Reinstatement and the Requirement of Maximal Specificity in Argument Systems*

Gustavo A. Bodanza[1] and Claudio Andrés Alessio[2]

[1] Universidad Nacional del Sur and CONICET, Argentina
ccbodanz@criba.edu.ar
[2] Universidad Católica de Cuyo and CONICET, Argentina
claudioalessio@uccuyo.edu.ar

Abstract. An argument is reinstated when all its defeaters are in turn ultimately defeated. This is a kind of principle governing most argument systems in AI. Nevertheless, some criticisms to this principle have been offered in the literature. Assuming that reinstatement is prima facie acceptable, we analyze some counterexamples in order to identify common causes. As a result, we found that the problem arises when arguments in a chain of attacks are related by specificity. We argue that the reason is that non-maximally specific arguments can be reinstated originating fallacious justifications. Following old intuitions by Carl Hempel about inductive explanations, we propose a requirement of maximal specificity on defeasible arguments and introduce "undermining defeaters" which, in essence, facilitate the rejection of those arguments which do not satisfy the requirement. This ideas are formally defined using the DeLP system for defeasible logic programming.

1 Introduction: Problems with Argument Reinstatement

Argument reinstatement is at the core of most argument systems, especially those which can be treated as instances of Dung's argumentation frameworks ([4]). The intuition is that an argument should be reinstated when all its possible defeaters are in turn defeated outright (cf. [1]). The example below, introduced by Dung as a motivation for his admissibility semantics, illustrates the rationale of reinstating argument A given argument C.

Example 1.
A: (Agent 1:) My government cannot negotiate with your government because your gorvernment doesn't even recognize my government.
B: (Agent 2:) Your government doesn't recognize my government either.
C: (Agent 1:) But your government is a terrorist government.
Then accepting that A and B are mutually attacking arguments and that C attacks B (but not the converse), the reinstatement of A by C makes sense.

On the other hand, other examples suggest that reinstatement cannot be taken as a general principle:

* Partially supported by SeCyT of the Universidad Nacional del Sur, and Universidad Católica de Cuyo, Argentina.

U. Kohlenbach et al. (Eds.): WoLLIC 2014, LNCS 8652, pp. 81–93, 2014.

Example 2.
A: Tweety flies because it is a bird, and birds tend to fly.
B: Tweety does not fly because it is a penguin, and penguins tend not to fly.
C: Tweety flies because it is a magic penguin, and magic penguins tend to fly.

Horty ([11]) has argued against reinstatement using a similar example. If A and C are jointly admitted, then a sound conclusion (Tweety flies) could be justified on basis of a weak reason (that flies because it is a bird). Clearly, a stronger reason is that Tweety has a skill that specifically magic penguins have. The acceptance of A would be unsound if the model is intended to offer the best explanation for the conclusion it yields. A would be acceptable only if all subclasses of birds (including penguins and magic penguins) are equally plausible to fly; but that is not the case here. The fact that C reinstates A's conclusion (which is also C's own conclusion) cannot be a reason for C to reinstate the whole argument A, because A does not meet that criterion.

A worse situation arises when the conclusion of the reinstated argument is stronger than that of the reinstating argument, as in the following case (also introduced by Horty):

Example 3.
A: Beth is millionaire because he is a Microsoft employee, and they tend to be millionaire.
B: Beth has less than half a million because he is a new Microsoft employee, and they tend to have less than half a million.
C: Beth has at least half a million because he is a new Microsoft employee in department X, and they have at least half a million.

Here A's conclusion is stronger than C's conclusion, in the sense that the last one is logically implied by the first one but not vice versa; that looks counterintuitive. Argument A would be reasonably accepted just in case that being millionaire be equally plausible for any subclass of the class of Microsoft employees.

Curiously, all the counterexamples to reinstatement that we found in the literature involve arguments that can be compared by specificity. That motivated the present study, which tries to show that the problem is the way in which specificity is used to establish defeat rather than a problem of the reinstatement principle.

The specificity criterion has been widely discussed in Philosophy of Science. Hempel ([9]) defended a *requirement of maximal specificity* as a condition for the acceptance of probabilistic/inductive-statistical explanations. Early applications of specificity in non-monotonic reasoning in AI were also aware of the intuition that only maximally specific explanations should be accepted, so from the argumentative point of view ([16]) as from the defeasible inheritance networks point of view ([6], [7], [10], [13], [18]). On the other hand, a specificity-based preference criterion among arguments combined with a reinstatement-based warrant procedure was introduced in [19].

Prakken ([17]) argued that reinstatement cannot be applied when statistical reasoning is at stake because more general arguments (like A in the above example) just cannot be constructed in a right representation, since the pertinent defaults must be blocked, and so only the most specific arguments remain; hence –Prakken concludes– the problem here is not about reinstatement but one of representation. In our opinion, while finding

general principles of representation could be a hard enterprise, the problem can instead be solved by finding general conditions under which the arguments can compete and defeat among them. Accordingly, we will argue that maximally specific arguments "undermine" less specific arguments when their conclusions are not plausible given the total evidence.

The paper is organized as follows. In section 2 a requirement of maximal specificity is formally introduced in terms of the defeasible logic programming language DeLP ([8]). Sections 3 and 4 introduce "undermining" defeaters and their role in a skeptical warrant procedure. Section 5 discuss the view of other authors through more examples, and our conclusions are offered in section 6.

2 The Requirement of Maximal Secificity in Rule-Based Argumentation Systems: The Case of DeLP

We introduce here the requirement of maximal specificity as a demarcation criterion for the acceptance of arguments. As such, it should be used to filter the arguments which are not maximally specific w.r.t. their conclusions as they can leave room for irrelevant explanations. We will formally define our criterion in the context of the particular rule-based argument system DeLP ([8]), where specificity is formally defined as a criterion for argument comparison.

DeLP is based on a first-order language $\mathcal{L}$ that is partitioned in three disjoint sets: a set of facts, a set of strict rules and a set of defeasible rules. *Facts* are literals, i.e. ground atoms (L) or negated ground atoms ($\sim L$, where '$\sim$' represents the classical negation); facts represent particular knowledge. Both *strict* and *defeasible rules* are program rules. Syntactically, strict rules are sequents of the form $L \leftarrow L_1, \ldots, L_n$ and defeasible rules are sequents of the form $L \multimap L_1, \ldots, L_n$, where $L, L_1, \ldots, L_n$ are literals. Strict rules represent general, non-defeasible knowledge while defeasible rules represent tentative, defeasible knowledge. A *defeasible logic program* (de.l.p.) $\mathcal{P}$ is a pair (Π, Δ) where Π is a set partitioned in two subsets Π_F, containing only facts, and Π_G, containing only strict rules, and Δ is a set of defeasible rules. Given a de.l.p. $\mathcal{P} = (\Pi, \Delta)$ we say that a literal L is a *defeasible derivation* from Γ in $\mathcal{P}$, in symbols, $\Gamma \mathrel{|\!\sim}_{\mathcal{P}} L$ iff $\Gamma \subseteq \Pi \cup \Delta$ and there exists a sequence of ground (instantiated) literals $L_1, \ldots, L_n$ such that $L_n = L$ and for each L_i, $1 \leq i \leq n$, either $L_i \in \Gamma$ or there exists either a strict rule $(L \leftarrow L_1, \ldots, L_k)$ or a defeasible rule $(L \multimap L_1, \ldots, L_k)$ in Γ such that $\{L_1, \ldots, L_k\} \subseteq \{L_1, \ldots, L_{i-1}\}$. If all the rules used in the derivation of A are strict then we say that L is a *strict derivation* from Γ, in symbols, $\Gamma \vdash_{\mathcal{P}} L$. (From now on, we will write '$\mathrel{|\!\sim}$' and '$\vdash$' instead of '$\mathrel{|\!\sim}_{\mathcal{P}}$' and '$\vdash_{\mathcal{P}}$', respectively, when the referenced de.l.p. is obvious.)

Definition 1. *(Argument structure ([8])) Given a de.l.p. $\mathcal{P} = (\Pi, \Delta)$, an* argument structure *(in $\mathcal{P}$) is a pair $\langle T, h \rangle$, where $T \subseteq \Delta$ and h is a literal (the argument's* conclusion*), and*

1. $\Pi \cup T \mathrel{|\!\sim} h$,
2. $\Pi \cup T \not\mathrel{|\!\sim} \bot$,
3. $\nexists T' \, (T' \subset T \ \wedge \ \Pi \cup T' \mathrel{|\!\sim} h)$.

Definition 2. *(Subargument ([8])) An argument structure* $\langle T, h\rangle$ *is a* subargument structure *of an argument structure* $\langle T', h'\rangle$ *if* $T \subseteq T'$.

Definition 3. *(Strictly more specific ([8])) Let* $\mathcal{P}=(\Pi, \Delta)$ *be a de.l.p. and let* F *be the set of all literals that have a defeasible derivation from* $\mathcal{P}$. *Let* $\langle T_1, h_1\rangle$ *and* $\langle T_2, h_2\rangle$ *be two argument structures obtained from* $\mathcal{P}$. $\langle T_1, h_1\rangle$ *is* strictly more specific than $\langle T_2, h_2\rangle$, *in symbols,* $\langle T_1, h_1\rangle \succ_{\mathbf{spec}} \langle T_2, h_2\rangle$ *iff*

1. for all $H \subseteq F$, *if* $H \cup \Pi_G \cup T_1 \mathrel{|\!\sim} h_1$ *and* $H \cup \Pi_G \nvdash h_1$ *then* $H \cup \Pi_G \cup T_2 \mathrel{|\!\sim} h_2$, *(every* H *that "activates"* h_1 *also 'activates'* h_2*), and*

2. there exists $H \subseteq F$ *such that* $H \cup \Pi_G \cup T_2 \mathrel{|\!\sim} h_2$, $H \cup \Pi_G \nvdash h_2$ *and* $H \cup \Pi_G \cup T_1 \not\mathrel{|\!\sim} h_1$ *(some* H *"activates"* h_2 *but not* h_1*).*

Using this same specificity criterion, Poole [16] proposes to choose the most specific explanations, i.e. those arguments which are maximal with respect to $\succ_{\mathbf{spec}}$. In this way, Poole leaves no room for reinstatement among arguments compared by specificity. This criterion is near to what we will propose here, but so stated it can have the effect of precluding acceptable arguments even when less specific arguments are not in conflict with the maximally specific ones.

Example 4. Let $\mathcal{P} = (\Pi, \Delta)$ be a de.l.p. representing the knowledge that all lapwings are birds, birds tend to fly, lapwings tend to nest on the ground and Pedro is a lapwing:
$\Pi = \{\ bird(x) \leftarrow lapwing(x),\ lapwing(pedro)\ \}$
$\Delta = \{\ flies(x) \prec\!\!- bird(x),\ nests\_on\_the\_ground(x) \prec\!\!- lapwing(x)\ \}$
Then we have the argument structures:
$A = \langle \{flies(pedro) \prec\!\!- bird(pedro)\},\ flies(pedro)\rangle$,
$B = \langle \{nests\_on\_the\_ground(pedro) \prec\!\!- lapwing(pedro)\},$
$\quad nests\_on\_the\_ground(pedro)\rangle$
Since $B \succ_{\mathbf{spec}} A$, choosing only the maximal elements of $\succ_{\mathbf{spec}}$ precludes the acceptable argument A and its conclusion $flies(pedro)$.

Indeed, selecting just the maximal elements of $\succ_{\mathbf{spec}}$ does not seem to be a good approach to the requirement of maximal specificity as proposed in Philosophy of Science for inductive-probabilistic explanations. The intuition in [9] is that what is inferred in a maximally specific explanation about a class G taking into account the total evidence must also be inferred about any subclass H of G with the same probability. Though extrapolating this criterion to defeasible argumentation is difficult since inferences are not obtained with probability measures, we propose that a maximally specific defeasible argument about a class G should at least not be contradictory with the defeasible conclusions obtained about any subclass H of G, considering the total evidence, i.e. the information represented in Π. In terms of DeLP, this means that maximally specific arguments should not have "proper defeaters" as these are indicative of "undermining" evidence.

Definition 4. *(Proper defeater ([8])) An argument structure* $\langle S, j\rangle$ *is a* proper defeater *of an argument structure* $\langle T, h\rangle$ *if for some sub-argument* $\langle T', h'\rangle$ *of* $\langle T, h\rangle$, $\langle S, j\rangle \succ_{\mathbf{spec}} \langle T', h'\rangle$ *and* $\Pi \cup \{j, h'\} \vdash \bot$. *Given a set of argument structures* S *we also define* $def_{prop}(S) =_{df} \{(A, B) : A, B \in S$ *and* A *is a proper defeater of* $B\}$.

Definition 5. *(Undermining evidence) Given a de.l.p. $\mathcal{P} = (\Pi, \Delta)$, a subset F of Π_F is* undermining evidence *of an argument structure $\langle T, h\rangle$ if $F \cup T' \cup \Pi_G \not\vdash h'$ for some proper defeater $\langle T', h'\rangle$ of $\langle T, h\rangle$ (i.e. F "activates" some proper defeater of the argument).*

Note that having a proper defeater is a sufficient condition for having undermining evidence, though other conditions could be also found (more on this in section 5). Now we can formally state the property of maximal specificity as follows:

Definition 6. (Maximal Specificity (MS)) *Given a de.l.p $\mathcal{P}$, we say that an argument structure $\langle T, h\rangle$ is* maximally specific (w.r.t. its conclusion h) *in $\mathcal{P}$ iff there exists no undermining evidence of $\langle T, h\rangle$ in $\mathcal{P}$.*

Requiring MS as a condition for argument warrant implies to reject any argument structure which has some proper defeater. Note that it does not matter whether proper defeaters are defeated or not to reject a non-maximally specific argument; that is why we prefer to highlight the undermining evidence and to use proper defeaters just as a way for detecting it.

Example 5. (Example 4 revisited) Both A and B satisfy MS, A w.r.t. $flies(pedro)$ and B w.r.t. $nests\_on\_the\_ground(pedro)$.

Example 6. (De.l.p. for representing Example 3) Let $\mathcal{P} = (\Pi, \Delta)$ a de.l.p. such that

$$\begin{aligned}\Pi = \{\ & has\_at\_least\_half\_a\_million(x) \leftarrow millionaire(x),\\ & ms\_employee(x) \leftarrow new\_ms\_employee(x),\\ & new\_ms\_employee(x) \leftarrow new\_ms\_employee\_dept\_x(x),\\ & new\_ms\_employee\_dept\_x(beth)\}\end{aligned}$$

$$\begin{aligned}\Delta = \{\ & millionaire(x) -\!\!< ms\_employee(x),\\ & \sim has\_at\_least\_half\_a\_million(x) -\!\!< new\_ms\_employee(x),\\ & has\_at\_least\_half\_a\_million(x) -\!\!< new\_ms\_employee\_deptX(x)\}\end{aligned}$$

Then we have the argument structures:

$$\begin{aligned}A &= \langle\{millionaire(beth) -\!\!< ms\_employee(beth)\}, millionaire(beth)\rangle,\\ B &= \langle\{\sim has\_at\_least\_half\_a\_million(beth) -\!\!< new\_ms\_employee(beth)\},\\ &\quad \sim has\_at\_least\_half\_a\_million(beth)\rangle,\\ C &= \langle\{has\_at\_least\_half\_a\_million(beth) -\!\!< new\_ms\_employee\_deptX(beth)\},\\ &\quad has\_at\_least\_half\_a\_million(beth)\rangle.\end{aligned}$$

Then C satisfies MS w.r.t. $has\_at\_least\_half\_a\_million(beth)$, and neither A nor B satisfy MS because $\{new\_ms\_employee\_dept\_x(beth)\}$ is undermining evidence for them.

3 Undermining Defeaters

Systems in which arguments interact *only* through proper defeaters can lead to the acceptance of non-maximally specific arguments if the warrant procedure satisfies the

reinstatement principle. But this does not necessarily imply that reinstatement is invalid. Those argument systems in which different kinds of defeat are used —including proper defeaters— can be amended to sanction only maximally specific arguments. Our proposal is simple and consists in the introduction of "undermining defeaters", which are based on the main result derived from the notion of 'undermining evidence':

Lemma 1. *Let $\langle T, h\rangle$ and $\langle T', h'\rangle$ be two argument structures such that $\langle T, h\rangle$ is a proper defeater of $\langle T', h'\rangle$. If H is undermining evidence for $\langle T, h\rangle$ then H is undermining evidence for $\langle T', h'\rangle$.*

Proof. Let $\langle T, h\rangle$ be a proper defeater of $\langle T', h'\rangle$ and let H be undermining evidence for $\langle T, h\rangle$. Then H activates some proper defeater $\langle S, j\rangle$ of $\langle T, h\rangle$. Since $\langle S, j\rangle$ is more specific than $\langle T, h\rangle$, H activates $\langle T, h\rangle$. And since $\langle T, h\rangle$ is more specific than $\langle T', h'\rangle$, H also activates $\langle T', h'\rangle$. Then, by Definition 5, H is undermining evidence for $\langle T', h'\rangle$. □

Definition 7. *(Undermining defeater) Given two arguments structures $\langle T, h\rangle$ and $\langle T', h'\rangle$, we say that $\langle T, h\rangle$ is an* undermining defeater *of $\langle T', h'\rangle$ iff for any subset of facts $F \subseteq \Pi_F$, if $F \cup T \cup \Pi_G \mathrel{|\!\sim} h$ then $F \cup S \cup \Pi_G \mathrel{|\!\sim} j$ for some proper defeater $\langle S, j\rangle$ of $\langle T', h'\rangle$ (i.e. if F activates $\langle T, h\rangle$ then F also activates some proper defeater of $\langle T', h'\rangle$). We also define $def_{und}(S) =_{df} \{(A, B) : A, B \in S$ and A is an undermining defeater of $B\}$.*

Undermining defeaters can be viewed as a kind of undercutting defeaters, at least indirectly, since their acceptance implies the use of total evidence which gives the reason that makes the conclusion of the defeated argument not inferable. They are clearly not rebutting defeaters since it cuould be the case that the joint acceptance of both an argument an its undermining defeater does not yield contradiction (for instance, in Example 5 argument C is an undermining defeater of argument A, but it is not a rebutting defeater. See, e.g., [15] for more on the distinction undercutting/rebutting defeater).

Clearly, from Lemma 1 and Definition 7 we have that undermining evidence 'propagate' through a chain of proper defeaters.

Lemma 2. *Let $\langle T, h\rangle$ be a proper defeater of $\langle T', h'\rangle$. Then for every proper defeater $\langle S, j\rangle$ of $\langle T, h\rangle$, $\langle S, j\rangle$ is an undermining defeater of $\langle T', h'\rangle$.*

Proof. Immediate from Definition 7 and Lemma 1. □

As a consequence, in cycles of proper defeaters the ensuing undermining defeaters are indicative of undermining evidence for all the arguments involved in the cycle, including themselves (Fig. 1). Finally, the previous lemmata lead immediately to the following equation.

Theorem 1. *Let S be any set of argument structures, and $def_{prop}(S)^{tr}$ be the transitive closure*[1] *of $def_{prop}(S)$. Then $def_{und}(S) = def_{prop}(S)^{tr}$.*

Proof. Immediate from Lemma 1 and Lemma 2. □

In the next section, the above result will enable us to think of different ways of representing the rejection of non-maximally specific arguments.

[1] The *transitive closure* of a binary relation R is the minimal (w.r.t. $\subseteq$) transitive relation R' such that $R \subseteq R'$.

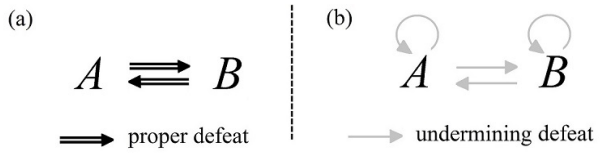

Fig. 1. Argumentation framework (a) with proper defeaters only and (b) with undermining defeaters

4 Undermining Defeaters and a Warrant Procedure Satisfying Reinstatement

Non-maximally specific arguments can be rejected through a warrant procedure satisfying reinstatement, which means that reinstatement can be saved from the before mentioned criticisms as a principle of defeasible argumentation. In DeLP, warrant can be determined through a dialectical analysis represented by a two-party game, where a proponent tries to defend an argument and an opponent tries to refute it (we define this game as in [2]). Given the set $Args$ of all the argument structures that can be constructed in a de.l.p. $\mathcal{P}$, and once all the defeat relations over $Args$ are established, argument warrant can be analyzed through a Dung's style argumentation framework ([4]).

Definition 8. *(Argumentation framework associated with a de.l.p.) Given a de.l.p. $\mathcal{P}$, the* argumentation framework associated with $\mathcal{P}$ *is the pair $(Args, attacks)$ where $Args$ is the set of all the argument structures obtained from $\mathcal{P}$ and $attacks = \bigcup DEF(Args)$, where $DEF(Args) = \{def_1, \ldots, def_k\}$, is the set containing every defeat criterion $def_i \subseteq Args \times Args$ ($1 \leq i \leq n$) defined on $Args$. (We will assume that $def_{prop}(Args) \in DEF(Args)$.)*

Definition 9. *(Argumentation game) An* argumentation game *on an argumentation framework $(Args, attacks)$ is a zero-sum extensive game in which:*
*1. There are two players, i and $-i$, who play the roles of **P** and **O**, respectively.*
2. A history *in the game is any sequence $A_0, A_1, A_2, \ldots, A_{2k}, A_{2k+1}, \ldots$ of choices of arguments in $Args$ made by the players in the game. A_{2k} corresponds to **P** and A_{2k+1} to **O**, for $k = 0, 1, \ldots$. At any history, A_0 is the argument that player **P** intends to defend.*
3. In a history, the choices by a player i at a level $k > 0$ are $C_i(k) = \{A \in Args : (A, A_{k-1}) \in attacks\}$.
4. A history of finite length K, $A_0, \ldots, A_K$, is terminal *if A_K corresponds to player j ($j = i$ or $j = -i$) and $C_{-j}(K + 1) = \emptyset$.*
*5. Payoffs are determined at terminal histories: at $A_0, \ldots, A_K$, **P**'s payoff is 1 (representing winning) if K is even (*i.e.*, **O** cannot reply to **P**'s last argument), and -1 (representing loosing) otherwise. In turn, **O**'s payoff at $A_0, \ldots, A_K$ is 1 if K is odd and -1 otherwise.*

Definition 10. *(Strategy) A* strategy *for a player i is a function that assigns an element $A_{l+1} \in C_i(l)$ at each non-terminal history $A_0, \ldots, A_l$ where A_l corresponds to player $-i$. A strategy of player i is said a* winning strategy *for i if for every strategy chosen by $-i$, the ensuing terminal history yields a payoff* 1 *for player i.*

Definition 11. *(Warrant) An argument A is* warranted *in $(Args, attacks)$ iff* ***P*** *has a winning strategy to defend A in the game associated to $(Args, attacks)$.*

Furthermore, different game protocols can be defined to obtain different behaviors. Since we are interested here in the refutation of non-maximally specific arguments, let us see how to do that in systems that incorporate undermining defeaters and in systems based only in proper defeaters. For the first approach we propose the following protocol:

(1) The game ends if, at any level k, a player i advances an argument A such that the argument B moved at level $k-1$ by player $-i$ is such that A is an undermining defeater of B (i wins).

(2) **P** is not allowed to advance an argument that was already advanced by either player in the same history.

Rule (1) says that once an undermining defeater is played the game ends (the player who moved the non-maximally specific argument loses). The purpose of this rule is to obligate the players to use only maximally specific arguments. Rule (2), in time, ensures finite, skeptical games. Let us call this protocol PU.

On the other hand, an obvious way to obtain the same behavior in argumentation frameworks where only proper defeaters are defined is by replacing the first rule as follows:

(1') The game ends if, at any level k, a player i advances an argument A such that the argument B moved at level $k-1$ by player $-i$ is such that there exists a sequence $A_1, \ldots, A_n$ where $A_1 = A$, $A_n = B$ and $(A_h, A_{h+1}) \in def_{prop}$ for every h, $1 \leq h < n$ (i wins).

Let us call this protocol PP. Then Theorem 1 clearly ensures the same behavior under protocol PU as under protocol PP.

Example 7. Given an argumentation framework $(Args, attacks)$ where $Args = \{A, B, C, D\}$ and $attacks = \{(C, D)\} \cup def_{prop}(Args)$, where $def_{prop}(Args) = \{(A, B),$

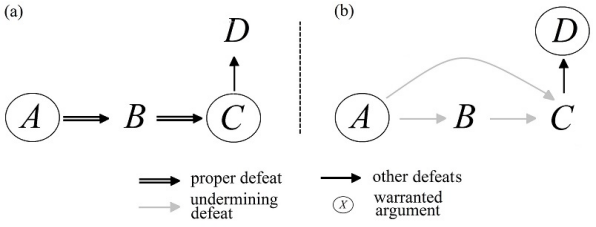

Fig. 2. Argumentation framework (a) without and (b) with undermining defeaters

$(B, C)\}$. Then $def_{und}(Args) = \{(A, B), (B, C), (A, C)\}$. Note that while **P** has winning strategies for defending both A and C in the game associated to $(Args, attacks)$ in the plain game (i.e. not having any added protocol) (Fig. 2, (a)), **P** has winning strategies for defending both A and D so in the game associated to $(Args, attacks)$ under protocol PP as in the game associated to $(Args, attacks \cup def_{und})$ under protocol PU (Fig. 2, (b)).

5 Discussion

The solution we have proposed here works well, in particular, for the simplest version of DeLP [19] where the attack relation is defined only in terms of blocking and proper defeaters. Given a de.l.p. $\mathcal{P} = (\Pi, \Delta)$, $\langle T, h\rangle$ is a *blocking defeater* of $\langle T', h'\rangle$ iff there exists some sub-argument $\langle T'', h''\rangle$ of $\langle T', h'\rangle$ such that $\Pi \cup \{h, h''\} \vdash \bot$, and $\langle T, h\rangle$ and $\langle T'', h''\rangle$ are not related by specificity. Define $def_{block} =_{df} \{(A, B) : A$ is a blocking defeater of $B\}$. Then def_{block} is clearly symmetric. Let us now analyze Example 7 in terms of this system. As the attack from C to D is not a case of proper defeater, it must be a case of blocking defeater. Then, by the symmetry of def_{block}, D also attacks C. But note that this does not change the resulting warrant of A and D under protocol PP. On the other hand, some dubious cases that can arise under other specifications of the attack relation are avoided in this system. For example, assume that A is a (non-proper) defeater of B and B is a proper defeater of C. Then it could seem reasonable the reinstatement of C by A, even when C is not maximally specific. Nevertheless, that could not happen in DeLP because the assumption that A is a non-proper defeater of B implies that A and B are blocking defeaters one of each other. Hence, it is easy to see that the reinstatement of C by A is impossible under protocol PP as **P** lacks of a winning strategy for A (**O** can repeat B to refute A, leaving **P** out of moves).

Similar examples would suggest that non-maximally specific arguments should be reinstated anyway. For instance, consider again Example 2 but where argument C is now: "It cannot be concluded that Tweety is a penguin since it was observed under deficient sight conditions during a blizzard". Now C could be seen as an undercutting defeater of B and B as a proper defeater of A, what would lead to the reinstatement of the non-maximally specific argument A. But note that the acceptance of C implies the treatment of 'Tweety is a penguin' not as evidence but as a questionable presumption, hence B should not be treated as a proper defeater of A strictly. Therefore, A is still a maximally specific argument and its reinstatement seems right.

For other cases where C is a (unidirectional, non-proper) defeater of B and B is a proper defeater of A, it is not clear whether A should be reinstated or not. Indeed, it is difficult for us to conceive such an example.

The last example was introduced by Prakken ([17]) to show that, unlike *direct reinstatement*, *indirect reinstatement* is valid. Direct reinstatement is when all three arguments are in conflict on their final conclusions (e.g. Example 2). Indirect reinstatement, on the other hand, is when the reinstating argument C defeats the 'middle' argument B on one of its intermediary conclusions (e.g. Example 3). But this distinction is not related to our solution as it does not focus on the kind of defeaters which are involved and the role of undermining evidence on them, which is the key in the MS property.

The following example was also introduced by Prakken to argue that reinstatement depends "on the nature of the domain, the kind of knowledge involved and the context in which this knowledge is used" ([17]: 93):

Example 8.
A: John will be imprisoned up to 6 years because for theft imprisonment up to 6 years is acceptable, and John has been found guilty of theft.
B: John will be imprisoned for no more than 3 years because for theft out of poverty imprisonment of more than 3 years is not acceptable, and evidence shows that John stole motivated by poverty.
C: John will be imprisoned for more than 4 years because he stole during riots, and for theft during riots, even when poverty is proved, only imprisonment of more than 4 years is acceptable.

Prakken argues that the reinstatement of A by C is valid here and leads to accept an imprisonment between 4 and 6 years. We disagree at this point since anyway, in our opinion, C is a proper defeater of B and B is a proper defeater of A, hence C is an undermining defeater of A. The total evidence considered in C about a more serious crime than theft out of poverty leads to put a minimum of 4 years of imprisonment, leaving the upper limit not established. Indeed, we can imagine even more serious crimes (e.g. murder) which occurrence together with theft would rise the top above 6 years. Hence we think that C is the only warranted argument and A should not be reinstated.

Nevertheless, there are still open problems to deal with. Our notion of undermining defeat is not completely characterized as it lies on a concept of undermining evidence for which we state sufficient but not necessary conditions. While having a proper defeater is a clear sign of an argument's undermining evidence, in other cases the total evidence should prevent some conclusion without sanctioning the contrary. Horty ([12]) analyses the following example. Assume that a population of ruffed finches, a kind of birds, is distributed among a couple of islands. Their nests are mostly but not entirely confined to Green Island, but there is a particular subspecies known as least ruffed finches whose nests are distributed almost evenly between Green Island and Sand Island. Now, consider a particular individual, Frank, who happens to be a least ruffed finch. What should we conclude about the location of Frank's nest? Though this situation cannot be represented in DeLP because disjunctions cannot occur in the head of a rule, we can adjust the information to the formalism by considering that Green Island and Sand Island conform a group of islands, call it 'Two Islands', so we can get the following representation:

Example 9. Let $\mathcal{P} = (\Pi, \Delta)$ be a de.l.p. representing the knowledge that all least ruffed finches are ruffed finches, ruffed finches tend to nest on Green Island, least ruffed finches tend to nest on Two Islands, nesting on Green Island implies nesting in Two Islands, and Frank is a least ruffed finch:

$$\begin{aligned}\Pi = \{\ & ruffed\_finch(x) \leftarrow least\_ruffed\_finch(x), \\ & nests\_on\_TwoIslands(x) \leftarrow nests\_on\_GreenIsland(x), \\ & least\_ruffed\_finch(frank)\ \}\end{aligned}$$

$$\Delta = \{\ nests\_on\_GreenIsland(x) \text{—}\!\prec ruffed\_finch(x),$$
$$nests\_on\_TwoIslands(x) \text{—}\!\prec least\_ruffed\_finch(x)\ \}$$

Then we have, among others, the argument structures:

$$A = \langle \{nests\_on\_GreenIsland(frank) \text{—}\!\prec ruffed\_finch(frank)\},$$
$$nests\_on\_GreenIsland(frank)\rangle$$
$$B = \langle \{nests\_on\_TwoIslands(frank) \text{—}\!\prec least\_ruffed\_finch(frank)\},$$
$$nests\_on\_TwoIslands(frank)\rangle$$

Though B is more specific than A it is not a proper defeater, hence the conclusion that Frank nests on Green Island is obtained. The formalisms incurs in the fallacy of exclusion, since the information that Frank is a least ruffed finch is obviated, treating Frank just as a ruffed finch. To solve the problem, Horty proposes to add a (meta-level) default expressing that cases of least ruffed finches exclude the application of the default that connects ruffed finches with nesting on Green Island (a kind of undercutting defeater). But this solution requires more representation, while we are inclined to less representation dependence. In our opinion, this must be solved by defining a new kind of undermining defeater which makes appropriate use of the total evidence, so that argument B (or the evidence on which B is built) undercuts, in some specified way, argument A.

More in the line of Horty's solution, the work by Modgil on hierarchical argumentation ([14]) offers another interesting turn to the problem of reinstatement introducing arguments for (meta-level) preference criteria. The model develops a form of meta-argumentation where, for example, if $A\ attacks\ B$ is established on basis of a preference criterion P1, and $B\ attacks\ A$ is established on basis of a preference criterion P2, an argument C supporting the preference of P1 over P2 poses an attack on $B\ attacks\ A$, A resulting reinstated. Note that, under this view, C is not attacking B but *the attack* of B over A. The example of Tweety observed during a blizzard can be interpreted in this terms assuming that the preference criterion is based on an ordering $>$ on the evidence, such that $bird(tweety) > penguin(tweety)$. Then, while B is a proper defeater of A, C expresses a preference of A over B based on $>$, so that C defends A. Examples like Example 9, on the other hand, cannot be solved unless, again, a special kind of undercutting defeater is defined.

This gives rise to the question of what kind of defeaters are undermining defeaters. We have argued that they qualify as undercutting defeaters. As undermining defeaters are based on a total-evidence requirement they can be considered a kind of –in Pollock's terms– subproperty defeaters, just the same as specificity (i.e. proper) defeaters. And subproperty defeaters are all undercutting defeaters. Pollock's words seem to confirm our opinion:

> To the best of my knowledge, there has never been an intuitive example of specificity defeat presented anywhere in the literature that is not an example of the operation of the total-evidence requirement in one of these special varieties of defeasible inference [statistical syllogism, direct inference, various kinds of legal and deontic reasoning], and the latter are all instances of undercutting defeat. Accordingly, I will assume that undercutting defeaters and rebutting defeaters are the only possible kinds of defeaters. ([15]: 236)

Finally, several principles have been introduced in order to validate the argumentation inference of rule-based argumentation systems, mainly consistency and closure ([3], [5]). A formal analysis of the relationship between maximal specificity and these principles is planned as future work.

6 Conclusion

The issue of reinstatement as a principle for argument systems was the subject of a serious criticism ([11]) while its defense (mainly that of [17]) has not been entirely satisfactory in our opinion. The criticism focuses only cases in which specificity is the comparison criterion among arguments. We argued here that the problem is that specificity based argument systems do not incorporate a precise way of defeating all non-maximally specific arguments. We proposed a formal criterion of maximal specificity which, in accordance with early researches about inductive explanations ([9]), is based on the total evidence represented in the knowledge base. Moreover, we introduced *undermining defeaters* and showed how they enable the warrant of only maximally specific arguments in the context of the DeLP system ([8]) defining particular argumentation game protocols.

Acknowledgements. We thank three anonymous referees who made helpful criticisms that improved this work.

References

1. Baroni, P., Giacomin, M.: Semantics of Abstract Argument Systems. In: Rahwan, I., Simari, G.R. (eds.) Argumentation in artificial intelligence, pp. 25–44. Springer Publishing Company (2009)
2. Bodanza, G., Thomé, F., Simari, G.: Argumentation Games for Admissibility and Cogency Criteria. In: Verheij, B., Szeider, S., Woltran, S. (eds.) Computational Models of Argument. Proceedings of COMMA 2012, pp. 153–164. IOS Press (2012)
3. Caminada, M., Amgoud, L.: On the Evaluation of Argumentation Formalisms. Artificial Intelligence 171(5-6), 286–310 (2007)
4. Dung, P.M.: On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games. Artificial Intelligence 77, 321–358 (1995)
5. Dung, P.M., Thang, P.M.: Closure and Consistency In Logic-Associated Argumentation. Journal of Artificial Intelligence Research 49, 79–109 (2014)
6. Etherington, D.: Formalizing Nonmonotonic Reasoning Systems. Artificial Intelligence 31, 41–85 (1987)
7. Etherington, D., Reiter, R.: On Inheritance Hierarchies with Exceptions. In: Proceedings of the Third National Conference on Artificial Intelligence (AAAI 1983), pp. 104–108 (1983)
8. García, A., Simari, G.: Defeasible Logic Programming: An Argumentative Approach. Theory and Practice of Logic Programming 4(1), 95–138 (2004)
9. Hempel, C.: Maximal Specificity and Lawlikeness in Probabilistic Explanation. Philosophy of Science 35(2), 116–133 (1968)

10. Horty, J.: Some Direct Theories of Nonmonotonic Inheritance. In: Gabbay, D., Hobber, C., Robinson, J. (eds.) Handbook of Logic in Artificial Intelligence and Logic Programming. Nonmonotonic Reasoning and Uncertain Reasoning, vol. 3, pp. 111–187. Oxford University Press (1994)
11. Horty, J.: Argument Construction and Reinstatement in Logics for Defeasible Reasoning. Artificial Intelligence and Law 9, 1–28 (2001)
12. Horty, J.: Reasons as Defaults. Oxford University Press (2012)
13. Horty, J., Thomason, R., Touretzky, D.: A Skeptical Theory of Inheritance in Nonmonotonic Semantic Networks. Artificial Intelligence 42, 311–348 (1990)
14. Modgil, S.: Value Based Argumentation in Hierarchical Argumentation Frameworks. In: Dunne, P., Bench-Capon, T. (eds.) Proc. of Computational Models of Argument, COMMA 2006, Liverpool, UK, September 11-12. Frontiers in Artificial Intelligence and Applications, vol. 144, pp. 297–308. IOS Press (2006)
15. Pollock, J.: Defeasible Reasoning with Variable Degrees of Justification. Artificial Intelligence 133, 233–282 (2001)
16. Poole, D.: On the Comparison of Theories: Preferring the Most Specific Explanation. In: Proc. of the Ninth IJCAI, Los Altos, pp. 144–147 (1985)
17. Prakken, H.: Intuitions and the Modelling of Defeasible Reasoning: Some Case Studies. In: Benferhat, S., Giunchiglia, E. (eds.) Proceedings of the 9th International Workshop on Non-Monotonic Reasoning (NMR 2002), Tolouse, France, April 19-21, vol. 2, pp. 91–102 (2002)
18. Reiter, R., Criscuolo, G.: On Interacting Defaults. In: Proceedings of the Seventh International Joint Conference on Artificial Intelligence (IJCAI 1981), pp. 270–276 (1981)
19. Simari, G., Loui, R.: A Mathematical Treatment of Defeasible Reasoning. Artificial Intelligence 53, 125–157 (1992)