Report Number: WUCS-89-06

1989-06-01

Defeat Among Arguments II

R. P. Loui

This technical report consists of three chapters from a larger manuscript that was a finalist in the 1988 Journal of Philosophy Johnsonian competition. These three chapters together represent a revised version of a paper that has been circulating under the title "Defeat Among Arguments II" since January 1988. "Defeat Among Arguments" updates my Computational Intelligence paper of 1987, which represented a novel way of formalizing defeasible reasoning, based on resolving competing arguments. "The Yale Shooting Problem" updates my Cognitive Science paper of 1987, and attempts a rebuttal of Hanks and McDermott's evaluation in their 1987 Artificial Intelligence paper. The... **Read complete abstract on page 2.**

Recommended Citation

Loui, R. P., "Defeat Among Arguments II" Report Number: WUCS-89-06 (1989). *All Computer Science and Engineering Research.*
https://openscholarship.wustl.edu/cse_research/719

Defeat Among Arguments II

R. P. Loui

Complete Abstract:

This technical report consists of three chapters from a larger manuscript that was a finalist in the 1988 Journal of Philosophy Johnsonian competition. These three chapters together represent a revised version of a paper that has been circulating under the title "Defeat Among Arguments II" since January 1988. "Defeat Among Arguments" updates my Computational Intelligence paper of 1987, which represented a novel way of formalizing defeasible reasoning, based on resolving competing arguments. "The Yale Shooting Problem" updates my Cognitive Science paper of 1987, and attempts a rebuttal of Hanks and McDermott's evaluation in their 1987 Artificial Intelligence paper. The last section, "Conventionalism and Non-Monotonicity," is a brief consideration of Touretzky, Thomason, and Horty's "Clash of Intuitions," and the prospects for choosing among languages for representing defeasible knowledge.

DEFEAT AMONG ARGUMENTS II

R. P. Loui

WUCS-89-06

June 1989

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
Saint Louis, MO 63130-4899

ABSTRACT

This technical report consists of three chapters from a larger manuscript that was a finalist in the 1988 Journal of Philosophy Johnsonian competition. These three chapters together represent a revised version of a paper that has been circulating under the title "Defeat Among Arguments II" since January 1988. "Defeat Among Arguments" updates my Computational Intelligence paper of 1987, which represented a novel way of formalizing defeasible reasoning, based on resolving competing arguments. "The Yale Shooting Problem" updates my Cognitive Science paper of 1987, and attempts a rebuttal of Hanks and McDermott's evaluation in their 1987 Artificial Intelligence paper. The last section, "Conventionalism and Non-Monotonicity," is a brief consideration of Touretzky, Thomason, and Horty's "Clash of Intuitions," and the prospects for choosing among languages for representing defeasible knowledge.

Chapter 4. Defeat Among Arguments

4.1. DEFEAT.

4.1.1. Form.

AI work on formalisms for non-monotonic reasoning has borne little resemblance to the philosophers' work. This is unfortunate, since there are things that knowledge representation people apparently want to do that are done quite easily in the way that philosophers do them. This includes the implicit preference among multiple extensions, and the ability to draw conclusions even though not all of the possible arguments have been considered.

Because of the form in which their reasoning principles are specified, epistemologists have not had to carry the cross of a century's work on deductive logic every time they have wanted to discuss defeasible reasoning. This frees them from model-theoretic issues. It makes their systems more flexible. It allows emphasis to be placed on the mechanisms for defeat, on the reasons for choosing among non-monotonic arguments, on what one wants to achieve, instead of on the most logic-like means of achieving it.

One idea is to identify arguments and counter-arguments relevant to establishing some proposition of interest; discovering which, if any, survive defeat; and finally, whether there is, among the survivors, an undefeated, uncompromised argument for the proposition. What remains is to figure out what kind of syntactic entities constitute arguments, and what considerations would lead us to prefer one over another.

To be fair, formalisms for this kind of reasoning are relatively new among philosophers, too; I am thinking of the systems of probabilistic reasoning due to Kyburg [Kyburg61, Kyburg74, Kyburg82], and Pollock [Pollock83, Pollock84].

In these systems, assertions of probability are meta-linguistic, e.g., for Kyburg, "Prob" is a two-place meta-linguistic relation between a term denoting an object-level sentence, "x is in V", and an interval "[.3, .3]". Direct inference arguments are based on inference structures: sequences of terms denoting those object language entities that are used in the argument. They are like proofs, which are also sequences of meta-linguistic entities, but there are two notable differences. Inference structures are demonstrations of claims only when they are undefeated. In contrast, a proof is a demonstration regardless of what other proofs there are. Also, the claims that arguments demonstrate are meta-linguistic assertions, not object-level assertions. There are various meta-linguistic relations between inference structures, e.g., "permits-a-stronger-conclusion-than", and "dominates". Defeasible inference is achieved by allowing inference to succeed only when these relations hold between the right sorts of things in the right sort of way. Kyburg and Pollock offer axioms that essentially say that the probability relation holds when there exist appropriate supporting arguments and there do not exist would-be defeating or interfering arguments.

In defeasible inference qua inference, i.e., as non-monotonic inference instead of defeasible inference to probability values, the "Prob" relation is replaced with a relation such as "is-a-rational-belief", or "is-a-defeasible-conclusion". The aim becomes one of determining which sentences bear this relation (that is, which terms denoting sentences can be so predicated). One can strain to think of all this meta-linguistic machinery as an unusual proof-theory. The "proof theories" to be considered are unlike the usual proof theories. Still, it is useful to think of the two-place meta-linguistic relation "allows-the-defeasible-conclusion" and the two-place meta-linguistic relation "$\vdash$" as kin.

When an inference system takes this form, there is no issue of semantics. The semantics of the first-order object language are standard. The semantics of the first-order meta-language are standard. The "is-a-defeasible-conclusion" predicate is mapped to a set. Meta-linguistic terms denoting object-level sentences are mapped to individuals. The two-place relation "is-a-reason-for" is mapped to a set of tuples. There would be semantical somersaults if "is-a-defeasible-conclusion" were equated with "is-true". They are not equated. Instead, theories of action are required to make use of those sentences that are predicated by "is-a-defeasible-conclusion."

4.1.2. Considerations.

In qualitative non-monotonic reasoning, what would count as reasons for one argument to be better than another? For one argument to interfere with another?

One popular consideration is specificity. Suppose

A is reason for C, i.e. $A >\!\!— C$, and
A & B is reason for $\neg$C, i.e., $(A \& B) >\!\!— \neg C$,

(eventually, $>\!\!—$ will become a two-place meta-linguistic relation).

If A & B are both accepted, many want to conclude ¬C ([Nute85], [Poole85], and [Glymour, Thomason85] ; also researchers on network inheritance, e.g. [Horty, Thomason, Touretzky87] and [Sandewall85]; recently [Geffner, Pearl87]; apparently [Etherington87]; and the early attempt of [Rich83]).

Superior specificity has been confused with superior evidence. The two are distinct, and I believe the latter is more important. To determine superior specificity, consider the logical strength of the antecedents of the non-monotonic rules employed in each argument. To determine superior evidence, consider the logical strength of the knowledge taken into consideration by each argument.

Suppose one line of reasoning uses

{ A;
A >— (B & C); and
(B & C) >— H
},

in the obvious way. Another line of reasoning uses

{ A;
A >— (B & C);
(B & C) ⊢ B;
B >— ¬H
}.

The former uses a rule with a more specific antecedent (B & C is stronger than B), so it exhibits superior specificity. But it does not exhibit more evidence, since each argument uses only the knowledge of A. Conversely, there can be superior evidence without superior specificity. Compare

{ A;
B;
A >— C;
B >— D;
(C & D) >— H
};

with

$$\begin{array}{l} \{\quad A; \\ \qquad A >\!\!— E; \\ \qquad E >\!\!— \neg H \\ \}. \end{array}$$

E does not entail C & D, nor vice versa: neither is logically stronger, hence neither antecedent is more specific. There are no other interesting antecedents to compare. So neither argument is more specific. Nevertheless, the former argument uses more of the evidence: $A, B \vdash A$.

Directness of arguments is another consideration. An argument is more direct than another if it is more closely attached to the evidence. If

$$\begin{array}{l} \{\quad A; \\ \qquad A >\!\!— H \\ \} \end{array}$$

is one argument, it defeats

$$\begin{array}{l} \{\quad A; \\ \qquad A >\!\!— B; \\ \qquad B >\!\!— \neg H \\ \}. \end{array}$$

Intuitions on directness fade in complicated cases, but many cases are clear, for instance,when all paths from evidence to conclusion in one argument shortcut the paths in the other argument, it is obvious that the former argument is more compelling.

We should bring to bear in our arguments as much relevant evidence as we can. We argue that Opus does not traverse vast territory because Opus' being a Penguin is a good reason for Opus' not flying, which is a good reason for Opus' not traversing vast territory. This argument is superior to the argument that Opus' being a bird is a good reason for Opus' traversing vast territory. We find the former argument more compelling even if we are not privy to the knowledge that Opus' being a Penguin is itself a good reason for Opus' not traversing vast territory. For suppose that Opus is merely a bird of type P, and Opus' being a bird of type P is good reason for Opus' not flying. We still choose the former argument, regardless of whether we know that being a bird of type P is itself a good reason for not traversing vast territory.

The argument just considered is superior because it uses more evidence, even though it is not as direct. When one argument has superior directness and the other has superior specificity, neither prevails. But when superior directness stands toe-to-toe with superior evidence, evidence emerges victorious. Evidence considerations are favored because the other considerations are structural: they have only to do with the structure of arguments; whereas, evidence considerations are empirically significant: they have to do with how well the argument is grounded in experience.

One more consideration is preferred premises. This has to do with the intermediary sentences used in arguments. When

{ A;
A >— C;
C >— H
}

is compared with

{ A;
A >— B;
B >— ¬C;
¬C >— ¬H
},

The argument for H is more compelling than the argument for ¬H because the sub-argument for C is more compelling than the sub-argument for ¬C. Preference among intermediary premises should determine preference among arguments based thereupon. In an earlier treatment [Loui87], I take "has-preferred-premises" to be a bona fide relation between two arguments, i.e., a basic consideration in determining whether one argument defeats another. Here, instead, our preference of arguments with preferred sub-arguments will be a consequence of how relations among arguments are aggregated, which I will borrow from Pollock. The argument for ¬H is defeated by the counter-argument {A; A >— C}, for C, which attacks the premise ¬C in the argument for ¬H. So the argument for H stands intact. I.

Here are four basic considerations that influence our evaluation of rival arguments (evidence; specificity and directness; preferred premises). Perhaps there are others that should be added; for example, an argument is self-defeating if too many premises intercede between evidence and conclusion. It should be obvious how additional considerations could be integrated into the formalism that follows, at least in principle. Defining defeasible inference in terms of meta-linguistic relations among arguments, as we do here, results in systems that are flexible. To alter the formalism is just to change some of these meta-linguistic definitions. Since defeasible conclusion and assertion are not conflated, that is, since "is-true" and "is-a-defeasible-conclusion" are not equated, there is no model-theoretic clarification required when there is alteration.

4.1.3. Defeater Defeaters.

Are defeater defeaters reinstaters? If G_2 defeats G_1 which is the only defeater of G_0, does G_2 thus reinstate G_0 (fig. 4.1). Obviously, if argument G_2 defeats argument G_1, while G_1 defeats G_0, G_2 could also defeat G0. So the answer is a resounding "not always." It isa more interesting when G_2 does not defeat G0. Still, the defeater defeater need not reinstate. It depends on the kind of defeat involved.

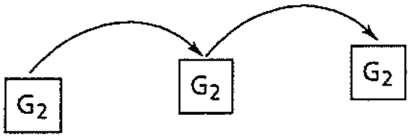

fig. 4.1. G_2 is a defeater defeater.

There ae two kinds of defeat. An argument interferes with another argument if it can prevent that argument from supporting its putative conclusion,that is, if it takes away the argument's power to support. An argument defeats another argument in the narrow sense if it eliminates the force of that argument: if it takes away the argument's power to interfere, as well as its power to support. Counter-arguments can be interfering or defeating.

Again, suppose G_0 is defeated by G_1, which is defeated in turn by G_2. If G_2 interferes with G_1, then there is prima facie reason to reinstate G_0. This happens regardless of whether G_1 defeats G_0 in the narrow sense, or merely interferes with G_0. But if G_2 merely interferes with G_1, G_1 is still a prima facie reason to doubt G_0.

Thus, an argument can be prevented from supporting its conclusion, but can still interfere with another argument. For concreteness, let G_1 be an argument for P based on A; and G_2 be an argument for ¬P based on B (fig. 4.2). G_1 and G_2 interfere with each other. Consider two elaborations of this situation.

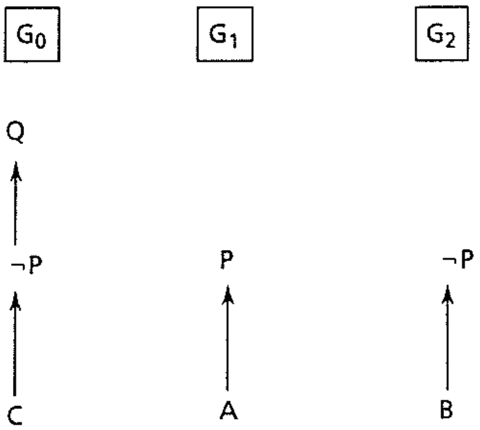

fig. 4.2. G_1 continues to interfere with G_0.

1. there is an argument, G_0, for Q based on C, which uses ¬P as a premise. G_2 interferes with G_1, so that G_1 does not effectively support P. G_1 does not establish P. But it still provides a reason to doubt G_0.

2. instead, suppose that there is an argument, G_3, which defeats G_2 in the narrow sense. It might be that G_2 uses the premise B′, on evidence B, in order to establish ¬P defeasibly. G_3 might argue for ¬B′ on evidence stronger than B, e.g., B & D. Then G_2 is simply a bad argument, in light of G_3, and the force of G_1 should not be affected by G_2. G_3 thus reinstates G_1 (fig. 4.3).

Pollock takes all defeater defeaters to be reinstating because he treats defeat differently. Whenever one argument defeats another in the narrow sense, Pollock presumes the existence of a third argument. G_3 defeats G_2 in the narrow sense.

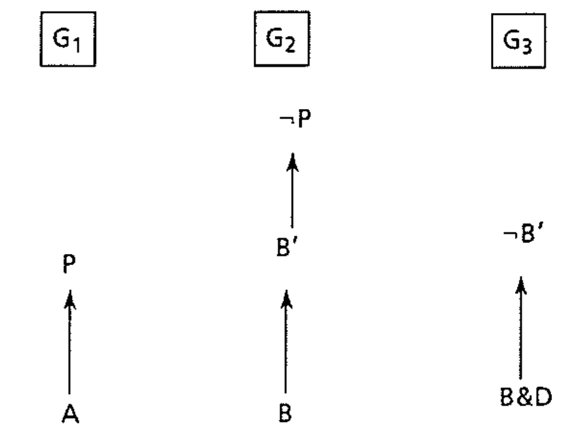

fig. 4.3. G_3 reinstates G_1.

Pollock would instead say G_3 and G_2 defeat each other, and would suppose some G_4, which defeats G_2. In this way, G_4 defeats G_2, thereby reinstating G_3 (fig. 4.4).

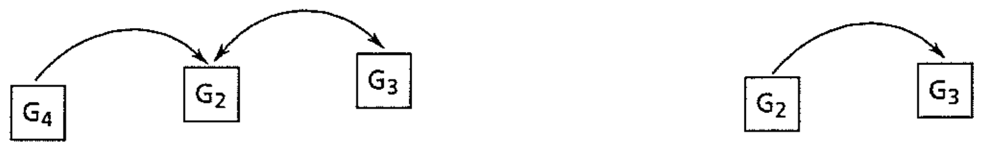

fig. 4.4. Pollock introduces a third argument when there is asymmetric defeat.

In contrast, in Kyburg's system, no defeater defeater is reinstating. If there is an inference structure based on the statistical knowledge that

$$\%(A, V) = [.3, .3],$$

it is defeated by the inference structure based on the knowledge that

$$\%(A \,\&\, B, V) = [.5, .7].$$

This in turn is defeated by the inference structure based on

$$\%(A \,\&\, B \,\&\, C, V) = [.2, .6].$$

If x is in A & B & C, then Kyburg takes Prob("x in V") to be [.2, .6]. Why not [.3, .3]? After all, its only defeater has been defeated. But the inference structure that leads to [.3, .3] does not itself dominate all inference structures with which it disagrees, which is Kyburg's rule. Battle among inference structures is to be done individual against individual, not as tag-team. Apparently, an inference structure, once sullied by unreflected disagreement, remains sullied.

The two possibilities here are related to those considered by [Touretzky, Horty, Thomason87] in their discussion of skeptical inheritance reasoners and the intersection of extensions of credulous reasoners.

4.1.4. Non-Supporting Reasons.

In Kyburg's system, there is a defeater based on relative strength. If

$$\%(A, V) = [.3, .3] \text{ and}$$
$$\%(A \,\&\, B, V) = [.1, .9],$$

then for x in A & B, Prob("$x \in V$") is [.3, .3]. As Kyburg views this situation, even though the [.1, .9] interval is based on more of the information about x, it is not in disagreement with [.3, .3], since it contains [.3, .3] as a sub-interval. Since [.3, .3] is stronger information, it is used instead. This is the source of much of the disagreement with Bayesian conditionalization [Levi80]. Bayesians are committed to the [.1, .9] interval unless they can state independence or irrelevance assumptions.

In most systems of qualitative inference, there is an implicit strength defeater. Suppose

$$A >\!\!- C, \text{ and}$$
$$(A \,\&\, B) >\!\!- (C \vee \neg C).$$

On the force of A & B, usually there is no problem concluding C. $C \vee \neg C$ is compatible with C, so there is apparently no reason to doubt the argument to C. However, it might be the case that B undercuts the relation between A and C, in virtue of which A is a reason for C. B does not thereby provide a reason for $\neg C$; it merely interferes with A's being a reason for C.

One way to write this is

$$A \succ\!\!- (C \,\&\, norm_1), \text{ and}$$
$$A \,\&\, B \succ\!\!- \neg norm_1.$$

Note that for this to have the desired behavior, $\succ\!\!-$ could not be right-weakenable:

$$A \succ\!\!- (C \,\&\, norm_1)$$

could not entail

$$A \succ\!\!- C \text{ and}$$
$$A \succ\!\!- norm_1.$$

The latter would allow an argument from A to C with which A & B could not interfere.

Another way to write that B interferes with $A \succ\!\!- C$ is to write the meta-linguistic axiom:

$$(A \succ\!\!- C) \text{ iff } \neg(\vdash (A \,\&\, B)).$$

This would be a coherence constraint on the assertability of $\succ\!\!-$ relatedness. Viledicet, one could not assert that A is a reason for C if one already accepted A & B.

A third way is due to Nute [Nute85] (and is related to the explicit censors of Pearl [Pearl87c]). It requires introducing a different kind of reason. Let it be possible for A to be a prima facie reason for C that is too weak to support arguments for C, but that is reason enough to doubt arguments for contraries of C. We write A X— C. Arguments can be constructed that use X— relations. But unlike arguments that use only $\succ\!\!-$ and $\vdash$ relations, these arguments are impotent. They can defeat other arguments, but they cannot be used to support the acceptance of a sentence.

Several choices present themselves regarding the behavior of X— relations. Should arguments be allowed to use more than one X— relation, or should X—

relations necessarily be terminal steps in arguments? If indeed, X— relations are non-supporting reasons, then perhaps it is the case that having argued

{ A;
A X— B
},

one cannot continue with

{ A;
A X— B;
B >— C
},

to produce an argument for C: not even as an interfering argument. Or perhaps, instead, (X—)-relatedness simply means that arguments that use X— relations, any number of X— relations, cannot support the acceptance of sentences. I formalize the former here, requiring that X— relations terminate arguments.

Are arguments based on non-supporting relations strong enough to reinstate? Suppose a non-supporting argument for ¬C defeats an argument for C. The argument for C interferes with an argument for H, which relies on a sub-argument for ¬C. A X— ¬C is supposed to interfere with arguments for C. But does it interfere in a way that reinstates supporting arguments for C? Can non-supporting arguments defeat other non-supporting arguments, or do they all linger as potential sources of interference? I formalize non-supporting arguments as potential reinstaters, and as arguments that can defeat each other in the narrow sense.

The various choices lead to various different meanings of X—. Which choice we make will be a function of which kind of X— we are most comfortable using, and how good we are at using it. This discussion is pursued in chapter 8.

4.1.5. Permitting a Reason and Being a Reason.

It will not be the case that A >— C whenever A ⊢ C. This is important because ⊢ is left-strengthenable, while >— is not. So

(A & B) ⊢ C whenever A ⊢ C.

But

it is not the case that (A & B) $>\!\!-$ C whenever A $>\!\!-$ C.

If the $>\!\!-$ relation were left-strengthenable, one could no longer evaluate how much evidence is actually taken into consideration by an argument.

If A is a reason for C, should (A & B) a reason for C? If A is a reason for C, then (A & B) permits a reason for C: A is a reason for C, and part of what it means to accept (A & B) is the acceptance of A. But that is different from (A & B) being a reason for C. This is an important distinction. We do not say that (A & B) is a reason for C if it is only a reason for C because A is a reason for C and (A & B) entails A. That would be misleading. It is quite a different situation when (A & B) is genuinely a reason for C. If A is a reason for C, B could still be a defeater of C, so that presupposing (A & B) should not yield the inference that C. In fact, (A & B) could be a reason for ¬C. Opus' being a bird is reason for Opus' flying. But Opus' being a penguin is not reason for Opus' flying. Rather, it provides a reason for Opus' flying, on account of Opus' being a bird.

The notation is overloaded. We could have written "A is reason for C even if B," or Reason(A, C, B). A system might even take Reasons to be four-place relations, Reason(A, C, B, D), where A is reason for C even if B unless D. The triad

A $>\!\!-$ C ,
(A & B) $>\!\!-$ C, and
(A & B & D) X— ¬C,

would follow from Reason(A, C, B, D). But it is not clear whether (A & D) X— ¬C is implied by the more compact notation.

4.2. ARGUMENTS.

This section presents a variation on the formalism originally presented in [Loui87]. I do not consider the differences between this formalism and the latter to be crucial. The original was essentially Kyburgian in its approach to reinstatement, while this system advances Pollock's ideas about reinstatement.

The idea is to take all of the reasons that can be used given the knowledge that is to be considered evidence, and determine what collections of them argue for what sentences. What is to be considered an argument is restricted to those parts of the evidence and those defeasible rules that are essential to the demonstration.

Arguments will be connected acyclic digraphs with a single sink, with sources corresponding to essential evidence, and with links corresponding to monotonic or non-monotonic reasons. Next, relations between arguments are defined, such as "defeats" and "uses-more-evidence-than." Finally, arguments are aggregated to determine whih can be made without interference from other undefeated arguments. This determines which sentences are justified and consequently deserve to be treated as defeasible knowledge, that is, as knowledge that can be used in subsequent inquiry and decision, but which can be retracted in light of new evidence.

Detailed examples are provided after the formalism, but it may be useful to consult them while digesting the definitions.

4.2.1. Formalism.

Let there be an object language L, and a meta-language, ML. Knowledge is given in the form 'p ∈ EK', and 'p >— q', where p and q are in SnL, the set of terms in the meta-language that denote sentences of the object language, i.e., they are quoted object-language sentences. The former asserts that sentence p belongs to the corpus of evidential knowledge. The latter asserts that p is a reason for q. The task is to define conditions under which it can be asserted that 'p ∈ DK', i.e., that p is in the corpus of defeasible knowledge. For example, given

"(Penguin Opus)" ∈ EK
"(FORALL ?x (OR (Bird ?x) (NOT (Penguin ?x))))" ∈ EK
"(Bird Opus)" >— "(Flies Opus)"
"(Penguin Opus)" >— "(NOT (Flies Opus))"

or more comprehensively,

(x) '(Bird x)' >— '(Flies x)'

(I will use the logic programming notation for the object language, in order to separate it from the meta-language. 'x' is the Quine quotation of x, but I will use it sparingly because it is distracting).

The question of interest is whether

"(Flies Opus)" $\in$ DK .

Axiom (consistency of EK):
$\neg$(EK $\vdash$ "(NEQUAL ?x ?x)") .

Axiom (monotonic closure of EK):
(p)(q)(r) . if p, q $\vdash$ r and p $\in$ EK and q $\in$ EK then r $\in$ EK .

Axiom (biconditional substitution in reasons):
(p)(q)(p1)(q1) . if
1. p >— q and
2. $\vdash$ '(IFF p p1)' and
3. $\vdash$ '(IFF q q1)'
then p1 >— q1 .

Defn. If P is a set of sentences, Conjoin(P) is the conjunctive concatentation, i.e.,
Conjoin("(Bird Opus)", "(NOT (Bird Opus))") is
"(AND (Bird Opus) (NOT (Bird Opus)))" .

Defn. S is EK-consistent iff
$\neg$(S, EK $\vdash$ "(NEQUAL ?x ?x)") .

Defn. p is weaker than q (and q is stronger than p) iff
q $\vdash$ p and $\neg$(p $\vdash$ q) .

Defn. $p >\!\!- q$ in R iff
$p >\!\!- q$ and $<p, q> \in R$.

Defn. An argument for p is <G, R> such that

1. R is a set of tuples $\{<p_i, q_i>\}$ s.t. for all i, $p_i >\!\!- q_i$.
2. G is <z, N, E, Label> where <N, E> is a graph, and all of
   - 2.1. N is the set of nodes .
   - 2.2. E is the set of edges .
   - 2.3. G is connected and acyclic .
   - 2.4. Label is a function from N to Sn_L, the set of object language sentences .
   - 2.5. Label(z) = p and z is the unique sink .
   - 2.6. Defn. Support(x) is $\{$Label(y) s.t. $<y, x> \in E\}$.
   - 2.7. Defn. Sources(G) is $\{$Label(x) s.t. Support(x) = $\{\}\}$.
   - 2.8. Defn. l is a Radius of G iff
     l = <p> or l = <Label(n_1), . . ., Label(z)> where
     <n1, . . ., z> is a directed path in G
     from some source n_1 to sink z .
   - 2.9. (nodes name equivalence classes)
     for all nodes x and y, if x, y $\in$ N and $\neg$(x = y)
     then '(IFF Label(x) Label(y))' is
     not EK-consistent .
   - 2.10. (nodes are consistent)
     Conjoin($\{$Label(n) s.t. n $\in$ N$\}$) is EK-consistent .
   - 2.11. (sources correspond to evidence)
     for every s $\in$ Sources(G) then s $\in$ EK .
   - 2.12. (support corresponds to reasons)
     for every node x, if $\neg$(Support(x) = $\{\}$) then
     Conjoin(Support(x)) $>\!\!-$ Label(x) in R
     or
     Support(x), EK $\vdash$ Label(x)
     or
     (x = z and
     Conjoin(Support(x)) $\mathrm{X}\!\!-$ p in R) .
   - 2.13. (sources are minimal)
     for all nodes x and y, if Support(x) = $\{\}$ and $<x, y> \in$ E then

Conjoin(Support(y)) >— Label(y) in R .

2.14. (every tuple in R is used)

if <p, q> in R, then for some $x \in N$,

Conjoin(Support(x)) >— Label(x) .

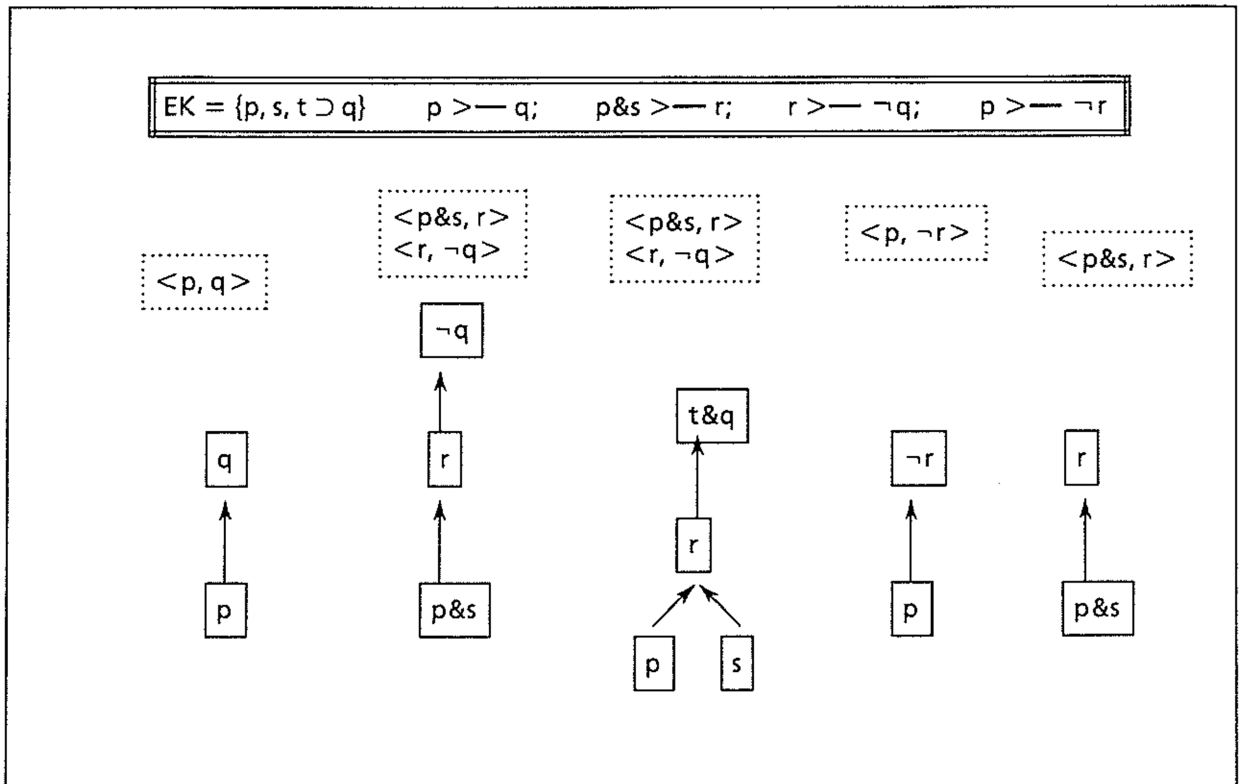

fig. 4.5. some arguments.

Defn. A is an argument if it is an argument for some $p \in Sn_L$.

Defn. <<z, N, E, Label>, R> is a supporting argument iff

(no X— relations are used)

Conjoin(Support(z)) >— Label(z) in R

or

Support(z) = {} .

Defn. A is an interfering argument iff

it is an argument and it is not a supporting argument.

Defn. Argument $<G_1, R_1>$ uses more evidence than argument $<G_2, R_2>$ iff
Conjoin(Sources(G_1)) is stronger than Conjoin(Sources(G_2)) .

Defn. Argument $<<z_1, \ldots>, R_1>$ has specificity over
argument $<<z_2, \ldots>, R_2>$ iff
for some rule in R_1, $a_1 >\!\!-\!\!- c_1$,
and some rule in R_2, $a_2 >\!\!-\!\!- c_2$,
c_1 and c_2 are not EK-consistent and
a_1 is stronger than a_2 .

Defn. Sequence l is a sub-path of sequence m iff
1. l = m or
2. (deletion)
m = concat(s, <A>, r) and l is a sub-path of concat(s, r) or
3. (weakening)
m = concat(s, <A>, r) and l is a sub-path of concat(s, <B>, r),
where $A \vdash B$.

Defn. Sequence concat(l_1, <r>) d-short-cuts sequence concat(l_2, <'¬r'>),
for $r \in Sn_L$, iff
l_1 is a sub-path of l_2 and l_2 is not a sub-path of l_1 .

Defn. Argument $<G_1, R_1>$ has directness over argument $<G_2, R_2>$ iff
for some Radius l_1 of G_1 and some Radius l_2 of G_2,
l_1 d-short-cuts l_2 .

Defn. A_1 SD-reflects A_2 iff
A_1 has specificity over A_2 or
A_1 has directness over A_2 .

Defn. A_1 reflects A_2 iff
1. A_1 uses more evidence than A_2
or
2. all of
¬(A_2 uses more evidence than A_1) and
A1 SD-reflects A_2 and

$\neg(A_2$ SD-reflects $A_1)$.

(Reflection is Kyburg's term: A_1 reflects the considerations of A_2).

Defn. A_1 and A_2 disagree iff
A_1 is an argument for p and
A_2 is an argument for q and
{p, q} is not EK-consistent .

Defn. A is a counter-argument of <<z, N, E, Label>, R> iff
A is an argument for p and
for some $n \in N$,
{Label(n), p} is not EK-consistent .

Defn. S = <<z', N', E', Label>, R'> is a sub-argument of
argument A = <<z, N, E, Label>, R> iff
S is an argument and
$z' \in N$ and
$N' \subseteq N$ and
$E' \subseteq E$ and
$R' \subseteq R$.

Defn. A defeats <<z, N, E, Label>, R> iff
A is a counter-argument of <<z, N, E, Label>, R>
and
for every sub-argument S = <<z', N', E', Label>, R'> of
<<z, N, E, Label>, R>,
if A and S disagree, then A reflects S .

(The next four definitions are adapted from Pollock's definitions for level(n) arguments)

Defn. A is a level(0) I-argument iff A is an argument.

Defn. A is a level(0) S-argument iff A is a supporting argument.

Defn. Supporting argument A is a level(n+1) S-argument iff
  there is no A′ s.t.
    A′ is a level(n) I-argument and
    A′ is a counter-argument of A.

Defn. Argument A is a level(n+1) I-argument iff
  there is no A′ s.t.
    A′ is a level(n) I-argument and
    A′ defeats A .

Defn. p is justified iff
  there is a m s.t. for every $n > m$,
    there is a level(n) S-argument for p .

Defn. $p \in DK$ iff
  p is justified
    and
  there is no minimal inconsistent subset of SnL, S, that includes p s.t.
    for every $s \in S$, s is justified .

4.2.2. Examples.

Example 1 (Tweety).
  Given

    'Bird x' >— 'Flies x'
    "Bird Tweety" $\in$ EK
    "$\neg$Flies Tweety" $\in$ EK .

There is no argument for "Flies Tweety" since arguments must be EK-consistent. There is an argument for "$\neg$Flies Tweety", i.e. the single node labled "$\neg$Flies Tweety", and the singleton tuple { < "Bird Tweety", "Flies Tweety" > }. This argument has no counter-arguments, so "$\neg$Flies Tweety" is justified, and in DK.

Example 2 (Opus).

Given

'Bird x' >— 'Flies x'
'Penguin x' >— '$\neg$Flies x'
"Penguin x $\supset$ Bird x" $\in$ EK
"Penguin Opus" $\in$ EK .

The argument for "Flies Opus", A_1, uses the reasoning

{ "Penguin Opus" $\in$ EK;
"Penguin Opus" $\vdash$ "Bird Opus";
"Bird Opus" >— "Flies Opus"
},

while the argument for "$\neg$Flies Opus", A_2, uses the reasoning

{ "Penguin Opus" $\in$ EK;
"Penguin Opus" >— "$\neg$Flies Opus"
}.

A_2 has directness over A_1, since <"Penguin Opus"> is a sub-path of <"Penguin Opus", "Bird Opus">. It also has specificity over A_1, since <"Penguin Opus", "$\neg$Flies Opus"> is used in A_2, <"Bird Opus", "Flies Opus"> is used in A_1, and "Penguin Opus" $\vdash$ "Bird Opus". Nothing recommends A_1 over A_2. So A_2 SD-reflects and reflects A_1. They disagree, and A_2 defeats A_1. Beyond level(1), A_2 is an S-argument and an I-argument, but A_1 is neither. So "$\neg$Flies Opus" is justified, and in DK.

Example 3 (Nixon's Pacifism).

"Quaker Nixon & Republican Nixon" $\in$ EK
'Quaker x' >— 'Pacifist x'
'Republican x' >— '$\neg$Pacifist x'

There is an argument for "Pacifist Nixon", A_1, and an argument for "$\neg$Pacifist Nixon", A_2. A_1 has the Sources {"Quaker Nixon"} and uses the set of rule tuples

{<"Quaker Nixon", "Pacifist Nixon">}. A_2 has the Sources {"Republican Nixon"} and the rule tuple {<"Republican Nixon", "¬Pacifist Nixon">}. Note that there is no argument A_3, with Sources {"Republican Nixon", "Quaker Nixon"} or {"Republican Nixon & Quaker Nixon"} and rule {<"Republican Nixon", "¬Pacifist Nixon">}, since 2.13 requires that sources be minimal for some rule in R. Between A_1 and A_2, there is no reason to suppose one is better; neither reflects the other, and neither defeats the other. But they are in disagreement. So each prevents the other from being an S-argument at each level, though both are I-arguments at every level. Neither "Pacifist Nixon" nor "¬Pacifist Nixon" finds its way into DK.

Example 4 (Nixon's Anti-militarism).

As above, but add

'Republican x' >— 'FootballFan x'
'FootballFan x' >— '¬AntiMilitary x'
'Pacifist x' >— 'AntiMilitary x'

Is Nixon "¬AntiMilitary Nixon" in DK? The argument for this sentence, A^-, is based on

{ "Republican Nixon" ∊ EK;
"Republican Nixon" >— "FootballFan Nixon";
"FootballFan Nixon" >— "¬AntiMilitary Nixon"
}.

But there are arguments A_1 and A_2, as above, for "Pacifist Nixon" and "¬Pacifist Nixon", respectively. Furthermore, there is the argument, A^+, for "AntiMilitary Nixon", based on

{ "Quaker Nixon" ∊ EK;
"Quaker Nixon" >— "Pacifist Nixon";
"Pacifist Nixon" >— "AntiMilitary Nixon"
}.

A^+ and A^- disagree, and they neither reflect nor defeat each other. A_2 is a counterargument of A^+ because A^+ contains the node labeled "Pacifist Nixon".

Incidentally, the sub-argument of A^+ that disagrees with A_2 is exactly the argument A_1.

At level(0), all arguments are S-arguments and I-arguments.

	A_1	A_2	A^+	A^-
level(0)	I,S	I,S	I,S	I,S
level(1)	I	I	I	I
level(2)	I	I	I	I

At level(1), none is an S-argument, since each has a counter-argument that is a level(0) I-argument. All remain I-arguments at subsequent levels because there is no defeat in the narrow sense. So neither "AntiMilitary Nixon" nor "¬AntiMilitary Nixon" becomes justified; neither belongs to DK.

Example 5 (reinstatement).

As above, but adding

"Quaker Nixon & Republican Nixon" >— "¬Pacifist Nixon" .

Now there is argument A_0, with Sources {"Quaker Nixon", "Republican Nixon"} (or with Sources {"Quaker Nixon & Republican Nixon"}; it does not matter here), and with rule tuple {<"Quaker Nixon & Republican Nixon", "¬Pacifist Nixon">}. A_0 reflects and defeats A_1 because of specificity and superior evidence. It is also a counter-argument of A^+. And since A^+'s sub-argument that disagrees with A_0 is A_1, A_0 defeats A^+. Now the aggregation is as follows:

	A_0	A_1	A_2	A^+	A^-
has counter-args:	A_1	A_0,A_2	A_1	A^-,A_0	A^+
is defeated by:		A_0		A_0	
level(0)	I,S	I,S	I,S	I,S	I,S
level(1)	I		I		I
level(2)	I,S		I,S		I,S
level(3)	I,S		I,S		I,S

At level(1), none is an S-argument, since each has a counter-argument at level(0). But the defeated arguments are not even I-arguments at level(1). So at level(2), the undefeated arguments become S-arguments again, and they remain that way. Now "Pacifist Nixon" and "¬AntiMilitary Nixon" are both justified and both find their way into DK.

4.3. RELATED WORK.

4.3.1. Non-Monotonicity based on M-operators.

Among the striking differences between this approach to defeasible reasoning and those that exist in the AI literature is that the present approach allows non-monotonicity in computation, in addition to non-monotonicity in evidence. The latter, non-monotonicity in evidence, is the familiar desideratum. But non-monotonicity in computation is just as interesting.

In systems of the present kind, conclusions can be drawn relative to a non-exhaustive set of arguments. If only one argument of many arguments can be constructed under some search strategy, then that argument has no detractors, stands undefeated, and justifies its conclusion. As soon as another argument is found, which interferes with or defeats the earlier argument, this conclusion must be retracted. The definitions of level(n) I-arguments and S-arguments can easily be modified so that all quantifiers range over a set ARGS, of arguments that have been identified. Thus, not all possible arguments need be considered when determining which arguments are undefeated and therefore justify their claims. One can consider just those arguments one can compute. Membership in DK is thus non-monotonic in EK, and also non-monotonic in ARGS.

Systems that use checks for consistency, such as non-monotonic logic [McDermott, Doyle80] and default logic [Reiter80] require proofs of consistency; it is insufficient merely to feign ignorance of inconsistency. Officially,

"Bird x & M Flies x ⊃ Flies x"

cannot be used just because the attempt to prove "¬Flies x" fails on limitations of resource. Of course, concluding "Flies x" on failing to prove "¬Flies x" is exactly what one wants to do, even though negation-as-failure is incomplete, or worse, as a proof procedure.

Defeating arguments is like implicitly preferring extensions. In default logic, this is often done by using non-normal defaults:

Bird x : M (Flies x & ¬Penguin x) / Flies x
Penguin x : M ¬Flies x / ¬Flies x

achieves

Bird x >— Flies x
Penguin x >— ¬Flies x .

But the specialization of default rules is inconvenient. Adding conjuncts under the scope of M unduly complicates the rule. Moreover, the use of non-normal defaults does not easily capture the behavior of directness and evidence considerations, or the notion of defeater defeaters.

4.3.2. Comparison with Nute and with Poole.

Nute [Nute85] and Poole [Poole85] have each produced systems that are based on intuitions similar to those used here, and that allow more specific arguments to defeat less specific ones.

Nute offers a proof theory instead of meta-linguistic machinery. His notion of specificity is much more limited than the one used here, and he is concerned with the specificity consideration exclusively. Roughly, Nute first determines which rules cannot be used: they have the form "A is a reason for C", and there is a rule of the form "(A & B) is a reason for not-C", where "A & B" is presupposable. Then he constructs proofs using only those rules that can be used. One advantage of this approach is that he retains the locality of deductive logics. Once a proof has been found, it does not matter what other proofs there are; the claim has been demonstrated once and for all.

Poole talks about aggregates of non-monotonic rules as "theories," and asks which theory is the most specific. When he finds this ultimately specific theory, he allows all the non-monotonic conclusions. This is different from the present conception of checking only sub-parts of "theories" against each other when they legitimize conflicting conclusions. So if T_1 and T_2 are both theories, T_1 and T_2 both allow the non-monotonic conclusion p, but neither is more specific, then Poole still will not conclude p. I take this to be merely an oversight on Poole's part, but it is legitimately an alternative way of looking at things.

Poole also has a different conception of superior specificity, and neither does his subsume mine nor vice versa. In my terminology, Poole tests to see whether, by adjusting the contents of EK, one argument can always be made whenever the other can. If so, then it has inferior specificity. It is possible to insert Poole's superior specificity in place of anti-symmetric reflection, i.e., as a replacement for superior evidence, superior directness, and superior specificity considerations.

4.3.3. Comparison with Touretzky, Sandewall, and Horty et. al.

Touretzky [Touretzky86] considers paths in inheritance networks, and those paths function like arguments. He defines preclusion and interference relations among paths, which are the defeat relations among arguments. His mediated paths are similar to d-short-cut radii. Touretzky's use of mediated paths enforces specificity as well as directness. If Opus ISA Penguin and Penguin ISA Bird, for instance, then inheritance from the Penguin class instead of the Bird class is a choice based on specificity.

For the purposes of comparison with the present system, the proposals of Sandewall [Sandewall86] and Horty, et. al., [Horty87] are merely variations on Touretzky's rules, equally different from those used here.

Inheritance networks are less expressive than first-order languages, and do not distinguish defeasible and indefeasible ISA links. They ignore the difference between

"A $\supset$ C" $\in$ EK and
"A" >— "C" .

In inheritance networks, there is no way to say that "Penguin Opus" is reason for "Cat Garfield"; defeasible rules must be restricted to the form 'P x' >— 'Q x' if they are to be translated into inheritance networks. But having imposed this restriction, Touretzky's rules for mediated paths correspond closely to the evidence, directness, and specificity relations.

Equally striking is the similarity between the inductive definition of "permitted" paths, in Horty, et. al., and the determination of level(i) I- and S-arguments.

Permitted paths in a network are like S-arguments that survive at indefinitely high levels. But there is a difference. Consider the paths

$$x \longrightarrow a \longrightarrow v$$
$$x \longrightarrow b \not\longrightarrow v$$

Horty, et. al. say that the network does not permit either path. The corresponding arguments interfere with each othe: both arguments are I-arguments, but not S-arguments, at non-zero levels. this much is the same. Now consider whether the path

$$x \longrightarrow c \longrightarrow d \longrightarrow y$$

is permitted, when the network also contains

$$x \longrightarrow a \longrightarrow v \longrightarrow f \not\longrightarrow y$$

which is like a counter-argument. Is this counter-argument strong enough to interfere with the <x, c, d, y> path, even though it passes through the contentious node, v? In the Horty, et. al. scheme, the latter path is not permitted, so it cannot interfere with the conclusion that x is a y. But in my manner, all three arguments that include v are I-arguments, and the fact that the <x, a, v, f, ¬y> argument is an I-argument is enough to prevent the <x, c, d, y> argument from remaining an S-argument.

Still, it is useful to think of the Horty-Touretzky-Thomason-Sandewall work as, roughly, what results by restricting the present system to inheritance relations.

4.3.4. Comparison with Geffner and Pearl.

Geffner and Pearl [Geffner, Pearl87] conceive of the project differently. They start with default rules that satisfy a consistency requirement. We cannot write, for instance, in their system, that

"A_1 is reason for C"
"A_2 is reason for C"
. . .

"A_{100} is reason for C"
"$A_1 \vee A_2 \vee ... \vee A_{100}$ is reason for $\neg$C"

even though there may be situations in which high probability relations would move us to do so (there exists a probability distribution in which $\text{Prob}(C \mid A_i) > .9$, for each A_i, and also $\text{Prob}(\neg C \mid A_1 \vee . . . \vee A_{100}) > .9$). This kind of situation is a problem for defeasible reasoning systems with specificity in the first place, since reasoning by cases will introduce artificial specificity; presuming each of the A_i in turn, the rule for C defeats the rule for $\neg$C. Perhaps, then, it is a good idea to exclude such rules.

More to the point, one cannot write in their system that

"A is reason for C"
"A is reason for $\neg$C"
"A" .

Clearly, this kind of input is ineffectual, and it is reasonable, if not desirable, to have systems that object to such input. This is Geffner and Pearl's virtue. My system will actually allow such input to be formalized. It represents an ambiguous or confused state from which no inference about C is to be drawn. It is desirable, too, to have robust systems in which one can recover from such ambiguous directives, since their treatment -- they lead to arguments that interfere with each other -- is no different from the treatment of the input that

"A is reason for C"
"B is reason for $\neg$C"
"A & B" .

To those who want non-monotonic systems that can be given close probabilistic interpretations, Geffner and Pearl's approach will seem more intuitive. The present system is more concerned with mimicking pre-analytic competition among reasons. It formalizes dialectic.

Chapter 5. The Yale Shooting Problem

5.1. THE YALE SHOOTING PROBLEM.

The Yale Shooting problem was supposed to be the death knell for non-monotonic reasoning.

Steve Hanks and Drew McDermott [Hanks&McDermott86] describe a temporal projection problem that involves reasoning about a gun, known to be loaded at a time, fired at a person at a later time. They want to know if the person ceases to live. They are willing to assume that if the gun remained loaded, then the firing was effectively fatal. But the kind of reasoning that permits concluding that the gun remained loaded until it was fired would also appear to allow reasoning to the conclusion that the person remained alive, even after the firing. One choice is to reason that the property of being alive persists, and hence, the property of being loaded does not. The other choice is to reason that the property of being loaded persists and the property of being alive does not.

Symbolically, (see fig. 5.1)

loaded@0
alive@0
fired@1
(1) if *fired*@1 and *loaded*@1 then not-*alive*@2
(d1) *alive*@0 >— *alive*@1
(d2) *alive*@1 >— *alive*@2
(d3) *loaded*@0 >— *loaded*@1 .

Hanks and McDermott think it is intuitive that an adequate defeasible reasoning system will allow not-*alive*@2 to be concluded, i.e., that the property of being loaded persisted and therefore the property of being alive did not. They correctly note that there is a temporal asymmetry between (d2) and (d3). Namely, (d3) refers to earlier times. Hanks and McDermott next propose that defeasible conclusions be ordered according to a "temporally forward" priority. Earlier defeasible conclusions defeat later defeasible conclusions, if the conclusions are contraries. They mention the work of their Yale colleague, Yoav Shoham [Shoham86], who justifies this approach on the basis of reflections on causality.

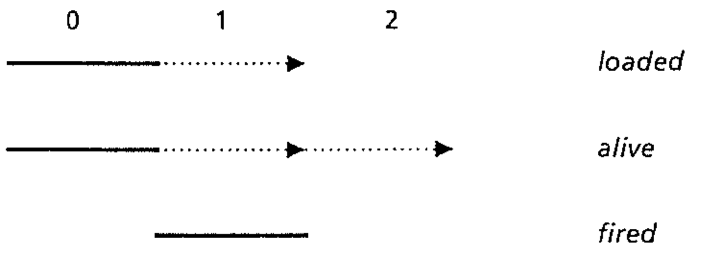

fig. 5.1. The property of being loaded tends to persist. So does the property of being alive. Because of the firing at 1, the tendencies to persist conflict.

So *loaded*@1 and *alive*@1 can be concluded, but *alive*@2 cannot be, because one has already committed to *loaded*@1. In the choice between *loaded*@1 and *alive*@2, the former is preferred.

5.1.1. Is the Intuition Correct?

The reasoning situation is not a problem for existing non-monotonic or defeasible reasoning systems if one rejects the Hanks-McDermott intuition that not-*alive*@2 is the mandatory, albeit defeasible, conclusion.

We can believe that *loaded*@1, therefore, not-*alive*@2. Or else we can believe *alive*@1, therefore *alive*@2, therefore not-*loaded*@1. It is indeed odd to say that being alive at a time caused the gun to be unloaded at an earlier time. But we are not committed to saying that. Causal laws may be involved, but the reasoning need not be from causes to effects.

Hanks and McDermott most likely have in mind the forward-chaining part of a planner. It is supposed to reason that the system can achieve its goal of not-*alive*@2 by performing an act, the effects of which guarantee *fired*@1. Is it desirable to reason that firing at time 1 will achieve the goal? It is, if at least one of the following is true:

a) actions can be performed that (acceptably) guarantee that the gun will not be unloaded between times 0 and 1; or

b) the possibility of unloading between times 0 and 1 is not considered a serious possibility by the planner.

If (a), then Hanks and McDermott have no problem. Just cite those actions (that guarantee *loaded*@1 if *loaded*@0) as part of the plan. (b) is more interesting. Rich

Pelavin's dissertation [Pelavin86] discusses the problem of "airtight" planning in non-deterministic worlds. His relevant contribution here is the idea that possibilities that could subvert plans must be antecedently defined. If Hanks and McDermott want to consider explicitly the possibility of unloading, then they cannot expect the planner to reason that firing at time 1 will be sufficient. On the other hand, suppose they do not want to consider explicitly the possibility of unloading. Then the is-loaded persistence rule should not be defeasible. In either case, either the problem is misrepresented, or there is no problem.

As the problem is stated, there is nothing wrong with concluding that there could have been an unloading, hence *alive*@2.

Believing that not-*alive*@2 is mandatory for a defeasible inference system, or even desirable, given the knowledge that is explicitly represented commits one to undesirable inferences, in related situations. Consider a business school student known to be registered for full-time studies at some time.

Example II.1. (see fig. 5.2)

MBA-student@1
(d1) *MBA-student*@1 >— *MBA-student*@2
(d2) *MBA-student*@1 >— *MBA-student*@0
(1) if *MBA-student*@0 then not-*MBA-student*@2.

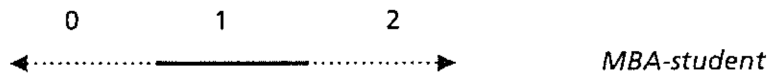

fig. 5.2. The property of being an *MBA-student* tends to persist both forward in time and backward in time. If the time of being a student is limited, then these two tendencies to persist conflict.

If (d1) is a "forward persistence axiom," then (d2) is a "backward persistence axiom." It seems in this example that backward persistence is as desirable, as probable, and as warranted as forward persistence. And it does not matter whether the task is historical reasoning or presentiment. A defeasible rule such as (d2) could be important in predictions about the future as well as the past. It may be, for instance, that what is at stake is future salary, which depends on when the MBA student actually graduated.

The problem with Hanks and McDermott's and Shoham's forward-marching solutions here is that they conclude *MBA-student*@0, and not-*MBA-student*@2. Given that we know our young corporate aspirant is in school, we are obliged to conclude that she is on the verge of graduation! Even keeping (d1) and discarding (d2), so that there is only a simple forward persistence axiom, the forward-marching solution is anomalous. It maintains that given the MBA student is matriculated, one must conclude (defeasibly) that she is a first-year student, or more radically, that she has just arrived.

Examples can be found at will that share exactly the same syntactic structure of the Hanks-McDermott problem, but do not seem to require the analogous conclusion. It does not seem to matter whether the laws involved are nomological, or reflect causal relations, or are just laws of association. In the two examples that follow, the strategy of drawing defeasible conclusions in a way that prefers earlier conclusions is repeatedly shown to be lacking. In fact, in the latter of the two examples that follow, the conclusion that is analogous to concluding *alive*@2 is the intuitively desirable one, contrary to Hanks and McDermott's pattern of reasoning.

Consider the extinction of the South American marsupial carnivore, Thylacosmilus. It is unknown whether Thylacosmilus had exclusively South American extent during the entire Tertiary period; certainly at the early Tertiary period that it did. But it could have migrated during the mid-Tertiary. Certainly if it did have such restricted geographic extent, then when the Panamanian isthmus rose, the better-developed northern species would have forced Thylacosmilus' extinction by the late Tertiary. Without the assumption of restricted-geography, the time of Thylacosmilus' extinction cannot be fixed. Had the species spread beyond the continent before the rise of the isthmus. Symbolically,

Example II.2. (see fig. 5.3)

SA-restricted@0
species-intact@0
rise-of-isthmus@1
(1) if *SA-restricted*@1 and *rise-of-isthmus*@1 then *species-extinct*@2
(d1) *SA-restricted*@0 >— *SA-restricted*@1
(d2) *species-intact*@0 >— *species-intact*@1

(d3) *species-intact*@1 >— *species-intact*@2.

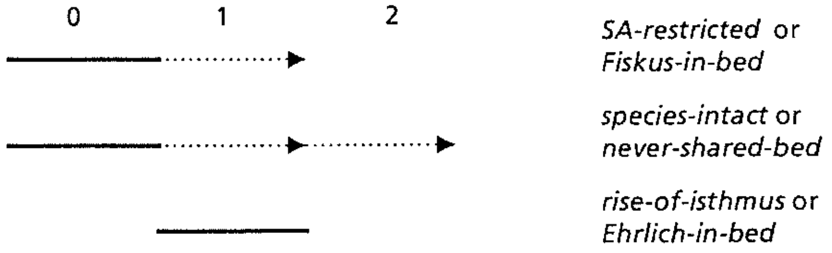

fig. 5.3. Two problems that we would like to represent exactly like the gun-firing problem, but which have different preferred conclusions.

Reasoning to not-*alive*@2 in the gun-firing problem seems to require reasoning to *species-extinct*@2 in this problem. But there is no reason to suppose extinction over intactness here. We deliberately posed an ambiguous reasoning situation.

Consider Drs. Fiskus and Ehrlich, medical residents who are on-call and must share a single resident's bed. Once in bed, each tends to remain there. But for reasons of propriety, they will not share the bed. To date, they have never shared the on-call bed. One night, Fiskus is in bed at midnight. Ehrlich is tired at midnight and he will be in the bed by the start of the late shift. If Fiskus remains in the bed, then the fact that Drs. Ehrlich and Fiskus have never shared the on-call bed will cease to persist. But it is more plausible that the prospect of sharing the bed with his colleague will cause Dr. Fiskus to get up and get out, preserving their proper professional relationship. Symbolically,

Example II.3. (see figure 5.3, again)

never-shared-bed@0
Fiskus-in-bed@0
Ehrlich-in-bed@1

(1) if *Ehrlich-in-bed*@1 and *Fiskus-in-bed*@1 then not-*never-shared-bed*@2

(d1) *Fiskus-in-bed*@0 >— *Fiskus-in-bed*@1

(d2) *never-shared-bed*@0 >— *never-shared-bed*@1 (d3) *never-shared-bed*@1 >— *never-shared-bed*@2.

In each case, we want to represent the problem with the sentences shown. But "temporally forward priority" permits an undesirable inference, unless we add some other information, which would defeat the unwanted inference.

On Hanks' and McDermott's side, there is reason to be suspicious of what has happened. There is an interdependence between representation and inference. As the inference rules and the meaning postulates of the language are changed, so too change the sentences that represent a given situation. Adopting new inference rules changes the way problems are represented. So perhaps the sentences above do not represent the situation as we understand it; perhaps the correct way to represent the situation is to add some sentence that represents the fact that we *do not* want the Hanks-McDermott conclusion here.

Moreover, non-monotonic inference mechanisms do not guarantee that their conclusions will be correct. Providing an example in which a non-monotonic rule guesses incorrectly does not refute the rule. It is not a refutation of a non-monotonic rule merely to show that there are situations in which the rule is a bad rule, irrespective of whether it correctly.

Nevertheless, the sentences above so naturally represent the problem described, and the Hanks-McDermott rule mandates a conclusion so egregiously unwarranted, that these cases must be taken as serious challenges to the plausibility of the rule.

So perhaps the Hanks-McDermott intuitions are wrong. The conclusion, not-*alive*@2, ought not be mandated, at least given the information represented. Nor ought it be mandated that Thylacosmilus was extinct by the late Tertiary period. Nor should be required the sharing of the on-call bed by our reluctant medical residents.

Hanks and McDermott are actually aware of situations in which the represented knowledge has identical form, but the intuition is reversed.

> In a richer temporal formalism the criterion [of] 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.

It is worth pausing to wonder about this last idea: "facts tend to stay true forever unless clipped . . .," as if with an inertial veracity.

5.2. THE PROPER REPRESENTATION.

Hanks and McDermott's second belief is that existing systems of non-monotonic reasoning cannot be made to mandate their preferred conclusion. If the represented facts are not augmented, if the situation is not better qualified, if the sentences stand as they are, then it is a virtue of existing systems that they abstain. Otherwise, the system would be vulnerable to the anomalies of reasoning just discussed.

But is there any way to force the not-*alive*@2 conclusion, possibly by altering the representation of the problem situation?

5.2.1. Implicit Orderings of Extensions.

One approach is to add another defeasible rule. It will encode knowledge that appeared to be reflected in the material conditional, but which actually needs to be explicit. Rule (1) ought to have been defeasible in the first place. The facts, *fired*@1 and *loaded*@1, do not guarantee not-*alive*@2. What if *firing-pin-removed*@1? Or *finger-in-front-of-hammer*@1? The rule comes with a natural set of "unless" conditions, i.e., simple defeaters. And it comes with a set of "even if" conditions, i.e., conditions which, in any combination, do not interfere with the association reflected in the defeasible rule. One "even if" condition is *wearing-after-shave*@1. Another is *sun-shining-in-Providence*@1. Another is *alive*@1. This last one is the most interesting.

The antecedent of the defeasible rule could be specialized in any of a number of ways, while still being a reason for the consequent. In particular,

(d4) *fired*@1 and *loaded*@1 and *alive*@1 >— not-*alive*@2

can be added to the knowledge base.

There were three theories for selecting among multiple extensions, which use specificity defeaters that existed when the shooting problem was posed: David Poole's [Poole85], Donald Nute's [Nute85, Nute86], and my own [Loui87]. Each theory says that in the choice between the following lines of reasoning, the latter is superior:

```
{
alive@0;
alive@0 >— alive@1;
    therefore alive@1
alive@1 >— alive@2;
    therefore alive@2
}
```

and

```
{
alive@0;
alive@0 >— alive@1;
    therefore alive@1;
loaded@0;
loaded@0 >— loaded@1;
    therefore loaded@1;
fired@1;
fired@1 and loaded@1 and alive@1 >— not-alive@2;
    therefore not-alive@2
}.
```

This superiority is determinable strictly on the basis of syntax. On Poole's account, the latter is "more specific" than the former. Either *alive*@0 or *alive*@1 will make the former "applicable" to *alive*@2, but will not make the latter "applicable" to not-*alive*@2. Meanwhile, the latter is made applicable only if it is unconditionally known that *alive*@1 or *alive*@0, and *fired*@1, and *loaded*@0 or *loaded*@1. That makes four possibilities. In each of the four, the former is made applicable too. Thus, according to Poole, the latter is more specific and hence is preferred.

In Nute's system, (1') is a "superior non-monotonic rule," because its antecedent, "*fired*@1 and *loaded*@1 and *alive*@1" entails "*alive*@1", which is the antecedent of its only challenger.

In [Loui87] or the system of chapter 4, the latter argument is superior for a couple of reasons. It has "superior unconditional evidence," i.e. {*alive*@0} is entailed by {*alive*@0; *loaded*@0; *fired*@1}. It has "superior specificity," i.e., {*alive*@1 >—

alive@2} is less specific than {*fired*@1 and *loaded*@1 and *alive*@1 >— not-*alive*@2}, and this comparison is crucial.

It is difficult to imagine a system for selecting among competing defeasible conclusions that would not favor the conclusion with superior evidence, superior specificity. Specializing the antecedent strengthens the rule in such a way that the rule now dictates what should be done regarding the conflict with the persistence rule.

The reason Hanks and McDermott want to clip being alive instead of being loaded is that intuitively, they know that *alive*@1 is one of the assertions that can be in the antecedent of (the defeasible version of) their rule. In short, they hold (2). What is interesting about the kinds of rules that they call "causal" is that they have implicit "even if" conditions. Suppose a rule is

(R) if Φ >— Ψ,

where Φ is a set of properties at some time, $\{\Phi_i@t_0\}$, and Ψ is a set of properties at a later time, $\{\Psi_i@t_1\}$. If (R) is a Hanks-McDermott "causal" rule, it should be the case that

for any $t^- \leq t_0$, and any Ψ_i s.t. $\Psi_i@t_1 \in \Psi$,
if $\{\Psi_i@t^-\} \cup \Phi$ >— Ψ.

This allows clipping of any property Ψ_i holding at a time t^-.

What is it about being alive that allows it to be clipped? It is not just its temporal relation to loaded-ness and the firing. It is the fact that it can appear in the antecedent. The rule that firing loaded guns results in dying takes into account the reasons for persistence of being alive. Consider properties that hold at subsequent times that can appear in the antecedents. These can be clipped too.

Example 5.3.1. (see fig. 5.4)

Well-to-do George Eastman was instrumental in the raising of Rochester into the city in the 20's that many thought would be the next great Eastern metropolis. Suppose that Eastman had most of his wealth in the stock market just before the great crash. We know that Eastman was rich in the 1910's and again by the 1930's, regardless of what might have happened in 1929 to his net worth. We would normally reason with backward persistence that (d1) *rich*@3 >— *rich*@2.

But we have the rule

(d2) *majority-of-personal-wealth-in-market*@1 and *market-crash*@1 >— not-*rich*@2.

And we know the antecedents are true:

majority-of-personal-wealth-in-market@1
stock-market-crash@1
rich@3.

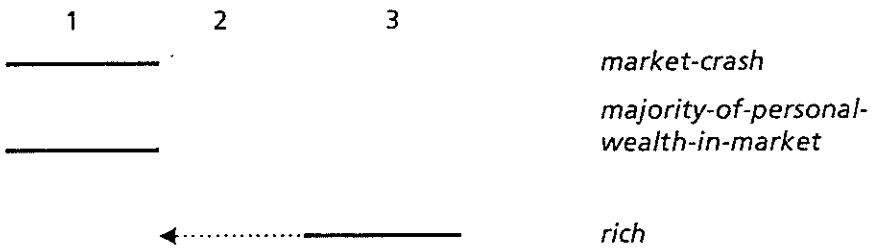

fig. 5.4. An example in which antecedent inclusiveness (or specificity) prevails, despite temporal order.

In this case, we will conclude not-*rich*@2, because what we really know is not just (d2), but also

(d2′) *majority-of-personal-wealth-in-market*@1 and *market-crash*@1 >— not-*rich*@2, even if *rich*@3

or

(d2″) *majority-of-personal-wealth-in-market*@1 and *market-crash*@1 and *rich*@3 >— not-*rich*@2.

We know that *rich*@3 can be included in the antecedent without disturbing the association. In short, we know that even in the presence of *rich*@3, we would conclude that Mr. Eastman lost his shirt in the 1929 crash.

The point is that it is not temporal precedence, but rather antecedent inclusiveness that is important. This should be clear if we alter the original gun example. Suppose we know not only that

loaded@0
alive@0
fired@1

(d1) *alive*@0 >— *alive*@1
(d2) *alive*@1 >— *alive*@2
(d3) *loaded*@0 >— *loaded*@1

but also the more "antecedent-specific" or "antecedent-inclusive" rule

(d4') if *alive*@1 and *fired*@1 and *loaded*@1 >— *alive*@2.

(discard the earlier (d4) rule). Thenthe conclusion is *alive*@2. (d4') implores us to conclude not-*alive*@2, despite loaded-gun-firing.

Consider the example of the extinction of Thylacosmilus augmented.
Rule

(d4') *species-intact*@1 and *SA-restricted*@1 and *rise-of-isthmus*@1 >—
species-extinct@2

forces the conclusion *species-extinct*@2. The only dissenting chain of reasoning uses the (d3) rule ("if *species-intact*@1 >— *species-intact*@2"). But it has a less specific antecedent, "*species-intact*@1," which we indicates deference to the rule with the more specific antecedent.

It is worth wondering whether (d4) should be derivable from (1) ("if *SA-restricted*@1 and *rise-of-isthmus*@1 then *not-species-extinct*@2"). (1) certainly entails the following rule:

(1′) if *species-intact*@1 and *SA-restricted*@1 and *rise-of-isthmus*@1 then species-*extinct*@2.

Is (1′) stronger than (d4)? Perhaps a material connection, an indefeasible connection, implies that there be a "defeasible" connection too. Apparently this is not the case. At least in Poole's, Nute's, and my systems, the defeasible conditional represents more than a poor man's version of the material connection. Such rules are taken to include implicit directives for choosing among competing defeasible extensions of the theory. Therefore, it must not be the case that material connections express certain connections and defeasible connections express high probability connections, making all material connections defeasible ones as well. Indefeasible connections are not improper or extreme versions of defeasible connections.

This proposal for arriving at the not-alive@2 conclusion amounts to an implicit encoding of preferred extensions. It is epistemologically no different from saying in a default formalism that there is knowledge that the not-alive@2 extension is preferred to the alive@2 extension; it is no different from saying that the models are ordered in such a way that not-alive@2 is preferentially entailed, in the sense of [Shoham87]. There is a drawback of writing down the preference of the not-alive@2 extension in this implicit form: not only is the rule

alive@1 >— alive@2 ,

defeated, but also rules

alive@1 & loaded@1 >— alive@2 or
alive@1 & fired@1 >— alive@2 or
loaded@1 & fired@1 >— alive@2 .

In the shooting example, it is difficult to imagine situations in which persistence is to be defeated without also defeating

alive@1 & fired@1 >— alive@2 , and so forth;

but perhaps in other situations, finer control is needed. If so, then we might actually need a language in which we can express explicit orderings over worlds, preferences over models, although such language would likely be cumbersome.

5.2.2. Reasoning about Events.

A different way of accomodating Hanks and McDermott's intuition introduces events and relations on events. McDermott's remarks about explanations in "Critique of Pure Reason," [McDermott87] suggest that the Hanks-McDermott intuition is based on reasoning such as:

1. fact: there was a FIRING event.
2. if not-alive@2, then there was a DYING event.
3. a DYING event could be EXPLAINED by the FIRING event.
4. if alive@2, then there must have been an UNLOADING event.
5. we don't know what would explain such an event.
6. so we prefer to reason that there was a DYING event rather than that there was an UNLOADING evet.
7. we can conclude not-alive@2.

If indeed this is the kind of reasoning that is desired, then it is appropriate to represent it.

1. unexplained-event(FIRING@1)
2. occurs(FIRING@1)
3. loaded@0
4. alive@0
5. (loaded@0 and not-loaded@1) iff occurs(UNLOADING@0)
6. (alive@0 and not-alive@1) iff occurs(DYING@0)
7. (alive@1 & not-alive@2) iff occurs(DYING@1)
8. unexplained-event(UNLOADING@0)
9. unexplained-event(DYING@0)
10. if (occurs(FIRING@1) and loaded@1) then (occurs(DYING@1) and explained-event(DYING@1)
11. if unexplained-event(x) >— not-occurs(x).

The crucial assertion is (8), which says that we do not know what would cause an unloading. It would be difficult to prove this from more primitive assertions. Apparently, closed-world reasoning or circumscription on explanations would be required. In the Yale shooting problem, it appears to be part of what is stated.

Now, in effect, the defeasible rule (11) minimizes the occurrence of events that are unexplained.

p1.	not-occurs(UNLOADING@0)	11, 8
p2.	not-loaded@0 or loaded@1	p1, 5
p3.	loaded@1	p2, 3
p4.	occurs(DYING@1)	p3, 2, 10
p5.	not-alive@2	p4, 7, 4.

Perhaps this approach makes the reasoning "domain-specific," which Hanks and McDermott would like to avoid. It certainly suggests a treatment of events that is not a part of the non-monotonic inference engine. Note that the persistence axioms, e.g.

if alive@0 then defeasibly alive@1

have been discarded. Treating events specially is no different from treating time specially, as a metaphysical notion external to the inferential mechanism.

Structure imposed on problem representation, reflecting metaphysical assumptions, is not a proper part of inference. It should not be surprising that non-monotonic inference systems leave some knowledge representation work to be done.

[Lifschitz87], [Haugh87], [Weber87] and [Morgenstern&Stein88] agree that it may be a good idea to state Hanks and McDermott's intuitions in terms of closed-world reasoning about events. This approach is no panacea, however. Events must be distinguished from non-events. In the present formulation, there is no syntactic reason why alive@1 & not-alive@2 indicates an event has occurred, while alive@1 & alive@2 does not. We could consider CONTINUING-TO-BE-ALIVE an event, and require explanation of it. At least, we must acknowledge that it is world knowledge that allows us to write that CONTINUING-TO-BE-ALIVE events normally do not need explanations. We also lack a theory of when events can account for other events (could the waiting have accounted for the unloading?). Finally, it is not clear that rational persons should minimize the number of unexplained events. However, if one must imbue reasoning with metaphysical assumptions about actions and events, this is the way to do it.

5.3. HANKS-McDERMOTT REBUTTAL.

Despite McDermott's positive initial reaction to the use of specificity in addressing the Yale shooting problem [McDermott87], Hanks and McDermott have decided that the approach is defective.

Nearly all of Hanks and McDermott's criticism approach refers to the use of Poole's system of defeasible inference. It has been pointed out that Poole's system has problems: the distinction between necessary and contingent facts does not seem to be tenable, yet it is critical; there is no rule stating in what way theories must disagree in order for specificity to be relevant for choosing among them; Poole himself discusses the anomaly that his system will do no chaining, i.e., it will not base preference of arguments on preference of sub-arguments. Hanks and McDermott also notice these problems. It is a fact that Poole's system can handle the original problem, but Hanks and McDermott construct elaborations of the shooting problem which Poole's system cannot handle because of its inferential limitations. None of this criticism applies to Nute's or my system, which I had noted were also species of reasoning systems with specificity defeaters.

Consider Hanks and McDermott's elaborations of the example. If the person lives, she goes out in the rain and gets wet. Otherwise, her corpse remains indoors and dry. So

if alive@2 then wet@3
if not-alive@2 then not-wet@3 .

In Nute's system, there is no non-monotonic proof of wet@3 because it would have to use a non-monotonic rule that is not a "superior non-monotonic rule":

alive@1 >— alive@2

would be required in the proof of wet@3; Meanshile, Nute can conclude not-wet@3:

alive@1 & fired@1 & loaded@1 >— not-alive@2 is a rule;

alive@2 and not-alive@2 conflict;
if alive@1 & fired@1 & loaded@1 then alive@1;
"alive@1 >— alive@2" is not a superior rule.
"alive@1 & fired@1 & loaded@1 >— not-alive@2" is a superior rule;

In my system, the argument for not-wet@3 has superior evidence and superior specificity, compared to the argument for wet@3, and it is even protected by undefeated counter-arguments. It is the better argument.

The second elaboration of Hanks and McDermott is

P@0 & Q@0 & R@2
P tends to persist, i.e.:
  P@0 >— P@1;
  P@1 >— P@2;
  P@2 >— P@3
Q tends to persist, i.e.:
  Q@0 >— Q@1;
  Q@1 >— Q@2;
  Q@2 >— Q@3
R tends to persist, i.e.:
  R@2 >— R@3
Q@0 >— not-P@1
P@2 & R@2 >— S@3

Hanks and McDermott would like to conclude not-P@1, and they do not want to conclude S@3. According to what is written, it is not clear that these are the rational inferences. However, they believe the conclusion cannot even be forced by adding

P@0 & Q@0 >— not-P@1 .

As they indicate, it is not clear what Poole will do here, though I believe that under the intended interpretation of when theory-selection should be applied, he should conclude not-P@1. Charity requires that we use Poole's rules only when comparing theories that establish contrary conclusions, and even then, only when comparing those parts of the theories required to establish those contrary conclusions. Otherwise the rules do not begin to make sense.

Regardless, Nute's system does what is desired. Nute takes the rule that concludes not-P@1 to be a superior rule, and the rule that concludes P@1 not to be. So Nute can non-monotonically prove not-P@1, and cannot prove S@3, non-monotonically or otherwise. My system will also do what is desired. It will find not-P@1 to be a justified defeasible conclusion, but will not find S@3 to be, since the argument for not-P@1 is an undefeated counter-argument to the argument for S@3.

Clearly a criticism of Poole's system will not succeed to impugn the idea that systems of inference with specificity defeaters can handle problems in which there are competing persistence rules. No doubt, Hanks and McDermott could find objections to my system and Nute's, as systems of inference, but there will be ways to accomodate those objections, while maintaining the thesis about representation.

Hanks and McDermott believe that I "provide no general method to do coding of temporal information in the causal rules so that theory-comparison will work". I disagree. I have suggested that what Hanks and McDermott may have in mind by a "causal rule" is a rule of which the following is true: the rule reports that properties holding at a time constitute a reason for a target property holding at a later time, and they do so regardless of whether the target property holds or fails to hold at earlier times. They are conditions under which, typically, future state is coerced to have this target property. Saying a rule is "causal" may just be a shorthand for saying that certain aspects of the past are irrelevant: that the force of the reason being reported is independent of whether the target property held in the immediate past. It may be shorthand for saying that it is a rule that should defeat all relevant persistence rules.

Hanks and McDermott claim the "proposed solution is singularly unsatisfying" because it "[does not depend] on issues of causality." I do not illuminate the relation between causality, time, and situation calculus.

Rules that genuinely deserve to be called causal have more to them. Causality has something to do with effective strategies or increased probability in homogeneous populations (see [Cartwright79] or [Salmon84]), or nomological necessity. In AI, when Pearl speaks of causality, he means implicit independence or irrelevance (e.g., [Pearl87c]). This is exactly what is captured by the strengthenability of antecedents of reasons.

There actually is a point worthy of disagreement here, regarding the role that causality plays in reasoning. In philosophical decision theory, there is a dispute over the role that causality plays in the proper analysis of some decisions. Causal decision theorists (e.g., [Gibbard, Harper77]) hold that the proper analysis of certain decision situations requires appeal to a causal notion: the constraint that events at future times cannot affect changes of state at earlier times.

Recall a version of Newcomb's Problem, which is the basis of the dispute. There is an opaque box next to a transparent box; it is clear to see that the latter holds \$1000. A clairvoyant decides to place \$1M in the opaque box at time t_0 iff she believes you will take only the opaque box. At time t_1, do you take just the opaque box, for maybe \$1M, or both boxes, for possibly \$1M+\$1000, or possibly just \$1000? Causal decision theorists think that both boxes should be taken because the clairvoyant has already acted. The causal information is relevant. Evidential decision theorists hold that if the information required to recommend this decision is actually known, then it is equally well stated as probabilistic information (e.g., [Kyburg80], [Levi75]). Suppose we go to another world, and the clairvoyant were an oracle instead. Then the evidence for the causal rule might be defeated by the knowledge that oracles can make the future affect the present, by reading the future. The causal analysis would falter, but the evidential analysis would not. The example need not be so outlandish; Nancy Cartwright puzzles over buying Diversified's Life Insurance, which advertises that policy holders happen to have a longer expected life [Cartwright79]. But her life expectancy is presumably independent of whether she holds Diversified's policy; buying a policy will not lengthen her life. Or will it? We presume that there is no causal relation, but if there were, say, because Diversified Life has an aggressive sales staff that occasionally hires assassins, then it is the probabilistic account that is required to discover that there is.

There is a large number of causal decision theorists. Apparently some people find the notion of cause indispensable and irreducible in their representation of knowledge.

But the Yale School's case for worrying about causality is weaker than the case for causal decision theory. In causal decision theory, causal considerations lead to statements of irrelevance: e.g., given that the clairvoyant has already acted, the choice among boxes is irrelevant to the contents of the opaque box. That seems a correct consequence of causality. In temporal non-monotonic reasoning, causal considerations are supposed to legitimize the postponement of abnormality: for example, if it is unclear whether one nucleus decayed at t_0 or instead a different nucleus decayed at t_1, because the nuclei are so spaced that the arrival of a single electron signals one or the other event (one nucleus is farther away than the other), then choose to believe the latter. Causality does not justify this reasoning.

It is also a mistake to relate the Yale School's chronological minimization rule to the work on temporal aspects of proximal possible worlds. In the context of counterfactuals, David Lewis has written about the difference between small miracles and large miracles, given fixed causal laws [Lewis79]. He argues that we should evaluate the truth of counterfactuals in the world in which abnormality is postponed rather than one in which abnormality occurs earlier. Because even if the miracle required is larger, it allows most of the past to be unchanged. Despite the similarities, none of this can be imported to support chronological minimization. In Lewis' work, counterfactuals are evaluated relative to the actual world. A fundamental assumption is that one knows in what world one is, when counterfactuals are being evaluated. Chronological minimization, as a principle of inference, purports to guess in what world one must be, given partial knowledge of the world.

Finally, consider problems with persistence rules that compete over the same times. The Yale School can do nothing with them, while non-monotonic systems with synchronically preferred extensions can. This is Weber's Trash-Can Problem, and is familiar to extended discussions of the frame problem. A robot is holding a trash-can in situation s_0, and moves from l_0 to l_1, thereby defining a new situation, s_1.

$s_1 = \text{result}(\text{move}(l_0, l_1), s_0)$
$\text{at}(\text{me}, \text{l0}, \text{s0})$
$\text{holding}(\text{trash-can}, s_0)$
$(x)(l)(l')(s) \,.\, \text{at}(x, l', \text{result}(\text{move}(l, l'), s)$
$(l)(s)(x) \,.\, [\text{at}(\text{me}, l, s) \,\&\, \text{holding}(x, s)] \supset \text{at}(x, l, s)$,

and the defeasible rules:

$\text{'holding}(x, s)\text{'} >\!\!— \text{'holding}(x, s)\text{'}$
$\text{'at}(x, l, s)\text{'} >\!\!— \text{'at}(x, l, \text{result}(a, s))\text{'}$

The question is whether holding the trash-can persisted between situations, so that the trash-can is at s_0 with the robot, or whether the trash-can's location persisted between situations, so that the trash-can remains at l_0. At the moment, either claim is possible. One argument runs

$\text{holding(trash-can, } s_0)$
$\text{holding(trash-can, } s_0) >\!\!- \text{holding(trash-can, } s_1)$
$\text{at(me, } l_1, s_1) \text{ \& holding(trash-can, } s_1) \supset \text{at(trash-can, } l_1, s_1)$.

The other argument is

$\text{at(trash-can, } l_0, s_0)$
$\text{at(trash-can, } l_0, s_0) >\!\!- \text{at(trash-can, } l_0, s_1)$.

The sentences here are the ones we would use for describing both those agents that continue to hold when they move, and those that drop things when they move. We can encode our knowledge of a preference by writing that holding persists even through moves. i.e.

$\text{'holding}(x, s) \text{ \& at}(x, l, s)\text{'} >\!\!- \text{'at}(x, l', \text{result(move}(l, l'), s))\text{'}$.

Now there is a stronger argument for the trash-can being at l_1, which defeats the argument for the trash-can remaining at l_0. This could not be done using temporally biased approaches, since the conflicting persistence rules refer to the very same temporal situations.

Of course, I have begged the epistemological problem of how we come to know that we want to prefer one extension over the other, and therefore, want to write down knowledge in a way that guarantees this.

McDermott says emphatically in [Pylyshyn87] that non-monotonic logic solves the frame problem. Energetic philosophers like Haugeland and Dennett who want to reflect on the nature of reasoning about changing worlds more generally are welcome to do so. They can worry about the epistemological problems. But they would be attacking the problem of reasoning about changing worlds, not the frame problem: the narrow, technical problem of finding a language in which we can easily specify what we know we want. Likewise, if the Yale Shooting Problem is the problem of competing frame axioms, then non-monotonic systems that prefer extensions syntactically solve the Yale Shooting Problem.

* page numbering skips here from 126 to 179.

Chapter 8. Conventionalism and Non-Monotonicity

8.1. THE CLASH OF INTUITIONS.

Touretzky, Horty, and Thomason have decided that there is an irremediable "clash of intuitions" on how inheritance reasoners should behave. Intuitions differ on how to prefer some arguments over others, some extensions over others, or some inheritance paths over others. There are numerous, plausible non-monotonic inference systems being proposed, which seem to disagree on what inferences are warranted, assuming a naive translation of rules and knowledge from system to system. Touretzky, Horty, and Thomason write: "Just as there are alternative logics, there may be no single 'best' approach to non-monotonic multiple inheritance." [Touretzky et.al.87] They are right, but there is more to be said.

We should avoid two types of anarchy. The first would result if we decided that there were no objective grounds for deciding among competing systems. There are numerous different systems being proposed. Although there may be no best approach, nor even a best class of approaches, nevertheless, there may exist standards for choosing one approach over another, and it may be possible to formalize some of those standards. The second anarchy would result if we allowed ourselves to propose increasingly complex systems of inference, based on increasingly expressive languages, without addressing the epistemological problem of how we come to know what to write down in those languages when we want to write down something in particular. This is the reason why we cannot begin to criticize the various approaches, whatever the degree of our malcontent. If we disagree with the mandates of a system of inference, the entrenched proponent of the system can always claim that we wrote down the wrong our knowledge in the language incorrectly. This kind of maneuver must be discouraged.

One way of stemming both of these currents is to confront the epistemological problem directly.

We normally think of writing down an uninterpreted formal system, such as a non-monotonic logic, as the first step, after which follows the use of the system. A language is defined, then users struggle to render their sentences consistent and struggle to develop proficiency. It takes time for some to learn what kinds of conditionals should be written as material conditionals in first-order logic; what instances of epistemic commitment correspond to the K operator in modal S4 , and which correspond to modal S5 ; what relations among blocks in the world should be represented by the "Supports" relation axiomatized in a program. It would take time for some to learn when to write down ISA links, and time to adjust from a first conception of ISA links, as axiomatized by [Touretzky86], for instance, to a later conception of ISA links, for instance, [Sandewall86].

There is a different way of thinking. Consider first the user of the language. Users in fact do write down certain sentences to represent certain situations, and empirically, when they think they are in some situations, represented by certain sentences, they may as well believe that they are in the situation represented by adding to those sentences. Most of the time, when one is moved to write down

a ISA p , and
p ISA q ,

one discovers one is in the kind of situation in which to write down

a ISA q .

There may be a logic to this kind of behavior, and we are responsible for formalizing it.

Imagine confiding in a young chess player your favorite maxim for effective play of mid-games, "Never trade for an undeveloped piece." Unfortunately, the young player is not as adept as you at recognizing a piece's development, and sometimes she mistakes an undeveloped piece for a piece quite ripe for trade. Still, when she does identify undeveloped pieces as you would, the rule is useful and saves many a game. This is like the user of a sophisticated defeasible reasoning or multiple inheritance system who often misuses ISA relations, but as well enjoys the power and succinctness of the language, for representation and inference. What factors affect the novice player's decision to follow the rule, or to forget it? Formal answers to this question are what we require.

8.2. CONVENTIONALISM IN NON-MONOTONIC REASONING.

The conventionalist view that has just been taken in decision theory can also * make sense of non-monotonic reasoning systems.

* .This is a reference to a Chapter 7, "Conventionalism and Decision Theory," which is not reproduced here.

8.2.1. Use

Non-monotonic systems are allowed to guess incorrectly, but not badly. So a priori criticism of a system often consists of (1) describing a situation allegedly representable in a particular way, (2) showing that the system mandates a particular conclusion, and (3) claiming that in the situation described, no one would want to guess that conclusion.

When something goes wrong with a non-monotonic system, one can blame the system or one can blame its use. In criticisms of the kind just described, defenders of the system can hold that adept users of the language would not have formally represented the informally described situation in the way that had been alleged. Since there is often no antecedent agreement on which sentences we want to use to represent which situations, the criticism is stymied.

There is no handbook for the use of non-monotonic logical languages. There are no assertability conditions for default rules, TMS justifications, or non-monotonic logic's rules: conditions under which observation of some empirical regularity moves us to write down a particular rule. For more complicated systems with fancier interaction among non-monotonic rules, the situation is worse. When should we write

"(CP (Bird x) (Flies x))" (Doyle) or
"Bird x & M Flies x ⊃ Flies x" (McDermott and Doyle) or
"Bird x $_s$—> Flies x" (Nute) or
"Bird x" >— "Flies x" (Loui) or
"Bird x > Flies x" (Glymour and Thomason),

especially if representing the information in way commits us to the inferential behavior that governs the formal symbols? Presumably the answer has something to do with conditional probability and implicature. We laud Geffner and Pearl [Geffner&Pearl87] because they give precise, probabilistic limit conditions for assertability, for most of their system. However, when they introduce the I(K) relation for irrelevance, they are in the same boat as the rest of us. Conventionalists would add that this boat is Neurath's, and extreme conventionalists would chide too that there is no handbook for the use of monotonic logical languages.

The remedy is to take actual linguistic use to be given and try to find non-monotonic language that effectively exploits whatever regularities there may be in the use of this language by users embedded in the world.

The disadvantage of the view that evaluates non-monotonic systems in an aprioristic way is that it cannot account for the lack of skill of the fallible agent who must assert defeasible connections, or who must make presuppositions. The advantage of the conventionalist view here is that reports of defeasible connections and defeasible conclusions are the starting point; what their origins are does not matter. The task is to systematize them in the most cogent, doubt-eliminating, and error-free way.

These are linguistic matters to be addressed with linguistic intuitions. The intuitions are no different from those that would move us to ask our friend to desist from pointing and declaring "*Bachelor*!" when she is wont to do so with no regular success of indicating an unmarried man.

8.2.2. Representation vs. Inference

The intimate connection between knowledge representation and inference arose in the context of my dispute with Hanks and McDermott. I held that knowledge represented in the way that Hanks and McDermott want to represent the shooting example is ambiguous between situations in which earlier defeasible conclusions are better, and situations in which later defeasible conclusions are better. Hanks and McDermott would hold conversely that if the situation represented is supposed to be ambiguous, then perhaps an explicit piece of knowledge should be added that directs the system *not* to prefer the earlier conclusion to the later conclusion.

The task of non-monotonic inference systems with non-monotonic rules is not merely to rearrange our language. The task is to encode regularities of the world so that when we think we are in a certain situation, in a world with certain regularities, it behooves us to believe that we are in a different situation. It behooves us because we have observed in the past that when we think we are in situation A, we are usually in situation A', and because we are willing to take acceptable chances. Note that we desire to encode this regularity independently of how we decide to describe situation A and situation A', and how we describe the regularities in the world.

However, maintaining the distinction between language rearrangement and world-regularity-prediction is difficult. Conventionalists hold that it is impossible.

In Touretzky's language, where

Chaplain ISA Man ;
Marine ISA Man ;
Chaplain NISA BeerDrinker ;

Man ISA BeerDrinker ;

serve as meaning postulates, one might describe a situation, A, with

George ISA Chaplain; George ISA Marine .

This leaves undetermined whether BeerDrinker should be inherited by George. In Sandewall's language, the very same situation is described by adding the link

Marine ISA BeerDrinker ,

the only effect of which is to prevent Sandewall's rules from inferring that Marine Chaplains are definitely not BeerDrinkers. Would Touretzky want to say that Sandewall has described A properly, but Sandewall's system misses a regularity about the world: that in A , it is not a good bet that George is a BeerDrinker? Or would Touretzky want to find no fault with Sandewall's empirical work; claiming rather that Sandewall's linguistic use is wrong: in the correct description of A in Sandewall's language, add

Marine ISA BeerDrinker ?

In non-monotonic and inductive logic, there are at least two consequence relations. For a set of sentences S , there are

$\mathrm{MCn}(S)$, the sentences monotonically inferrable from S , and
$\mathrm{NmCn}(S)$, the sentences non-monotonically inferrable from S .

$\mathrm{MCn}(S) \subseteq S$;
$\mathrm{MCn}(S) = \mathrm{MCn}(\mathrm{MCn}(S))$;
$\mathrm{MCn}(S \cup T) \supseteq \mathrm{MCn}(S)$;
usually $\mathrm{NmCn}(S) \supseteq \mathrm{MCn}(S)$ (augmentation) ;
usually $\mathrm{NmCn}(S) = \mathrm{MCn}(\mathrm{NmCn}(S))$ (closure) ;
usually $\mathrm{NmCn}(S) = \mathrm{NmCn}(\mathrm{MCn}(S))$ (indifference) ; and
sometimes, but not often, $\mathrm{NmCn}(\mathrm{NmCn}(S)) = \mathrm{NmCn}(S)$ (iteration).

If it makes sense to speak of models, then the models that satisfy NmCn(S) are the preferred models of S, in Shoham's sense [Shoham87].

There are two ways of looking at NmCn. For each S, there is the set

$$\{ S' : MCn(S') = NmCn(S) \} .$$

On one view, NmL appears to have rearranged L so that what each of $\{ S' \}$ used to name in L is now named by S. On the other view, when we used to write down S, we still write down S. NmCn simply says that in these situations, bet that the situation is identical to those in which we'd write NmCn(A).

There is a duality. We can think of being right in our use of L, and lucky that the world's regularities match NmL's predictions. Or we can think of being poor users of L, but being lucky that there is a regularity to our errors which NmL captures.

8.2.3. What Criteria?

In its purest form, suppose we entertain two different non-monotonic languages, L_1 and L_2. In L_1, non-monotonic reasons chain; in L_2, they do not. So writing

$$\{A >\!\!- B\ ;\ B >\!\!- C\} \text{ in } L_1$$

is equivalent to writing

$$\{A >\!\!- B\ ; B >\!\!- C\ ;\ A >\!\!- C\} \text{ in } L_2 .$$

We are happy using either language and can translate freely between them. Suddenly, we are inclined to write

$$\{X >\!\!- Y\ ; Y >\!\!- Z\ ; X\} .$$

Do we conclude Z ? All of a sudden, it matters which of the two languages we seem inclined to speak. It may turn out that in trying to use L_1, we really speak L_2. This may have psychological significance, but none epistemological: if we speak L_2 by trying to speak L_1, then apparently we speak L_2. What is important is that there is an empirical regularity about our language use. This regularity is discovered by applying inductive methods to the choice of non-monotonic logical language.

Suppose at a previous time, t, we made assertions such as

"Evidently p" and
"p >— q" ,

the collection of which is

ObsK(t) .

And at the present time, t_0 , we make assertions such as

"Evidently r" ,

which collectively are

ObsK(t_0) .

These are assertions at times, not assertions about properties at times; p , q , and r may contain temporal aspects, such as

p = "alive@t_0"

of no interest here.

For a non-monotonic logical language L , L's conclusions at the earlier time are sometimes corroborated and sometimes contradicted by descriptions of the world at the present time. Eventually, it will be appropriate to aggregate over various times of assertion (we will vary t). Of two logical languages, L_1 and L_2 , one may be better corroborated.

ObsK(t) permits non-monotonic conclusions under L , the collection of which is

$\text{NmCn}_L(\text{ObsK}(t))$.

It is compared with ObsK(t_0) . Consider

$\text{PrimaFacieK}_{L,t}$ =
$\text{NmCn}_L(\text{ObsK}(t)) \cup \text{ObsK}(t_0)$.

$\text{PrimaFacieK}_{L,t}$ is likely to be inconsistent. We might have asserted

> "Evidently p" and
> "p >— q" ,

from which we concluded

> "Evidently q" and
> "¬Evidently ¬q" ,

and later been confronted with evidence that led us to assert

> "Evidently ¬q" .

We might later have been led to assert "¬Evidently p" too.

Inconsistencies in $\text{PrimaFacieK}_{L,t}$ represent errors. They are either failures of our attempts to assert properly, or evidence of L's inability to capture empirical regularities; of course we cannot decide between the two. Intuitively, if errors are few, we are willing to accept them as evidence of our frailties as users and retain L . But if errors are numerous, it is unproductive to continue to use L and place all of the blame on the user; this may be evidence that a different logico-linguistic convention is in order.

Errors are charged to the user's ability to use various relations in the observation language, ">—" and "Evidently" , instead of charging them to the language L . If errors are negligible, much of $\text{PrimaFacieK}_{L,t}$ can be accepted as rational belief, as belonging to

$$K_{L,t} = \text{the set of rational beliefs} .$$

As errors grow, much less of $\text{PrimaFacieK}_{L,t}$ can be accepted into $K_{L,t}$. There is a tradeoff between a language being informative and a language avoiding error: the more risks taken, the more errors. Of course, taking better risks -- guessing informatively and correctly -- also results in fewer errors. The desire to populate $K_{L,t}$ forces languages that take risks, that populate $\text{PrimaFacieK}_{L,t}$. The fact that error prevents sentences in $\text{PrimaFacieK}_{L,t}$ from being transferred to $K_{L,t}$ forces languages that avoid big risks and bad guesses. By treating error in this way, the two forces oppose each other properly.

There may be many ways of rejecting a minimal set of sentences from PrimaFacie$K_{L,t}$ so that it becomes consistent. Consider each set in turn. For each rejection set, count how many assertions of ">—" must be considered erroneous, and how many assertions of "Evidently" must be taken to be erroneous. Count how many uses of each kind of assertion need not be considered erroneous. Through standard statistical methods, this leads to a probability of error for each type of assertion (and this probability may differ from context to context, again, using standard methods, e.g., depending on what it is that the p and q are about).

Knowing this probability to err for various assertions, we can determine at the appropriate level of acceptance which sentences in PrimaFacie$K_{L,t}$ can be copied into $K_{L,t}$. Do we copy "p >— q" ? It is the type of sentence that uses the ">—" relation on sentences concerning "Domain(p, q)" . If the probability of such an assertion being in error is too high, it does not get transferred. With those sentences that do survive the transfer

$$f\colon \text{PrimaFacie}K_{L,t} \rightarrow K_{L,t},$$

consider the conclusions licensed by the non-monotonic logic. These are the intersection of

$\text{NmCn}_L(mc_i(K_{L,t}))$,
where $mc_i(S)$ picks out the i-th maximal consistent subset of S .

As in the last chapter, some of these sentences are interesting and are counted, to determine the informational value ot the theory. This count is summed over all times, t . If the value for L_1 exceeeds the value for L_2 , we claim that L_1 is the better convention.

8.3. RECAPITULATION.

This procedure is an impractical computation. But it fixes in our minds plausible objective criteria for choosing, in principle, among competing non-monotonic systems: a way of avoiding the two sources of anarchy. It permits the conventionalist view to explain the clash of intuitions, without forcing relativism. A clash of intuitions? Yes, but somebody has got to be wrong.

What we have here, in short, is an abdication of our aprioristic responsibilities. There may be respected and entrenched conventions of the scientific community that impose constraint on inquiry and action a priori. But all such considerations are conventions masquerading as prior intuition. We have acquiesced to the ultimate pragmatism in our evaluation of theories, including our evaluation of normative theories. We expect nothing of the agent except that she use a language with which she is facile, that renders her experience cogent, that is informative, that does not force her to view the world as replete with error. All obligations are excused, as long as she projects future scratches and marks on the wall from past scratches and marks, according to our cherished inductive method.

Bibliography

Aitchison, J. "Discussion of Professor Dempster's Paper," *Annals of Mathemtaical Statistics 39*, 1968.

Allais, M. "The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the postulates and Axioms of the American School (1952)," in M. Allais and O. Hagen, eds., *Expected Utility and the Allais Paradox*, Reidel, 1979.

Allais, M. and Hagen, O., eds. *Expected Utility and the Allais Paradox*, Reidel, 1979.

Asher, N. "Linguistic Understanding and Non-Monotonic Reasoning," *Proc. AAAI Workshop on Non-Monotonic Reasoning*, New Paltz, 1984.

Barnett, J. "Computational Methods for a Mathematical Theory of Evidence," *Proc. IJCAI-81*, 1981.

Bassett, G. "Expected Utility with Perturbed Lotteries," *Theory and Decision 20*, 1986.

Bell, D. and Farquhar, P. "Perspectives on Utility Theory," *Operations Research 34*, 1986.

Birnbaum, A. "Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses," *Journal of the American Statistical Association 56*, 1961.

Blyth, C. "Approximate Binomial Confidence Limits," *Journal of the American Statistical Association 81*, 1986.

Blyth, C. and Still, H. "Binomial Confidence Intervals," *Journal of the American Statistical Association 78*, 1983.

Bolker, E. "The Simultaneous Axiomatization of Utility and Subjective Probability," *Philosophy of Science 34*, 1967.

Brachman, R. "I Lied About the Trees," *AI Magazine 6*, 1985.

Brown, F. M. "Towards the Automation of Set Theory and its Logic," *Artificial Intelligence 10*, 1978.

Brown, R. and Lindley, D. "Improving Judgement by Reconciling Incoherence," *Theory and Decision 14*, 1982.

Carnap, R. *Logical Foundations of Probability*, Chicago, 1950.

Carnap, R. "The Continuum of Inductive Methods," U. Chicago Press, 1952.

Carnap, R. "The Aim of Inductive Logic," in *Logic, Methodology and Philosophy of Science*, E. Nagel, et. al., ed. Stanford Univ. Press, 1962.

Cartwright, N. "Causal Laws and Effective Strategies," *Nous 13*, 1979.

Cartwright, N. *How The Laws of Physics Lie*, Oxford, 1983.

Charniak, E. and McDermott, D. *Introduction to AI*, Addison-Wesley, 1984.

Cheeseman, P. "In Defense of Probability," *Proc. IJCAI-85*, 1985.

Chew, S. "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica 51*, 1983.

Churchland, P. *Scientific Realism and the Plasticity of Mind*, Cambridge, 1979.

Cohen, M. and Nagel, E. *An Introduction to Logic and Scientific Method*, Harcourt, 1934.

Colson, G. and Zeleny, M. "Multicriterion Concept of Risk under Incomplete Information," *Computers and Operational Research 7*, 1980.

Davidson, B. and Pargetter, R. "In Defence of The Dutch Book Argument," *Canadian Journal of Philosophy 15*, 1985.

Davidson, D. "Hempel on Explaining Action," *Erkenntnis 10*, 1976.

Dempster, A. "A Generalization of Bayesian Inference," *Journal of the Royal Statistical Society 2*, 1968.

Diaconis, P. and Zabell, S. "Some Alternatives to Bayes' Rule," T.R. 205, Dept. of Statistics, Stanford University, 1983.

Dillard, R. "The Dempster-Shafer Theory Applied to Tactical Fusion in an Inference System," *Proc. Fifth MIT/ONR Workshop*, 1982.

Doyle, J. "A Truth Maintenance System," *Artificial Intelligence 12*, 1979.

Doyle, J. "Methodological Simplicity in Expert System Construction," *AI Magazine 4*, 1983a.

Doyle, J. "Some Theories of Reasoned Assumptions," CMU Dept. of Computer Science Technical Report CMU-CS-83-125, 1983b.

Edwards, W. "The Theory of Decision-Making," *Psychological Bulletin 51*, 1954.

Etherington, D. "Formalizing Nonmonotonic Reasoning Systems," *Artificial Intelligence 31*, 1987.

Etherington D. and Reiter, R. "On Inheritance Hierarchies With Exceptions," *Proc. AAAI-83*, 1983.

Fagin, R. and Halpern, J. "Belief, Awareness, and Limited Reasoning: Preliminary Report," *Proc. IJCAI-85*, 1985.

Fahlman, S. *NETL: A System for Representing and Using Real-World Knowledge*, MIT, 1979.

Fishburn, P. "Analysis of Decisions with Incomplete Knowledge of Probabilities," *Operations Research 13*, 1965.

Fishburn, P. "Subjective Expected Utility: A Review of Normative Theories," *Theory and Decision 13*, 1981.

Fishburn, P. "Non-Transitive Measurable Utility," *Journal of Economic Theory 31*, 1983.

Franke, G. "Expected Utility with Ambiguous Probabilities and Irrational Parameters," *Theory and Decision 9*, 1978.

Gaifman, H. "A Theory of Higher Order Probabilities," working paper, Hebrew University Mathematics Department, 1985.

Gale, W., ed. *Artificial Intelligence in Statistics*, Addison-Wesley, 1986.

Gärdenfors, P. "Epistemic Importance and Minimal Changes of Belief," *Australasian Journal of Philosophy 62*, 1984.

Gärdenfors, P. and Sahlin, N. "Unreliable Probabilities, Risk Taking, and Decision Making," *Synthese 53*, 1982.

Gärdenfors, P. and Sahlin, N. "Reply to Levi," *Synthese 53*, 1982.

Gärdenfors, P. and Sahlin, N. "Decision Making with Unreliable Probabilities," *British Journal of Mathematical and Statistical Psychology 36*, 1983.

Garvey, T., Lowrance, J., and Fischler, M. "An Inference Technique for Integrating Knowldge from Disparate Sources," *Proc. IJCAI-81*, 1981.

Geffner, H. and Pearl, J. "Sound Defeasible Reasoning," UCLA Cognitive Systems Laboratory TR-94, 1987.

Gibbard, A. and Harper, W. "Counterfactuals and Two Kinds of Expected Utility," in C. Hooker, J. Leach, and E. McLennan, eds., *Foundations and Applications of Decision Theory, v. I*, Reidel, 1977.

Ginsberg, M. "Non-Monotonic Reasoning using Dempster's Rule," *Proc. AAAI-84*, 1984.

Ginsberg, M. "Does Probability Have a Place in Non-Monotonic Reasoning?," *Proc. AAAI-86*, 1986.

Glymour, C. *Theory and Evidence*, Princeton, 1980.

Glymour, C. and Thomason, R. "Default Reasoning and the Logic of Theory Perturbation," *Proc. AAAI Workshop on Non-Monotonic Reasoning*, New Paltz, 1984.

Good, I. *Probability and the Weighing of Evidence*, Hafner, 1950.

Good, I. *Good Thinking: The Foundations of Probability and Its Applications*, Minnesota, 1983.

Good, I. and McMichael, A. "A Pragmatic Modification of Explicativity for the Acceptance of Hypotheses," *Philosophy of Science 51*, 1984.

Hacking, I. *The Logic of Statistical Inference*, Cambridge, 1965.

Hanks, S. & McDermott, D. "Default Reasoning, Nonmonotonic Logics, and The Frame Problem," *Proc. AAAI-86*, 1986.

Hanks, S. and D. McDermott. "Nonmonotonic logic and temporal projection,"*Artificial Intelligence* 33, 1987.

Harman, G. *Change in View*, MIT, 1986.

Harsanyi, J. "Bayesian Theory and its Opponents: Comments on John Watkins," working paper given at Turin, Italy, 1983.

Harsanyi, J. "Acceptance of Empirical Statements: A Bayesian Theory without Cognitive Utilities," *Theory and Decision 18*, 1985.

Haugh, B. "Simple Causal Minimizations for Temporal Persistence and Projection," *Proc. AAAI-87*, 1987.

Heilig, K. "Carnap and de Finetti on Bets and the Probability of Singular Events: The Dutch Book Argument Reconsidered," *British Journal for the Philosophy of Science 29*, 1978.

Hempel, C. "Deductive-Nomological versus Statistical Explanation," *Minnesota Studies in the Philosophy of Science 3*, Minnesota, 1962.

Hintikka, J. and Hilpinen, R. "Knowledge, Acceptance, and Inductive Logic," in Hintikka, J. and Suppes, P., eds., *Aspects of Inductive Logic*, North-Holland, 1966.

Hilpinen, R. "Rules of Acceptance and Inductive Logic," *Acta Philosophica Fennica 22*, 1968.

Horty, J., R. Thomason, and D. Touretzky. "A Skeptical Theory of Inheritance in Non-Monotonic Semantic Nets," *Proc. AAAI-87*, 1987.

Hurwicz, L. "Some Specification Problems and Applications to Econometric Models," *Econometrica 19*, 1951.

Isaacs, H. "Sensitivity of Decisions to Probability Estimation Errors," *Operations Research 11*, 1963.

Israel, D. "What's Wrong with Non-Monotonic Logic?", *Proc. AAAI-80*, 1980.

Jaynes, E. "Where Do We Stand on Maximum Entropy?" in Levine and Tribus, eds., *The Maximum Entropy Formalism*, MIT, 1979.

Jeffrey, R. *The Logic of Decision*, McGraw-Hill, 1965.

Jeffrey, R. "Dracula Meets Wolfman: Acceptance versus Partial Belief," in Swain, M., ed. *Induction, Acceptance, and Rational Belief*, Reidel, 1970.

Jeffrey, R. "Risk and Human Rationality," working paper delivered at the Boston Philosophy of Science Series, 1986.

Kautz, H. "A Logic of Persistence," *Proc. AAAI*-86, 1986.

Kennedy, R. and Chihara, C. "The Dutch Book Argument: Its Logical Flaws, Its Subjective Sources," *Philosophical Studies 36*, 1979.

Klein, P. "The Virtues of Inconsistency," *Monist 68*, 1985.

Krantz, D., Luce, R., Suppes, P., and Tversky A. *Foundations of Measurement*, Academic Press, 1971.

Kyburg, H. *Probability and the Logic of Rational Belief*, Wesleyan University Press, 1961.

Kyburg, H. "Recent Work in Inductive Logic," *American Philosophical Quarterly 1*, 1964.

Kyburg, H. "Conjunctivitis," in Swain, M., ed. *Induction, Acceptance, and Rational Belief*, Reidel, 1970.

Kyburg, H. *The Logical Foundations of Statistical Inference*, Reidel, 1974.

Kyburg, H. "All Acceptable Generalizations are Analytic," *American Philosophical Quarterly 14*, 1977.

Kyburg, H. "Subjective Probability: Criticisms, Reflections, and Problems," *Journal of Philosophical Logic 7*, 1978.

Kyubrg, H. "Acts and Conditional Probabilities," *Theory and Decision 12*, 1980.

Kyburg, H. "The Reference Class," *Philosophy of Science 50*, 1982.

Kyburg, H. "Rational Belief," *Behavioral and Brain Sciences 6*, 1983a.

Kyburg, H. *Epistemology and Inference*, Minnesota, 1983b.

Kyburg, H. *Theory and Measurement*, Cambridge, 1984.

Kyburg, H. "Bayesian and Non-Bayesian Evidential Reasoning," U.R. Dept. of Computer Science Technical Report 139, 1985.

Kyburg, H. *Science and Reason*, draft, 1986.

Kyburg, H. and Smokler, H. *Studies in Subjective Probability*, Wiley, 1964.

Lakatos, I. *Mathematics, Science, and Epistemology*, Cambridge, 1978.

Lehrer, K. "Justification, Explanation, and Induction," in Swain, M., ed. *Induction, Acceptance, and Rational Belief*, Reidel, 1970.

Lehrer, K. *Knowledge*, Oxford, 1974.

Lemmer, J. and Kanal, L., eds. *Uncertainty in AI*, North-Holland, 1986.

Lemmer, J., ed. *Uncertainty in AI v. II*, North-Holland, 1987.

Levi, I. "Corroboration and Rules of Acceptance," *British Journal for the Philosophy of Science 13*, 1963.

Levi, I. *Gambling with the Truth*, Knopf, 1967.

Levi, I. "On Indeterminate Probabilities," *Journal of Philosophy 71*, 1974.

Levi, I. "Newcomb's Many Problems," *Theory and Decision 6*, 1975.

Levi, I. "Acceptance Revisited," in Bogdan, ed., *Local Induction*, Reidel, 1976.

Levi, I. *The Enterprise of Knowledge*, MIT, 1980a.

Levi, I. "Potential Surprise," in Cohen, L. and Hesse, M., eds., *Applications of Inductive Logic*, Oxford, 1980b.

Levi, I. "Self-Profile," in *Henry Kyburg, Jr. and Isaac Levi*, R. Bogdan, ed., Reidel 1982a.

Levi, I. "Ignorance, Probability, and Rational Choice," *Synthese 53*, 1982b.

Levi, I. *Decisions and Revisions*, Cambridge, 1984.

Levi, I. "Imprecision and Indeterminacy in Probability Judgement," *Philosophy of Science 52*, 1985.

Levi, I. "The Paradoxes of Allais and Ellsberg," *Journal of Economics and Philosophy 2*, 1986.

Lewis, C. *An Analysis of Knowledge and Valuation*, Open Court Publishing Company, 1946.

Lewis, D. "Counterfactual Dependence and Time's Arrow," *Nous 13*, 1979.

Lewis, D. "Probabilities of Conditionals and Conditional Probabilities," in *Philosophical Papers, v. II*, Oxford, 1986.

Lifschitz, V. "Pointwise Circumscription," *Proc. AAAI*-86, 1986.

Lifschitz, V. "Formal Theories of Action," *Proc. IJCAI-87*, 1987.

Lopes, L. "Normative Theories of Rationality: Occam's Razor, Procrustes' Bed? Response to Kyburg," *Behavioral and Brain Sciences 6*, 1983.

Loui, R. "Defeat among Arguments: A System of Defeasible Inference," *Computational Intelligence 3*, 1987.

Lowrance, J. "Dependency-Graph Models of Evidential Support," Ph.D. Thesis, U. Massachusetts, Amherst, 1982.

Lu, S. and Stephanou, H. "A Set-Theoretic Framework for the Processing of Uncertain Knowledge," *Proc. AAAI-84*, 1984.

Luce, R. and Raiffa, H. *Games and Decisions*, Wiley, 1958.

Lukaszewicz, W. "Non-Monotonic Logic for Default Theories," *Proc. ECAI-84*, 1984a.

Lukaszewicz, W. "Considerations on Default Logic," *Proc. AAAI Workshop on Non-Monotonic Reasoning*, New Paltz, 1984b.

Lukaszewicz, W. "Formalization of Knowledge and Ignorance: Introduction to Non-Monotonic Reasoning," *Communication and Cognition – AI 3*, 1986.

Marschak. J. "Utilities, Values, and Decision Makers," in M. Allais and O. Hagen, eds., *Expected Utility and the Allais Paradox*, Reidel, 1979.

McCarthy, J. "Epistemological Problems of Aritificial Intelligence," *Proc. IJCAI-77*, 1977.

McCarthy, J. "Applications of Circumscription to Formalizing Common Sense Knowledge," *Proc. AAAI Workshop on Non-Monotonic Reasoning*, New Paltz, 1984.

McCarthy, J. and Hayes, P. "Some Philosophical Problems from the Standpoint of Artificial Intelligence," *Machine Intelligence 4*, 1969. Reprinted in Webber, B. and Nilsson, N., eds. *Readings in Artificial Intelligence*, Tioga, 1981.

McDermott, J. "Easy and Hard Problems in Artificial Intelligence," address to the Artificial Intelligence Society of New England, Brandeis, November 1985.

McDermott, D. "Critique of Pure Reason," *Computational Intelligence 3*, 1987.

McDermott, J. and Doyle, J. "Non-Monotonic Logic I," *Artificial Intelligence 13*, 1980.

McGee, V. "A Counterexample to Modus Ponens," *Journal of Philosophy 35*, 1985.

Mitchell, T. "Learning and Problem Solving," *Proc. IJCAI-83*, 1983.

Moore, R. "Semantical Considerations on Nonmonotonic Logic," *Artificial Intelligence 25*, 1985.

Moore, P. "A Dutch Book and Subjective Probabilities," *British Journal for the Philosophy of Science 34*, 1983.

Morgenstern, L. and L. Stein. "Why Things Go Wrong: a formal theory of causal reasoning," *Proc. AAAI-88*, 1988.

Nilsson, N. "Probabilistic Logic," SRI Technical Note 321, 1984.

Nozick, R. *Philosophical Explanations*, Harvard, 1981.

Nute, D. "Logical Relations," *Philosophical Studies 46*, 1984

Nute, D. "A Non-Monotonic Logic Based on Conditional Logic", working paper, Center for Advanced Computational Methods, Atlanta, 1985.

Nute, D. LDR: A logic for defeasible reasoning. ACMC Research Report 01 - 0013, Advanced Computational Methods Center, University of Georgia, Athens, Georgia, 1986.

Oppacher, F. "Recovering from Knowledge Base Inconsistency," *Communication and Cognition – AI 3*, 1986.

Pastre, D. "Towards the Automation of Set Theory and its Logic," *Artificial Intelligence 10*, 1978.

Pearl, J. "Legitimizing Causal Reasoning in Default Logics," UCLA Computer Science Technical Report CSD-86, R-69, 1986.

Pearl, J. "Probabilistic Semantics for Inheritance Hierarchies with Exceptions," UCLA Computer Science Technical Report CSD-87, R-93, 1987.

Pearl, J. "Embracing Causal Reasoning in Default Logics," *Proc. AAAI-87*, 1987.

Pelavin, R. "A Formal Logic that Permits Planning in Temporally Rich Domains," University of Rochester Department of Computer Science Ph.D. dissertation, forthcoming, 1986.

Poincare, H. *Science and Hypothesis*, (transl. 1905), Dover, 1952.

Pollock, J. *Knowledge and Justification*, Princeton, 1974.

Pollock, J. "A Theory of Direct Inference," *Theory and Decision 16*, 1983.

Pollock, J. "Foundations for Direct Inference," Theory and Decision 17, 1984.

Pollock, J. "Defeasible Reasoning," working paper, U. Arizona Dept. of Philosophy, 1987.

Poole, D. "On the Comparison of Theories: Preferring the Most Specific Explanation," *Proc. IJCAI-85*, 1985.

Potter, J. and Anderson, B. "Partial Prior Information and Decisionmaking," *IEEE Systems, Man, and Cybernetics 10*, 1980.

Pratt, J., Raiffa, H., and Schlaifer, R. *Introduction to Statistical Decision Theory*, McGraw-Hill, 1965.

Quine, W.V.O. *From a Logical Point of View*, Harvard, 1953.

Quinlan, J. "Consistency and Plausible Reasoning," *Proc. IJCAI-83*, 1983.

Raiffa, H. *Decision Analysis*, Addison-Wesley, 1970.

Reichenbach, H. *The Theory of Probability*. U.C. Berkeley Press, 1949.

Reiter, R. "On Closed World Data Bases," in Logic and Data Bases, Gallaire, H. and Minker, J. eds., Plenum, 1978a. Reprinted in Webber, B. and Nilsson, N., eds. *Readings in Artificial Intelligence*, Tioga, 1981.

Reiter, R. "On Reasoning by Default," Proceedings TINLAP, Urbana-Champaign, 1978b. Reprinted in Brachman, R. and Levesque, H., eds. *Readings in Knowledge Representation*, Morgan-Kaufman, 1985.

Reiter, R. "A Logic for Default Reasoning," *Artificial Intelligence 13*, 1980.

Rescher, N. and Manor, R. "On Inference From Inconsistent Premisses," *Theory and Decision 1*, 1970.

Rich, E. "Default Reasoning as Likelihood Reasoning," *Proc. AAAI-83*, 1983.

Roberts, F. "What if Utility Functions Do Not Exist?" *Theory and Decision 3*, 1972.

Roberts, R. and Goldstein, I. *The FRL Manual*. MIT AI Memo 409, AI Laboratory, MIT, 1977.

Sahlin, N. "Three Decision Rules for Generalized Probability Representations," *Behavioral and Brain Sciences 8*, 1985.

Salmon, W. *Scientific Explanation and the Causal Structure of the World*, Princeton, 1984.

Sandewall, E. "Nonmonotonic Inference Rules for Multiple Inheritance with Exceptions," *Proc. IEEE 74*, 1986.

Savage, L. *The Foundations of Statistics*, Wiley, 1954.

Scott, D. and Suppes, P. "Foundational Aspects of Theories of Measurement," *Journal of Symbolic Logic 23*, 1958.

Seidenfeld, T. "Why I Am Not an Objective Bayesian," *Theory and Decision 11*, 1979.

Seidenfeld, T. "Levi on the Dogma of Randomization in Experiments," in *Henry Kyburg, Jr. and Isaac Levi*, R. Bogdan, ed., Reidel, 1982.

Seidenfeld, T. "Decisions with Indeterminate Probabilities: Response to Kyburg," *Behavioral and Brain Sciences 6*, 1983.

Shafer, G. *A Mathematical Theory of Evidence*, Princeton, 1976.

Shafer, G. "Constructive Decision Theory," working paper, Dept. of Mathematics, U. Kansas, 1982.

Shafer, G. "The Empirical Content of the Normative Interpretation of Subjective Expected Utility," Working Paper 185, U. Kansas School of Business, 1986.

Shoham, Y. "Chronological Ignorance: Time, Nonmonotonicity, Necessity, and Causal Theories," *Proc. AAAI*-86, 1986.

Shoham, Y. "Non-Monotonic logics: meaning and utility," *Proc. IJCAI-87*, 1987.

Simon, H. *The Sciences of the Artificial*, MIT Press, 1969.

Simpson, E. "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society B 13*, 1951.

Snow, P. "Bayesian Inference without Point Estimates," *Proc. AAAI-86*, 1986.

Sowden, L. "The Inadequacy of Bayesian Decision Theory," *Philosophical Studies 45*, 1984.

Stalnaker, R. *Inquiry*, MIT, 1985.

Starr, M. "A Discussion of Some Normative Criteria for Decision-Making under Uncertainty," *Industrial Management Review 8*, 1966.

Strat, T. "Continuous Belief Functions for Evidential Reasoning," *Proc. AAAI-84*, 1984.

Swain, M. *Induction, Acceptance, and Rational Belief*, Reidel, 1970.

Szolovits, P. and Pauker, S. "Categorical and Probabilistic Reasoning in Medical Diagnosis," *Artificial Intelligence 11*, 1978.

Takeguchi, T. and Akashi, H. "Analysis of Decisions under Risk with Incomplete Knowledge," *IEEE Transactions of Systems, Man, and Cybernetics 14*, 1984.

Thompson, T. "Parallel Formulation of Evidential-Reasoning Theories," *Proc. IJCAI-85*, 1985.

Touretzky, D. "Implicit Orderings of Defaults in Inheritance Systems," *Proc. AAAI-84*, 1984.

Touretzky, D. *The Mathematics of Inheritance Systems*, Pitman, 1986.

Touretzky, D., J. Horty, and R. Thomason. "A Clash of Intuitions: the current state of nonmonotonic multiple inheritance systems," *Proc. IJCAI-87*, 1987.

Tversky, A. "On the Elicitation of Preferences: Descriptive and Prescriptive Considerations," International Institute for Applied Systems Analysis, Laxenburg, Austria, 1978. Also in Hagen, O. and Wenstøp, F., eds., *Progress in Utility and Risk Theory*, Reidel, 1984.

Tversky, A. and Kahneman, D. "Judgement under Uncertainty: Heuristics and Biases," *Science 185*, 1974.

Vickers, J. "Some Remarks on Coherence and Subjective Probability," *Philosophy of Science*, 1965.

Wald, A. *Statistical Decision Functions*, Wiley, 1950.

Watkins, J. *Science and Scepticism*, Princeton, 1984.

Weber, J. "Formal Theories of Action and the Ramification Problem," U. Rochester Dept. of Computer Science TR252, 1987.

Weirich, P. "Expected Utility and Risk," *British Journal of Philosophy of Science 37*, 1986.

Wesley, L. and Hanson, A. "The Use of an Evidential Based Model for Representing Knowledge and Reasoning about Images in the Vision System," *Proc. IEEE*, 1982.

Weyhrauch, R. "Prologemena to a Theory of Mechanized Formal Reasoning," *Artificial Intelligence 13*, 1980.

Whalen, T. "Decisionmaking under Uncertainty with Various Assumptions about Available Information," *IEEE Transactions on Systems, Man, and Cybernetics 14*, 1984.

White, C., Sage, A., and Scherer, W. "Decision Support with Partially Identified Parameters," *Large Scale Systems 3*, 1982.

Winograd, T. "Extended Inference Modes in Reasoning by Computer Systems," in Cohen, L. and Hesse, M., eds., *Applications of Inductive Logic*, Oxford, 1980a.

Winograd, T. "Extended Inference Modes in Reasoning by Computer Systems," *Artificial Intelligence 13*, 1980b.

Wolfenson, M. and Fine, T. "Bayes-Like Decision-Making with Upper and Lower Probabilities," *Journal of the American Statistical Association 77*, 1980.