ be a finitely axiomatized (first- or second-order) theory, whose predicate, function and constant letters include (but need not be limited to) those in $\mathbf{X}$; let R be a binary relation on tuples of type $\mathbf{X}$; let $\leq_R$ be the pre-order induced by R; and let M, M' be models of T. Then M is an $(\mathbf{X},R)$-*submodel* of M' (written $M \leq_{(\mathbf{X},R)} M'$) iff

(i) $|M| = |M'|$,

(ii) If t is a term and $t \notin \mathbf{X}$, then $|t|_M = |t|_{M'}$,

(iii) If $Q \notin \mathbf{X}$ is a predicate letter of T, then $|Q|_M = |Q|_{M'}$, and
(iv) $<|\mathbf{X}|_M, |\mathbf{X}|_{M'}> \in R$. ∎

Definition: $(\mathbf{X},R)$-Minimal Model

A model, M, of T is $(\mathbf{X},R)$-*minimal* iff T has no model, M', such that $M' \leq_{(\mathbf{X},R)} M$ and $\neg(M \leq_{(\mathbf{X},R)} M')$. ∎

Theorem 6.4 – Soundness

CLOSURE(T; $\mathbf{X}$; R) is satisfied by every $(\mathbf{X},R)$-minimal model of T.

The proof is similar to that of Theorem 6.1, except that the interpretations of each of the variable terms must also be set. ∎

Definition: Well-Foundedness

The theory, T, is *well-founded with respect to* $(\mathbf{X},R)$ iff every model of T has an $(\mathbf{X},R)$-minimal submodel. ∎

Theorem 6.9

If T is a universal theory, and $\mathbf{X}$, $\mathbf{P}$ are finite tuples of predicate letters, then T is well-founded with respect to $\leq_{(\mathbf{X},\mathbf{P})}$.

Proof

We show that any chain of submodels of a model of T has a lower bound among the submodels of that model. It follows by Zorn's lemma that every model has a minimal submodel.

Let M_0,... be a chain of models of T, ordered under the submodel relation. If the chain is finite, it has a lower bound, hence assume it is infinite.

Let $\{\mathbf{d}_1,...\}$ be the elements of $|M_0|$. Extend the language of T, L, to L' by adding a new constant symbol, d_i, for each $\mathbf{d}_i$. Let $T' = T \cup \{P\vec{d} \mid \text{for all } i, M_i \models P\vec{\mathbf{d}}\} \cup \{\neg P\vec{d} \mid \text{for some } i, M_i \not\models P\vec{\mathbf{d}}\}$.

Assume T' is inconsistent. Then, by compactness, so is a finite subset. But then some M_i must set each $P\vec{a}$ in this finite set accordingly, so $M_i \not\models T$, which is a contradiction, since the chain $\{M_i\}$ is ordered. Hence T' is consistent, so T' has a model, M'.

Now we can add the diagrams (over all ground terms of L') of the equality predicate and all fixed predicates from M_0 to T' to get T''. By the above argument, T'' must be consistent. Hence there is an M'' such that $M'' \models T''$. By virtue of the fact that M'' satisfies the diagram of the equality predicate from M_0, we can isomorphically embed the domain of M_0 into M''. (Because T'' contains the diagrams of the equality predicate over all ground terms of L', it is clear that the resulting substructure is closed under and preserves the functions.) Finally, since $T'' \supseteq T$, $M'' \models T$.

Since T is a universal theory, the restriction, M, of M'' to $|M_0|$ is a model of T. Clearly $M \leq_{(\mathbf{X},\mathbf{P})} M_i$, for all i, so M is the lower bound we require.

QED Theorem 6.9

Theorem 6.11

If T is well-founded with respect to $(\mathbf{X}, \mathbf{P})$; $P \in \mathbf{P}$ is an n-ary predicate; $\mathbf{X}$ is a set of predicate letters; and $\vec{\alpha}_1,\ldots,\vec{\alpha}_k$ are n-tuples of ground terms; then

$$CLOSURE(T;\mathbf{X};\mathbf{P}) \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k . \quad \blacksquare$$

Theorem 6.12

If T is well-founded with respect to $(\mathbf{X}, \mathbf{P})$; $\mathbf{X}$ is a set of predicate letters; $P \notin \mathbf{P} \cup \mathbf{X}$ is an n-ary predicate; and $\vec{\alpha}_1,\ldots,\vec{\alpha}_k$ are n-tuples of ground terms; then

(i) $CLOSURE(T;\mathbf{X};\mathbf{P}) \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k \Longleftrightarrow T \vdash P\vec{\alpha}_1 \vee \ldots \vee P\vec{\alpha}_k$, and

(ii) $CLOSURE(T;\mathbf{X};\mathbf{P}) \vdash \neg P\vec{\alpha}_1 \vee \ldots \vee \neg P\vec{\alpha}_k \Longleftrightarrow T \vdash \neg P\vec{\alpha}_1 \vee \ldots \vee \neg P\vec{\alpha}_k . \ \blacksquare$

Theorem 6.13

If T is well-founded for $(\mathbf{P},R)$ and T has a model with domain D, then so does $CLOSURE(T(\mathbf{P});\mathbf{P};R)$. $\blacksquare$

Theorem 6.14

If T is a first-order theory containing axioms which define the equality predicate, $=$, then $T \vdash CLOSURE(T;\mathbf{X};\{=\})$. $\blacksquare$

The proofs of Theorems 6.11, 6.12, 6.13, and 6.14 are essentially alphabetic variants of those of Theorems 5.4, 5.5, 5.10, and 5.6, respectively. We do not repeat them here.

Theorem 7.1 – Soundness

> Every instance of the revised domain circumscription schema for a theory, T, is true in all minimal models of T.

Proof

The proof is identical to that presented in [Davis 1980, p75], except that, in the proof of the lemma, the revised schema guarantees that D_0 is non-empty and hence N is well-defined.

QED Theorem 7.1

Theorem 7.3

> If T is a well-founded theory which contains axioms which define the equality predicate, $=$, and $\alpha_1,\ldots,\alpha_n$, $\beta_1,\ldots,\beta_n$ are ground terms, then
>
> (i) $T \vdash (\bigvee_{i=1}^{n} \alpha_i = \beta_i) \iff DC(T) \vdash (\bigvee_{i=1}^{n} \alpha_i = \beta_i)$
>
> (ii) $T \vdash (\bigvee_{i=1}^{n} \alpha_i \neq \beta_i) \iff DC(T) \vdash (\bigvee_{i=1}^{n} \alpha_i \neq \beta_i)$

Proof

Every model of a well-founded theory has a minimal submodel. Let M' be a model of T. Let $M \leq M'$ be a minimal submodel of M'. Thus M and M' agree on all ground terms, and M is the restriction of M' to a smaller domain. But then clearly they must have the same set of ground (in)equalities, since new equalities imply that M is *not* a restriction of M', and new inequalities imply that $|M|$ does not contain the interpretation of some of the ground terms (since M is a restriction of M'), which is false.

QED Theorem 7.3

Theorem 7.4 – Finitary Completeness

> If T is a finitely axiomatizable theory, and every model of T is finite, then only the minimal models of T satisfy every instance of the domain circumscription schema for T, $DC(T)$.

Proof

Assume every model of T is finite. Consider some non-minimal model, M. We assume that every instance of $DC(T)$ is true in M and arrive at a contradiction.
M is finite, with m elements in its domain. Since M is not minimal, there is a submodel, $N < M$,

with $n < m$ domain elements. Let Φx be $x = x_1 \vee ... \vee x = x_n$ where the x_i's are variables. We can instantiate these x_i's in M to be the n elements which survive the submodeling to N. Clearly $\exists x.\ \Phi x$ is true, as is $AXIOM(\Phi)$. A^Φ must be true, as follows: Consider an arbitrary expression, Ψx. $\forall x.\ \Psi x \supset [\forall x.\ \Phi x \supset \Psi x]$, and the existentials given by T must be satisfied in N (since N is a model). Furthermore, Φ is true for all of $|N|$. Thus $[\exists x.\ \Psi x] \in T$ will mean that $\exists x.\ \Phi x \wedge \Psi x$ will be true in M. But since $n < m$, $\forall x.\ \Phi x$ is clearly false in M, so we have a falsifying instance of the schema.

QED Theorem 7.4

Corollary 7.5

> If T is a finitely axiomatizable theory, and every model of $DC(T)$ is finite, then only the minimal models of T satisfy every instance of $DC(T)$.

Proof

$DC(T)$ is true in all minimal models, so there are no infinite minimal models. $DC(T)$ false in all infinite models, so only finite non-minimal models remain to be eliminated. Every finite model has a minimal submodel (there can't be an infinite chain of proper submodels). The argument for Theorem 7.4 serves to rule out non-minimal finite models.

QED Corollary 7.5

Theorem 8.2

If $T \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n$ and $T \vdash \alpha_i \neq \alpha_j$, for $i \neq j$ for ground terms $\alpha_1,...,\alpha_n$; and **X** includes all of the predicates of L; then those formulae true in every extension of $\Delta = \left[\left\{\frac{:\neg Px}{\neg Px}\right], T\right]$ are precisely those entailed by $CLOSURE(T; \mathbf{X}; \{P\})$.

Proof

Lemma 8.2.1 shows that every model for any extension of Δ is $(\mathbf{X},\{P\})$-minimal. Lemma 8.2.2 shows that every $(\mathbf{X},\{P\})$-minimal model of T is a model for some extension of Δ. From these the result follows.

QED Theorem 8.2

Lemma 8.2.1

If $T \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n$ for ground terms $\alpha_1,...,\alpha_n$; and **X** includes all of the predicates of L; then any model of any extension of $\Delta = \left[\left\{\frac{:\neg Px}{\neg Px}\right], T\right]$ is an $(\mathbf{X},\{P\})$-minimal model for T.

Proof

Any model, M for an extension, E, for Δ has domain $|M| = \bigcup_{i=1}^{n} \{|\alpha_i|_M\}$. Assume that M is not minimal. Then there is an $M' < M$. Without loss of generality, assume $|P|_M = \{\alpha_1,...,\alpha_k \mid 0 < k \leq n\}$, and $|P|_M{}' = \{\alpha_1,...,\alpha_r \mid 0 \leq r < k\}$. ($k>0$ or there is no $M' < M$.) Now, given the existence of M', it is clear that $E \not\vdash P\alpha_k$ so $\neg P\alpha_k$ must be in E, so $M \not\models E$, which is a contradiction. Hence, M is minimial.

QED Lemma 8.2.1

Lemma 8.2.2

If $T \vdash \forall x.\ x = \alpha_1 \vee ... \vee x = \alpha_n$ and $T \vdash \alpha_i \neq \alpha_j$, for $i \neq j$ for ground terms $\alpha_1,...,\alpha_n$; and **X** includes all of the predicates of L; then any $(\mathbf{X},\{P\})$-minimal model for T is a model of some extension of $\Delta = \left[\left\{\frac{:\neg Px}{\neg Px}\right], T\right]$.

Proof

We construct the extension, E, from the minimal model, M. Clearly $M \models T$. If $M \models \neg P\alpha_i$, put $\frac{: \neg P\alpha_i}{\neg P\alpha_i}$ in $GD(E,\Delta)$. Obviously, $T \cup CONSEQUENTS(GD(E,\Delta))$ then entails $P\alpha_j$ for each α_j such that $P\alpha_j \notin CONSEQUENTS(GD(E,\Delta))$. (Otherwise M is not minimal). The existence of M guarantees that $E \not\vdash P\alpha_i$ for the α_i's which make up GD. Thus $E = Th(T \cup CONSEQUENTS(GD(E,\Delta))$ is an extension for Δ. Clearly $M \models E$.

QED Lemma 8.2.2

Proposition 8.6

If T does not entail a domain-closure axiom, and $T \not\vdash \forall x.\ \neg Px$, then every extension for Δ has models which are not $(\mathbf{X}, \{P\})$-minimal.

The proof of this proposition lies in the observation that one can always set $\neg P\alpha$ for some domain element α which does not correspond to any term in the language. Since T does not entail a domain closure axiom, a model with such an element will always exist. ■

Theorem 8.7

There are theories, T, such that $T \vdash \forall x.\ x = \alpha_1 \vee \ldots \vee x = \alpha_n$ and $T \vdash \alpha_i \neq \alpha_j$, for $i \neq j$ and yet no combination of the extensions of $\Delta = \left[\left\{ \frac{: \neg Px}{\neg Px} \right\}, T \right]$ precisely characterizes the $(\mathbf{X}, \{P\})$-minimal models of T.

The proof of this theorem follows from Example 8.2. ■

Proposition 8.9

If there are no variable predicates ($Z = \{\ \}$), then $ECWA(T)$ adds to T every instance of the circumscription schema.

The proof of this follows directly from of the third corollary to Gelfond, Prsymusinska, and Prsymusinski's [1985] theorem 1. ■

APPENDIX B

Dictionary of Symbols

Symbol	**Definition**
$\in$	Set membership
$\notin$	Set non-membership
$\cup$	Set union
$\cap$	Set intersection
$\{\ \}$	The empty set
$-$	Set difference: $\Psi - \Gamma = \{\alpha \mid \alpha \in \Psi \text{ and } \alpha \notin \Gamma\}$
Δ	Symmetric set difference: $\Psi \ \Delta\ \Gamma \equiv (\Psi - \Gamma) \cup (\Gamma - \Psi)$
$\vdash$	First-order provability
$\nvdash$	First-order non-provability
$\models$	Logical entailment
$\not\models$	Logical non-entailment
$\supset$	Logical implication
$\neg$	Logical negation
$\wedge$	Logical and
$\vee$	Logical or
$\equiv$	Logical equivalence
$\exists$	Existential quantifier
$\forall$	Universal quantifier
.	Preceding quantifier's scope extends over 1st enclosing formula.
$\Box$	The null clause
$\perp$	Contradiction
Th	Logical closure operator
$\Rightarrow$	"It follows that" or "Implies"
iff , $\Longleftrightarrow$	If and only if
$\ll$	Strong precedence relation on *Literals* × *Literals*
$\underline{\ll}$	Weak precedence relation on *Literals* × *Literals*
$\mapsto$	Function mapping
L	The first-order language (*i.e.*, all well-formed formulae)
$\mathbf{N}$	The set of all Natural numbers
Literals	The set of all atomic formulae and their negations
■	Marks end of definition, example, or theorem

APPENDIX C

Useful Logical Definitions

Clause – A *clause* is a finite disjunction of literals.

Closed Forumula – A formula is *closed* iff it contains no free variables.

Ground – An expression (literal, term, or formula) is *ground* iff it contains no variables.

Herbrand Universe – If T is a universal theory, then the *Herbrand Universe* of T is $H(T) = \{f^n(t_1,...,t_n) \mid f^n$ is an n-ary function-letter of T, and $t_1,...,t_n \in H(T)\}$. (This is well-defined because the 0-ary function-letters (or constants) provide the base for the recursion.)

Herbrand Base – If T is a universal theory, then the *Herbrand Base* of T is $\hat{H}(T) = \{P^n(t_1,...,t_n) \mid P^n$ is an n-ary predicate-letter of T, and $t_1,...,t_n \in H(T)\}$.

Herbrand Interpretation – If T is a universal theory, then a *Herbrand Interpretation*, I, of T is a subset of T's Herbrand base, $\hat{H}(T)$. Those atomic formulae $P^n(t_1,...,t_n) \in I$ are interpreted as *true* in I, all others are interpreted as *false*.

Herbrand Model – If T is a universal theory, then a *Herbrand Model* of T is a Herbrand interpretation of T which satisfies every formula in T, according to the usual definition of satisfaction by an interpretation.

Horn – A set of clauses, T, is *Horn* iff every clause in T contains at most one positive literal.

Literal – A *literal* is an atomic formula or the negation of an atomic formula.

Skolemized form – The *Skolemized form* of a theory is the theory obtained by converting to prenex-normal form then progressively, from the right-most quantifier, replacing each existentially quantified variable by a *new* function-symbol taking as arguments each of the variables captured by quantifiers occurring further to the left. The process of obtaining the skolemized form of a theory is called *skolemization.*