

Corrigenda to Poole's Rules and A Lemma of Simari-Loui

R. Loui
K. Stiefvater
Department of Computer Science
Washington University
St. Louis¹

Abstract. This note corrects a lemma in the recent paper [1] of one of the authors by first correcting problems with Poole's rule for specificity of arguments. It also responds to the criticism of Touretzky, et al. [9]

1

As noted as Simari-Loui [1] went to press, the computationally most important lemma of the article is wrong. It states that checking specificity is equivalent to checking that the antecedents of the less specific theory can be derived from the antecedents of the more specific theory. This is not even true for the system restricted to Horn clause-like rules.

Getting this lemma right is crucial because computing Poole's specificity [2] as implied by the definitions is impossible. This is the issue that McDermott raised in [3]:

If this can really be made to work, then it is the only direct rebuttal
... However, it is hard to tell ... exactly how Poole's idea is to work.
... In the definitions ..., proving part (a) would require looking for
an arbitrary subset of facts Proving part (b) is even worse

and it is one reason why David Poole's THEORIST [4] does not attempt to order theories according to Poole's own rule (see also Prakken's implementational concerns in [5]).

The lemma states that a theory T_1 is more specific than T_2 just in case for every x , an antecedent of a rule used in T_2 , x can be defeasibly derived from K_N , the necessary evidence, T_2 's rules, and the antecedents of T_1 's rules:

$$(\forall x \in An(T_2))(K_N \cup An(T_1) \cup T_2 \vdash x).$$

In the case of

$$\langle \{Penguin(O) \multimap \neg Flies(O)\}, \neg Flies(O) \rangle$$

versus

$$\langle \{Bird(O) \multimap Flies(O)\}, Flies(O) \rangle,$$

¹Work supported by NSF R-9008012 and NSF CDA-9102090.

for example, $Bird(O)$ can be defeasibly derived from $Penguin(O)$. And in

$$\langle \{Cat(G) \succsim Aloof(G), \\ Aloof(G) \succsim \neg LikesPeople(G)\}, \neg LikesPeople(G) \rangle$$

versus

$$\langle \{Cat(G) \succsim LikesPeople(G)\}, LikesPeople(G) \rangle,$$

$Cat(G)$ defeasibly derives both $Cat(G)$ and $Aloof(G)$.

This last example, seen in reverse, is also a trivial counterexample to the lemma. If the lemma were right, then not only is the theory for $LikesPeople(G)$ more specific than the theory for $\neg LikesPeople(G)$, but also vice versa. Not only is this intuitively undesirable, but it also violates the antisymmetry of specificity. The required relationship between antecedents is necessary for specificity, but not sufficient.

Where the alleged proof goes wrong is quite easy to see. When there is specificity, every antecedent of the weaker theory can be derived from every antecedent of the stronger theory, but not necessarily from an activator of the stronger theory (an activator is a sentence of the proscribed contingent kinds, which allows a theory's conclusion to be derived, using the necessary evidence, K_N , and the theory's rules). That is, it may be possible to activate the theory without allowing *all* of its antecedents to be defeasibly derived, which is just plain to see.

Correcting this error requires first that the definition of specificity be repaired. As it stands, it is vulnerable to some counter-intuitive behavior. This counter-intuitive behavior plagues Poole's rule generally, not just our use of it.

The argument, for example,

$$\begin{array}{l} A \prec B \wedge C \\ \quad B \prec D \\ \quad C \prec E \end{array}$$

should be more specific than

$$\begin{array}{l} \neg A \prec B \\ \quad B \prec D. \end{array}$$

But it is not more specific according to the definition in the paper, because of two separate flaws in the definition.

The first reason is that $E \wedge (B \vee \neg C)$ allows the former theory to be activated without activating the latter. This prevents the desired conclusion that the former theory is more specific. What is basically wrong here is that the weaker theory should be allowed to use the defeasible rule $E \succsim C$, in order to derive (defeasibly) B .

To see the second flaw, consider $E \wedge (A \vee \neg C)$, which is also disjunctive, but this time uses the disjunction to derive the theory's ultimate conclusion. This

is essentially a side-stepping of the non-triviality condition for activators. The non-triviality condition should be strengthened.

Fixing the flaws in Poole's rule is important because this kind of comparison is the comparison used in the Yale Shooting Problem arguments [6], as exhibited among the examples in the paper by Simari-Loui.

The rule also suffers in an example reminiscent of Royal Elephants [7]: Consider

$$\begin{array}{l} D \prec B \wedge C \\ B \wedge C \prec A \end{array}$$

compared with what should be an inferior theory:

$$\begin{array}{l} \neg D \prec B \\ B \prec E. \end{array}$$

Neither is more specific by Poole's rule. The example appears to require right-weakening of rules: allowing rules to be derived from rules by weakening the consequent:

$$\text{If } A \succ B \wedge C, \text{ then } A \succ C;$$

then the theory

$$\begin{array}{l} D \prec B \wedge C \\ B \prec E \\ C \prec A \end{array}$$

would be the defeater of the weaker theory. The first theory does not defeat the third, but with right-weakening, it allows the construction of a third theory which directly accounts for the considerations used in the weaker theory, and which ought to be more specific.

But right-weakening is notoriously problematic. When there are competing arguments, such as

$$B \prec A$$

versus

$$\neg B \prec A,$$

a rule can be right-weakened with an arbitrary dilution, such as

$$A \succ B, \text{ therefore, } A \succ B \vee C,$$

allowing an argument for C . This cannot be done without right-weakening, since arguments must have consistent intermediate claims. This particular case is not so bad, since the argument

$$\begin{array}{l} C \prec (B \vee C) \wedge \neg B \\ B \vee C \prec A \\ \neg B \prec A \end{array}$$

has counterargument

$$B \multimap A,$$

but if $A \wedge D \multimap \neg B$, the arbitrary dilution can actually be supported. Instead, fix the rule for specificity so that it treats the problem properly without requiring right-weakening.

The proposed development for specificity, which fixes both kinds of counter-intuitive behavior due to disjunction, which allows proper treatment of the last example, and which allows a proper pruning lemma, is as follows.

A rule, R , is a *top rule* of argument $\langle T, h \rangle$ just in case its consequent, $Con(R)$, is not needed for the derivation of anything but the argument's conclusion. Because the rules used in arguments are a minimal set, that is equivalent to saying that the antecedent of any rule can be derived (from evidence) using rules other than this top rule.

Definition. **Top**($R, \langle T, h \rangle$) iff for every r in T , $An(r)$ can be defeasibly derived from $K_N \cup K_C$ using $T - \{R\}$.

Example. $B \multimap C$ is a top rule in the argument from A to C , using $A \multimap B$, and $B \multimap C$.

Let Δ be a set of defeasible rules, let h be a sentence in the language, L , and let A be a set of sentences in L .

A finite sequence of sentences, $\langle B_1, \dots, B_n \rangle$ is a *consistent defeasible derivation (CD-derivation)* of h from A using rules Δ just in case h is derived by A 's activating ground instances of rules, and the set of all intermediate sentences is consistent in L .

Definition. $\langle B_1, \dots, B_n \rangle$ is a **CD-derivation** iff

1. $B_n = h$;
2. For each B_i , either
 - a. $\{B_j \mid j < i\} \cup A \vdash B_i$; or
 - b. for some ground instance of a rule R in Δ ,
 $An(R) = B_j$ for some $j < i$
and $B_i = Con(R)$;
3. $\{B_i \mid i \leq n\} \not\vdash \perp$.

Example (continued). $\langle A, B, C \rangle$ is a CD-derivation of C from A using $A \multimap B$ and $B \multimap C$.

Definition. There is a **CD-derivation** of h from A using Δ with all **top rules** of $\langle T, h \rangle$ just in case

1. there is a CD-derivation of h from A using Δ ;
2. for each x such that $Top(x, \langle T, h \rangle)$, there is no CD-derivation of h from A using $\Delta - \{x\}$.

$\langle T_1, h_1 \rangle$ is more specific than $\langle T_2, h_2 \rangle$ just in case some legitimate sentence activates T_2 for h_2 without activating T_1 for h_1 , using CD-derivations from *the two theories' combined set of rules*, using *every top rule* of T_2 ; and there is no such *asymmetric activator* of T_1 for h_1 that does not also activate T_2 for h_2 . That is,

Definition. e is an **asymmetric activator** of $\langle T_1, h_1 \rangle$ **but not** $\langle T_2, h_2 \rangle$ just in case

1. there is some CD-derivation of h_2 from $K_N \cup \{e\}$ using $T_1 \cup T_2$ with all top rules of $\langle T_2, h_2 \rangle$;
2. there is no CD-derivation of h_1 from $K_N \cup \{e\}$ using $T_1 \cup T_2$ with all top rules of $\langle T_1, h_1 \rangle$.

Definition. $\langle T_1, h_1 \rangle$ is **more specific** than $\langle T_2, h_2 \rangle$ just in case

1. there is some asymmetric activator in S_C of $\langle T_2, h_2 \rangle$ but not of $\langle T_1, h_1 \rangle$, and
2. there is no asymmetric activator in S_C of $\langle T_1, h_1 \rangle$ but not $\langle T_2, h_2 \rangle$.

The corresponding pruning lemma is developed as follows:

For a CD-derivation of h from A using T , construct a digraph showing which rules depended on other rules in the derivation. Given a CD-derivation, $\langle B_1, \dots, B_k \rangle$, it is conceptually simple to construct the *rule-dependency digraph*, though the definition is not as simple as the concept. Vertices of the digraph are the rules in the digraph; edges show dependencies.

Definition. For $\langle B_1, \dots, B_k \rangle$ a CD-derivation using Δ , let **min-parents**(B_i) be any minimal set of predecessors of B_i in the sequence such that

1. $min-parents(B_i) \vdash B_i$ or
2. $min-parents(B_i)$ is a singleton, $\{An(r)\}$, for some rule $r \in \Delta$, and $B_i = Con(r)$.

Consider the digraph $\langle \{B_i\}, \{\langle x, y \rangle \mid x \in \text{min-parents}(y)\} \rangle$. Construct a new digraph, $\langle V, E \rangle$ by taking any pair of vertices in the old digraph, of the form $An(r)$ and $Con(r)$ and such that $\{An(r)\} = \text{min-parents}(Con(r))$, and letting that be a vertex in the new digraph. V is the set of such vertices formed from pairs. For $x, y \in V$, $\langle x, y \rangle$ is an edge in the new digraph, i.e. $\langle x, y \rangle \in E$, just in case there is a path from $Con(x)$ to $An(y)$ in the old digraph, and no $An(z)$ for $z \in V$, $z \neq y$, occurs on the path.

Example (continued). The rule-dependency digraph for the CD-derivation above has two vertices, ‘ $A \succ B$ ’, and ‘ $B \succ C$ ’. There is an edge from the former to the latter.

Definition. Let $\mathbf{RDD}(\langle T, h \rangle, A) = \langle T, E \rangle$ be the rule-dependency digraph for some canonical CD-derivation of h from A using T .

It won’t matter which CD-derivation is chosen if there are many.

Consider $\mathbf{RDD}(\langle T, h \rangle, K_N \cup K_C)$, which is essentially the preferred way of drawing an argument for h from evidence, using T . A cutset of this digraph is a set of nodes that separates all top rules (which must be sinks, having out-degree zero) from rules that are sources (have in-degree zero). That is, Let C , a subset of T , be a minimal set such that every path from some source to some sink in $\mathbf{RDD}(\langle T, h \rangle, K_N \cup K_C)$ contains at least one member of C . Let $\mathbf{CSETS}(\mathbf{RDD}(\langle T, h \rangle, K_N \cup K_C)) = \mathbf{CS}(\langle T, h \rangle)$ be the set of all such cutsets.

Antecedents of rules in cut-sets are activators of arguments. They are not the only activators, but the search for asymmetric activators can be confined to these cut-sets.

Definition. Let $\mathbf{ACS}(\langle T, h \rangle) = \{\text{Conjoin}(\{An(i) \mid i \in S\}) \mid S \in \mathbf{CS}(\langle T, h \rangle)\}$. That is, $s \in \mathbf{ACS}(\langle T, h \rangle)$ just in case s conjoins the antecedents of all the rules in some cutset of an appropriate digraph.

Lemma (pruning). For any arguments $\langle T_1, h_1 \rangle$ and $\langle T_2, h_2 \rangle$, there exists an asymmetric activator of $\langle T_1, h_1 \rangle$ but not $\langle T_2, h_2 \rangle$ just in case there exists among $\mathbf{ACS}(\langle T_1 \cup T_2, h_1 \rangle)$ an asymmetric activator of $\langle T_1, h_1 \rangle$ but not $\langle T_2, h_2 \rangle$.

Proof. Let e be an asymmetric activator of $\langle T_1, h_1 \rangle$ but not $\langle T_2, h_2 \rangle$. We construct e' , an asymmetric activator among $\mathbf{ACS}(\langle T_1 \cup T_2, h_1 \rangle)$ as follows.

If e asymmetrically activates $\langle T_1, h_1 \rangle$ but not $\langle T_2, h_2 \rangle$, then there is some CD-derivation of h_1 from e using $T_1 \cup T_2$, and some

$RDD(< T_1 \cup T_2, h_1 >, e)$. The set of sources of this digraph is a minimal set of rules whose antecedents would allow a derivation of h_1 using $T_1 \cup T_2$. So $S = \{An(x) \mid x \text{ is a source in } RDD(< T_1 \cup T_2, h_1 >, e)\}$ is sufficient evidence to allow a CD-derivation of h_1 from $T_1 \cup T_2$. But since $Conjoin(S) = e'$ is weaker than e , there is no CD-derivation of h_2 from e' using $T_1 \cup T_2$, if there is none from e . Consequently, e' , which is among $ACS(< T_1 \cup T_2, h_1 >)$ is an asymmetric activator of $< T_1, h_1 >$ if e is, but not $< T_2, h_2 >$ if e isn't.

2

Touretzky, Horty, and Thomason have also attacked the Pollock-based part [8] of the system [9]. We are sympathetic to their concerns but feel no revision is required. They hold that reinstatement is inappropriate when the original argument, its counterargument, and the putative reinstater all contend the same proposition. They have no dispute with reinstatement that occurs when the reinstater attacks a subargument of the counterargument. At issue is a set of rules such as:

birds tend to fly;

chickens tend not to fly;

wild chickens tend to fly.

There is no dispute that a wild chicken should fly, but it is unusual in this example to say the argument that a wild chicken flies reinstates the argument that it flies in virtue of being a bird, by eliminating its only counterargument, that it does not fly in virtue of being a chicken. Clearly

birds tend to fly-because-they-are-birds

will not be reinstated by

wild chickens tend to fly-because-they-are-wild-chickens.

Part of the oddness of reinstatement here is that the reasons on which arguments are based imply explanations; it is difficult to regard them simply as the basis for inference. But even restricting attention to the idea that an argument is reinstated if its counterarguments are defeated, which makes no claims about explanations, the wild chicken example gives pause. Wildness has nothing to do with the reason chickens are exceptional birds.

Consider on the other hand,

Porsches tend to be fun;

Porsches co-developed with VW are not fun;

refined Porsches co-developed with VW are fun.

Refined co-developed Porsches are fun because they're Porsches, not because they are refined. There are statistical examples of both kind: a team that usually wins, but loses at night, wins at night in well-lit stadia. In this case, there is reinstatement. But if it wins at night with a particular pitcher, this apparently does not address the reason why the subclass is exceptional.²

Contrary to the implicit concerns of Touretzky, Horty, and Thomason, a system is useable by the wary knowledge representer whether it has reinstatement or not. If the system reinstates automatically, then the following set of rules prevents unwanted reinstatement:

birds are things-that-fly-because-they-are-birds, which tend to fly;

chickens tend not to fly;

wild chickens tend to fly, but are not things-that-fly-because-they-are-birds.

If, on the other hand, the system does not automatically reinstate, reinstatement can be enabled with the following representation:

Porsches are fun;

Porsches co-developed with VW are the-kind-of-thing-that-is-no-fun,
which are things that tend not to be fun;

refined Porsches co-developed with VW are fun, and are not the kind-
of-thing-that-is-no-fun.

Touretzky, Thomason, and Horty are once again pointing to a mere clash of intuitions. We see no need to revise the Pollock-based part of the system in response to their essay.

²William Chen offers this example:

animals are uncivilized;

people are civilized;

murderers are uncivilized.

There is reinstatement, because murderers aren't people; they're animals!

3

We correct just the Poole-based part of the system and retain Pollock-based reinstatement. The corrections are unfortunate considering the original authors' hope that the rules and their form would exhibit more permanence than those of competing systems. Poole's rule seemed to be a tidy replacement for the potentially numerous meta-reasons for superiority of one argument over another (for example, as advocated by Loui [10], and more deeply voiced by Konolige and Pollack [11]). Poole's rule has largely escaped pointed criticism of its behavior. However, its appeal was largely based on its tidiness, not on its obviousness (we still have seen no exact text of Popper, though Poole makes the attribution). The merit of simple syntactic rules for argument preference may depend on the adequacy of the repair reported here.

4 References

- [1] Simari, G. and R. Loui. "A Mathematical Treatment of Defeasible Reasoning and its Implementation," *Artificial Intelligence* 52, 1992.
- [2] Poole, D. "A logical framework for default reasoning," *Artificial Intelligence* 36, 1988.
- [3] McDermott, D. "AI, logic, and the frame problem," in *The Frame Problem in AI*, Frank Brown, ed., pp. 105-118, Morgan Kaufman, 1987.
- [4] Poole, D. "On the comparison of theories: preferring the most specific explanation," *IJCAI*, 1985.
- [5] Prakken, H. "A tool in modelling disagreement in law: preferring the most specific argument," *Proc. of the Third Conference on AI and Law*, ACM Press, 1991.
- [6] Hanks, S. and D. McDermott. "Nonmonotonic logic and temporal projection," *Artificial Intelligence* 33, 1987.
- [7] Sandewall, E. "Non-monotonic inference rules for multiple inheritance with exceptions," *Proc. IEEE* 74, 1986.
- [8] Pollock, J. "Defeasible reasoning," *Cognitive Science* 11, 1987.
- [9] Touretzky, D., R. Thomason, and J. Horty. "A skeptic's menagerie: conflictors, preemptors, reinstaters, and zombies in nonmonotonic inheritance," *Proc. IJCAI*, 1991.
- [10] Loui, R. "Defeat among arguments: a system of defeasible inference," *Computational Intelligence* 3, 1987.
- [11] Konolige, K. and M. Pollack. "Ascribing plans to agents," *Proc. IJCAI*, 1989.