Default Reasoning, Nonmonotonic Logics, and the Frame Problem

Steve Hanks and Drew McDermott[1]

Department of Computer Science, Yale University
Box 2158 Yale Station
New Haven, CT 06520

Abstract

Nonmonotonic formal systems have been proposed as an extension to classical first-order logic that will capture the process of human "default reasoning" or "plausible inference" through their inference mechanisms just as *modus ponens* provides a model for deductive reasoning. But although the technical properties of these logics have been studied in detail and many examples of human default reasoning have been identified, for the most part these logics have not actually been applied to practical problems to see whether they produce the expected results.

We provide axioms for a simple problem in temporal reasoning which has long been identified as a case of default reasoning, thus presumably amenable to representation in nonmonotonic logic. Upon examining the resulting nonmonotonic theories, however, we find that the inferences permitted by the logics are *not* those we had intended when we wrote the axioms, and in fact are much weaker. This problem is shown to be independent of the logic used; nor does it depend on any particular temporal representation. Upon analyzing the failure we find that the nonmonotonic logics we considered are inherently incapable of representing this kind of default reasoning. Finally we discuss two recent proposals for solving this problem.

1 Introduction

Logic as a representation language for AI theories has always held a particular appeal in the research community (or in some parts of it, anyway): its rigid syntax forces one to be precise about what one is saying, and its semantics provide an agreed-upon and well-understood way of assigning meaning to the symbols. But if logic is to be more than just a concise and convenient notation that helps us in the task of writing programs, we somehow have to validate the axioms we write: are the conclusions we can draw from our representation (*i.e.*, the inferences the logic allows) the same as the ones characteristic of the reasoning process we are trying to model? If so, we've gone a long way toward validating our theory.

The limitation of classical logic as a representation for human knowledge and reasoning is that its inference rule, *modus ponens*, is the analogue to human *deductive* reasoning, but for the most part everyday human reasoning seems to have significant non-deductive components. But while certain aspects of human reasoning (*e.g.* inductive generalization and abductive explanation) seem to be substantially different from deduction, a certain class of reasoning, dubbed "default reasoning," resembles deduction more closely. Thus it was thought that extensions to first-order logic might result in formal systems capable of representing the process of default reasoning.

While it is still not clear *exactly* what constitutes default reasoning, the phenomenon commonly manifests itself when we know what conclusions should be drawn about *typical* situations or objects, but we must jump to the conclusion that an observed situation or object *is* typical. For example, I may know that I typically meet with my advisor on Thursday afternoons, but I can't deduce that I will actually have a meeting *next* Thursday because I don't know whether next Thursday is typical. While certain facts may allow me to deduce that next Thursday is *not* typical (*e.g.* if I learn he will be out of town all next week), in general there will be no way for me to deduce that it *is*. What we want to do in cases like this is to jump to the conclusion that next Thursday is typical based on two pieces of information: first that most Thursdays *are* typical, and second that we have no reason to believe that this one is not. Another way to express the same notion is to say that I know that I have meetings on typical Thursdays, and that the only *atypical* Thursdays are the ones that I *know* (can deduce) are atypical.

Research on nonmonotonic logics[2], most notably by McCarthy (in [8] and [9]), McDermott and Doyle (in [12]) and Reiter (in [14]) attacked the problem of extending first-order logic in a way that captured the intuitive meaning of statements of the form "lacking evidence to the contrary, infer α" or more generally "infer β from the inability to infer α."[3] But since that first flurry of research the area has developed in a strange way. On one hand the logics have been subjected to intense technical scrutiny (in the papers cited above, and also, for example, in Davis [2]) and have been shown to produce counterintuitive results under certain circumstances. At the same time we see in the literature practical representation problems such as story understanding (Charniak [1]), social convention in conversation (Joshi, Webber, and Weischedel [6]), and temporal reasoning (McDermott [11] and McCarthy [9]), in which default rules would *seem* to be of use, but in these cases technical details of the formal systems are for the most part ignored.

The middle ground—whether the technical workings of the logics correctly bear out one's intentions in representing practical default-reasoning problems—is for the most part empty, though the work of Reiter, Etherington, and Criscuolo, in [3], [4], and elsewhere, is a notable exception. Logicians have for the most part ignored practical problems to focus on technical details, and "practitioners" have used the default rules intuitively, with the hope (most often unstated) that the proof theory or semantics of the logics can eventually be shown to support those intuitions.

We explore that middle ground by presenting a problem in temporal reasoning that involves default inference, writing axioms in nonmonotonic logics intended to represent that reasoning process,

[1] This work was supported in part by ONR grant N00014-85-K-0301. Many thanks to Alex Kass and Yoav Shoham for discussing this work with us, and for reading drafts of the paper.

[2] So called because of the property that inferences allowed by the logic may be *disallowed* as axioms are added. For example I may jump to the conclusion that next Thursday is typical, thus deduce that I will have a meeting. If I later come to find out that it is *atypical*, I will have to *retract* that conclusion. In first-order logic the addition of new knowledge (axioms) to a theory can never diminish the deductions one can make from that theory, thus it is never necessary to retract conclusions.

[3] This sounds much more straightforward than it is: consider that the theorems of a logic are defined in terms of its inference rules, yet here we are trying to define an inference rule in terms of what is or is not a theorem.

then analyzing the resulting theory. Reasoning about time is an interesting application for a couple of reasons. First of all, the problem of representing the tendency of facts to endure over time (the "frame problem" of McCarthy and Hayes [10] or the notion of "persistence" in McDermott [11]) has long been assumed to be one of those practical reasoning problems that nonmonotonic logics would solve. Second, one has strong intuitions about how the problem *should* be formalized in the three logics, and even stronger intuitions about what inferences should then follow, so it will be clear whether the logics have succeeded or failed to represent the domain correctly.

In the rest of the paper we discuss briefly some technical aspects of nonmonotonic logics, then go on to pose formally the problem of temporal projection. We then analyze the inferences allowed by the resulting theory and show that they do *not* correspond to what we intended when we wrote the axioms. Finally we point out the (unexpected) characteristics of the domain that the logics were unable to capture, and discuss proposed solutions to the problem.

2 Nonmonotonic inference

Since we are considering the question of what inferences can be drawn from a nonmonotonic theory, we should look briefly at how inference is defined in these logics. We will concentrate on Reiter's default logic and on circumscription, but the discussion and subsequent results hold for McDermott's nonmonotonic logic as well.

2.1 Inference in default logic

Reiter in [14] defines a *default theory* as two sets of rules. The first consists of sentences in first-order logic (and is usually referred to as **W**), and the second is a set of *default rules* (referred to as **D**). Default rules are supposed to indicate what conclusions to jump to, and are of the form

$$\frac{\alpha \ : \mathbf{M}\,\beta}{\gamma}$$

where α, β, and γ are first-order sentences. The intended interpretation of this rule is "if you believe α, and it's consistent to believe β, then believe γ," or, to phrase the idea more like an inference rule, "from α and the *inability* to prove $\neg\beta$, infer γ." (But recall our note above about the futility of trying to define inference in terms of inference.)

In order to discuss default inference we must introduce the concept of an *extension*—a set of sentences that "extend" the sentences in **W** according to the dictates of the default rules. A default theory defines zero or more extensions, each of which has the following properties: (1) any extension **E** contains **W**, (2) **E** is closed under (monotonic) deduction, and (3) **E** is faithful to the default rules. By the last we mean that if there's a default rule in the theory of the form $\frac{\alpha \ : \ M\,\beta}{\gamma}$, and if $\alpha \in \mathbf{E}$, and $(\neg\beta) \notin \mathbf{E}$, then $\gamma \in \mathbf{E}$. The extensions of a default theory are all the minimal sets **E** that satisfy these three properties. Extensions can be looked upon as internally consistent and coherent states of the world, though the union of two extensions may be inconsistent.

Finding a satisfying definition of default inference—what sentences can be said to follow from a default theory—is tricky. Reiter avoids the problem altogether, focusing on the task of defining extensions and exploring their properties. He expresses the view that default reasoning is really a process of selecting *one* extension of a theory, then reasoning "within" this extension until new information forces a revision of one's beliefs and hence the selection of a new extension.

This view of default reasoning, while intuitively appealing, is infeasible from a practical standpoint: there is no way of "isolating" a single extension of a theory, thus no procedure for enumerating or testing theoremhood within an extension. So any definition of default reasoning based on discriminating among extensions is actually beyond the expressive power of default logic. Reiter does provide a proof procedure for asking whether a sentence is a member of *any* extension, but, as he points out, this is not a satisfying definition of inference since both a sentence and its negation may appear in different extensions.

Our view in this paper is that *some* notion of inference is necessary to judge the representational power of the logic. A logic that generates one intuitive extension and one unintuitive extension does not provide an adequate representation of the problem, since there is no way to distinguish between the two interpretations. For that reason we will define inference in the manner of McDermott's logic: a sentence δ can be inferred from a default theory just in case δ is in *every* extension of that theory. (This definition is also consistent with circumscriptive inference as described in the next section.)

While there is no general procedure for determining how many extensions a given theory has, as a practical matter it has been noted that theories with "conflicting" default rules tend to generate multiple extensions. For example, the following default theory

$$\mathbf{W} = \{Q(N), R(N)\}, \quad \mathbf{D} = \{\frac{Q(x) \ : \ \mathbf{M}\,P(x)}{P(x)}, \frac{R(x) \ : \ \mathbf{M}\,\neg P(x)}{\neg P(x)}\}$$

has two rules that would have us jump to contradictory conclusions. But note that applying one of the rules means that the other cannot be applied, since its precondition is not met. This default theory has two extensions: $\mathbf{E}_1 = \{Q(N), R(N), P(N)\}$ and $\mathbf{E}_2 = \{Q(N), R(N), \neg P(N)\}$ that correspond to the two choices one has in applying the default rules. (One interpretation of this theory reads Q as "Quaker," R as "Republican," P as "Pacifist," and N as "Nixon.") Thus the above theory entails only the sentences in **W**, plus tautologies (for example $P(N) \vee \neg P(N)$).

We are not claiming that this admission of multiple extensions is a fault or deficiency of the logic—in this particular example it's hard to imagine how the logic could license any *other* conclusions. Our point is that when a theory generates multiple extensions it's generally going to be the case that only weak inferences can be drawn. Further, if one extension captures the *intended* interpretation but there are other different extensions, it will not be possible to make only the intended inferences.

2.2 Inference in circumscription

To describe inference in circumscribed theories we will have to be rather vague: there are several versions of the logic, defined in [8], [9] and elsewhere, and we will not spend time discussing these differences.

We will speak generally of *predicate circumscription*, in which the intent is to minimize the extension of a predicate (say P) in a set of first-order axioms. Using terms like those we used in describing default logic, we might say that when we circumscribe axioms over P we intend that "the only individuals for which P holds are those individuals for which P *must* hold," or alternatively we might phrase it as "believe '*not P*' by default."

To circumscribe a set of axioms **A** over a predicate P one adds to **A** an axiom (the exact form of which is not important for our discussion) that says something like this: "any predicate P' that satisfies the axioms **A**, and is at least as strong as P, is *exactly* as strong as P." The intended effect is (roughly) that for any individual x,

if $\mathbf{A} \not\vdash P(x)$ then
$\mathrm{Circum}(\mathbf{A}, P) \vdash \neg P(x)$

where Circum(**A**, P) refers to the axioms **A** augmented by the circumscription axiom for P.

To talk about circumscriptive inference, we should first note that since Circum(**A**, P) is a first-order theory we want to know what *deductively* follows, but we are interested in characterizing these deductions in terms of the original axioms **A**. In brief, the results are these: if a formula φ is a theorem of Circum(**A**, P) then φ is true in all models of **A** *minimal in* P (this property is called *soundness*), and if a formula φ is true in all models of **A** minimal in P then φ is a theorem of Circum(**A**, P) (this property is called *completeness*). Completeness does not hold for all circumscribed theories, but it does hold in certain special cases—see Minker and Perlis [13].

Minimal models, the model-theoretic analogue to default-logic extensions, are defined as follows: a model $\mathcal{M}$ is minimal in P just in case there is no model $\mathcal{M}'$ that agrees with $\mathcal{M}$ on all predicates *except* for P, but whose extension of P is a *subset* of $\mathcal{M}$'s extension of P.

As with default logic, there is no effective procedure for determining how many minimal models a theory has. And note that the converse of the soundness property says that if φ is *not* true in all models minimal in P it does *not* follow from Circum(**A**, P). So once again, if we have multiple minimal models, what we can deduce are only those formulas true in all of them. Because of the obvious parallels between extensions and minimal models (and "NM fixed points" in McDermott's logic, which we will not discuss here), we will use the terms interchangeably when the exact logic or object doesn't matter.

3 The temporal projection problem

The problem we want to represent is this: given an initial description of the world (some facts that are true), the occurence of some events, and some notion of causality (that an event occuring can cause a fact to become true), what facts are true once all the events have occured?

We obviously need some temporal representation to express these concepts, and we will use the *situation calculus* [10]. We will thus speak about *facts* holding true in *situations*. A fact syntactically has the form of a first-order sentence, and is intended to be an assertion about the world, such as $SUNNY$, $LOADED(GUN\text{-}35)$, or $\forall\, x.HAPPY(x)$. Situations are individuals denoting intervals of time over which facts hold or do not hold, but over which *no* fact changes its truth value. This latter property allows us to speak unambiguously about what facts are true or false in a situation. To say that a fact f is true in a situation s we assert $T(f,\, s)$, where T is a predicate and f and s are terms.

Events are things that happen in the world, and the occurence of an event may have the effect of changing the truth value of a fact. So we think of an event occuring in a situation and causing a transition to another situation—one in which the event's effects on the world are reflected. The function $RESULT$ maps a situation and an event into another situation, so if S_0 is a situation and $WAKEUP(JOHN)$ is an event, then $RESULT(WAKEUP(JOHN),\, S_0)$ is also a situation—presumably the one resulting from $JOHN$ waking up in situation S_0. We might then want to state that $JOHN$ is awake in this situation:

$$T(\,AWAKE(JOHN),\ RESULT(\,WAKEUP(JOHN),\ S_0))$$

or more generally we might state that

$$\forall\, p,\, s.\ T(AWAKE(p),\ RESULT(WAKEUP(p),\ s)).$$

A problem arises when we try to express the notion that facts tend to *stay* true from situation to situation as irrelevant events occur. For example, is $JOHN$ still awake in the state S_2, where

$$S_2 = RESULT(EAT\text{-}BREAKFAST(JOHN),\ RESULT(WAKEUP(JOHN),\ S_0))?$$

Intuitively we would like to assume so, because it's typically the case that eating breakfast does not cause one to fall asleep. But given the above axioms there is no way to deduce

$$T(\,AWAKE(JOHN),\ S_2).$$

We could add an axiom to the effect that if one is awake in a situation then one is still awake after eating breakfast, but this seems somewhat arbitrary (and will occasionally be false). And in any reasonable description of what one might do in the course of a morning there would have to be a staggering number of axioms expressing something like "if fact f is true in a situation s, and e is an event, then f is still true in the situation $RESULT(e,\, s)$." McCarthy and Hayes (in [10]) call axioms of this kind "frame axioms," and identified the "frame problem" as that of having to explicitly state many such axioms. Deductive logic forces us into the position of assuming that an event occurence may potentially change the truth value of all facts, thus if it does *not* change the value of a particular fact in a particular situation we must explicitly say so. What we would like to do is assume just the opposite: that most events do *not* affect the truth of most facts under most circumstances.

Intuitively we want to solve the frame problem by assuming that in general an event happening in a situation is *irrelevant* to a fact's truth value in the resulting situation. Or, to make the notion a little more precise, we want to assume "by default" that

$$T(f,\, s) \supset\ T(f,\ RESULT(e,\, s))$$

for all facts, situations, and events. But note the quotes: the point of this paper is that formalizing this assumption is not as straightforward as the phrase might lead one to believe.

McCarthy's proposed solution to the frame problem (described in [9]) involves extending the situation calculus a little, to make it what he calls a "simple abnormality theory." We state that all "normal" facts persist across occurences of "normal" events:

$$\forall\, f,\, e,\, s.\ T(f,\, s)\ \wedge\ \neg\, AB(f,\, e,\, s) \supset\ T(f,\ RESULT(e,\, s))$$

where $AB(f,\, e,\, s)$ is taken to mean "fact f is abnormal with respect to event e occuring in state s," or, "there's something about event e occuring in state s that causes fact f to stop being true in $RESULT(e,s)$." We would expect, for example, that it would be true that

$$\forall\, p,\, s.\ AB(AWAKE(p),\ GOTOSLEEP(p),\ s)$$

and we would have to add a specific axiom to that effect.

Of course we still haven't solved the frame problem, since we haven't provided any way to deduce $\neg\, AB(f,e,s)$ for most facts, events, and situations. As an alternative to providing myriad frame axioms of this form, McCarthy proposes that we circumscribe over the predicate AB, thus "minimizing" the abnormal temporal individuals. The question is whether this indeed represents what we intuitively mean by saying that we should "assume the persistence of facts by default," or "once true, facts tend to stay true over time."

As an illustration of what we *can* infer from a very simple situation-calculus abnormality theory, consider the axioms of Figure 1. For simplicity we have restricted the syntactic form of facts and events to be propositional symbols, so the axioms can be interpreted as referring to a single individual who at any point in time (situation) can be either $ALIVE$ or $DEAD$ and a gun that can be either $LOADED$ or $UNLOADED$. At some known situation S_0 the person is alive (Axiom 1), and the gun becomes loaded any time a $LOAD$ event happens (Axiom 2). Axiom 3 says that any time the person is shot with a loaded gun he becomes dead, and furthermore that being shot with a loaded gun is abnormal with respect to staying alive. Or to use our definition from above: there is something about a $SHOOT$ event occuring in a situation in which the gun is loaded that causes the fact $ALIVE$ to stop being true in the situation resulting from the shot. Axiom 4 is just the assertion we made above that "normal" facts persist across the occurence of "normal" events.

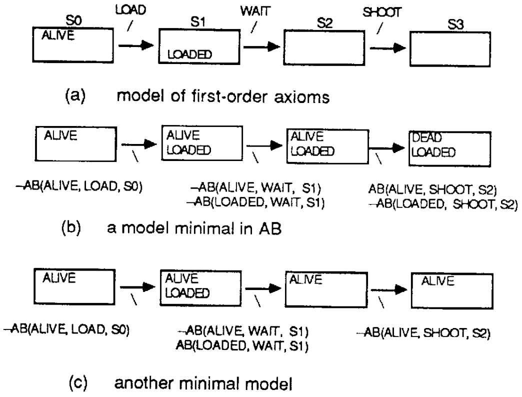
(1) $T(ALIVE, S_0)$

(2) $\forall s.\ T(LOADED, RESULT(LOAD, s))$

(3) $\forall s.\ T(LOADED, s) \supset AB(ALIVE, SHOOT, s) \wedge$
$\quad T(DEAD, RESULT(SHOOT, s))$

(4) $\forall f, e, s.\ T(f, s) \wedge \neg AB(f, e, s) \supset T(f, RESULT(e, s))$

Figure 1: Simple situation-calculus axioms



(a) model of first-order axioms

(b) a model minimal in AB

(c) another minimal model

Figure 2: Three models of the Figure 1 axioms

Then we circumscribe Axioms 1–4 over *AB*, recalling that the deducible formulas will be those that are true in all models minimal in *AB*. As above we will refer to Axioms 1–4 as **A**, and to the circumscribed axioms as Circum(**A**, *AB*).

Now consider the problem of projecting what facts will be true in the following situations:

S_0
$S_1 = RESULT(LOAD, S_0),$
$S_2 = RESULT(WAIT, S_1),$ and
$S_3 = RESULT(SHOOT, S_2)$
$\quad = RESULT(SHOOT, RESULT(WAIT, RESULT(LOAD, S_0))).$

In other words, our individual is initially known to be alive, then the gun is loaded, then he waits for a while, then he is shot with the gun. The projection problem is to determine what facts are true at the situations S_i. The event *WAIT* is supposed to signify a period of time when nothing of interest happens. Since according to the axioms no fact is abnormal with respect to a *WAIT* event occuring, we intend that every fact true before the event happens should also be true *after* it happens.

One interpretation of the Figure 1 axioms is shown in Figure 2a. This first picture represents facts true in all models of **A** (thus what we can deduce if we *don't* circumscribe). From Axioms 1 and 2 we can make the following deductions:

$T(ALIVE, S_0), \qquad T(LOADED, S_1),$

but we can deduce nothing about what is true in S_2 or in S_3. We also cannot deduce any "abnormalities" nor their negations. But this is pretty much as expected: the *ALIVE* fact did not persist because we could not deduce that it was "not *AB*" with respect to loading the gun, and the gun being loaded did not persist through the *WAIT* event because we could not deduce that it was "not *AB*" with respect to waiting.

Intuitively we would like to reason about "minimizing abnormalities" like this: we know *ALIVE* must be true in S_0, and nothing compels us to believe $AB(ALIVE, LOAD, S_0)$, so we assume its negation. From this assumption and from Axiom 4 we deduce $T(ALIVE, S_1)$. Reasoning along the same lines, we can deduce $T(LOADED, S_1)$ and we are free to make the assumptions $\neg AB(ALIVE, WAIT, S_1)$ and $\neg AB(LOADED, WAIT, S_1)$ so we do so and go on to deduce $T(ALIVE, S_2)$ and $T(LOADED, S_2)$. Again moving forward in time, we can deduce from Axiom 3 that $AB(ALIVE, SHOOT, S_2)$ so we can't assume its negation, but we *can* assume $\neg AB(LOADED, SHOOT, S_2)$. At that point we have $T(DEAD, S_3)$ and $T(LOADED, S_3)$.

This line of reasoning leads us to the interpretation of Figure 2b. We can easily verify that this interpretation is indeed a model of **A** (that Axioms 1 through 4 are satisfied). Furthermore, this model is minimal in *AB*: any submodel would have to have an empty extension for *AB*, which cannot be the case.[5]

The interesting question now is whether the model of Figure 2b—our *intended* model of **A**—is the *only* minimal model, or, more to the point, whether $T(DEAD, S_3)$ and $T(LOADED, S_2)$ are true in all minimal models. Because if they're not true in all minimal models they can't be deduced from Circum(**A**, *AB*).

Consider the situation in Figure 2c. The picture describes a state of affairs in which the gun ceases to be loaded "as a result of" waiting. Then the individual *does not* die as a result of the shot, since the gun is not loaded. Of course this state of affairs directly contradicts our stated intention that since nothing is explicitly "*AB*" with respect to waiting everything should be "not *AB*" with respect to waiting. Does this interpretation describe a minimal model? First recall that there can be no models having a null extension for *AB*, so if this interpretation is a model at all it must be minimal in *AB*.

One can "build" this model in much the same way we constructed the model of Figure 2b, except this time instead of starting at S_0 and working forward in time, we will start at S_3 and work backward. In other words, we will start by noting that $T(ALIVE, S_2)$ must be true, then consider what must have been true at S_2 and in earlier situations for that to have been the case.

The first abnormality decision to make is whether $AB(ALIVE, SHOOT, S_2)$ is true. Since we haven't made a decision to the contrary, we will assume its negation. But then from the contrapositive of Axiom 3, we we can deduce $\neg T(LOADED, S_2)$. But if that is the case, and since it must also be the case that $T(LOADED, S_1)$, we can deduce from Axiom 4 that $AB(LOADED, WAIT, S_1)$. The rest of Figure 2c follows directly, since we can assume that *ALIVE* is "not AB" with respect to *LOAD* and *WAIT*, thus deduce that it is true in both S_1 and S_2.

What, then, can be deduced from the (circumscribed) abnormality theory? It's fairly easy to verify that the two models we have presented are the only two minimal in *AB*, so the theorems of Circum(**A**, *AB*) are those common to those two models. So we can deduce that *ALIVE* and *LOADED* are true in S_1, that *ALIVE* is true in S_2, but we can say nothing about what is true in S_3except for statements like $T(ALIVE, S_3) \vee T(DEAD, S_3)$. What we can deduce from Circum(**A**, *AB*) is therefore considerably weaker than what we had intended.

4 How general is this problem?

The question now arises: how dependent is this result on the specific problem and formulation we just presented? Does the same problem arise if we use a different default logic or a different temporal formalism?

[5] To see why this is true, consider that in any model of **A** either $T(ALIVE, S_2)$ is true, or it's false. If it's true we can immediately deduce an abnormality from Axiom 3. But if it's false then either $AB(ALIVE, LOAD, S_0)$ or $AB(ALIVE, WAIT, S_1)$ would have to be true. In either case we must have at least one abnormality.

We can easily express the theory above in Reiter's logic: we use the same first-order axioms from Figure 1, but instead of circumscribing over AB we represent "minimizing abnormal individuals" with a class of (normal) default rules of the form

$$D = \{\frac{:\mathbf{M}\neg AB(f,e,s)}{\neg AB(f,e,s)}\}$$

(where any individual may be substituted for f, e, and s). Recall that extensions are defined proof-theoretically instead of in terms of models, so we must translate the minimal models shown in Figure 2 (b and c) into sets of sentences; the question becomes whether the following sets are default-logic extensions:

$\mathbf{E_a}$	$\mathbf{E_b}$
$T(ALIVE, S_0)$	$T(ALIVE, S_0)$
$\neg AB(ALIVE, LOAD, S_0)$	$\neg AB(ALIVE, LOAD, S_0)$
$T(ALIVE, S_1)$	$T(ALIVE, S_1)$
$T(LOADED, S_1)$	$T(LOADED, S_1)$
$\neg AB(ALIVE, WAIT, S_1)$	$\neg AB(ALIVE, WAIT, S_1)$
$\neg AB(LOADED, WAIT, S_1)$	$AB(LOADED, WAIT, S_1)$
$T(ALIVE, S_2)$	$T(ALIVE, S_2)$
$T(LOADED, S_2)$	
$AB(ALIVE, SHOOT, S_2)$	$\neg AB(ALIVE, SHOOT, S_2)$
$\neg AB(LOADED, SHOOT, S_2)$	
$T(DEAD, S_3)$	$T(ALIVE, S_3)$
$T(LOADED, S_3)$	

Of course these are partial descriptions of extensions. Each set also contains **A** and all tautologies, and in $\mathbf{E_a}$, for example, we also include all sentences of the form $\neg AB(f, e, s)$ for all individuals (f, e, s) *except* $(ALIVE, SHOOT, S_2)$.

We will omit the proof that both $\mathbf{E_a}$ and $\mathbf{E_b}$ are extensions, though in our longer paper, [5], we carry it out in some detail. It should be easy to convince oneself that both sets satisfy the three conditions we set down in Section 2: they contain **A**, they are closed under deduction, and they are faithful to the default rule in the sense defined previously. To verify that they are both minimal, note that in both cases all the sentences except the default-rule assumptions indeed follow from the default assumptions and the axioms in **A**.

So circumscription is not the culprit here—Reiter's proof-theoretic default logic has the same problem. We can also express the same problem in McDermott's nonmonotonic logic and show that the theory has the same two fixed points.

Nor is the situation calculus to blame: in a previous paper [5] we use a simplified version of McDermott's temporal logic and show that the same problem arises, again for all three default logics. In the next section we will show what characteristics of temporal projection lead to the multiple-extension problem, and why it appears that the three default logics are inherently unable to represent the domain correctly.

5 A minimality criterion for temporal reasoning

We noted above that default-logic theories often generate multiple extensions. But characteristic of all the usual examples, like the one we used in Section 2, is the fact that the default rules of these theories were mutually exclusive: the application of one rule rendered other rules inapplicable by blocking their preconditions.

Thus it comes as somewhat of a surprise that the temporal projection problem should exhibit several extensions. How can there be conflicting rules in the same way we saw above when our theory has only a single default rule? It turns out that conflict between rules arises in our domain in a different, more subtle, manner. To see how, recall how we built the first minimal model (that of Figure 2b). The idea was that we assumed one "normality," then went on to make all possible deductions, then assumed another "normality," and so on. The picture looks something like this:

$$\neg AB(LOADED, WAIT, S_1) \Rightarrow T(LOADED, S_2) \Rightarrow \ldots \Rightarrow AB(ALIVE, SHOOT, S_2)$$

where the conflict to notice is that as a result of assuming a "normality" we could deduce an *abnormality*. The same thing happened when we build the model in Figure 2c, except the picture looks like this instead (reading from right to left):

$$AB(LOADED, WAIT, S_1) \Leftarrow \ldots \Leftarrow \neg T(LOADED, S_2) \Leftarrow \neg AB(ALIVE, SHOOT, S_2).$$

The only difference between the two models is that in the first case we started at the (temporally) earliest situation and worked our way forward in time, and in the second case we started at the latest point and worked our way backward in time. Another way to express the idea is that in the first model we always picked the "earliest possible" (f, e, s) triple to assume "normal" and in the second model we always picked the latest.

So the class of models we want our logic to select is not the "minimal models" in the set-inclusion sense of circumscription, but the "chronologically minimal" models (a term due to Yoav Shoham): those in which normality assumptions are made in chronological order, from earliest to latest, or, equivalently, those in which abnormality occurs as late as possible.

(In a richer temporal formalism the criterion chronological minimality might not be the right one. If several years had lapsed between the $WAIT$ and the $SHOT$, for example, it would be reasonable to assume that the gun was no longer loaded. But chronological minimality *does* correctly represent our simple notion of persistence: that facts tend to stay true (forever) unless they are "clipped" by a contradictory fact.)

There appears to be no way represent this criterion, either in published versions of circumscription[6] or in the logics of Reiter or McDermott. The concept of minimality in circumscription is intimately bound up with the notion of set inclusion, and chronological minimality cannot be expressed in those terms. As far as Reiter and McDermott's logics go, what we need is some way to mediate *application* of default rules in building extensions or fixed points, which is beyond the expressive power of (Reiter's) default rules or of NML sentences involving the M operator.

6 Potential solutions

Two lines of work have been proposed as solutions to this problem. Yoav Shoham in [15] presents a logic that directly addresses the problem of representing causation in terms of "time flowing forward." Rather than trying to extend existing nonmonotonic logics so that they capture this new minimality criterion, he instead starts with a precise description of the chronologically minimal models. He then demonstrates that when a certain restricted class of first-order theories are minimized with respect to how much is *known* about each situation (instead of minimizing what is *true* in each situation) the resulting theory has a unique chronologically minimal model. While Shoham's logic handles the specific case of causal or temporal

[6] These include predicate circumscription and joint circumscription [8], formula circumscription and prioritized circumscription [9]. But see the note on pointwise circumscription below.

reasoning, his solution is obviously not an answer to the question we pose about the general relationship between default reasoning and nonmonotonic logics.

A second proposal, due to Vladimir Lifschitz in [7], involves a reformulation of and extension to predicate circumscription called ***pointwise circumscription***, in which one minimizes a predicate one point at a time (in our example a point would be a *(fact, event, situation)* triple). The order in which points are minimized is specified by an object-language formula that can express the concept of "temporally earlier" and "temporally later." Thus one is able to say something to the effect "minimize abnormalities, but favoring chronologically earlier ones." Pointwise circumscription contains predicate circumscription as a special case, and has been shown to solve a simple example of interacting defaults that we presented in [5].

But what benefits do we realize from these new, more expressive, more complex versions of circumscription? The problem is that the original idea behind circumscription, that a simple, problem-independent extension to a first-order theory would "minimize" predicates in just the right way, has been lost along the way. Instead, a complex, problem-specific axiom must be found to rationalize a set of inferences which must themselves be justified *on completely separate grounds*. The real theory of reasoning is the minimality criterion. In this example it was Shoham's chronological minimality; for other cases of default reasoning there will be other criteria for adding deductively unwarranted conclusions to a theory. It contributes little to our understanding of the problem that these criteria can be expressed as a second-order circumscription axiom; the criteria are justifying the axiom rather than the other way around.

The situation might be different if the second-order axiom were "productive," that is, if further, perhaps unforeseen conclusions could be drawn from it, mechanically or otherwise. But it can be very hard to characterize the consequences of the circumscription axioms for a reasonably large and complex theory, and when the consequences *are* understood, they may not be at all what we intended. The upshot is that no one really wants to know what follows from circumscription axioms; they usually wind up as hopefully harmless decorations to the actual theory.

7 Conclusion

We have presented a problem in temporal reasoning—causal or temporal projection—that involves defeasible inference of the sort normally associated with nonmonotonic logics. But upon writing axioms that describe temporal projection in an intuitive way, we found that the inferences licensed by the logics did not correspond to our intentions in writing the axioms. There seem to be two reasons for this: that conflicting default rule *instances* lead to unexpected multiple fixed points (minimal models), and that our preference of one extension over another (our criterion for minimality) depends on an ordering of individuals that cannot be expressed by circumscribing over any predicate or set of predicates, or by the default rules in the other nonmonotonic logics.

At this point we need to re-evaluate the relationship between nonmonotonic logics and human default reasoning. We can no longer engage in the logical "wishful thinking" that led us to claim that circumscription solves the frame problem [9], or that "'consistent' is to be understood in the normal way it is construed in nonmonotonic logic.[1]" From a technical standpoint, there *is* no "normal way" to understand the *M* operator, or the Reiter default rules, or a theory circumscribed over some predicate, apart from the proof- or model theory of the chosen logic.

The term "consistent," has too often used informally by researchers (*e.g.* in [6]) as if it had an intuitive and domain-independent meaning. We have shown that in at least one case a precise definition of the term is much more complex than intuition would have us believe, and that the definition is tightly bound up with the problem domain. As such, the claim implicit in the development of nonmonotonic logics—that a simple extension to classical logic would result in the power to express an important class of human nondeductive reasoning—is certainly called into question by our result.

References

[1] Charniak, Eugene "Motivation Analysis, Abductive Unification, and Non-Monotonic Equality", *Cognitive Science*, to appear.

[2] Davis, Martin, "The Mathematics of Non-Monotonic Reasoning", *Artificial Intelligence*, vol. 13 (1980), pp. 73–80.

[3] Etherington, David W., "Formalizing Non-Monotonic Reasoning Systems", Computer Science Technical Report No. 83-1, University of British Columbia.

[4] Etherington, David W. and Raymond Reiter, "On Inheritance Hierarchies with Exceptions", *Proceedings AAAI-83*, pp. 104–108.

[5] Hanks, Steven and Drew McDermott, "Temporal Reasoning and Default Logics", Computer Science Research Report No. 430, Yale University, October 1985.

[6] Joshi, Aravind, Bonnie Webber and Ralph Weischedel, "Default Reasoning in Interaction", *Proceedings of the Non-Monotonic Reasoning Workshop*, AAAI, October 1984, pp. 151–164.

[7] Lifschitz, Vladimir "Pointwise Circumscription", unpublished, draft of March 11, 1986.

[8] McCarthy, John, "Circumscription – A Form of Non-Monotonic Reasoning", *Artificial Intelligence*, vol. 13 (1980), pp. 27–39.

[9] McCarthy, John, "Applications of Circumscription to Formalizing Common Sense Knowledge", *Proceedings of the Non-Monotonic Reasoning Workshop*, AAAI, October 1984, pp. 295–324.

[10] McCarthy, John, and P. J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in: B. Meltzer and D. Michie (eds.), *Machine Intelligence 4*, Edinburgh University Press, 1969, pp. 463–502.

[11] McDermott, Drew V., "A Temporal Logic for Reasoning About Processes and Plans", *Cognitive Science*, vol. 6 (1982), pp. 101–155.

[12] McDermott, Drew V. and Jon Doyle, "Non-Monotonic Logic I", *Artificial Intelligence*, vol. 13 (1980), pp. 41–72.

[13] Perlis, Donald, and Jack Minker, "Completeness Results for Circumscription", Computer Science Technical Report TR-1517, University of Maryland.

[14] Reiter, Raymond, "A Logic for Default Reasoning", *Artificial Intelligence*, vol. 13 (1980), pp. 81–132.

[15] Shoham, Yoav, "Time and Causation from the Standpoint of Aritificial Intelligence", Computer Science Research Report, Yale University, forthcoming (1986).