Chapter 15

Nonmonotonic Logic

John F. Horty

15.1. Introduction

The goal of a logic is to define a consequence relation between a set of formulas Γ and, in most cases, an individual formula A. This definition generally takes one of two forms. From a proof theoretic standpoint, A is said to be a consequence of Γ whenever there is a deduction of A from the set Γ, viewed as a set of premises; from a model theoretic standpoint, A is said to be a consequence of Γ whenever A holds in every model that satisfies each formula in Γ.

Although the detailed inferences sanctioned by particular logics vary widely depending on the connectives present and the properties attributed to them, certain abstract features of the consequence relation are remarkably stable across logics. Among these is the property of ***monotonicity***: if A is a consequence of Γ, then A is a consequence of $\Gamma \cup \{B\}$. What this means is that any conclusion drawn from a set of premises will be preserved as a conclusion even if the premise set is supplemented with additional information – that the set of conclusions grows monotonically as the premise set grows.

The monotonicity property flows from assumptions that are deeply rooted in both the proof theory and the semantics, not only of classical logic, but of most philosophical logics as well. From the proof theoretic standpoint, monotonicity follows from the fact that any derivation of the formula A from the premise set Γ also counts as a derivation of that formula from the expanded premises set $\Gamma \cup \{B\}$; the addition of further premises cannot perturb a derivation, since standard inference rules depend only on the presence of information, not its absence. The verification of monotonicity is, if anything, even more immediate from the model theoretic standpoint: since every model of $\Gamma \cup \{B\}$ is a model of Γ, it follows at once, if the formula A holds in every model of Γ, that it must hold also in every model of $\Gamma \cup \{B\}$.

A ***nonmonotonic logic*** is simply one whose consequence relation fails to satisfy the monotonicity property – where the addition of further premises can lead to the retraction of a conclusion already drawn, so that the conclusion set need not increase monotonically with the premise set. Although certain philosophical logics, such as relevance logic [see chapter 13], could be classified as nonmonotonic in this sense, the phrase is generally reserved for a family of logics originating in the field of artificial intelligence (AI), and aimed at formalizing the patterns of ***default reasoning*** that seem to guide much of our intelligent behavior.

Without attempting anything like a formal definition, one can think of default reasoning, very roughly, as reasoning that relies on the absence of information as well as its presence, often mediated by rules of the general form: given *P*, conclude *Q* unless there is information to the contrary. It is easy to see why a logical account of this kind of reasoning requires a nonmonotonic consequence relation. Suppose, for example, that the generic truth 'Birds fly' is taken to express such a default: given that *x* is a bird, conclude that *x* flies unless there is information to the contrary. And suppose one is told that Tweety is a bird. Taken alone, these two premises – that birds fly, and that Tweety is a bird – would then support the conclusion that Tweety flies, since the premise set contains no information to the contrary. But now, imagine that this premise set is supplemented with the additional information that Tweety does not fly (perhaps Tweety is a penguin, or a baby bird). In that case, the original conclusion that Tweety flies would have to be withdrawn, since the default leading to this conclusion relied on the absence of information to the contrary, but the new premise set now contains such information.

The field of nonmonotonic logic began in the late 1970s as an attempt to represent this kind of reasoning within a general logical framework. Since then, the area has been the focus of intense activity, giving rise to hundreds of conference and journal papers, most of which, however, are still confined to the AI literature. At this point, it would be impossible to provide a balanced survey of the field in anything less than a full-length monograph. The present chapter is intended, instead, only as an introductory presentation of two of the main lines of approach – a fixed-point theory and a model-preference theory – in a way that is accessible to a philosophical audience, with an emphasis on conceptual rather than implementational issues.

15.2. Some Motivating Problems

Here are some of the problems that led to development of nonmonotonic logics, namely, the ***frame problem***, first noticed by McCarthy and Hayes (1969), what is known as the ***qualification problem***, and the problems of ***closed-world reasoning*** and ***defeasible inheritance reasoning***.

15.2.1. The frame problem

One of the most important reasoning tasks studied within AI is that of planning – the problem of finding, in the simplest case, a sequence of actions to achieve a specified goal from a specified initial state. Within a logical framework, the planning problem is often studied from the standpoint of the situation calculus, a first-order formalism containing expressions of the form $H[\phi, s]$ to represent the fact that the proposition ϕ holds in the situation s, and allowing also for a description of the effects of various actions.

To illustrate the use of this formalism, imagine that four blocks – A, B, C, and D – are arranged on a table, with blocks A, C, and D set on the table's surface, block B stacked on top of block A, and none of the others having anything on top of them. If this situation is referred to as $s1$, some of the relevant facts from the situation might be depicted through the formulas

$$\begin{aligned} &H[\mathit{On}(B, A), s1] \\ &H[\mathit{Clear}(B), s1] \\ &H[\mathit{Clear}(C), s1] \\ &H[\mathit{Clear}(D), s1] \end{aligned} \tag{15.1}$$

which state that the proposition that block B is on block A holds in the situation $s1$, as do the propositions that the blocks B, C, and D are clear. Note that expressions like $\mathit{On}(B, A)$ and $\mathit{Clear}(B)$ are treated grammatically as complex terms referring to propositions or facts, not as sentences.

Suppose then that these blocks must be manipulated using a robot arm that can perform only two primitive actions: stacking one block on another and unstacking one block from another (and placing it on the table). Let $\mathit{Stack}(X, Y)$ and $\mathit{Unstack}(X, Y)$ represent the actions of stacking X on Y and unstacking X from Y, the effects of these actions can be captured through the axioms

$$\begin{aligned} &(H[\mathit{Clear}(X), s] \wedge H[\mathit{Clear}(Y), s] \wedge X \neq Y) \supset H[\mathit{On}(X, Y), \mathit{Res}(\langle \mathit{Stack}(X, Y)\rangle, s)] \\ &(H[\mathit{On}(X, Y), s] \wedge H[\mathit{Clear}(X), s]) \supset H[\mathit{Clear}(Y), \mathit{Res}(\langle \mathit{Unstack}(X, Y)\rangle, s)] \end{aligned} \tag{15.2}$$

in which it is assumed that all variables are universally quantified. Where α is a sequence of actions, the expression $\mathit{Res}(\alpha, s)$ denotes the situation that results when the actions in α are executed in turn, beginning with situation s. What the first of these two axioms says, then, is that, as long as the distinct blocks X and Y are both clear in the situation s, the situation that results from s when X is stacked on Y is one in which X is on Y; the second axiom says that, if X is on Y and X is clear in s, then Y is clear in the situation that results from s by unstacking X from Y.

Of course, these two axioms define the effects only of action sequences containing a single action, the base case. The effects of longer sequences can be defined inductively by stipulating that

$$\mathit{Res}(\langle A_1, \ldots, A_n\rangle, s) = \mathit{Res}(\langle A_n\rangle, \mathit{Res}(\langle A_1, \ldots, A_{n-1}\rangle, s) \tag{15.3}$$

when n is greater than one; the result of executing a sequence of n actions in a situation s is equivalent to the result of executing the last of these actions in the situation that results from executing all but the last.

Now suppose that Γ is a set of sentences containing a description of some initial situation *s*, as well as axioms specifying the effects of the available actions and perhaps some bookkeeping material, such as the inductive definition of the *Res* function; and let ϕ represent the proposition desired as a goal. Then the planning problem is the problem of finding an action sequence α whose execution in the initial state *s* can be proved from the information in Γ to yield a state in which the goal proposition ϕ holds – more formally, a sequence α for which it can be shown that

$$\Gamma \vdash H[\phi, Res(\alpha, s)]$$

where $\vdash$ is the classical consequence relation.

As a concrete example, imagine that *s*1 above is the initial state, and that Γ contains the statements (15.1)–(15.3): the four sentences describing the initial state, the axioms describing the *Stack* and *Unstack* actions, and the inductive specification of the *Res* function. Now suppose the goal is to achieve a situation in which block *A* is stacked on top of block *C* – that is, a situation in which the statement *On*(*A*, *C*) holds. In this simple case, it is easy to find an appropriate plan: first unstack *B* from *A*, then stack *A* on *C*. More formally, the appropriate plan appears to be ⟨*Unstack*(*B*, *A*), *Stack*(*A*, *C*)⟩, and it seems intuitively – just thinking about how this sequence of actions should work – that it should be possible to verify the correctness of this plan by establishing that

$$\Gamma \vdash H[On(A, C), Res(\langle Unstack(B, A), Stack(A, C)\rangle, s1)]$$

showing that the plan achieves its goal.

In fact, however, this result cannot be established, and it is important to see why. Because Γ contains the statements *On*(*B*, *A*) and *Clear*(*B*), one can indeed conclude from the *Unstack* axiom that

$$H[Clear(A), Res(\langle Unstack(B, A)\rangle, s1)]$$

which states that the block *A* is clear in the situation that results from *s*1 when *B* is unstacked from *A*. And because Γ contains *H*[*Clear*(*C*), *s*1], one knows that the block *C* was already clear in the initial state. Since *A* is now clear as well, it is reasonable to think that a goal state could now be achieved simply by stacking block *A* onto block *C* – that is, that the *Stack* axiom could be used to derive

$$H[On(A, C), Res(\langle Stack(A, C)\rangle, Res(\langle Unstack(B, A)\rangle, s1))]$$

from which the desired conclusion would then follow by the definition of the *Res* function. Unfortunately, this application of the *Stack* axiom would require one to know, not just that *C* is clear in the original state, but that *C* remains clear also in the state that results from the *Unstack*(*B*, *A*) action – that is, one would need to be able to establish

$$H[Clear(C), Res(\langle Unstack(B, A)\rangle, s1)] \tag{15.4}$$

as an intermediate step.

Of course, this intermediate step seems perfectly natural from the standpoint of one's ordinary reasoning about actions: since C is clear in the initial state, it is natural to suppose that it would remain clear even after B is unstacked from A. In fact, however, nothing in Γ allows this intermediate step to be derived – and indeed, the step should not be derivable as a matter of logic, for it is always possible, at least, that the removal of B from A does interfere with the fact that C is clear. (Perhaps blocks B and D are connected by a wire in such a way that removing B from A causes D to be pulled to the top of C; this possibility is consistent with the information in Γ.) What one has here is the notorious *frame problem*, originally noticed by McCarthy and Hayes (1969). When an action is performed, some facts change and some do not. How can one tell which are which, and in particular, how does one propagate those facts that do not change from the original to the resulting situation in a natural way?

15.2.2. *The qualification problem*

Look again at the axiom governing the *Stack* action. Notice that it does not state that X will be on Y in any situation that results from a $\mathit{Stack}(X, Y)$ action, but only that X will be on Y as long as X and Y are distinct blocks that are both clear in the original situation. These qualifications are necessary, of course, because the robot arm cannot reach blocks that are not clear, and because it is impossible to stack a block on top of itself.

But once these qualifications are in place, is the *Stack* axiom then correct? Well, no. What if the block X is so slippery that the robot arm cannot pick it up? What if X is so heavy that it will crush the block Y? What if Y is a bomb that will explode if another block is placed on top of it? The difficulty suggested by these peculiar considerations is known as the *qualification problem*: how does one arrive at an accurate, suitably qualified formulation of the axioms governing actions?

One might respond to this problem by deciding simply to fold all the various possible qualifications into the antecedent of the axioms, either explicitly or implicitly. In the present case, for example, one might introduce a new propositional constant *Weird* to represent the occurrence of a weird circumstance that would interfere with the *Stack* action, and then modify the axiom governing this action with the further precondition that no such weird circumstances occur:

$$
\begin{aligned}
&(H[\mathit{Clear}(X), s] \wedge H[\mathit{Clear}(Y), s] \wedge X \neq Y \wedge \neg\, \mathit{Weird})\\
&\supset H[\mathit{On}(X, Y), \mathit{Res}(\langle \mathit{Stack}(X, Y)\rangle, s)]
\end{aligned}
\tag{15.5}
$$

The interfering circumstances imagined in the previous paragraph could then be classified, quite naturally, as weird:

$$
\begin{aligned}
&\mathit{Slippery}(X) \supset \mathit{Weird}\\
&\mathit{Heavy}(X) \supset \mathit{Weird}\\
&\mathit{Bomb}(Y) \supset \mathit{Weird}
\end{aligned}
\tag{15.6}
$$

There are, however, two problems with this suggestion. The first – to which I know of no solution – is that the list of circumstances that might interfere with a stacking action is open-ended. No conceivable list of possible interfering circumstances could be complete. What if a meteor hits the laboratory and destroys the robot? Then the stack action would not be successful. What if there is an evil demon in the room that does not want to see X on Y and will knock X out of the hand of the robot arm as it approaches Y?

The second problem is more subtle, and would arise even if there was a relatively exhaustive list of qualifications. The point of placing preconditions in the antecedent of an action axiom is that one must verify that the preconditions are satisfied before concluding that the action is successful. And it does seem reasonable, in the case of the *Stack* axiom, that one should have to verify that the blocks X and Y must both be clear before one can know that the result of stacking X on Y is successful. But it seems less reasonable to suppose that one must actually have to verify that all of the various weird circumstances that might interfere with this action do not occur – that there is no bomb, no meteor, no evil demon, and so on. It would be better to be able simply to assume that weird circumstances like these do not occur unless there is information to the contrary.

15.2.3. *Closed-world reasoning*

Suppose I ask my travel agent if United Airlines has a direct flight from Washington to Barcelona. The travel agent has access to a database containing flight information. From a logical standpoint, one can think of this database as a set of sentences of the form

$$\begin{aligned}&\mathit{Connects}(\mathit{UA354}, \mathit{Baltimore}, \mathit{Boston})\\&\mathit{Connects}(\mathit{UA750}, \mathit{Washington}, \mathit{London}) \qquad (15.7)\\&\mathit{Connects}(\mathit{UA867}, \mathit{London}, \mathit{Barcelona})\end{aligned}$$

and so on; the travel agent answers my question by drawing inferences from these sentences. Suppose I am told: No, there is no direct flight from Washington to Barcelona. How can the travel agent reach this conclusion? The airline database only says what cities are connected by what flights; it does not list the cities that are not connected, and certainly this kind of negative information does not follow as an ordinary logical consequence from the positive information provided.

The answer is that the travel agent's reasoning is governed by a convention known as the *closed-world assumption* (Reiter, 1978), which states, in the simplest case, that all relevant positive information is explicitly listed. Because of this convention, it is legitimate to conclude that a positive proposition is false whenever it is not explicitly present in the database; the travel agent can legitimately conclude, for example, that there is no direct flight between Washington and Barcelona simply because no such flight is listed.

The closed-world assumption applies, of course, not only to the airline database, but to any number of situations in which positive information is overwhelmed by negative information. When I look at a list of people invited to a party, I can conclude, if I am not on the list, that I am not invited to the party; when I look at my desk calendar, I can conclude, if there is no doctor's appointment listed for Thursday at 3:00, that I have no doctor's appointment at that time. Reasoning based on the closed-world assumption exemplifies the general pattern of default reasoning as relying on the absence of information: lacking information to the contrary, one can assume that there is no direct flight between two cities; an entry in the database provides information to the contrary.

15.2.4. Defeasible inheritance reasoning

Returning to the initial example: birds fly, Tweety is a bird, therefore Tweety flies. Reasoning like this is known in AI as inheritance reasoning, and was originally developed in response to the need for an efficient way of representing and accessing taxonomic information. Rather than having to list explicitly the properties of each individual, it is imagined that classes and properties are arranged in a taxonomic hierarchy, and that individuals inherit their properties from the classes to which they belong. It is not necessary to state explicitly that Tweety flies, since this property is inherited from the general class of birds.

This kind of taxonomic reasoning has been familiar since Aristotle, and was explored in some detail by medieval philosophers; what is new in AI is the idea that – again, for reasons of efficiency – the taxonomy is often allowed to represent defeasible as well as strict information. An example of such a defeasible inheritance network is provided in Figure 15.1, known as the Tweety Triangle. Here, strict links are represented by the strong arrow ⇒ and defeasible links by the weak arrow →, so that the displayed network provides the following information: Tweety is a penguin; penguins are birds; as a rule, birds tend to fly, and penguins tend not to.

When these defeasible inheritance networks were first introduced, they were supplied only with a 'procedural' semantics, according to which the meaning of the representations was supposed to be specified implicitly by the inference algorithms operating on them. It was soon realized, however, that these algorithms could lead

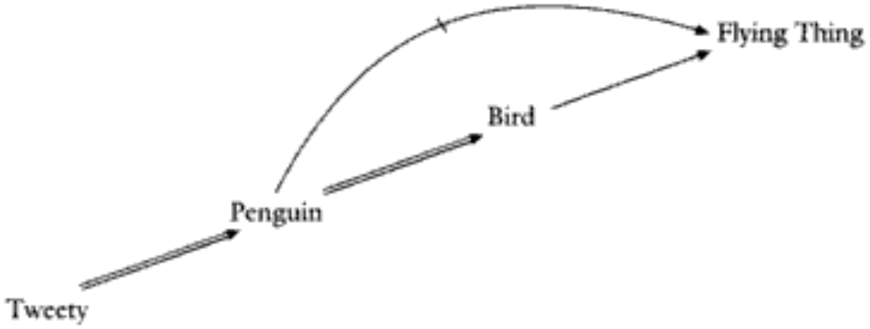

Figure 15.1 The Tweety Triangle

to bizarre and unintuitive results in complicated cases, and researchers felt the need to provide an implementation independent account of the meaning of these network formalisms. One natural idea involved providing a logical interpretation of the networks – interpreting the individual links in the network as logical formulas, and so the entire network as a collection of formulas, whose meaning could then be specified by the appropriate logic. The logical interpretation of strict links, of course, presents no problems: a link like *Tweety* $\Rightarrow$ *Penguin*, for example, could naturally be represented as an atomic statement, such as Pt, and a link like *Penguin* $\Rightarrow$ *Bird* as a universal statement of the form $\forall x(Px \supset Bx)$. But there is nothing in ordinary logic to represent the defeasible links *Bird* $\rightarrow$ *Fly* and *Penguin* $\not\rightarrow$ *Fly*, carrying the intuitive meaning birds fly and that penguins do not.

15.3. A Fixed-Point Approach: Default Logic

Perhaps the best known and most widely applied formalism for nonmonotonic reasoning is ***default logic***, introduced by Reiter (1980). This formalism results from supplementing ordinary logic with new rules of inference, known as ***default rules***, and then modifying the standard notion of logical consequence to accommodate these new rules.

15.3.1. Basic ideas

An ordinary rule of inference (with a single premise) can be depicted simply as a premise/conclusion pair, such as (A/B); this rule commits the reasoner to B once A has been established. By contrast, a default rule is a triple, of the form $(A : C/B)$. Very roughly, such a rule commits the reasoner to B once A has been established and, in addition, C is consistent with the reasoner's conclusion set. The formula A is referred to as the ***prerequisite*** of this default rule, B as its ***consequent***, and C as its ***justification***.[1] A ***default theory*** is a pair $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$, in which $\mathcal{W}$ is a set of ordinary formulas and $\mathcal{D}$ is a set of default rules.

Before characterizing the new notion of logical consequence defined by Reiter, consider how default logic might be used to represent the initial example, in which one is told that Tweety is a bird and that birds fly. The generic statement that birds fly can reasonably be taken to mean something like: once one learns of an object x that it is a bird, one should conclude that x flies unless there is information to the contrary – unless, that is, this conclusion is inconsistent with one's beliefs. What this suggests is that the generic statement should be represented as a sort of universally quantified default rule, perhaps of the form $\forall x(Bx : Fx/Fx)$, but unfortunately it is no more meaningful to quantify a default rule than it is to quantify an ordinary rule of inference. To avoid this problem, Reiter allows open formulas to occur in defaults, so that the generalization concerning birds can be expressed as $(Bx : Fx/Fx)$. However, to avoid the resulting complexities – involving the application of these open defaults to yield closed formulas – the somewhat simpler approach of representing these defeasible generalizations, not by open defaults, but instead by appropriate instance of these defaults for each object in the domain is adopted here. In the present case, where Tweety is the only object of concern, the only default necessary is $(Bt : Ft/Ft)$, which says that if Tweety is a bird, one should conclude that Tweety flies as long as this is consistent with what is known. The information from this initial example can then be represented through the default theory $\Delta_1 = \langle \mathcal{W}_1, \mathcal{D}_1 \rangle$, where $\mathcal{W}_1 = \{Bt\}$ and $\mathcal{D}_1 = \{(Bt : Ft/Ft)\}$.

In this example, because one knows that Bt, and because Ft is consistent with one's knowledge, the default rule justifies drawing the conclusion Ft. The appropriate conclusion set based on Δ_1 therefore seems to be $Th(\{Bt, Ft\})$, the logical closure of what one is told to begin with, together with the conclusions of the applicable defaults. If one is told, in addition, that Tweety does not fly, one moves to the default theory $\Delta_2 = \langle \mathcal{W}_2, \mathcal{D}_2 \rangle$, with $\mathcal{D}_2 = \mathcal{D}_1$ and $\mathcal{W}_2 = \mathcal{W}_1 \cup \{\neg Ft\}$. Here the default rule $(Bt : Ft/Ft)$ can no longer be applied, because its justification is now inconsistent with one's knowledge and so the appropriate conclusion set based on Δ_2 is simply $Th(\mathcal{W}_2)$.

15.3.2. Extensions

The discussion of this example illustrates the kind of conclusion sets desired from particular default theories. The task of arriving at a general definition of this notion, however, is not trivial; the trick is to find a way of capturing the meaning of the new component – the justification – present in default rules.

In ordinary logic, the conclusion set associated with a set of formulas $\mathcal{W}$ is simply $Th(\mathcal{W})$, the logical closure of $\mathcal{W}$. It might seem, then, that the conclusion set associated with a default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ should be

$$\mathcal{E} = Th(\mathcal{W}) \cup \{C : (A : B/C) \in \mathcal{D},\ A \in Th(\mathcal{W}),\ \neg B \notin Th(\mathcal{W})\}$$

the closure of $\mathcal{W}$ together with the consequents of those default rules whose prerequisites are entailed by and whose justifications are consistent with $\mathcal{W}$. A moment's thought, however, shows that this suggestion is inadequate. For one thing, the set $\mathcal{E}$ defined in this way is not even closed under logical consequence: the addition of the consequent from some default rule into the set $\mathcal{E}$ may trigger new logical implications that should, intuitively, be included in the conclusion set, or worse still, the addition of the consequent from one default rule may trigger the firing of another. As an example, consider the default theory $\Delta_3 = \langle \mathcal{W}_3, \mathcal{D}_3 \rangle$ in which $\mathcal{W}_3 = \{A\}$ and $\mathcal{D}_3 = \{(A : B/C), (C : D/E)\}$. The above definition correctly adds the consequent C of the first default rule into the conclusion set $\mathcal{E}$. It seems, though, that the presence of C should then trigger the firing of the second rule, resulting also in the addition of E to the conclusion set, but this statement is not included.

What this example suggests is that the definition of the appropriate conclusion set for a default theory should be iterative. Perhaps one should take the conclusion set of the default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ to be

$$\mathcal{E} = \bigcup_{i=0}^{\infty} \mathcal{E}_i$$

with

$$\mathcal{E}_0 = \mathcal{W}$$

$$\mathcal{E}_{i+1} = Th(\mathcal{E}_i) \cup \{C : (A : B/C) \in \mathcal{D},\ A \in Th(\mathcal{E}_i),\ \neg B \notin Th(\mathcal{E}_i)\}$$

This suggestion responds to the previous concern, giving $Th(\{A, C, E\})$ as the conclusion set for the default theory Δ_3, as desired. Now, however, there is a new problem, illustrated by the theory $\Delta_4 = \langle \mathcal{W}_4, \mathcal{D}_4 \rangle$, with $\mathcal{W}_4 = \{A, B \supset \neg C\}$ and $\mathcal{D}_4 = \{(A : C/B)\}$. Tracing through the iteration, one can see that the rule $(A : C/B)$ is applicable at the first stage, since its prerequisite belongs to $Th(\mathcal{W}_4)$ and its justification is consistent with this set; hence one has B in $\mathcal{E}_1$. Just a bit of additional reasoning then shows that $\neg C$ must belong to $\mathcal{E}_2$, and so to $\mathcal{E}$, since this formula is a logical consequence of the information contained in $\mathcal{E}_1$. The rule $(A : C/B)$ seems initially to be applicable, since, prior to its application, there is no reason to conclude $\neg C$; but once the rule has been applied, the information it provides does allow us to derive $\neg C$. The rule thus seems to undermine its own applicability.

Of course, a chain of reasoning like this showing that some default rule is undermined can be arbitrarily long; and so one cannot really be sure that a default rule is applicable in some context until one has applied it, along with all the other rules that seem applicable, and then one has surveyed the logical closure of the result. Because of this, the conclusion set associated with a default theory cannot be defined in the usual iterative way, by successively adding to the original data the conclusions of the applicable rules of inference, and then taking the limit of this process.

Instead, Reiter is forced to adopt a fixed-point approach in specifying the appropriate conclusion sets of default theories – which are described as *extensions*. In fact, he actually offers two characterizations of the concept of an extension. The first considered here, although not the official definition, is both more intuitive and more useful in practice. The idea behind this particular characterization is that, given a default theory, one first conjectures a candidate extension for the theory, and then – using this candidate – defines a sequence of approximations to some conclusion set. If this approximating sequence has the original candidate as its limit, the candidate is then certified as an extension for the default theory.

Definition 15.1 The set $\mathcal{E}$ is an *extension* of the default theory $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ iff (if and only if) there exists a sequence of sets $\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2, \ldots$ such that

$$\mathcal{E} = \bigcup_{i=0}^{\infty} \mathcal{E}_i$$

$$\mathcal{E}_0 = \mathcal{W}$$

$$\mathcal{E}_{i+1} = Th(\mathcal{E}_i) \cup \{C : (A : B/C) \in \mathcal{D},\ A \in Th(\mathcal{E}_i),\ \neg B \notin \mathcal{E}\}$$

Here, of course, the set $\mathcal{E}$ is the candidate, which is certified as a true extension of Δ if it turns out that $\mathcal{E}$ coincides with the union of the approximating sequence

$\mathcal{E}_0, \mathcal{E}_1, \mathcal{E}_2$. . . . Note that $\mathcal{E}$ figures in the definition of $\mathcal{E}_{i+1}$: the approximating sequence is defined in terms of the original candidate.

The fixed-point nature of extensions is more apparent in Reiter's official definition, which relies on an operator Γ that uses the information from a particular default theory to map formula sets into formula sets.

Definition 15.2 Where $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ is a default theory and $\mathcal{S}$ is some set of formulas, $\Gamma_\Delta(\mathcal{S})$ is the minimal set satisfying three conditions:

1 $\mathcal{W} \subseteq \Gamma_\Delta(\mathcal{S})$
2 $Th(\Gamma_\Delta(\mathcal{S})) = \Gamma_\Delta(\mathcal{S})$
3 For each $(A : B/C) \in \mathcal{D}$, if $A \in \Gamma_\Delta(\mathcal{S})$ and $\neg B \notin \mathcal{S}$, then $C \in \Gamma_\Delta(\mathcal{S})$.

The first two conditions in this definition simply state that $\Gamma_\Delta(\mathcal{S})$ contains the information provided by the original theory, and that it is closed under logical consequence; the third condition states that it contains the conclusions of the default rules applicable in $\mathcal{S}$; and the minimality constraint prevents unwarranted conclusions from creeping in. Where $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ is a default theory, the operator Γ_Δ maps any formula set $\mathcal{S}$ into the minimal superset of $\mathcal{W}$ that is closed under both ordinary logical consequence and the default rules from $\mathcal{D}$ that are applicable in $\mathcal{S}$. The official definition of extensions – here presented as a theorem – then identifies the extensions of a default theory as the fixed points of this operator.

Theorem 15.1 The set $\mathcal{E}$ is an *extension* of the default theory Δ iff $\Gamma_\Delta(\mathcal{E}) = \mathcal{E}$.

As the reader can verify, the default theories Δ_1 and Δ_2 have, as desired, the respective sets $Th(\{Bt, Ft\})$ and $Th(\{Bt, \neg Ft\})$ as their extensions. It should be clear that the notion of an extension defined here is a conservative generalization of the corresponding notion of a conclusion set from ordinary logic: the extension of a default theory $\langle \mathcal{W}, \mathcal{D} \rangle$, in which $\mathcal{D}$ is empty, is simply $Th(\mathcal{W})$. And it can be shown also that default rules themselves cannot introduce inconsistency: any extension of a default theory $\langle \mathcal{W}, \mathcal{D} \rangle$ will be consistent as long as the ordinary component $\mathcal{W}$ of that theory is consistent.

15.3.3. Default consequence

In contrast to the situation in ordinary logic, however, not every default theory leads to a single extension, a single set of appropriate conclusions. Some default theories have no extensions; Δ_4 is an example. The easiest way to see that this theory has no extensions is to work with the Definition 1 of the notion, and then to suppose that Δ_4 did have an extension – say, $\mathcal{E}$. Evidently, one would then have either $\neg C \in \mathcal{E}$ or $\neg C \notin \mathcal{E}$. Suppose, first, that $\neg C \in \mathcal{E}$. Well, since $\neg C \notin \mathcal{E}_0$, and under the supposition that $\neg C \in \mathcal{E}$ it is easy to see from the definition of the approximating sequence that $\neg C \notin \mathcal{E}_1$, that $\neg C \notin \mathcal{E}_2$, and so on. But since $\mathcal{E}$ is simply the union of $\mathcal{E}_0$, $\mathcal{E}_1$, $\mathcal{E}_2$, and so on, it follows, contrary to assumption, that $\neg C \notin \mathcal{E}$. Next, suppose

$\neg C \notin \mathcal{E}$. In that case, it is easy to see that $\neg C \in \mathcal{E}_2$, and since $\mathcal{E}_2$ is a subset of $\mathcal{E}$, that $\neg C \in \mathcal{E}$, which again contradicts the assumption.

Default theories without extensions are often viewed as incoherent, and can perhaps be dismissed simply as anomalous. But there are also perfectly coherent default theories that allow multiple extensions. A standard example arises when one tries to encode as a default theory the inheritance network depicted in Figure 15.2, known as the Nixon Diamond, and representing the following set of facts:

- Nixon is a Quaker.
- Nixon is a Republican.
- Quakers tend to be pacifists.
- Republicans tend not to be pacifists.

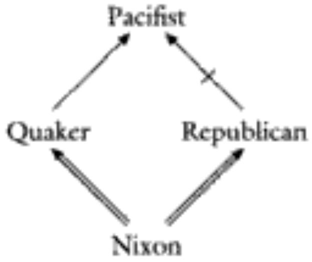

Figure 15.2 The Nixon Diamond

If one instantiates for Nixon the general statements expressed here about Quakers and Republicans, the resulting theory is $\Delta_5 = \langle \mathcal{W}_5, \mathcal{D}_5 \rangle$, with

$$\mathcal{W}_5 = \{Qn, Rn\}$$

and

$$\mathcal{D}_5 = \{(Qn : Pn/Pn), (Rn : \neg Pn/\neg Pn)\}$$

This theory allows both $Th(\mathcal{W}_5 \cup \{Pn\})$ and $Th(\mathcal{W}_5 \cup \{\neg Pn\})$ as extensions. Initially, before drawing any new conclusions, both of the default rules from $\mathcal{D}_5$ are applicable, but once one adopts the conclusion of either, the applicability of the other is blocked.

In cases like this, when a default theory leads to more than one extension, it is difficult to decide what conclusions a reasoner should actually draw from the information contained in the theory, and several options have been discussed in the literature. One option is to suppose that the reasoner should arbitrarily select one of the theory's several extensions and endorse the conclusions contained in it; a second option is to suppose that the reasoner should be willing to endorse a conclusion as long as it is contained in some extension of the default theory. These first two options are sometimes said to reflect a ***credulous*** reasoning strategy. A third option, sometimes described as ***skeptical***, is to suppose that the reasoner should endorse a conclusion only if it is contained in every extension of the default theory.[2]

The first of these options – pick an arbitrary extension – really does seem to reflect a rational policy for reasoning in the face of conflicting information: often, given such information, one simply adopts some internally coherent point of view in which the conflicts are resolved in some particular way, regardless of the fact that there are other coherent points of view available in which the conflicts are resolved in a different way. Still, although this reasoning policy is rational, it is hard to see how such a policy could be codified as a formal consequence relation. If the choice of extension really is arbitrary, different reasoners could easily select different extensions, or the same reasoner might select different extensions at different times. Which extension, then, would represent the consequence set of the theory?

The second option – endorse a conclusion whenever it is contained in some extension of the default theory – can indeed be codified as a consequence relation, but it would be a peculiar one. According to this policy, the consequence set of a default theory need not be closed under standard logical consequence, and, in fact, might easily be inconsistent. The consequence set of Δ_5, for example, would contain both Pn and $\neg Pn$, since each of these formulas belongs to some extension of the default theory, but it would not contain $Pn \wedge \neg Pn$. This second option seems to provide a characterization, not so much of the formulas that should be believed on the basis of a default theory, but instead of the formulas that are believable.[3]

Only the third, skeptical option – endorse a conclusion whenever it is contained in every extension of the default theory – results in a natural consequence relation, as follows.

> **Definition 15.3** Let $\Delta = \langle \mathcal{W}, \mathcal{D} \rangle$ be a default theory and A a formula. Then A is a *skeptical consequence* of Δ – written, $\Delta \vdash A$ – just in case $A \in \mathcal{E}$ for each extension $\mathcal{E}$ of Δ.

And it is worth noting explicitly, now that a formal consequence relation has been defined, that it is indeed nonmonotonic in two ways: both adding new factual information to the $\mathcal{W}$-component of a default theory and adding new default information to the $\mathcal{D}$-component can force one to abandon consequences previously supported. The first possibility can be illustrated by referring back to the default theories Δ_1 and Δ_2. Here, $\Delta_1 \vdash Ft$, but it is not the case that $\Delta_2 \vdash Ft$ even though Δ_2 is obtained by adding the new factual information that $\neg Ft$ to the $\mathcal{W}$-component of Δ_1. To illustrate the second case, consider the default theory $\Delta_6 = \langle \mathcal{W}_6, \mathcal{D}_6 \rangle$, where $\mathcal{W}_6 = \mathcal{W}_5$ and $\mathcal{D}_6 = \{(Qn : Pn/Pn)\}$; this theory is like the Nixon Diamond Δ_5, except without the default that Republicans tend not to be pacifists. It is easy to see that Δ_6 has $Th(\mathcal{W}_6 \cup \{Pn\})$ as its only extension, so that $\Delta_6 \vdash Pn$. The theory Δ_5, however, has two extensions, one of which does not contain Pn; so it is not the case that $\Delta_5 \vdash Pn$, even though Δ_5 results from the addition of the new default information $(Rn : \neg Pn/\neg Pn)$ to the $\mathcal{D}$-component of Δ_6.

15.3.4. Examples and non-normal defaults

Now, how can the motivating examples from section 15.2 be handled from the perspective of default logic?

To begin with, the frame problem appears to have a straightforward solution that results when one supplements the standard logical description of the initial situation and the available actions with default rules which simply say that facts tend to persist. To illustrate, one might encode the problem from section 15.2.1 into the default theory $\Delta_7 = \langle \mathcal{W}_7, \mathcal{D}_7 \rangle$, as follows. First, the factual component $\mathcal{W}_7$ contains the formulas (15.1) through (15.3), describing the initial situation, the axioms characterizing the effects of the *Stack* and *Unstack* actions, and the inductive description of sequences of actions. Second, the default component $\mathcal{D}_7$ contains all instances of the default rule schema

$$(H[\phi, s] : H[\phi, Res(\alpha, s)]/H[\phi, Res(\alpha, s)])$$

which states that: whenever a fact ϕ holds in a situation s, if it is consistent to conclude that ϕ still holds after the performance of the action α, then one should conclude by default that ϕ still holds after the performance of α.

It is easy to verify that this default theory has a single extension containing the formula (15.4), which is, of course, the intermediate step that was not derivable earlier without the help of frame axioms. Although the proposition that block C is still clear even after B is unstacked from A does not follow from the factual information contained in (15.1) through (15.3) alone, it can be derived with the help of the default rule which says to conclude, unless there is information to the contrary, that facts tend to persist.[4]

Turning to the qualification problem, again a partial solution can be found using default logic by supplementing the statement of the axioms governing actions with default rules which say simply that peculiar circumstances that might interfere with these actions tend not to occur. In the case of the example from section 15.2.2, the relevant information might be formulated through the theory $\Delta_8 = \langle \mathcal{W}_8, \mathcal{D}_8 \rangle$, in which $\mathcal{W}_8$ contains, in addition to the appropriate background information, the modified *Stack* axiom (15.5) as well as the specifications from (15.6) of the various weird circumstances that might interfere with that action, and in which $\mathcal{D}_8$ contains the single default

$$(\top : \neg Weird/\neg Weird)$$

which says to assume, absent information to the contrary, that no such weird circumstances occur ($\top$ stands for the universally true proposition). Of course, this representation does not help to resolve the first of the two issues presented by the qualification problem – that the list of conditions that might interfere with the *Stack* action is open-ended. The representation does, however, offer a resolution to the second of these issues. Given a list of various peculiar conditions that might conceivably interfere with the *Stack* action, one no longer actually verifies that each of these conditions fails in order to conclude that *Stack* has the desired effects; the default rule allows one simply to assume that these conditions fail unless there is information to the contrary.

Like the frame and qualification problems, the difficulties presented by closed-world reasoning also seem to be amenable to a solution based on default logic. As an initial suggestion, one might represent the information from section 15.2.3, for example, through the default theory $\Delta_9 = \langle \mathcal{W}_9, \mathcal{D}_9 \rangle$, with $\mathcal{W}_9$ containing the factual data from (15.7) and $\mathcal{D}_9$ containing each instance of the default rule schema

$$(\top : \neg Connects(x, y, z)/\neg Connects(x, y, z))$$

which says that, in the absence of information to the contrary, one should assume that cities are not connected by a direct flight. This theory will then have a single extension, allowing one to conclude (under reasonable assumptions, such as that all existing flights are named) that there is no direct flight between Baltimore and Barcelona.

Now, step back and notice a common feature in our default logic representation of these various examples illustrating the frame problem, the qualification problem, and closed-world reasoning, as well as in our representation of the Nixon Diamond. Each of these cases relied entirely on default rules of the special form $(A : B/B)$, in which the same formulas occurs as both justification and conclusion. Such default rules are known as ***normal defaults***, and theories containing only normal defaults as ***normal default theories***. As shown in Reiter (1980), normal default theories possess a number of attractive properties that are not shared by default theories in general – most notably, normal theories are guaranteed to have extensions. Because of these attractive properties, and because, as has been seen, many important examples can be coded into normal theories, Reiter originally conjectured that the full expressive power of default logic might not be needed in realistic applications, and it could be limited to normal theories.

This conjecture, however, was soon seen to be incorrect, as is illustrated by considering the final example – the Tweety Triangle from section 15.2.4. Considering only normal defaults, the information from the Tweety Triangle is naturally represented in the theory $\Delta_{10} = \langle \mathcal{W}_{10}, \mathcal{D}_{10} \rangle$ with $\mathcal{W}_{10}$ containing the sentences Pt and $\forall x(Px \supset Bx)$, stating that Tweety is a penguin and that all penguins are birds, and with $\mathcal{D}_{10}$ containing the defaults $(Bt : Ft/Ft)$ and $(Pt : \neg Ft/\neg Ft)$, instantiating for Tweety the generic truths that birds tend to fly and that penguins tend not to. This default theory, like the representation of the Nixon Diamond as Δ_4, contains two conflicting default rules, and so leads to two extensions:

$$Th(\mathcal{W}_{10} \cup \{Ft\}) \qquad \text{and} \qquad Th(\mathcal{W}_{10} \cup \{\neg Ft\})$$

But is this right? In the case of the Nixon Diamond, the multiple extensions are reasonable, since the defaults concerning Quakers and Republicans appear to carry equal weight. But in the case of the Tweety Triangle, it really does seem that the default concerning penguins should be preferred to the default concerning birds, since penguins are a specific kind of bird, and it is always best to reason on the basis of the most specific information available. One way of capturing such preferences among defaults – first explored by Etherington and Reiter (1983) – is to modify the representation so that the reasons that might override the application of a default rule are explicitly built into the statement of that rule. Following this approach, the default concerning birds from the Tweety Triangle, for example, could be represented, not by the normal default rule $(Bt : Ft/Ft)$, but instead by the non-normal rule

$(Bt : [Ft \wedge \neg Pt]/Ft)$

What this rule says is that, once it is known that Tweety is a bird, if it is consistent with what is known that Tweety flies and that he is not a penguin, then one should presume that he flies.

This appeal to non-normal rules solves the initial problem presented by the Tweety Triangle: when this new, non-normal default is substituted for its normal predecessor in the previous Δ_{10}, the resulting theory now has only the single extension $Th(\mathcal{W}_{10} \cup \{\neg Ft\})$, which states unambiguously that Tweety does not fly. Only the default rule $(Pt : \neg Ft/\neg Ft)$ can be applied. The new default $(Bt : [Ft \wedge \neg Pt]/Ft)$ does not come into play, since Pt is known.

Unfortunately, in solving the previous problem, the strategy of using non-normal rules to express preferences among competing defaults from defeasible inheritance networks now introduces a new difficulty: the new mapping of information from inheritance networks into default rules is holistic – the translation of a particular statement can vary depending on the context in which it is embedded. To illustrate, suppose one was to supplement the Tweety Triangle with the additional information that another class of birds – say, very young birds – does not fly. Of course, one would then have to add to the representation the formula $\forall x(\Upsilon x \supset Bx)$, which states that young birds are birds, as well as the default $(\Upsilon t : \neg Ft/\neg Ft)$, instantiating for Tweety the statement that young birds tend not to fly. But in addition, since there is now another possible reason present for overriding the default that birds tend to fly, the previous representation of that default must also be replaced with the new rule

$(Bt : [Ft \wedge \neg Pt \wedge \neg \Upsilon t]/Ft\ )$

From a computational point of view, this consequence is unattractive because it makes the process of updating a body of information extremely complicated, involving, not only the representation of new information, but also the reformulation of information that was already represented. From a philosophical point of view, the consequence is unattractive for much the same reason that holism is generally unattractive: the meaning of the statement that birds tend to fly seems not to vary from context to context, and so it is odd that its translation should vary.

15.4. A Model-Preference Approach: Circumscription

It was noted in the introduction that the monotonicity property reflects both proof theoretic and model theoretic assumptions of ordinary logic. Default logic results from a modification of the usual proof theoretic assumptions, introducing rules of inference that depend on the absence as well as the presence of information. This section now turns to a theory that results from a modification of the usual semantic assumptions.

Typically, a formula A is said to be a semantic consequence of a set of formulas Γ – written, $\Gamma \Vdash A$ – when A is true in every model of Γ. For many applications, however, one does not really care about all the models of Γ, but only about certain *preferred* models, and it then seems reasonable to modify the usual notion of consequence so that A is said to be a consequence of Γ whenever A is true in all the preferred models of Γ. The theory of circumscription, originally formulated by McCarthy (1980), results from this general preferential framework when the preferred models are defined as those in which certain predicates have minimal extensions.

15.4.1. Predicate circumscription

Taking a model as a pair $\mathcal{M} = \langle \mathcal{D}, \nu \rangle$, with $\mathcal{D}$ a domain and ν an interpretation of some fixed background language over that domain, begin by defining more precisely the preference ordering on models that forms the semantic background for the theory of circumscription. The general idea is that one model is at least as preferable as another just in case, while agreeing on everything else, the first assigns to some particular predicate P an extension at least as small as that assigned by the second.

Definition 15.4 Where $\mathcal{M}_1 = \langle \mathcal{D}_1, \nu_1 \rangle$ and $\mathcal{M}_2 = \langle \mathcal{D}_2, \nu_2 \rangle$ are models and where P is a predicate, then $\mathcal{M}_1 \preceq_P \mathcal{M}_2$ just in case

(i) $\mathcal{D}_1 = \mathcal{D}_2$
(ii) $\nu_1(Q) = \nu_2(Q)$ for every linguistic symbol Q other than P, and
(iii) $\nu_1(P) \subseteq \nu_2(P)$.

It should be clear that the weak preference relation $\preceq_P$ is a partial ordering, so that a corresponding strong preference relation is definable in the standard way.

Definition 15.5 Where $\mathcal{M}_1$ and $\mathcal{M}_2$ are models and where P is a predicate, then $\mathcal{M}_1 \prec_P \mathcal{M}_2$ just in case $\mathcal{M}_1 \preceq_P \mathcal{M}_2$ but $\mathcal{M}_1 \neq \mathcal{M}_2$.

And one can then define the minimal elements in a class of models – the most preferred elements – as those models from the class for which the class contains no model that is more preferred.

Definition 15.6 Let $\mathcal{K}$ be a set of models and P a predicate. Then $\mathcal{M}$ is *P-minimal* in $\mathcal{K}$ just in case $\mathcal{M} \in \mathcal{K}$ and there is no $\mathcal{M}' \in \mathcal{K}$ such that $\mathcal{M}' \prec_P \mathcal{M}$.

Suppose $|\Gamma|$ is the model class of Γ, the set of models that satisfies each member of Γ. Having identified the minimal, or most preferred, models in a class, one can now define McCarthy's original notion of preferential, or minimal, consequence by focusing only on the minimal models of a theory, defining a formula as a consequence of the theory whenever it is true in all those models.

Definition 15.7 Where Γ is a set of formulas, P a predicate, and A a formula, A is said to be a *P-minimal consequence* of Γ – written $\Gamma \Vdash_P A$ – just in case $\mathcal{M} \vDash A$ for every model $\mathcal{M}$ that is P-minimal in the set $|\Gamma|$.

And it is easy to see that this notion of minimal consequence is nonmonotonic. As an example, take $\Gamma_1 = \{Pa, a \neq b\}$. Then $\Gamma_1 \Vdash_P \neg Pb$, since the P-minimal models of Γ are those in which P holds only of the single element a, but of course one does not have $\Gamma_1 \cup \{Pb\} \Vdash_P \neg Pb$.

In addition to defining the notion of minimal consequence, McCarthy provides a sound second-order syntactic characterization of the idea through the axiom of circumscription, for which some preliminary notation is needed. Where P and Q are n-ary predicates, take $P \leq Q$ as an abbreviation of the formula

$$\forall x_1 \cdots x_n(Px_1 \ldots x_n \supset Qx_1 \ldots x_n)$$

Likewise, $P < Q$ abbreviates

$$P \leq Q \wedge \neg(Q \leq P)$$

and $P = Q$ abbreviates

$$P \leq Q \wedge Q \leq P$$

Where Γ is a finite theory, $\overline{\Gamma}$ stands for the conjunction of the members of Γ, and $\Gamma^{P/Q}$ stands for the result of substituting the predicate P for the predicate Q throughout Γ.

Using this notation, the ***circumscription formula*** for the predicate P in the theory Γ – abbreviated $Circ[\Gamma; P]$ – can be expressed quite simply through the second-order sentence

$$\overline{\Gamma} \wedge \neg\exists P'[\overline{\Gamma^{P'/P}} \wedge P' < P]$$

Any model $\mathcal{M}$ that satisfies the first conjunct of this formula, of course, is a model of Γ. But what does the second conjunct say? Well, if there were another model $\mathcal{M}'$ also satisfying Γ and such that $\mathcal{M}' \prec_P \mathcal{M}$, one could then use the value assigned by $\mathcal{M}'$ to the predicate P to show that $\mathcal{M}$ satisfies the formula $\exists P'[\overline{\Gamma^{P'/P}} \wedge P' < P]$. The force of the second conjunct, then, is simply that there is no such model $\mathcal{M}'$, and so together, what the two conjuncts say is that $Circ[\Gamma; P]$ is satisfied by exactly the P-minimal models of Γ.

Theorem 15.2 Let Γ be a finite set of sentences, P a predicate, and $\mathcal{M}$ a model. Then $\mathcal{M} \models Circ[\Gamma; P]$ just in case $\mathcal{M}$ is P-minimal in $|\Gamma|$.

From this result, the soundness of circumscription with respect to minimal consequence follows at once.

Theorem 15.3 Let Γ be a finite set of sentences, P a predicate, and A a formula. Then $\Gamma \Vdash_P A$ whenever $Circ[\Gamma; P] \vdash A$.

The argument is again straightforward. To say that $\Gamma \Vdash_P A$ is to say that every P-minimal model of Γ satisfies A, so let $\mathcal{M}$ be such a model. From the preceding result, it is known that $\mathcal{M} \vDash \mathit{Circ}[\Gamma;\ P]$. Since $\mathit{Circ}[\Gamma;\ P] \vdash A$, the soundness of second-order logic says that $\mathit{Circ}[\Gamma;\ P] \Vdash A$, and so one can conclude that $\mathcal{M} \vDash A$.

Of course, circumscription is not complete with respect to minimal consequence; not every minimal consequence of a theory can be derived from the circumscription formula. But this failure is no surprise, following from the incompleteness of second-order logic itself. It was also noticed early on that the result of circumscribing certain predicates even in consistent theories might lead to inconsistency; a simple example, due to Etherington et al. (1985), results when one considers the theory Γ_2, containing the sentences

$$\exists x[Nx \wedge \forall y(Ny \supset x \neq s(y))]$$

$$\forall x(Nx \supset Ns(x))$$

$$\forall xy(s(x) = s(y) \supset x = y)$$

Any model $\mathcal{M}$ of Γ_2 must assign to N an extension containing a series isomorphic to the natural numbers (with s interpreted as successor); and one can then define another model $\mathcal{M}'$ of Γ_2 simply by deleting from the extension of N the initial element of this series. Evidently, then, $\mathcal{M}' <_N \mathcal{M}$, and so the model class of Γ_2 has no N-minimal elements. Since, as has been seen, $\mathit{Circ}[\Gamma_2;\ N\}$ is satisfied by all and only the N-minimal elements of this model class, it follows that the result of circumscribing the predicate N in the theory Γ_2 is not satisfiable.

To illustrate the use of the circumscription formula, consider how circumscribing the predicate P in the earlier example of Γ_1 allows one to derive $\neg Pb$. To begin with, it is most convenient to express the circumscription formula $\mathit{Circ}[\Gamma_1;\ P]$, not exactly in the fashion displayed above, but instead in the logically equivalent form

$$\overline{\Gamma}_1 \wedge \forall P'[(\overline{\Gamma_1^{P'/P}} \wedge P' \leq P) \supset P' = P]$$

The second conjunct of this formula can then be instantiated by identifying P' with the predicate $\lambda x(x = a)$, in which case it is easy to see from the ordinary logic of identity that both the formulas $\overline{\Gamma_1^{P'/P}}$ and $P' \leq P$ are derivable from $\overline{\Gamma}_1$. The second conjunct therefore allows us to derive the formula $P' = P$ – that is, $\forall x(\lambda x(x = a)x \equiv Px)$ – and from this one can conclude at once that $\neg Pb$, since Γ_1 contains the information that $a \neq b$.

15.4.2. Variable circumscription

The inference relation defined by the theory of predicate circumscription allows one, for example, to formalize the kind of closed-world reasoning illustrated in section 5.2.3 by circumscribing the extension of the predicate *Connects*; one could then conclude that there is no direct flight connecting Washington and Barcelona. It turns out, however, that this theory is of severely limited applicability for the simple reason that it never allows new positive conclusions to be drawn by default.

This failure can be illustrated by returning again the initial example. Given the information that Tweety is a bird and that birds fly, how could one use the theory of circumscription to reach the conclusion that Tweety flies? It was suggested by McCarthy that defaults might naturally be represented in the theory through an appeal to explicit abnormality predicates. Where the predicate AB stands for abnormality with respect to flying, for example, the statement that birds fly might be represented through the formula $\forall x((Bx \wedge \neg ABx) \supset Fx)$ – saying that all birds that are not abnormal in this respect fly. Suppose Γ_3 contains this statement as well as Bt then it might seem that one should be able to reach the conclusion Ft simply by circumscribing the predicate AB, ensuring that there are no more abnormal birds than necessary.

In fact, this is a reasonable idea, but it fails for technical reasons, as can be seen by considering the model $\mathcal{M} = \langle \mathcal{D}, v \rangle$, with $\mathcal{D} = \{t\}$, $v(B) = \{t\}$, $v(AB) = \{t\}$, and $v(F) = \emptyset$. Of course, $\mathcal{M}$ does not support the statement Ft, but it turns out that it is an AB-minimal model of Γ_3. The only way of decreasing the extension of the predicate AB, while still modeling Γ_3, would result in increasing the extension of the predicate F – but this violates clause (ii) of definition 15.4, which tells us that models involved in a preference ordering with respect to a particular predicate must agree in their treatment of all other predicates.

Because of this problem, McCarthy (1986) elaborated the basic theory of predicate circumscription into a more flexible theory of variable circumscription, which orders models with respect to a pair of predicates, P and Z. The idea is that those models are preferred that minimize the extension of P while agreeing on everything else, with the possible exception of the predicate Z, whose extension is allowed to vary.

Definition 15.8 Where $\mathcal{M}_1 = \langle \mathcal{D}_1, v_1 \rangle$ and $\mathcal{M}_2 = \langle \mathcal{D}_2, v_2 \rangle$ are models and where P and Z are distinct predicates, then $\mathcal{M}_1 \preceq_{P,Z} \mathcal{M}_2$ just in case

(i) $\mathcal{D}_1 = \mathcal{D}_2$,
(ii) $v_1(Q) = v_2(Q)$ for every linguistic symbol Q other than P and Z, and
(iii) $v_1(P) \subseteq v_2(P)$.

This weak preference ordering is reflexive and transitive, but it is not anti-symmetric, since it is possible for distinct models, agreeing in their interpretation of every predicate but Z, to bear the $\preceq_{P,Z}$ relation to one another. Still, one can define a corresponding strong preference ordering between models by requiring the weak ordering to hold in only one direction.

Definition 15.9 Where $\mathcal{M}_1$ and $\mathcal{M}_2$ are models and where P and Z are distinct predicates, then $\mathcal{M}_1 \prec_{P,Z} \mathcal{M}_2$ just in case $\mathcal{M}_1 \preceq_{P,Z} \mathcal{M}_2$ and it is not the case that $\mathcal{M}_2 \preceq_{P,Z} \mathcal{M}_1$.

And then the pattern set out above can be followed in defining the P,Z-minimal models in a class, and the corresponding notion of consequence.

Definition 15.10 Let $\mathcal{K}$ be a set of models and P and Z distinct predicates. Then $\mathcal{M}$ is ***P;Z-minimal*** in $\mathcal{K}$ just in case $\mathcal{M} \in \mathcal{K}$ and there is no $\mathcal{M}' \in \mathcal{K}$ such that $\mathcal{M}' <_{P;Z} \mathcal{M}$.

Definition 15.11 Where Γ is a set of formulas and A a formula and P and Z are distinct predicates, A is a ***P;Z-minimal consequence*** of Γ – written $\Gamma \Vdash_{P;Z} A$ – just in case $\mathcal{M} \vDash A$ for every $\mathcal{M}$ that is $P;Z$-minimal in the set $|\Gamma|$.

These ideas can be illustrated by returning once again to the initial example. As already seen, the formula Ft is not an AB-minimal consequence of Γ_3, since the model $\mathcal{M}$ defined above is AB-minimal in the model class of Γ_3 but does not support this statement. One can now, however, define the model $\mathcal{M}' = \langle \mathcal{D}', v' \rangle$, like $\mathcal{M}$ except that $v'(AB) = \emptyset$ and $v'(F) = \{t\}$. It is then easy to see that $\mathcal{M}' <_{AB;F} \mathcal{M}$, so that $\mathcal{M}$ is not $AB;F$-minimal, that $\mathcal{M}'$ is itself $AB;F$-minimal, and that every $AB;F$-minimal model of Γ_3 supports the statement Ft, so that now $\Gamma_3 \Vdash_{AB;F} Ft$.

As before, a sound second-order syntactic characterization of the notion of $P;Z$-minimal consequence can be provided through the following circumscription formula, abbreviated $Circ[\Gamma; P; Z]$ and expressing the result of circumscribing the predicate P in the theory Γ while allowing Z to vary:

$$\overline{\Gamma} \wedge \neg\exists P', Z'[\overline{\Gamma^{P'/P\ Z'/Z}} \wedge P' < P]$$

And again, the variable circumscription formula $Circ[\Gamma; P; Z]$ can be seen to hold in exactly the $P;Z$-minimal models of the theory Γ, from which it follows immediately that variable circumscription is sound with respect to $P;Z$-minimal consequence.

Theorem 15.4 Let Γ be a finite set of sentences, P and Z distinct predicates, and $\mathcal{M}$ a model. Then $\mathcal{M} \vDash Circ[\Gamma; P; Z]$ just in case $\mathcal{M}$ is $P;Z$-minimal in $|\Gamma|$.

Theorem 15.5 Let Γ be a finite set of sentences, P and Z distinct predicates, and A a formula. Then $\Gamma \Vdash_{P;Z} A$ whenever $Circ[\Gamma; P; Z] \vdash A$.

The application of this new variable circumscription formula can be illustrated through the initial example, deriving Ft from Γ_3 by circumscribing AB while allowing F to vary. As before, begin by rewriting $Circ[\Gamma_3; AB; F]$ as

$$\overline{\Gamma}_3 \wedge \forall P'Z'[(\overline{\Gamma_3^{P'/AB\ Z'/F}} \wedge P' \leq AB) \supset P' = AB]$$

Then, the second conjunct of this formula can be instantiated by identifying P' with the empty predicate $\lambda x(x \neq x)$ and identifying Z' with $\lambda x(x = t)$. It is a straightforward matter, using the information from $\overline{\Gamma}_3$, to verify both $\overline{\Gamma_3^{P'/AB\ Z'/F}}$ and $P' \leq AB$, and so one can conclude that $P' = AB$ – i.e., that $\forall x(\lambda x(x \neq x)x \equiv ABx)$. From this it follows at once, of course, that $\neg ABt$, which allows one to conclude, again using the information from $\overline{\Gamma}_3$, that Ft.

15.4.3. Parallel and prioritized circumscription

The theory of circumscription set out here has been generalized in a number of ways. Two are sketched – parallel circumscription, which allows several predicates to be circumscribed at once, while several others vary; and prioritized circumscription, which allows some predicates to be circumscribed with higher priority than others.

In fact, the theory of parallel circumscription is best seen simply as a notational elaboration of the previous theory. Suppose that, while allowing $X \subseteq Y$ to carry its usual meaning when X and Y are sets, this notation is generalised so that, when $X = X_1, \ldots, X_n$ and $Y = Y_1, \ldots, Y_n$ are n-tuples of sets, $X \subseteq Y$ means that $X_i \subseteq Y_i$ for each i between 1 and n. Suppose also that, where $P = P_1, \ldots, P_n$ is a tuple of predicates, $\nu(P)$ represents the tuple $\nu(P_1), \ldots, \nu(P_n)$ of extensions assigned to these predicates by the interpretation ν. And finally, suppose that, where $P = P_1, \ldots, P_n$ and $Q = Q_1, \ldots, Q_n$ are n-tuples of predicates, with each P_i taking the same number of arguments as the corresponding Q_i, let $P \leq Q$ mean $P_1 \leq Q_1 \wedge \cdots \wedge P_n \leq Q_n$, and take $P < Q$ and $P = Q$ to be defined as before.

Once these notational enhancements are in place, the theory of parallel circumscription can be presented just as before – in definition 15.8 through theorem 1.5.5 – with the sole exception that now P and Z must be disjoint tuples of predicates instead of distinct individual predicates: rather than looking at models in which the individual predicate P is circumscribed, look at models in which the various predicates belonging to the tuple P are circumscribed in parallel.

To illustrate this theory, return to the Nixon Diamond from figure 15.2, here represented through the theory Γ_4, containing the statements Qn and Rn, saying that Nixon is a Quaker and a Republican, as well as the statements

$$\forall x((Qx \wedge \neg AB_1x) \supset Px)$$

and

$$\forall x((Rx \wedge \neg AB_2x) \supset \neg Px)$$

saying that Quakers that are normal in one respect are pacifists, and that Republicans normal in an another respect are not. To decide whether to conclude that Nixon is a pacifist, it seems reasonable to minimize both sorts of abnormality in parallel, while allowing the predicate P to vary – focusing, that is, on the AB_1, AB_2;P-minimal models. The reader can then verify that Γ_4 has one AB_1, AB_2;P-minimal model that assigns an empty extension to AB_1 and supports the conclusion Pn, as well as another that assigns an empty extension to AB_2 and supports the conclusion $\neg Pn$. Since neither Pn nor $\neg Pn$ is supported by all AB_1, AB_2;P-minimal models of Γ_4, one can conclude that neither formula is an AB_1, AB_2;P-minimal consequence of this theory. And by the soundness of circumscription with respect to minimal consequence, one can conclude also that neither Pn nor $\neg Pn$ can be derived from the parallel circumscription formula $Circ[\Gamma_4; AB_1, AB_2; P]$.

In the case of the Nixon Diamond, it does seem reasonable to minimize the abnormalities associated with Quakers and Republicans in parallel; but in other cases, when defaults have different degrees of strength, it is more natural to assign a higher priority to the minimization of some abnormalities than others. An example is provided by the Tweety Triangle, from figure 15.1, which can be represented through the theory Γ_5, containing the statements Pt and $\forall x(Px \supset Bx)$, saying that Tweety is a penguin and that all penguins are birds, as well as the statements

$$\forall x((Bx \wedge \neg AB_1 x) \supset Fx)$$

and

$$\forall x((Px \wedge \neg AB_2 x) \supset \neg Fx)$$

saying that birds normally fly but that penguins normally do not. Here, if one minimizes the two abnormalities in parallel, again, as in the Nixon Diamond, there are some minimal models supporting the formula Ft supported and others supporting $\neg Ft$, so that one is unable to draw any conclusions. It seems more natural, however, to minimize the abnormality associated with penguins with a higher priority than that associated with birds, so that all minimal models then support the desired conclusion $\neg Ft$.

To develop the theory of prioritized circumscription leading to this result, first define the relation

$$\langle X_1, X_2 \rangle \sqsubseteq \langle Y_1, Y_2 \rangle$$

to mean that

(i) $X_1 \subseteq Y_1$ and
(ii) if $X_1 = Y_1$ then $X_2 \subseteq Y_2$.

Although this new relation can actually be taken – using the enhanced notation just introduced in connection with parallel circumscription – as holding between pairs of tuples of sets, things can be kept simple by reading it as a relation between pairs of sets, and use it to define the following preference ordering on models.

> **Definition 15.12** Where $\mathcal{M}_1 = \langle \mathcal{D}_1, v_1 \rangle$ and $\mathcal{M}_2 = \langle \mathcal{D}_2, v_2 \rangle$ are models and where P, Q and Z are distinct predicates, then $\mathcal{M}_1 \preceq_{P>Q;Z} \mathcal{M}_2$ just in case
>
> (i) $\mathcal{D}_1 = \mathcal{D}_2$
> (ii) $v_1(R) = v_2(R)$ for every linguistic symbol R other than P, Q, or Z, and
> (iii) $\langle v_1(P), v_1(Q) \rangle \sqsubseteq \langle v_2(P), v_2(Q) \rangle$.

The idea behind this weak prioritized ordering is that those models are preferred that minimize the extensions assigned to both the predicates P and Q while allowing Z to vary, but that minimizing P is assigned a higher priority than minimizing Q.

Once this weak prioritized preference ordering has been defined, the development of the theory follows the pattern set out earlier. A corresponding strong ordering can be introduced as in definition 15.9, with $\mathcal{M}_1 <_{P>Q;Z} \mathcal{M}_2$ taken to mean that $\mathcal{M}_1 \preceq_{P>Q;Z} \mathcal{M}_2$ and it is not the case that $\mathcal{M}_2 \preceq_{P>Q;Z} \mathcal{M}_1$. The minimal elements of a class of models can then be defined as in definition 15.10, with $\mathcal{M}$ taken as $P > Q$; Z-minimal in the class $\mathcal{K}$ whenever $\mathcal{M}$ belongs to $\mathcal{K}$ and there is no $\mathcal{M}'$ from $\mathcal{K}$ such that $\mathcal{M}' <_{P>Q;Z} \mathcal{M}$. And the appropriate notion of consequence can be defined as in definition 15.11, with A taken to be a $P > Q$; Z-minimal consequence of Γ – written, $\Gamma \Vdash_{P>Q;Z} A$ – whenever $\mathcal{M} \vDash A$ for each $P > Q$; Z-minimal model $\mathcal{M}$ from $|\Gamma|$. With these definitions in hand, the reader can then verify that $\Gamma_5 \Vdash_{AB_2>AB_1;F} \neg Ft$ – i.e., that the statement $\neg Ft$ follows as a consequence of Γ_5 when the predicate AB_2 is minimized with a higher priority than AB_1, allowing F to vary.

Turning to the proof theory for prioritized circumscription, begin by defining $\langle P_1, P_2\rangle \leq \langle Q_1, Q_2\rangle$ as an abbreviation of the statement

$$P_1 \leq Q_1 \wedge (P_1 = Q_1 \supset P_2 \leq Q_2)$$

and then taking $\langle P_1, P_2\rangle < \langle Q_1, Q_2\rangle$ to mean that

$$\langle P_1, P_2\rangle \leq \langle Q_1, Q_2\rangle \wedge \neg(\langle Q_1, Q_2\rangle < \langle P_1, P_2\rangle)$$

The circumscription formula for minimizing P with higher priority than Q in the theory Γ while allowing Z to vary, abbreviated as $Circ[\Gamma;\ P > Q;\ Z]$, can now be expressed through the second-order statement

$$\overline{\Gamma} \wedge \neg\exists P', Q', Z'[\overline{\Gamma^{P'/P\, Q'/Q\, Z'/Z}} \wedge \langle P_1, P_2\rangle < \langle Q_1, Q_2\rangle]$$

Analogues to theorems 15.4 and 15.5 can be established, saying that $Circ[\Gamma;\ P > Q;\ Z]$ holds in exactly the $P > Q;Z$-minimal models of Γ, and therefore, that prioritized circumscription is sound with respect to the appropriate prioritized notion of minimal consequence. And the interested reader can verify that $\neg Ft$ is indeed derivable from the formula $Circ[\Gamma_5;\ AB_2 > AB_1;\ F]$.

It should be clear that the theories presented here of parallel and prioritized circumscription can be combined and generalized, so that groups of predicates can be minimized in parallel, but all with higher priority than other groups of predicates. One could, for example, speak of the $P_1, P_2 > P_3 > P_4, P_5; Z_1, Z_2$-minimal models as those obtained by minimizing the predicates P_1 and P_2 in parallel with higher priority than P_3, which is itself minimized with higher priority than P_4, and P_5, all the while allowing Z_1 and Z_2 to vary. Note, however, that – just as with default logic – it is still necessary to specify the preferences among various competing defaults by hand, in this case by explicitly tailoring the priorities involved in the minimization ordering, rather than coding these preferences into non-normal default rules.

Suggested further reading

Many of the original papers on nonmonotonic logic are reprinted in Ginsberg (1987). A more recent collection is Gabbay et al. (1994), which contains several valuable survey articles on different approaches. There have been a number of variations on the general themes introduced in Reiter's default logic; the most readable and comprehensive presentation of these is Delgrande et al. (1994). Another fixed-point theory of nonmonotonic reasoning, closely related to default logic, is the modal approach of McDermott and Doyle (1987 [1980]). This modal approach was refined in Moore (1985); relations to default logic are established in Konolige (1988 [1987]). The best general survey of the theory of circumscription is Lifschitz (1994). Different model-preference approaches, based on different preference orderings can be found in Kautz (1986) and Shoham (1988). A general study of nonmonotonic consequence relations, with a special emphasis on model preference logics, was initiated by Makinson (1989) and Kraus et al. (1990).

Notes

[1] Just as ordinary inference rules allow multiple premises, default rules allow multiple prerequisites and also multiple justifications; we limit our attention to default rules in which prerequisites and justification are unique for ease of exposition.

[2] The use of the *credulous/skeptical* terminology to characterize these two broad reasoning strategies was first introduced in Touretzky et al. (1987), but the distinction is older than this; it was noted already by Reiter, and was described in McDermott (1982) as the distinction between *brave* and *cautious* reasoning.

[3] Reiter provides a proof procedure, sound and complete under certain conditions, for determining whether a formula is believable in this sense on the basis of a default theory. A different interpretation of this second credulous option is provided in Horty (1994), which interprets default logic as a deontic logic allowing for moral conflicts.

[4] Unfortunately, although the treatment of the frame problem suggested here does seem to work for the simple example set out in section 15.2.1, it was shown in Hanks and McDermott (1987) that this straightforward kind of nonmonotonic approach delivers anomalous results in situations that are only slightly more complicated. Since then, a number of more sophisticated encodings of actions and their effects in various nonmonotonic logics have been explored, such as those of Lifschitz (1994) and Morgenstern and Stein (1988), as well as renewed attempts to resolve the frame problem in ordinary monotonic logics, such as that of Reiter (1991). The field is now an area of active research; a recent survey can be found in Shanahan (1997).

References

Delgrande, J., Schaub, T. and Jackson, W. K. 1994: "Alternative Approaches to Default Logic," *Artificial Intelligence*, 70, 167–237.

Etherington, D. and Reiter, R. 1983: "On Inheritance Hierarchies with Exception," in *Proceedings of AAAI-83*, (William Kaufman, Los Altos, CA), 104–8.

Etherington, D., Mercer, R. and Reiter, R. 1985: "On the Adequacy of Predicate Circumscription for Closed-World Reasoning," *Computational Intelligence*, 1, 11–15.

Gabbay, D. M., Hogger, C. and Robinson, J. A. (eds.) 1994: *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume 3: Nonmonotonic Reasoning and Uncertain Reasoning*, (Oxford University Press, Oxford).

Ginsberg, M. (ed.) 1987: *Readings in Nonmonotonic Reasoning*, (Morgan Kaufmann, Los Altos, CA).

Hanks, S. and McDermott, D. 1987: "Nonmonotonic Logic and Temporal Projection," *Artificial Intelligence*, 33, 379–412.

Horty, J. 1994: "Moral Dilemmas and Nonmonotonic Logic," *Journal of Philosophical Logic*, 23, 35–65.

Kautz, H. 1986: "The Logic of Persistence," in *Proceedings of AAAI-86*, (Morgan Kaufmann, Los Altos, CA), 401–5.

Konolige, K. 1988: "On the Relation between Default Theories and Autoepistemic Logic," *Artificial Intelligence*, 35, 343–82; also in Ginsberg (1987, 195–226).

Kraus, S., Lehman, D. and Magidor, M. 1990: "Nonmonotonic Reasoning, Preferential Models, and Cumulative Logics," *Artificial Intelligence*, 44, 167–207.

Lifschitz, V. 1994: "Circumscription," in Gabbay et al. (1994, 297–352).

Makinson, D. 1989: "General Theory of Cumulative Inference," in *Proceedings of the Second International Workshop on Nonmonotonic Reasoning*, M. Reinfrank, J. de Kleer, M. Ginsberg and E. Sandewall, eds., Springer-Verlag Lecture Notes in Artificial Intelligence, 346, 1–18.

McCarthy, J. 1980: "Circumscription – A Form of Non-Monotonic Reasoning," *Artificial Intelligence*, 13, 27–39.

McCarthy, J. 1986: "Applications of Circumscription to Formalizing Commonsense Knowledge," *Artificial Intelligence*, 28, 89–116.

McCarthy, J. and Hayes, P. 1969: "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence, volume 4*, B. Meltzer and D. Michie, eds., (Edinburgh Press, Edinburgh), 463–503.

McDermott, D. 1982: "A Temporal Logic for Reasoning about Processes and Plans," *Cognitive Science*, 6, 101–55.

McDermott, D. and Doyle, J. 1987: "Non-Monotonic Logic – I," *Artificial Intelligence*, 13 (1980), 41–72; reprinted in Ginsberg (1987, 111–26).

Moore, R. 1985: "Semantical Considerations on Nonmonotonic Logic," *Artificial Intelligence*, 25, 75–94.

Morgenstern, L. and Stein, L. 1988: "Why Things Go Wrong: A Formal Theory of Causal Reasoning," in *Proceedings of AAAI-88*, (Morgan Kaufmann, Los Altos, CA), 518–23.

Reiter, R. 1978: "On Closed World Data Bases," in *Logic and Data Bases*, H. Gallaire and J. Minker, eds., (Plenum Publishing Corp., New York), 119–40.

Reiter, R. 1980: "A Logic for Default Reasoning," *Artificial Intelligence*, 13, 81–132.

Reiter, R. 1991: "The Frame Problem in the Situation Calculus: A Simple Solution (Sometimes) and a Completeness Result for Goal Regression," in *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*, V. Lifschitz, ed., (Academic Press, Boston), 359–80.

Shanahan, M. 1997: *Solving the Frame Problem*, (The MIT Press, Cambridge, MA).

Shoham, Y. 1988: *Reasoning about Change*, (The MIT Press, Cambridge, MA).

Touretzky, D., Horty, J. and Thomason, R. 1987: "A Clash of Intuitions: The Current State of Nonmonotonic Multiple Inheritance Systems," in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87)*, (Morgan Kaufmann, Los Altos, CA), 476–82.