Defeat Among Arguments: A System of Defeasible Inference

Ronald P. Loui
Department of Computer Science and
Department of Philosophy
The University of Rochester
Rochester, NY 14627

TR 190 (revised)
December 1986

Abstract

This paper presents a system of non-monotonic reasoning with defeasible rules. The advantage of such a system is that many multiple extension problems can be solved without additional explicit knowledge; ordering competing extensions can be done in a natural and defensible way, via syntactic considerations. The objectives closely resemble Poole's objectives.

But the logic is different from Poole's. The most important difference is that this system allows the kind of chaining that many other non-monotonic systems allow. Also, the form in which the inference system is presented is quite unusual. It mimics an established system of inductive logic, and it treats defeat in the way of the epistemologist-philosophers.

The contributions are both of content and of form: (content) the kinds of defeat that are considered, and (form) the way in which defeat is treated in the rules of inference.

The author is supported by a grant from the U.S. Army Signals Warfare Center.

Defeat Among Arguments: A System of Defeasible Inference

R. P. Loui
University of Rochester
Depts. of Computer Science and Philosophy
Draft of 12/2/86

Abstract.

This paper presents a system of non-monotonic reasoning with defeasible rules. The advantage of such a system is that many multiple extension problems can be solved without additional explicit knowledge; ordering competing extensions can be done in a natural and defensible way, via syntactic considerations. The objectives closely resemble Poole's objectives.

But the logic is different from Poole's. The most important difference is that this system allows the kind of chaining that many other non-monotonic systems allow. Also, the form in which the inference system is presented is quite unusual. It mimics an established system of inductive logic, and it treats defeat in the way of the epistemologist-philosophers.

The contributions are both of content and of form: (content) the kinds of defeat that are considered, and (form) the way in which defeat is treated in the rules of inference.

keywords. defeat, defeasible, non-monotonic, multiple extension, inference.

subject category. knowledge representation.

1. Motivations and Directives.

This paper presents inference rules for a new system of defeasible inference. In rough comparison to existing non-monotonic logics, it is presumptive; it often chooses among multiple extensions via relatively bold syntactic considerations.

The first interesting aspect of this system is the form in which its inference rules are presented. They explicitly mention defeat and enumerate the various kinds of defeat. Defeat is much more complicated than being a member of an exception list. So the form of presentation will already be something new to A.I., borrowing from an established style among epistemologists. The original idea was to copy the defeat mechanisms in Kyburg's and Pollock's theories of direct inference [Kyburg83, Pollock83] and export them to a logic of conditionals. What resulted instead was a

way of evaluating defeasible arguments, based in part on the structure of the arguments.

The next interesting things about the system have to do with the varieties of defeat.

Let a $>\!\!-$ b mean b is inferrable when a is established, unless there is defeat; i.e., a $>\!\!-$ b means that a is a defeasible reason for b. I'll use the $>\!\!-$ symbol when I mean to refer to some abstract non-monotonic rule, whether it is a connective or a meta-linguistic relation, and whatever its particular formal behavior turns out to be.

Most have noticed that something like a "specificity defeater" is needed for non-monotonic inference systems. If $e_1 \wedge e_2$ is evidence, and there are non-monotonic rules $e_1 >\!\!- h$ and $e_1 \wedge e_2 >\!\!- \neg h$, then infer $\neg h$. The latter inferential connection defeats the former. In the simple choice between $e_1 >\!\!- h$ and $e_1 \wedge e_2 >\!\!- \neg h$, the latter rule is both more specific and appeals to more evidence.

Specificity can be distinguished from superior evidence. Specificity has to do with whether the antecedents of the non-monotonic rules in one argument are more specific than the antecedents of the rules in another argument. Usually, we say that $g_1 >\!\!- h$ is more specific than $g_2 \wedge g_3 >\!\!- \neg h$ if g_1 entails $g_2 \wedge g_3$. Evidence has to do with the monotonically derived premises that are used in non-monotonic arguments. We say an argument uses more evidence than another if the monotonically derived premises of the first entail those of the second. This is irrespective of the form of the arguments themselves.

Consider the following conflicting chains of reasoning: superior evidence can hold when there is not superior specificity. Let

$e_1 \wedge e_2 >\!\!- g_1$; $\quad g_1 >\!\!- h$;

$e_1 >\!\!- g_2 \wedge g_3$; $\quad g_2 \wedge g_3 >\!\!- \neg h$;

and let $e_1 \wedge e_2$ be evidence. The choice to chain through the first two non-monotonic rules relies on more evidence. But on the respective last steps of the chains, where the chains are led to contrary conclusions, the rule in the first chain does not have a more specific antecedent than the rule in the second chain. We don't know that g_1 entails $g_2 \wedge g_3$.

Conversely, superior specificity can hold when there is not superior evidence. Let

$e >\!\!- g_1 \wedge g_2$; $\quad g_1 >\!\!- h$;

$e >\!\!- g_1 \wedge g_2$; $\quad g_1 \wedge g_2 >\!\!- \neg h$;

where e is the evidence. In the chain that leads to $\neg h$, there is more specificity in the antecedent of the critical inference. But of the two chains, neither uses more evidence.

There is another defeater based on the directness of the non-monotonic argument from evidence to conclusion. Among the two chains below, where the evidence is e, the former is more direct.

$e >\!- g_1;$ $\quad g_1 >\!- h;$

$e >\!- g_1;$ $\quad g_1 >\!- g_2;$ $\quad g_2 >\!- \neg h.$

Among the next two chains, neither is more direct:

$e >\!- g_1;$ $\quad g_1 >\!- h;$

$e >\!- g_2;$ $\quad g_2 >\!- g_3;$ $\quad g_3 >\!- \neg h.$

Directness relies on there being a subset of intermediary conclusions. Unlike a "shortest path" rule, directness does not hold just because there are fewer intermediary conclusions.

The last defeater allows non-monotonic arguments to be chained. If one conclusion is preferred to another, then other things being equal, conclusions based on the first should be preferred to conclusions based on the second. Among the two chains below, where the evidence is e, the former makes use of a preferred intermediary proposition, or "preferred premise"; we decided above that h_1 is preferred to $\neg h_1$. So h_2 is preferred to $\neg h_2$.

$e >\!- g_1;$ $\quad g_1 >\!- h_1;$ $\quad\quad\quad h_1 >\!- h_2$

$e >\!- g_1;$ $\quad g_1 >\!- g_2;$ $\quad g_2 >\!- \neg h_1;$ $\quad \neg h_1 >\!- \neg h_2.$

In the logic, using more evidence is better than having more specificity, being more direct, or having a preferred premise. When one chain uses more evidence and another has more specificity, the one with more evidence prevails. When directness stands toe-to-toe with superior evidence, evidence always emerges victorious. The same is true of arguments with preferred premises that challenge arguments with more evidence. When any two of directness, specificity, and preferred premise compete, neither prevails; the arguments interfere with each other.

I will not attempt a long justification of these defeaters here. I think they are based on intuitions that clearly exist, though the intuitions may not be clear themselves. There may even be more defeaters that some would like to include. It should be obvious how to integrate them in what follows.

We should bring as much evidence to bear, in our arguments. Thus, we have the evidence and specificity defeaters. We argue that Opus doesn't traverse vast territory because being a Penguin is a good reason for not flying, which is a good reason for not traversing vast territory. This argument is superior to the counter-argument that being a bird is a good reason for being able to traverse vast territory, even though this argument is more direct.

Next, we have the directness defeater because conclusions should be tied closely to the evidence. Suppose we argue that Garfield doesn't like people because Garfield is a cat, and cats are generally aloof to people, and aloofness is an indicator of dislike. We expect to be refuted by the argument that cats generally like people.

Finally, we should, within reason, be prepared to take the conclusions of our arguments and use them in further argumentation. Suppose the only reason to conclude that Opus can spend the season in Sydney is that there's an argument that he can traverse vast territory, which is reason for possible seasonal saucing in Sydney. But as we saw, there's a better argument that Opus can't traverse vast territory. And the inability to traverse vast territory is reason for not being in Sydney. So the argument that Opus cannot spend the season in Sydney is superior.

2. Formalism.

We will write defeasible rules as assertions in the meta-language. $\mathrel{|\!\sim}$ is going to be an infix, two-place, meta-linguistic relation. Symbolically, it has the same status as $\vdash$. Meta-linguistic assertions involving this relation are supposed to be supplied by the user. So this system requires that knowledge be supplied in both an object language and a meta-language.

L is a language defined as usual. Call the sentences of L, $\mathbf{Sn}_L$.

$\Phi \mathrel{|\!\sim} \Psi$ reads "Φ in the absence of defeaters, is reason for Ψ,"
or just "Φ is a defeasible reason for Ψ".
$\Psi, \Phi \in \mathrm{Sn}_L$. Φ is the antecedent of the rule. Ψ is the consequent of the rule.
Sentences of this form belong to the meta-linguistic class: $\mathbf{D\text{-}rules}_{ML}$.

Do not suppose that there are any interesting rules that govern $\mathrel{|\!\sim}$. It won't be transitive, it won't be left-adjunctive (it's not the case that "A" $\mathrel{|\!\sim}$ "C" entails "A $\wedge$ B" $\mathrel{|\!\sim}$ "C") or right-disjunctive (it's not the case that "A" $\mathrel{|\!\sim}$ "C" entails "A" $\mathrel{|\!\sim}$ "B $\vee$ C"). They are supplied externally; with the following exceptions: if $\Phi \vdash \Gamma$ and $\Gamma \vdash$

Φ (i.e., are logically equivalent) and if $\Psi \vdash \Omega$ and $\Omega \vdash \Psi$ (i.e., are logically equivalent) and $\Phi \mathrel{|\!\!\sim} \Psi$, then $\Gamma \mathrel{|\!\!\sim} \Omega$.

The logic of sentences with the new relation is not claimed to analyze a locution such as "if A, then subjunctively conditionally, B", or "if A, evidently B", or "A is a prima facie reason for B". Rather, the new relation and its logic are axiomatic and are supposed to be useful for knowledge representation, and non-monotonic inference therefrom.

A **database** is any pair $<EK, R>$ where
$EK \subseteq Sn_L$, the "**evidential knowledge**," is supplied; and
$R \subseteq$ D-rules$_{ML}$, the set of "**defeasible rules**," which must also be supplied.

For each database, we define a defeasible extension,
$DK_{(<EK, R>)} \subseteq Sn_L$, the "**defeasible knowledge.**"
We leave off the subscript when it is unambiguous what database it extends.

Eventually, we'll define the membership of DK in terms of EK and R. DK will not be monotonic with repect to monotonic growth of EK or R.

DK and EK taken together are supposed to contain knowledge for subsequent action and practical deliberation. EK is assumed consistent with respect to $\vdash$.

The goal is to define the set of non-monotonic conclusions:

$P \in DK_{(<EK, R>)}$ iff
some argument for P has no good counter-arguments . . .,

where P is in Sn_L. The present concern, therefore, is defining what kinds of defeat and arguments there are.

I'll use single quotes when asserting that a sentence belongs to R, e.g., ' "$A \wedge B$" $\mathrel{|\!\!\sim}$ "$C \vee D$" ' $\in$ R. That just says that "$A \wedge B$" $\mathrel{|\!\!\sim}$ "$C \vee D$" is a defeasible rule supplied. R is closed under instantiation of open variables. This is just a representational shortcut. If R contains $\ulcorner Px \urcorner \mathrel{|\!\!\sim} \ulcorner Qx \urcorner$, where $x \in Var_L$, then R contains $\ulcorner Pa \urcorner \mathrel{|\!\!\sim} \ulcorner Qa \urcorner$, where $a \in Term_L$.

Conjoin(Φ) is the sentence obtained by conjoining all the elements of Φ.

$\Phi \mathrel{|\!\!\sim}_R \Psi$ holds just in case
$\Phi \in Sn_L$ and $\ulcorner \Phi \mathrel{|\!\!\sim} \Psi \urcorner \in R$, or
$\Phi \subseteq Sn_L$ and $\ulcorner$Conjoin(Φ) $\mathrel{|\!\!\sim} \Psi \urcorner \in R$.

$\Phi \; {}_{N}\vdash \Psi$ iff
$\Phi \vdash \Psi$ without redundancy or inconsistency, i.e.,
Φ is consistent and for no proper subset, ξ of Φ, $\xi \vdash \Psi$.

Consider connected, acyclic digraphs with a unique sink,
with nodes labeled by sentences, where no two nodes have the same label.
Let **nl**(P) be the *node labeled* P, $P \in Sn_L$, and
Label(n) $\in Sn_L$, be the label of node n.
(Note that I use capital letters for entities that are subsets or members of Sn_L).
An **internal node** is a node which is neither a source, nor the sink.
A **radius** is a path from a source to the sink that contains no cycle.

The **Sources** of a graph G, Sources(G) is the set of sentences
{Label(m) : m is a source of G}.
The **Support** of a node n of a graph G, Support(n), is the set
{Label(m) : <m, n> is an edge in G}.
In figure 1a, Support(nl("A")) = { "B", "C" }.

G, e.g. figure 1a, is an **argument for** P (in <EK, R>) iff G
is such a graph, with sink labeled P; and
Conjoin({Label(n) : n is a node in G}) is consistent with EK; and
support corresponds to a defeasible rule or monotonic entailment, i.e.,
for all non-source nodes x, either
a. Support(x) ${}_{R}\!\!\rightarrow$ Label(x) or
b. Support(x) ${}_{N}\vdash$ Label(x);
and sources correspond to non-redundant evidence, i.e.,
if s is a source of G, then EK $\vdash$ Label(s) and
for every <s, n> an edge of G, either
a. Support(n) ${}_{R}\!\!\rightarrow$ Label(n) or
b. there's no ξ s.t. Label(s) $\vdash \xi$ and not[$\xi \vdash$ Label(s)] and
$\{\xi\} \cup$ [Support(n) – Label(s)] $\vdash$ Label(n).

Note that if EK $\vdash$ P, the single node labeled P is an argument for P.
For any $P \in Sn_L$, there are potentially many arguments for P in <EK, R>.

G is an **argument** iff for some P, G is an argument for P.

I'll define some relations among arguments, which will be used as means of defeat. G_1 and G_2 are arguments, for arbitrary propositions.

Suppose:
"D" ⊢ "B ∧ E";
"B ∧ C" ⊢ "A";
EK ⊢ "C"; EK ⊢ "D";
EK ⊢ "F".

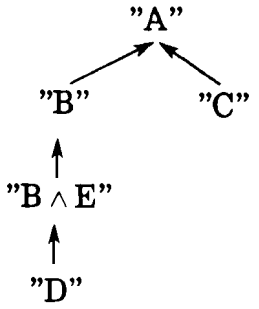
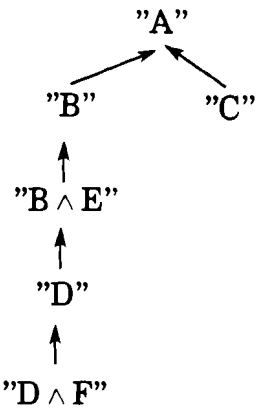
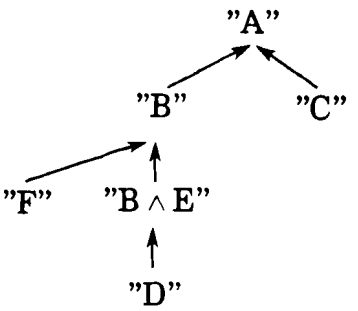


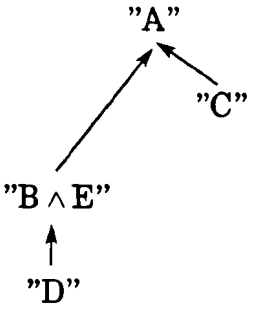
fig. 1a. An argument for "A".



nl("D ∧ F") is a source, but uses unnecessary evidence; i.e. "D ∧ F" entails "D" which by itself entails the label of the node it supports.



nl("F") is redundant in the link between nl("B ∧ E") and nl("B").



It is not the case that "B ∧ E", "C" are R⊢ - related or ⊢ - related to "A".

figs. 1b. Some graphs that are not arguments for "A".

G_1 uses **as much evidence** as G_2 iff
Sources(G_1) is consistent and
for every P ∈ Sources(G_2), Sources(G_1) ⊢ P.

G_1 uses **more evidence** than G_2 iff
G_1 uses as much evidence as G_2 and
it is not the case that G_2 uses as much evidence as G_1.

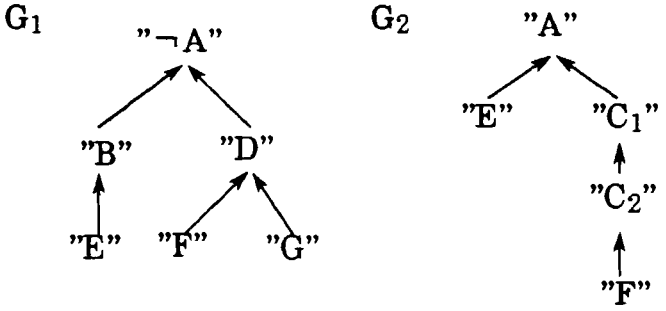


fig. 2. G_1 uses more evidence than G_2.

G_1 is **as specific** as G_2 iff
there is a node n_1 in G_1 and a node n_2 in G_2 s.t.
Label(n_1) and Label(n_2) are inconsistent and
Support (n_1) $\vdash$ Support (n_2).

G_1 is **more specific** than G_2 iff
G_1 is as specific as G_2 and
it is not the case that G_2 is as specific as G_1.

path l_1 is **as d-short** as path l_2 iff
$\vdash$ ⌜Label(origin(l_1)) ≡ Label(origin(l_2))⌝ and
$\vdash$ ⌜Label(destination(l_1))) ≡ Label(destination(l_2))⌝
(i.e., are logically equivalent); and
for every edge in l_1, <nl(P), nl(Q)>,
there is a path, l_3, s.t. l_3 is a sub-path of l_2 and
$\vdash$ ⌜P ≡ Label(origin(l_3))⌝ and
$\vdash$ ⌜Q ≡ Label((destination(l_3))⌝.

l_1 is **d-shorter than** l_2 iff
l_1 is as d-short as l_2 and it is not the case that l_2 is as d-short as l_1.

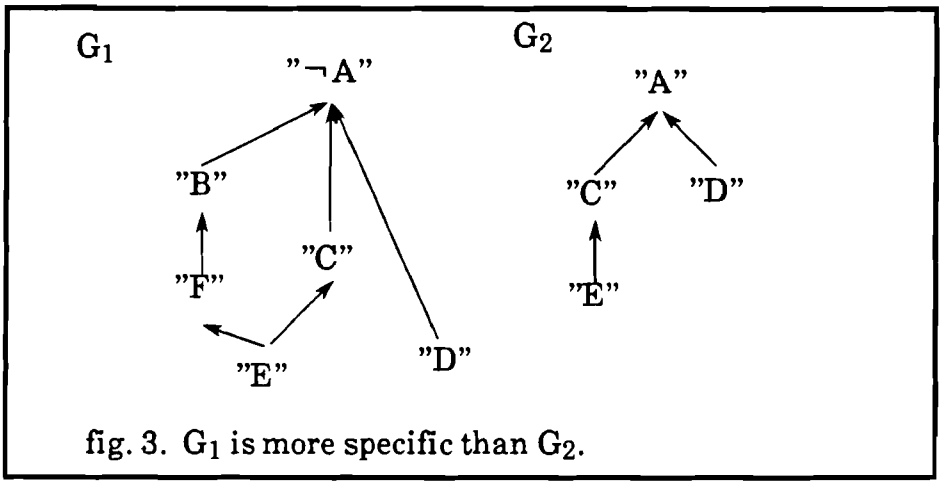


fig. 3. G_1 is more specific than G_2.

The path **negate-end-of**(l) is the path just like l, except for the last node, which has been replaced with a node labeled by its negation.
So if l is $<nl(p_1), \ldots, nl(p_k)>$,
then negate-end-of(l) is $<nl(p_1), \ldots, nl(\ulcorner \neg p_k \urcorner)>$.

G_1 is **as direct** as G_2 iff
for some radii, l_1, in G_1, and l_2, in G_2,
negate-end-of(l_1) is d-shorter than l_2.

G_1 is **more direct** than G_2 iff
G_1 is as direct as G_2 and
it is not the case that G_2 is as direct as G_1.

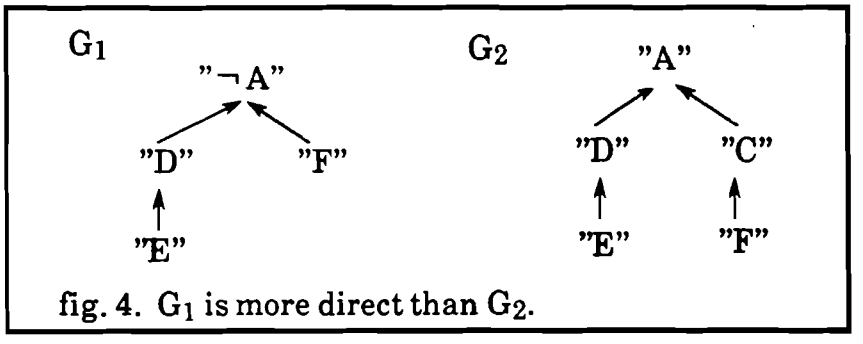


fig. 4. G_1 is more direct than G_2.

The next relation mentions defeat, which is defined below. It should be non-circular since it refers to proper sub-graphs of the original graphs.

G_1 **has a preferred premise**, compared to G_2 iff
there exist some non-sink nodes: n_2 in G_2 and n_1 in G_1, s.t.
for every proper sub-graph S_2 of G_2 that is an argument for n_2
there is a proper sub-graph S_1 of G_1 that is an argument for n_1 s.t.
S_1 defeats S_2.

G_1 **has preferred premises**, compared to G_2 iff
G_1 has a preferred premise, compared to G_2 and
it is not the case that G_2 has a preferred premise, compared to G_1.

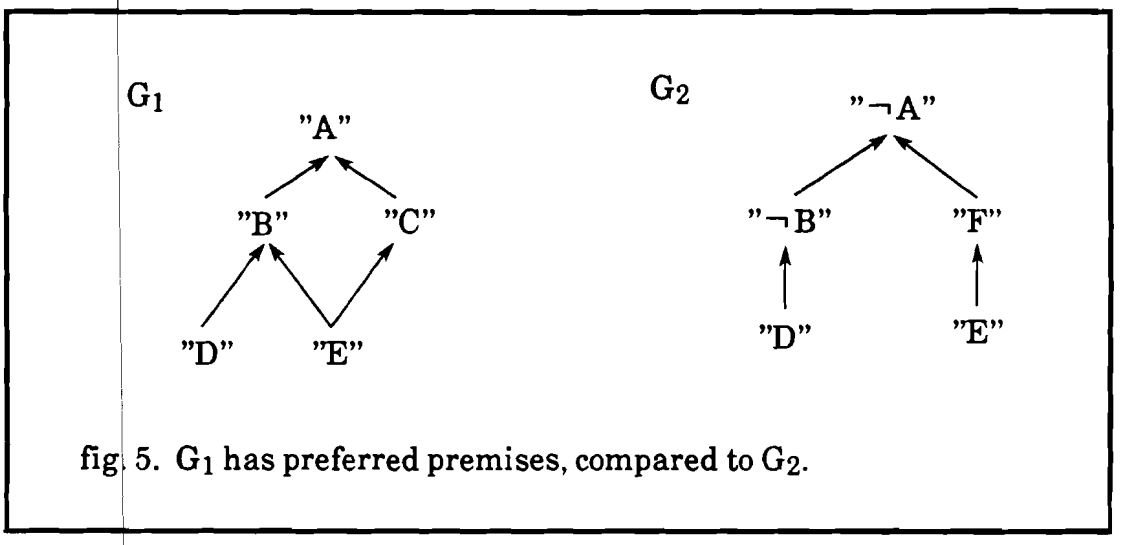


fig. 5. G_1 has preferred premises, compared to G_2.

Argument G_1 **interferes with** argument G_2 iff
Label(sink(G_1)) and Label(sink(G_2)) are inconsistent and either
1. G_1 uses more evidence than G_2; or
2. It is not the case that G_2 uses more evidence than G_1
(note the reversed order) and
a. G_1 is more specific than G_2; or
b. G_1 is more direct than G_2; or
c. G_1 has preferred premises, compared to G_2.

Argument G_1 **defeats** argument G_2 iff
G_1 interferes with G_2 and it is not the case that G_2 interferes with G_1.

Argument G is **undefeated** (in <EK, R>) iff
There is no argument, G' (in <EK, R>) s.t. G' defeats G.

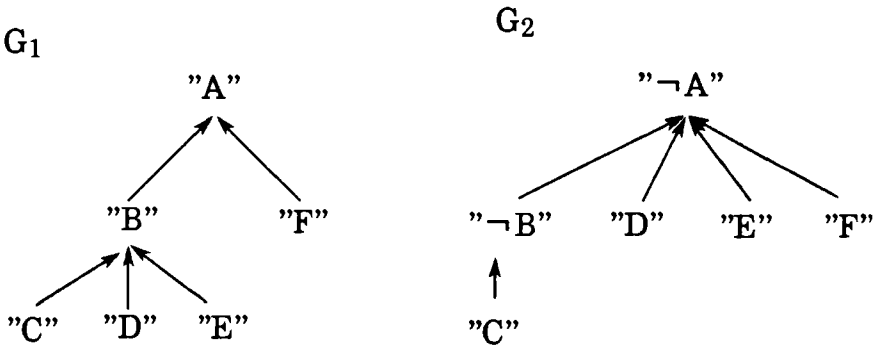


fig. 6. G_1 interferes with G_2 because of the preferred premise, "B" (preferred because of more evidence). G_2 interferes with G_1 because of directness. So neither defeats the other.

G_1 is a **counter-argument** of G_2 iff
for some node, n, in G_2, Label(sink(G_1)) and Label(n) are inconsistent.

G **justifies** P (in $<EK, R>$) iff
G is an undefeated argument for P and
for every counter-argument of G, there is an argument G' s.t. G' defeats G.

P is **justified** (in $<EK, R>$) iff
for some G, G justifies P.

$P \in DK_{<EK, R>}$ iff
P is justified and
there is no set $S \subseteq Sn_L$ s.t. S is inconsistent and each member of S is justified.

3. Observations.

The definitions proposed are supposed to be only suggestive, not legislative. So I won't boast any theorems; I'm more concerned with the basic ideas. However, here are a few observations and conjectures.

The first have to do with the form of arguments. Roughly, the idea is that syntactic variations don't matter.

Obsv. Choices between logically equivalent but syntactically different node labels are moot.

Consider that none of the relations is sensitive to such syntactic variations.

Argument G_1 **is at least as strong as** G_2 iff
the set of arguments that G_1 defeats is a superset of the arguments that G_2 defeats.

Argument G' **is monotonically smaller than** G iff

1) G' can be obtained from G by deleting an internal node n, and adding edge $<m_1, m_2>$ for every pair of edges $<m_1, n>$ and $<n, m_2>$ in G; or
2) G' can be obtained from G by deleting a node n from G, finding a node n' s.t. Label(n) is logically equivalent to Label(n'), and adding edges $<m_1, n'>$ and $<n', m_2>$, respectively, for every pair of edges $<m_1, n>$ and $<n, m_2>$ in G; or
3) G' can be obtained from G by removing other such nonsense.

Conjct. If G' is monotonically smaller than G, then G' is at least as strong as G.

The next has to do with the properties of the relations among arguments.

Obsv. Defeat is incomplete and anti-symmetric but not acyclic, hence not transitive.

Incompleteness and anti-symmetry are obvious. Here's a cycle:
Consider arguments G_1, G_2, and G_3, based respectively on

$D_1 = \{"E_1 \wedge E_2" \vdash "B"; "E_5" \vdash "\neg D"; "B \wedge \neg D" \vdash "A_1"\}$;
$D_2 = \{"E_1" \vdash "\neg B"; "E_3 \wedge E_4" \vdash "C"; "\neg B \wedge C" \vdash "A_2"\}$; and
$D_3 = \{"E_3" \vdash "\neg C"; "E_5 \wedge E_6" \vdash "D"; "\neg C \wedge D" \vdash "A_3"\}$;

where the E_i are all in EK.
A_1, A_2, and A_3 are collectively inconsistent. Because of the preferred premise in each, G_1 defeats G_2, G_2 defeats G_3, and G_3 defeats G_1.

The last and most important have to do with the properties of EK and DK.

Obsv. If $P \notin EK$ then $P \in DK$.

The argument consisting of P itself cannot be defeated. (Note this would not be true if "$P \wedge Q$" $\vdash$ "$\neg P$" could be used as an argument, in which case more evidence could defeat pure evidence!).

Obsv. If G_P justifies P then for every node n in G_P, $\ulcorner \neg$Label(n)$\urcorner \notin DK$.

Suppose not. Then some undefeated G justifies $\ulcorner \neg$Label(n)$\urcorner$. But that

means that G is an undefeated counter-argument of G_P; hence G_P does not justify P. This is a contradiction.

Obsv. G_P can justify P even if for some node n in G_P, Label(n) $\notin$ DK.

Let G_P be the argument based on D_P = {"E_1" |~ "Q"; "$E_2 \wedge Q$" |~ "P"}, where E_1 and E_2 are in EK. There is a counter-argument, $G_{\neg Q}$, based on $D_{\neg Q}$ = {"$E_1 \wedge E_2$" |~ "A"; "A" |~ "$\neg$Q"}. It doesn't defeat G_P; suppose G_P is undefeated. Suppose it doesn't prevent P from being in DK because it in turn is defeated by G_A, which is based on $D_{\neg A}$ = {"$E_1 \wedge E_2 \wedge E_3$" |~ "$\neg$A"}. Now the argument for Q is G_Q, which is just "E_1" |~ "Q". But it is defeated by $G_{\neg Q}$. Hence, G_Q does not justify Q. There need not be any other arguments for Q, in which case, Q would be excluded from DK.

Obsv. DK is not strongly closed; i.e., if DK $\vdash$ Q, still Q might not be in DK.

Suppose "E_1" |~ "A" and "E_1" |~ "B" justify "A" and "B" respectively, via G and G_B. The argument for "A $\wedge$ B", G_{AB}, combines the two individual arguments, by adding a node nl("A $\wedge$ B") supported by nl("A") and nl("B"). There's an argument for "$\neg$(A $\wedge$ B)", G_{NOT}, based on "$E_1 \wedge E_2$" |~ "Q"; "Q" |~ "$\neg$(A $\wedge$ B)". It does not interfere with G_A or G_B, but it defeats G_{AB}. Of course, G_{NOT} had better not justify "$\neg$(A $\wedge$ B)" or else all of the justifying arguments are ineffective (because "A", "B", and "$\neg$(A $\wedge$ B)" are collectively inconsistent). So suppose there is an argument based on "E_3" |~ "$\neg$Q". This does not defeat G_{NOT}, so it adds nothing to the case for "A $\wedge$ B". But it does prevent G_{NOT} from being justifying. Hence, "A" $\in$ DK, "B" $\in$ DK, but neither "A $\wedge$ B" nor "$\neg$(A $\wedge$ B)" is in DK.

Obsv. DK is not even weakly closed; i.e., if P $\in$ DK and P $\vdash$ Q, still Q might not be in DK.

Let G_P be the linear graph based on "E_1" |~ "A" and "A" |~ "P". A counter-argument, G_{NOT}, which is simply "E_1" |~ "$\neg$Q", is defeated by the more complex argument, G_{QR}, based on "$E_1 \wedge E_2$" |~ "B" and "B" |~ "Q $\wedge$ R". So "P" is in DK. But is "Q"? "P" entails "Q", so G_Q could be G_P extended by the node labeled "Q". But G_{NOT} defeats G_Q on directness. There's another argument for "Q", which extends G_{QR} by a node labeled "Q". But it can be defeated by an argument based on "$E_1 \wedge E_2 \wedge E_3$" |~ "$\neg$B", without changing any of the above. So "Q" is not in DK.

Obsv. DK is strongly consistent; i.e., Conjoin(DK) is consistent.

Since this is guaranteed by the definition of DK, we could just define **Closed-DK**: P $\in$ Closed-DK iff P $\in$ DK or DK $\vdash$ P.

4. Comparison with Poole.

This work is closely related to David Poole's work [Poole85] in terms of goals and intuitions. It was, however, developed independently of his work. At the present time, it seems that a superior system would share some of his ideas (e.g., paying close attention to the actual rules used in an argument) and some of mine (e.g., paying close attention to the structure of the argument).

Poole's system is much simpler.

His rules look like this:

1. Find two non-monotonic conclusions to choose between: h_1 and h_2.
   e.g., h_1 = "flies(edna)"; h_2 = "$\neg$flies(edna)".
2. Find two sets of defeasible rules, two *theories*. Together with the evidence, the two theories *allow* h_1 and h_2 respectively, via (just what anyone would expect) monotonic inference and non-monotonic modus ponens.
   e.g., D_1 = {bird(x) >– flies(x)}; D_2 = {emu(x) >– $\neg$flies(x)},
   given "emu(edna)" and "bird(x) $\vee$ $\neg$emu(x)".
3. Find an assertion, p_1 (which does not monotonically entail h_1), s.t. D_1 is solely *applicable*, i.e.,
   a. D_1, the "necessary" facts, and p_1 allow h_1.
   b. D_2, the "necessary" facts, and p_1 do not allow h_1 and do not allow h_2.
   e.g., p_1 = "bird(edna)".
4. Fail to find an assertion, p_2 (which does not monotonically entail h_2), s.t. D_2 is solely applicable, i.e.,
   a. D_2, the "necessary" facts, and p_2 allow g_2.
   b. D_1, the "necessary" facts, and p_2 do not allow g_2 and do not allow g_1.
   e.g., the only p_2 satisfying 4a is "emu(edna)". But "bird(x) $\vee$ $\neg$emu(x)" is a necessary fact. Thus, D_1, p_2, and the necessary facts allow "bird(edna)", hence "flies(edna)". 4b is violated. Therefore, 4 is satisfied.
5. Declare that D_2 is better than D_1 and therefore g_2 is preferred to g_1.

My basic concepts of more evidence, directness, and specificity are all implied by Poole's rules, for simple cases.

1. More evidence: Contingently, e_1 and e_2. $D_1 = \{e_1 >\!\!- h\}$. $D_2 = \{e_1 \wedge e_2 >\!\!- \neg h\}$. e_1 makes D_1 applicable, but not D_2. Anything entailing $e_1 \wedge e_2$ makes D_2 applicable, must also entail e_1, thus makes D_1 applicable. So $\neg h$ is preferred.

2. Directness: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- \neg h\}$. g_1 makes D_1 applicable, but not D_2. Anything that makes D_2 applicable must entail e_1, which makes D_1 applicable. So $\neg h$ is preferred.

3. Specificity: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- g_1; e_1 >\!\!- g_2; g_1 \wedge g_2 >\!\!- \neg h\}$. g_1 makes D_1 applicable, but not D_2. Anything that makes D_2 applicable entails e_1, or entails $g_1 \wedge g_2$; in either case, D_1 is also made applicable. So $\neg h$ is preferred.

But there are more complicated examples whereupon we disagree.

4. Redundant Defeasible Connections: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; e_1 >\!\!- g_2; e_1 >\!\!- g_3; g_2 \wedge g_3 >\!\!- g_1; g_1 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- g_2; e_1 >\!\!- g_3; g_2 \wedge g_3 >\!\!- \neg h\}$. g_1 makes D_1 applicable, but not D_2. Anything that makes D_2 applicable entails e_1, or entails $g_2 \wedge g_3$; in either case, D_1 is also made applicable. So $\neg h$ is Poole-preferred. D_1 permits the argument $\{e_1 >\!\!- g_1; g_1 >\!\!- h\}$. Between this argument and the one in D_2, neither is defeating, so neither is preferred by my system.

5. Cyclic Redundancies: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- e_1; g_1 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- g_1; g_1 >\!\!- e_1; e_1 \wedge g_1 >\!\!- \neg h\}$. I prefer D_2 and $\neg h$ on specificity. Poole can't choose either, since D_1 and D_2 are applicable at the same times.

6. Directness: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- g_2; g_2 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- g_2 \wedge g_3; g_2 \wedge g_3 >\!\!- \neg h\}$. g_2 makes D_1 applicable, but not D_2. Anything that makes D_2 applicable must yield g_2, which makes D_1 applicable. So $\neg h$ is Poole-preferred. But I don't consider this directness because the path $<e_1, g_2 \wedge g_3, \neg h>$ is not d-shorter than $<e_1, g_1, g_2, h>$. $<e_1, g_2, \neg h>$ would have been d-shorter. But there is a difference between g_2 and $g_2 \wedge g_3$. Neither is preferred by my system.

Poole could amend his rules so that he considers only minimal sets of defaults, i.e., those sets that allow their conclusion, each of which has no subset that allows the conclusion. Then we would not differ over cases 4 and 5.

I should probably yield over case 6. But it seems that $D_2' = \{e_1 >\!\!- g_2; g_2 >\!\!- \neg h\}$ ought to be the defeater of D_1. Certainly $D_2'' = \{e_1 >\!\!- g_2; e_1 >\!\!- g_2; g_2 \wedge g_3 >\!\!-$

$\neg h\}$ ought not to defeat D_1. Is D_2 more like D_2' or D_2''? It's plausible that $e_1 >\!\!- g_2 \wedge g_3$ should lead to $e_1 >\!\!- g_2$. But it is clear that $g_2 \wedge g_3 >\!\!- h$ shouldn't lead to $g_2 >\!\!- h$. So D_2 isn't as strong as D_2'. What about $D_2''' = \{e_1 >\!\!- g_2 \wedge g_3; g_2 >\!\!- \neg h\}$? That leads to a better argument, and my definition of d-shorter might be revised to allow it to defeat D_1. My definition stands because I want to draw attention to those places where hard choices need to be made.

Poole and I further diverge over the primacy of evidence compared to specificity and directness.

7. Evidence versus Directness: Contingently, e_1 and e_2. $D_1 = \{e_1 >\!\!- g_2; g_2 >\!\!- h\}$. $D_2 = \{e_1 \wedge e_2 >\!\!- D; D >\!\!- \neg h\}$. g_2 makes D_1 applicable, but not D_2. D makes D_2 applicable but not D_1. Poole won't choose between these, but I'll take D_2 and $\neg h$.

8. Evidence versus Specificity: Contingently, e_1 and e_2. $D_1 = \{e_1 >\!\!- g_2 \wedge g_3 ; g_2 \wedge g_3 >\!\!- h\}$. $D_2 = \{e_1 \wedge e_2 >\!\!- g_3; g_3 >\!\!- \neg h\}$. e_1 makes D_1 applicable, but not D_2. g_3 makes D_2 applicable but not D_1. Poole again won't choose between these; again, I'll take D_2 and $\neg h$.

Most importantly, we disagree over chaining. I have the preferred premises defeater. Poole is more cautious.

9. Chaining: Contingently, e_1. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- g_2; g_2 >\!\!- h\}$. $D_2 = \{e_1 >\!\!- \neg g_2; \neg g_2 >\!\!- \neg h\}$. Both Poole and I prefer $\neg g_2$ to g_2. g_2 makes D_1 solely applicable. $\neg g_2$ makes D_2 solely applicable. I will allow $\neg h$, but Poole will abstain.

Poole could iterate his theory-selection mechanism. At each stage, the non-monotonic conclusions would be treated as evidence for the next iteration of theory-selection. Then he would effectively get my evidence defeater and something very much like the preferred premises defeater. There would be complications. We would still differ.

He could also consider extensions of more specific theories to be preferred to extensions of less specific theories, other things being equal. Again, there would be complications.

Poole says that his system is sometimes counterintuitive. He considers the situation:

10. Closure: Contingently e_1 and e_2. $D_1 = \{e_1 >\!\!- g_1; g_1 >\!\!- h_1; e_2 >\!\!- g_2; g_2 >\!\!- h_2\}$. $D_2 = \{e_1 >\!\!- \neg h_1; e_2 >\!\!- \neg h_2\}$. We also know that $h_1, h_2 \vdash h_3$ and $\neg h_1, \neg h_2 \vdash \neg h_3$. He says he can infer $\neg h_1$ and $\neg h_2$, but not $\neg h_3$. I can infer all three, since I can consider the argument for h_3.

I get the conclusion Poole finds intuitive. Poole could again make amendments. He could restrict p_1 and p_2 in steps 3 and 4 of his rules (above) to antecedents of defeasible rules.

4. Conclusion.

I don't pretend that my definitions of directness, specificity, and preferred premises are now a writ on biblical stone, or that this paper constitutes the last word on the subject of formalizing defeasible inference. Looking at the definitions of directness and specificity, and of interference, it's clear that there are many plausible alternatives (for instance, in figure 6, should G_2 be more specific than G_1; i.e., should "$\neg$B" be treated like "B" for these purposes?).

In fact, I have given a detailed comparison of my system to David Poole's system because I think dialogue should be opened on the subject. My conviction suffers whenever our systems disagree (e.g., when we differ over directness in case 6), and I expect that his does as well (esp. in cases 4 and 5). The same should be true of the others who have attempted similar systems, including Touretzky, Nute, and Sandewall.

It seems a better treatment would start with Poole's idea of a theory, and then construct canonical graphs from theories. This would simplify the definition of an argument and make possible direct comparison of the defeasible rules actually used in the argument.

I do expect that the present system will be a benchmark for the adequacy of future systems of defeasible inference, and will lead to useful programs of defeasible inference in practice.

Appendix 1. Relativization to Limited Computation.

It's very easy to relativize this logic to some set of arguments that are computable or salient. This relativization is required in implementation. Let Ω be the computed arguments, e.g., by forward chaining. Then in the definitions, just restrict attention to those arguments in Ω. For example,

G justifies P (in Ω) iff
G is an argument for P, undefeated in Ω, and
for every counter-argument of G in Ω, there is an argument in Ω,G',
s.t. G' defeats G.

Another natural relativization is to the set of inferences that are performed, rather than to the full obligations of $\vdash$.

These relativizations make inferences non-monotonic not only in evidence and rules, but also in computation.

Appendix 2. Non-Supporting Interference Relations.

Until now, all defeat has come from conflicting consequents. It was not possible to write explicit defeat of defeasible rules. If $\mid\!\sim$ had been a connective, we could have written "$A \mid\!\sim \neg(B \mid\!\sim C)$" or "$A \supset \neg(B \mid\!\sim C)$" directly in Sn_L (the latter uses the standard material, truth-functional connective $\supset$). At present, the only way "A" can defeat "B" $\mid\!\sim$ "C" is if "A" is a reason to infer "$\neg$C", defeasibly or otherwise.

But sometimes in the presence of "A", the connection between "B" and "C" is simply defeated, and no conflicting alternative is suggested.

McCarthy's way of solving this is to tag rules with their exceptions, as conjuncts in the consequents. Instead of "B" $\mid\!\sim$ "C", write "B" $\mid\!\sim$ "$C \wedge \neg Ab_A$"; then write "$A \supset Ab_A$", which now conflicts with the defeasible rule. The drawback of McCarthy's method is that rules in R will have to be modified every time an exception is added. If one of the above forms could be used, these exceptions could simply be added to EK, monotonically, without threat of subsequent revision.

Also, there is still no way to defeat a chain of non-monotonic reasoning without defeating one of the individual links. Suppose "A" is reason for "B" and "B" is reason for "C", but "A" is not reason for "C". How do we represent this?

The best way to take care of these situations in this logic is to follow Nute's system. Let there be a negating-version of $\mid\!\!\sim$, $\times\!\!\sim$. This relation can be used to construct invalidating arguments, but cannot be used to construct supporting arguments. $R\times\!\!\sim$ is just like $R\mid\!\!\sim$, with the obvious substitutions. There are now two kinds of arguments. One kind uses only $R\mid\!\!\sim$ and $\vdash$. The other kind uses one $R\times\!\!\sim$, in combination with any number of $R\mid\!\!\sim$. Call the two kinds of arguments, respectively, **supporting and interfering arguments.** An argument is a supporting argument or an interfering argument. The relations among arguments (e.g. more specific than) remain the same. The definition of justification becomes:

G justifies P (in <EK, R>) iff
G is an undefeated supporting argument for P and
for every counter-argument of G, there is an argument G' s.t. G' defeats G.

Now we can defeat the connection between "A" and "C" in the chain by asserting "A" $\times\!\!\sim$ "$\neg$C". This doesn't interfere with the connections "A" $\mid\!\!\sim$ "B" and "B" $\mid\!\!\sim$ "C". It can interfere with arguments for "C", but can't be used to argue for "$\neg$C".

As for defeating the connection between "B" and "C" in the presence of "A", we can assert "A $\wedge$ B" $\times\!\!\sim$ "$\neg$C".

Appendix 3. Defeasibility at the Meta-Level.

What about something like "$\mid\!\!\sim$(B $\mid\!\!\sim$ C)" and "A $\mid\!\!\sim$ $\neg$(B $\mid\!\!\sim$ C)"? These would be simple to write if $\mid\!\!\sim$ had been a connective. But it's easier to see what's going on meta-linguistically.

We can countenance complex assertions about membership in R. For "A" to defeat the connection between "B" and "C", write

if '$\vdash$ "A" ' is true *then* '"B" $\mid\!\!\sim$ "C"' $\notin$ R.

This sentence must use the material conditional connective and the predicate "is true" from the meta-meta-language. This is because membership and non-membership in R already require the naming of sentences in the meta-language.

But these assertions are not defeasible. Sometimes we need a way to say that "B" $\mid\!\!\sim$ "C" is in R, defeasibly.

Consider a new relation in the meta-meta-language, which is to $\mid\!\!\sim$ as $\wedge$ is to &, i.e., it's just the meta-meta-linguistic analogue of the meta-lingusitic $\mid\!\!\sim$. Let's use the symbol $\overset{\text{Meta}}{\mid\!\!\sim}$ for this relation. We can define interference and defeat to govern $\overset{\text{Meta}}{\mid\!\!\sim}$, just as they were defined for $\mid\!\!\sim$. Then we write, respectively,

$\overset{\text{Meta}}{\mid\!\!\sim}$ ['"B" $\mid\!\!\sim$ "C"' $\in$ R], and
[' $\vdash$ "A" ' is true] $\overset{\text{Meta}}{\mid\!\!\sim}$ ['"B" $\mid\!\!\sim$ "C"' $\notin$ R],

and include them in the meta-level analogue of R, Meta-R.

The latter rule says: If "A", (i.e., "A" is in EK, i.e., $\vdash$ "A", i.e., ' $\vdash$ "A" ' is true) and if there is no undefeated way of getting "B" $\mid\!\!\sim$ "C" into R, then it isn't in R.

There would be problems if "A" in DK led to ' $\vdash$ "A" ' is true. We have to restrict what is meant by ' $\vdash$ "A" ' is true to "A" $\in$ EK, not "A" $\in$ EK $\cup$ DK.

Meta-linguistic defeasibility seems to be required for mimicking reasoning by cases. "B" $\mid\!\!\sim$ "C" and "A" $\mid\!\!\sim$ "C" may be in R, and "A $\vee$ B" in EK. Still, we don't get "C" in DK. We can't just allow reasoning by cases because "A $\vee$ B" $\mid\!\!\sim$ "$\neg$C" could also be in R. Instead, we defeasibly infer "A $\vee$ B" $\mid\!\!\sim$ "C" from:

['"A" $\mid\!\!\sim$ "C"' $\in$ R and '"B" $\mid\!\!\sim$ "C"' $\in$ R] $\overset{\text{Meta}}{\mid\!\!\sim}$ ['"A $\vee$ B" $\mid\!\!\sim$ "C"' $\in$ R].

Acknowledgements.

Thanks to the referees for looking past lapses in professionalism in the draft; the treatment was uneven in part because Poole's work was discovered after this work was in draft. Rich Pelavin fixed the recursive definitions. Henry Kyburg suggested the meta-linguistic relation. Thanks also to Henry Kautz and Dave Sher. David Israel also provided encouragement. I owe the typesetter a beer.

Bibliography.

Etherington, D. and Reiter, R. "On Inheritance Hierarchies with Exceptions," *Proceedings of AAAI-83*, 1983.

Kyburg, H. "The Reference Class," *Philosophy of Science 50*, 1983.

McCarthy, J. "Applications of Circumscription to Formalising Common Sense Knowledge," Proceedings of AAAI Workshop on Non-Monotonic Reasoning, New Paltz, NY, 1984.

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

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

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

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

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