Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning

David Poole's Specificity Revised*

Claus-Peter Wirth and **Frieder Stolzenburg**
FB AI, Harz University of Applied Sciences, 38855 Wernigerode, Germany
wirth@logic.at, *fstolzenburg@hs-harz.de*

Abstract

In the middle of the 1980s, David Poole introduced a semantical, model-theoretic notion of specificity to the artificial-intelligence community. Since then it has found further applications in non-monotonic reasoning, in particular in defeasible reasoning. Poole's notion, however, turns out to be intricate and problematic, which — as we show — can be overcome to some extent by a closer approximation of the intuitive human concept of specificity. Besides the intuitive advantages of our novel specificity ordering over Poole's specificity relation in the classical examples of the literature, we also report some hard mathematical facts: Contrary to what was claimed before, we show that Poole's relation is not transitive. Our new notion of specificity is transitive and also monotonic w.r.t. conjunction.

Keywords: specificity, defeasible reasoning, argumentation.

1 Introduction

A possible explanation of how humans manage to interact with reality — in spite of the fact that their information on the world is partial and inconsistent — mainly consists of the following two points: Humans use a certain amount of rules for default reasoning and are aware that some arguments relying on these rules may be defeasible. In case of the frequent conflicting or even contradictory results of their reasoning, they prefer more specific arguments to less specific ones. An intuitive concept of specificity plays an essential rôle in this explanation, which seems to be highly successful in practice. On the long way approaching this proven intuitive human concept of specificity, the first milestone marks the development of a semantical, model-theoretic notion of specificity having passed first tests of its usefulness and empirical validity. Indeed, at least as the first step, a semantical, model-theoretic notion will probably offer a broader and better basis for applications in systems for common sense reasoning than notions of specificity that depend on peculiarities of special calculi or even on extra-logical procedures. This holds in particular if the results of these systems are to be accepted by humans.

David Poole (1985) has sketched such a notion as a binary relation on arguments and evaluated its intuitive validity with some examples. Poole's notion of specificity was given a more appropriate formalization in (Simari and Loui 1992). The properties of this formalization were examined in detail in (Stolzenburg et al. 2003).

In this paper, before we give a specification of the formal requirements on any reasonably conceivable relation of specificity in Sect. 4, we present a detailed analysis of the intentional motivation of our *intuition that Poole's specificity is a first step on the right way* (Sect. 3). Moreover, in Sect. 5, we clearly disambiguate Poole's specificity from slightly improved versions such as the one in (Simari and Loui 1992), and introduce a novel specificity relation ($\lesssim_{\mathrm{CP1}}$), which presents a major correction of Poole's specificity because it removes a crucial shortcoming of Poole's original relation ($\lesssim_{\mathrm{P1}}$) and its slight improvements ($\lesssim_{\mathrm{P2}}$, $\lesssim_{\mathrm{P3}}$), namely their lack of transitivity. Finally, in Sect. 6, we present several examples that will convince every carefully contemplating reader of the superiority of our novel specificity relation $\lesssim_{\mathrm{CP1}}$ w.r.t. human intuition, including monotonicity w.r.t. conjunction. We briefly discuss related works in Sect. 7, and conclude with Sect. 8.

2 Basic Notions and Notation

Let us narrow the general logical setting of specificity down to the concrete framework of *defeasible logic with the restrictions of logic programming*, as found e.g. in (Stolzenburg et al. 2003; Chesñevar et al. 2003).

A *literal* is an atom, possibly prefixed with the symbol "$\neg$" for negation. A *rule* is a literal, but possibly suffixed with a reverse implication symbol "$\Leftarrow$" that is followed by a conjunction of literals, consisting of one literal at least. Let Π be a set of rules. The *theory of* Π is the set $\mathfrak{T}_\Pi$ inductively defined to contain all instances of literals from Π and all literals L for which there is a conjunction C of literals from $\mathfrak{T}_\Pi$ such that $L \Leftarrow C$ is an instance of a rule in Π. For $\mathfrak{L} \subseteq \mathfrak{T}_\Pi$, we also say that Π *derives* $\mathfrak{L}$, and write $\Pi \vdash \mathfrak{L}$. Π is *contradictory* if there is an atom A such that $\Pi \vdash \{A, \neg A\}$; otherwise Π is *non-contradictory*.

*Supported by the DFG grant Sto 421/5-1.

Throughout this paper, we will assume a set of literals and two sets of rules to be given: A set Π^{F} of literals meant to describe the *Facts* of the concrete situation under consideration, a set Π^{G} of *General rules* meant to hold in all possible worlds, and a set Δ of *defeasible* (or default) rules meant to hold in most situations. The set $\Pi := \Pi^{\mathrm{F}} \cup \Pi^{\mathrm{G}}$ is the set of *strict* rules that — contrary to the defeasible rules — are considered to be safe and are not doubted in any concrete situation. There is no difference in derivations between the strict rules from Π and the defeasible rules from Δ. If a contradiction occurs, however, we will narrow the defeasible rules from Δ down to a subset $\mathcal{A}$ of its *ground* instances, i.e. instances without free variables, such that no further instantiation can occur. Such a subset, together with the literal whose derivation is in focus, is called an *argument*. With implicit reference to the fixed sets of rules Π and Δ, we formally define:

Definition 2.1 (Argument)
$(\mathcal{A}, L)$ is an *argument* if $\mathcal{A}$ is a set of ground instances of rules from Δ and $\mathcal{A} \cup \Pi \vdash \{L\}$.

For ease of distinction, we will use the special symbol "$\leftarrow$" as a syntactical sugar in concrete examples of defeasible rules from Δ, instead of the symbol "$\Leftarrow$", which — in our concrete examples — will be used only in strict rules.

Example 2.2 (Poole 1985, Example 1)

$$\Pi^{\mathrm{F}}_{2.2} := \{\ \mathsf{bird}(\mathsf{tweety}),\ \mathsf{emu}(\mathsf{edna})\ \},$$
$$\Pi^{\mathrm{G}}_{2.2} := \left\{ \begin{array}{l} \mathsf{bird}(x) \Leftarrow \mathsf{emu}(x), \\ \neg\mathsf{flies}(x) \Leftarrow \mathsf{emu}(x) \end{array} \right\},$$
$$\Delta_{2.2} := \{\ \mathsf{flies}(x) \leftarrow \mathsf{bird}(x)\ \},$$
$$\mathcal{A}_2 := \{\ \mathsf{flies}(\mathsf{edna}) \leftarrow \mathsf{bird}(\mathsf{edna})\ \}.$$

We have $\mathfrak{T}_{\Pi_{2.2}} = \{\mathsf{bird}(\mathsf{tweety}), \mathsf{emu}(\mathsf{edna}), \mathsf{bird}(\mathsf{edna}), \neg\mathsf{flies}(\mathsf{edna})\}$, $\mathfrak{T}_{\Pi_{2.2} \cup \Delta_{2.2}} = \{\mathsf{flies}(\mathsf{edna}), \mathsf{flies}(\mathsf{tweety})\} \cup \mathfrak{T}_{\Pi_{2.2}}$. It is intuitively clear here that we prefer the argument $(\emptyset, \neg\mathsf{flies}(\mathsf{edna}))$ to the argument $(\mathcal{A}_2, \mathsf{flies}(\mathsf{edna}))$, simply because the former is more specific. We will further discuss this in Example 5.17.

We will use several binary relations comparing arguments according to their specificity. For any relation written as $\lesssim_N$ ("being more or equivalently specific w.r.t. N"), we set

$$\gtrsim_N \ := \ \{ (X, Y) \mid Y \lesssim_N X \} \quad \left(\text{"less or equivalently specific"}\right),$$
$$\approx_N \ := \ \lesssim_N \cap \gtrsim_N \quad (\text{"equivalently specific"}),$$
$$<_N \ := \ \lesssim_N \setminus \gtrsim_N \quad (\text{"properly more specific"}),$$
$$\leq_N \ := \ <_N \cup \{ (X, X) \mid X \text{ is an argument} \} \quad (\text{"more specific or equal"}),$$
$$\vartriangle_N \ := \ \left\{ (X, Y) \;\middle|\; \begin{array}{l} X, Y \text{ are arguments with} \\ X \not\lesssim_N Y \text{ and } X \not\gtrsim_N Y \end{array} \right\} \quad (\text{"incomparable w.r.t. specificity"}).$$

A *quasi-ordering* is a reflexive transitive relation. An *(irreflexive) ordering* is an irreflexive transitive relation. A *reflexive ordering* (also called: "partial ordering") is an anti-symmetric quasi-ordering. An *equivalence* is a symmetric quasi-ordering.

Corollary 2.3
If $\lesssim_N$ is a quasi-ordering, then $\approx_N$ is an equivalence, $<_N$ is an ordering, and $\leq_N$ is a reflexive ordering.

3 An Intuitive Notion of Specificity

It is part of general knowledge that a criterion is [properly] more specific than another one if the *class of candidates that satisfy it* is a [proper] subclass of that of the other one. Analogously — taking logical formulas as the criteria — a formula A is [properly] more specific than a formula B, if the model class of A is a [proper] subclass of the model class of B, i.e. if $A \models B$ [and $B \not\models A$].

To enable a closer investigation of the critical parts of a defeasible derivation, we have to isolate its defeasible parts. Abstracting from the concrete derivation of a literal L, let us take the set $\mathcal{A}$ of the ground instances of the defeasible rules that are actually applied in the derivation, and form the pair $(\mathcal{A}, L)$, which we already called an *argument* in Definition 2.1.

If we want to classify a derivation with defeasible rules according to its specificity, then we have to isolate the defeasible part of the derivation and look at its input. In our setting, the input consists of the set of those literals on which the defeasible part of the derivation is based, called the *activation set* for the defeasible part of the derivation. In our framework of defeasible logic programming, the only relevant property of an activation set can be the conjunction of its literals which is immediately represented by the set itself.

Because all literals of an activation set have been derived from the given specification, it does not make sense to compare activation sets w.r.t. the models of the entire specification. Indeed, only a comparison w.r.t. the models of a sub-specification can show any differences between them.

It is clear that we want to have the *entire* set Π^{G} for our comparison of activation sets, simply because we want to base our specificity classification on our specification, namely on its general and strict part. We have to exclude Π^{F} from this comparison, however. This exclusion makes sense because the defeasible rules are typically default rules not written in particular for the given concrete situation that is formalized by Π^{F}, and because the inclusion of Π^{F} would trivially admit every activation set. Moreover, as we want to compare the defeasible parts of derivations, we should exclude the defeasible rules from Δ from this comparison. We conclude that Π^{G} is that part of our specification according to which activation sets are to be compared.

Very roughly speaking, if we have fewer activation sets, then we have fewer models, which again means to have a higher specificity. Accordingly, a first sketch of a notion of specificity can now be given as follows:

> An argument $(\mathcal{A}_1, L_1)$ is [properly] *more specific than* an argument $(\mathcal{A}_2, L_2)$ if, for each activation set H_1 for $(\mathcal{A}_1, L_1)$, there is an activation set $H_2 \subseteq \mathfrak{T}_{H_1 \cup \Pi^{\mathrm{G}}}$ for $(\mathcal{A}_2, L_2)$ [but not vice versa].

Note that this notion of specificity is preliminary, and that the notion of an activation set for argument has not been properly defined yet.

On the one hand, the argument $(\mathcal{A}, L)$ is a nice abstraction from the derivation of L, because it perfectly suits our model-theoretic intentions described in Sect. 1. By this abstraction, on the other hand, we lose the possibility to isolate the defeasible parts of the derivation more precisely.

Let us compare this set $\mathcal{A}$ with an *and-tree of the derivation*. Every node in such a tree is labeled with the conclusion of an instance of a rule, such that its children are labeled exactly with the elements of the conjunction in the condition of this instance.

An isolation of the defeasible parts of an and-tree of the derivation may proceed as follows:

- Starting from the root of the tree, we iteratively erase all applications of strict rules. This results in a set of trees, each of which has the application of a defeasible rule at the root.
- Starting now from the leaves of these trees, we again erase all applications of strict rules. This results in a set of trees where all nodes *all* of whose children are leaves result from an application of a defeasible rule.

In a first approximation, we may now take all leaves of all resulting trees as the activation set for the original derivation.

Note that in the set of trees resulting from the described procedure, there may well have remained instances of rules from Π^{G} connecting a defeasible root application with the defeasible applications right at the leaves. Thus — to cover the whole defeasible part of the derivation in our abstraction — we have to consider the set $\mathcal{A} \cup \Pi^{\mathrm{G}}$ instead of just the set $\mathcal{A}$.

More precisely, we have to include all proper rules (i.e. those with non-empty conditions) from Π^{G}, and may also include the literals in Π^{G} because they cannot do any harm. As a consequence, in the modeling via our abstraction $\mathcal{A}$, we cannot prevent the precisely isolated defeasible sub-trees resulting from the described procedure from using the rules from Π^{G} to grow toward the root and toward the leaves again. It is clear, however, that only the growth toward the leaves can affect our activation sets and our notion of specificity.

Let us have a closer look at the effects of such a growth toward the leaves in the most simple case. In addition to a given activation set $\{\mathsf{Q}(\mathsf{a})\}$, in the presence of a general rule

$$\mathsf{Q}(x) \Leftarrow \mathsf{P}_0(x) \wedge \cdots \wedge \mathsf{P}_{n-1}(x)$$

from Π^{G}, we will also have to consider the activation set $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\ \}$. This has two effects.

The first effect is that we immediately realize that every model of Π^{G} that is represented by the activation set $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\ \}$ is also represented by the activation set $\{\mathsf{Q}(\mathsf{a})\}$; simply because $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\ \}$ is added to the activation sets by the growth toward the leaves, such that an argument based on $\{\mathsf{Q}(\mathsf{a})\}$ becomes obviously less (or equivalently) specific than any argument that gets along with $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\ \}$.

3.1 Preference of the "More Concise"

The second effect is that an argument that gets along with $\{\mathsf{Q}(\mathsf{a})\}$ becomes even *properly* less specific than one that actually requires $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\ \}$ and does not get along with $\{\mathsf{Q}(\mathsf{a})\}$, simply because the former argument has the additional activation set $\{\mathsf{Q}(\mathsf{a})\}$. This effect is called *preference of the "more concise",* cf. e.g. (Stolzenburg et al. 2003, p. 94), (García and Simari 2004, p. 108). The standard example for the preference of the "more concise" is the following minor variation of Example 2.2.

Example 3.1 (Poole 1985, Example 2)
$\Pi^{\mathrm{F}}_{3.1} := \Pi^{\mathrm{F}}_{2.2} \quad \Pi^{\mathrm{G}}_{3.1} := \{\ \mathsf{bird}(x) \Leftarrow \mathsf{emu}(x)\ \}$,
$\Delta_{3.1} := \left\{ \begin{array}{l} \neg\mathsf{flies}(x) \leftarrow \mathsf{emu}(x), \\ \mathsf{flies}(x) \leftarrow \mathsf{bird}(x) \end{array} \right\}$,
$\mathcal{A}_1 := \{\ \neg\mathsf{flies}(\mathsf{edna}) \leftarrow \mathsf{emu}(\mathsf{edna})\ \}$,
$\mathcal{A}_2 := \{\ \mathsf{flies}(\mathsf{edna}) \leftarrow \mathsf{bird}(\mathsf{edna})\ \}$.
The argument $(\mathcal{A}_2, \mathsf{flies}(\mathsf{edna}))$ gets along with the activation set $\{\mathsf{bird}(\mathsf{edna})\}$, and thus it is properly less specific than the argument $(\mathcal{A}_1, \neg\mathsf{flies}(\mathsf{edna}))$, which actually requires the stronger $\{\mathsf{emu}(\mathsf{edna})\}$.

The problem now is that the statement

$$\mathsf{Q}(\mathsf{a}) \not\models \mathsf{P}_0(\mathsf{a}) \wedge \cdots \wedge \mathsf{P}_{n-1}(\mathsf{a}),$$

which is required to justify the the appropriateness of this effect, is not explicitly given by the specification via $(\Pi^{\mathrm{F}}, \Pi^{\mathrm{G}}, \Delta)$.

Nevertheless — if we do not just want to see it as a matter-of-fact property of notions of specificity in the style of Poole — the preference of the "more concise" can be justified by the habits of human specifiers as follows: If human specifiers write an implication in form of a rule $\mathsf{Q}(x) \Leftarrow \mathsf{P}_0(x) \wedge \cdots \wedge \mathsf{P}_{n-1}(x)$ into a specification Π of strict (i.e. non-defeasible) knowledge, then they typically intend that the implication is proper in the sense that its converse does not hold in general; otherwise they would have used an equivalence or equality symbol instead of the implication symbol, or replaced each occurrence of each $\mathsf{Q}(t)$ with $\mathsf{P}_0(t) \wedge \cdots \wedge \mathsf{P}_{n-1}(t)$, respectively. In particular, in our setting of logic programming — where disjunctive properties of the definition of a predicate are spread over several rules — the implications definitely tend to be proper. Therefore, if seasoned specifiers write down such a rule, then they do not want to exclude models where Q holds for some object a, but not all of the P_i do. This means that if we find such a rule in the strict and general part Π^{G} of a specification, then it is reasonable to assume that the implication is proper w.r.t. the intuition captured in the defeasible rules in Δ.

Thus, it makes sense to consider a defeasible argument based on $\{\mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n-1\}\}$ to be properly more specific than an argument that can get along with $\mathsf{Q}(\mathsf{a})$.

Finally, let us remark that our justification for the preference of the "more concise" does not apply if

$$\mathsf{Q}(x) \Leftarrow \mathsf{P}_0(x) \wedge \cdots \wedge \mathsf{P}_{n-1}(x)$$

is a *defeasible* rule instead of a strict one, because we then have the following three problems:

- the inclusion given by the rule is not generally intended (otherwise the rule should be a strict one),
- we cannot easily describe the actual instances to which the default rule is meant to apply (otherwise this more concrete description of the defeasible rule should be stated as a strict rule), and
- the direct treatment of a defeasible equivalence neither has to be appropriate as a default rule in the given situation, nor do we have any means to express a defeasible equivalence in the current setting.

Example 3.2 (Poole 1985, Example 3, renamed)

$$\Pi^{\mathrm{F}}_{3.2} := \{\ \mathsf{emu}(\mathsf{edna})\ \}, \qquad \Pi^{\mathrm{G}}_{3.2} := \emptyset,$$
$$\Delta_{3.2} := \left\{ \begin{array}{l} \neg\mathsf{flies}(x) \leftarrow \mathsf{emu}(x), \\ \mathsf{flies}(x) \leftarrow \mathsf{bird}(x), \\ \mathsf{bird}(x) \leftarrow \mathsf{emu}(x) \end{array} \right\},$$
$$\mathcal{A}_1 := \{\ \neg\mathsf{flies}(\mathsf{edna}) \leftarrow \mathsf{emu}(\mathsf{edna})\ \},$$
$$\mathcal{A}_2 := \left\{ \begin{array}{l} \mathsf{flies}(\mathsf{edna}) \leftarrow \mathsf{bird}(\mathsf{edna}), \\ \mathsf{bird}(\mathsf{edna}) \leftarrow \mathsf{emu}(\mathsf{edna}) \end{array} \right\}.$$

According to the above discussion, there is no clear reason why we should consider the argument $(\mathcal{A}_2, \mathsf{flies}(\mathsf{edna}))$ to be *properly* less specific than the argument $(\mathcal{A}_1, \neg\mathsf{flies}(\mathsf{edna}))$.

3.2 Preference of the "More Precise"

By an analogous argumentation, we can say that an argument that essentially requires an activation set $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in \{0, \ldots, n\}\ \}$ is *properly* more specific than an argument that gets along with a proper subset $\{\ \mathsf{P}_i(\mathsf{a}) \mid i \in I\ \}$ for some index set $I \subsetneq \{0, \ldots, n\}$. The effect of the assumption of this intention is called *preference of the "more precise"*.

There is, however, an exception to be observed where this analogy does not apply, namely the case that we actually can derive the set from its subset with the help of Π^{G}. In this case, the before-mentioned growth toward the leaves with rules from Π^{G} again implements the approximation of the subclass relation among model classes via the one among activation sets. This is demonstrated in the following example, which also nicely shows that a notion of specificity based only on single defeasible rules (without considering the context of the strict rules as a whole) cannot be intuitively adequate.

Example 3.3 (Stolzenburg et al. 2003, p. 95)

$$\Pi^{\mathrm{F}}_{3.3} := \{\ \mathsf{q}(\mathsf{a})\ \}, \quad \Pi^{\mathrm{G}}_{3.3} := \{\ \mathsf{s}(x) \Leftarrow \mathsf{q}(x)\ \},$$
$$\Delta_{3.3} := \left\{ \begin{array}{l} \mathsf{p}(x) \leftarrow \mathsf{q}(x), \\ \neg\mathsf{p}(x) \leftarrow \mathsf{q}(x) \wedge \mathsf{s}(x) \end{array} \right\},$$
$$\mathcal{A}_1 := \{\ \neg\mathsf{p}(\mathsf{a}) \leftarrow \mathsf{q}(\mathsf{a}) \wedge \mathsf{s}(\mathsf{a})\ \},$$
$$\mathcal{A}_2 := \{\ \mathsf{p}(\mathsf{a}) \leftarrow \mathsf{q}(\mathsf{a})\ \}$$

The arguments $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a}))$ and $(\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$ are equivalently specific because the minimal activation set of each of them is $\{\mathsf{q}(\mathsf{a})\}$.

Apart from this exception, however, there is again a problem, namely that it is not the case that

$$\bigwedge_{i \in I} \mathsf{P}_i(\mathsf{a}) \not\models \bigwedge_{i \in \{0,\ldots,n\}} \mathsf{P}_i(\mathsf{a})$$

would be explicitly given by the specification via $(\Pi^{\mathrm{F}}, \Pi^{\mathrm{G}}, \Delta)$.

Nevertheless — if we do not just want to see preference of the "more precise" as a matter-of-fact property of notions of specificity in the style of Poole — we can again justify that it is unlikely that a seasoned specifier would not have intended this non-consequence statement, namely by an argumentation analogous to the one we gave for the preference of the "more concise". Indeed, a seasoned specifier who wants to exclude the above non-consequence would just specify a rule like $\mathsf{P}_j(x) \Leftarrow \bigwedge_{i \in I} \mathsf{P}_i(x)$, for each $j \in \{0, \ldots, n\} \backslash I$.

Example 3.4 *(continuing Example 3.3)*
For an example of preference of the "more precise" let us modify Example 3.3 by setting $\Pi^{\mathrm{F}}_{3.3} := \{\mathsf{q}(\mathsf{a}), \mathsf{s}(\mathsf{a})\}$ and $\Pi^{\mathrm{G}}_{3.3} := \emptyset$. Then $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a}))$ becomes properly more specific than $(\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$, because the latter argument has the additional activation set $\{\mathsf{q}(\mathsf{a})\}$.

4 Requirements Specification

With implicit reference to specification via $(\Pi^{\mathrm{F}}, \Pi^{\mathrm{G}}, \Delta)$, let us designate Poole's relation of being more (or equivalently) specific by "$\lesssim_{\mathrm{P1}}$". Here, "P1" stands for "Poole's original version".

The standard usage of the symbol "$\lesssim$" is to denote a *quasi-ordering* (cf. Sect. 2). Instead of the symbol "$\lesssim$", however, Poole (1985) uses the symbol "$\leq$". The standard usage of the symbol "$\leq$" is to denote a *reflexive ordering* (cf. Sect. 2). We cannot conclude from this, however, that Poole intended the additional property of anti-symmetry; indeed, we find a concrete example specification in (Poole 1985) where the lack of anti-symmetry of $\lesssim_{\mathrm{P1}}$ is made explicit (see last three sentences of Sect. 3.2 in (Poole 1985, p.145)).

The possible lack of anti-symmetry of quasi-orderings — i.e. that different arguments may have an equivalent specificity — cannot be a problem because any quasi-ordering $\lesssim_N$ immediately provides us with its equivalence $\approx_N$, its ordering $<_N$, and its reflexive ordering $\leq_N$ (cf. Corollary 2.3).

By contrast to the non-intended anti-symmetry, *transitivity* is obviously a *conditio sine qua non* for any useful notion of specificity. Indeed, if we already have an argument $(\mathcal{A}_2, \mathsf{wine})$ that is more specific than another argument $(\mathcal{A}_3, \mathsf{vodka})$, and if we come up with yet another argument $(\mathcal{A}_1, \mathsf{beer})$ that is even more specific than $(\mathcal{A}_2, \mathsf{wine})$, then, by all means, $(\mathcal{A}_1, \mathsf{beer})$ should be more specific than the argument $(\mathcal{A}_3, \mathsf{vodka})$ as well. It is obvious that a notion of specificity without transitivity could hardly be helpful in practice.

A further *conditio sine qua non* for any useful notion of specificity is that the conjunctive combination of respectively more specific arguments results in a more specific argument. Indeed, if a square is more specific than a rectangle and a circle is more specific than an ellipse, then a square inscribed into a circle should be more specific than a rectangle inscribed into an ellipse. This property is called monotonicity of conjunction, which we discuss in Sect. 6.2.

5 Formalizations of Specificity

A generative, bottom-up (i.e. from the leaves to the root) derivation with defeasible rules can now be split into three phases of derivation of literals from literals. This splitting follows the discussion in Sect. 3 on how to isolate the defeasible parts of a derivation (phase 2) from strict parts that may occur toward the root (phase 3) and toward the leaves (phase 1):

(phase 1) First we derive the literals that provide the basis for specificity considerations.

In our approach we derive the set $\mathfrak{T}_{\Pi}$ here. Poole takes the set $\mathfrak{T}_{\Pi\cup\Delta}$ instead.

(phase 2) On the basis of

- a subset H of the literals derived in phase 1,
- the first item $\mathcal{A}$ of a given argument $(\mathcal{A}, L)$, and
- the general rules Π^{G},

we derive a further set of literals $\mathfrak{L}$: $H\cup\mathcal{A}\cup\Pi^{\mathrm{G}}\vdash\mathfrak{L}$.

(phase 3) Finally, on the basis of $\mathfrak{L}$, the literal of the argument is derived: $\mathfrak{L}\cup\Pi \vdash \{L\}$.

In Poole's approach, phase 3 is empty and we simply have $\mathfrak{L}=\{L\}$. In our approach, however, it is admitted to use the facts from Π^{F} in phase 3, in addition to the general rules from Π^{G}, which were already admitted in phase 2.

With implicit reference to our sets $\Pi = \Pi^{\mathrm{F}} \cup \Pi^{\mathrm{G}}$ and Δ, the phases 2 and 3 can be more easily expressed with the help of the following notions.

Definition 5.1 (Activation Set)
Let $\mathcal{A}$ be a set of ground instances of rules from Δ, and let L be a literal. H is a *simplified activation set for* $(\mathcal{A}, L)$ if $L \in \mathfrak{T}_{H\,\cup\,\mathcal{A}\,\cup\,\Pi^{\mathrm{G}}}$. H is an *activation set for* $(\mathcal{A}, L)$ if $L \in \mathfrak{T}_{\mathfrak{L}\,\cup\,\Pi}$ for some $\mathfrak{L} \subseteq \mathfrak{T}_{H\,\cup\,\mathcal{A}\,\cup\,\Pi^{\mathrm{G}}}$. H is a *minimal* [*simplified*] *activation set for* $(\mathcal{A}, L)$ if H is an [simplified] activation set for $(\mathcal{A}, L)$, but no proper subset of H is an [simplified] activation set for $(\mathcal{A}, L)$.

Roughly speaking, an argument is now more (or equivalently) specific than another one if, for each of its activation sets H_1, the same set H_1 is also an activation set for the other argument. Note that we have replaced here the option of some $H_2 \subseteq \mathfrak{T}_{H_1\cup\Pi^{\mathrm{G}}}$ of the first straightforward sketch for a notion of specificity displayed in Sect. 3 with the more restrictive $H_2 = H_1$. Indeed, this simplification applies here because all we consider from any activation set H in Definition 5.1 (such as H_2 in this case) is just the closure $\mathfrak{T}_{H\,\cup\,\mathcal{A}\,\cup\,\Pi^{\mathrm{G}}} = \mathfrak{T}_{\mathfrak{T}_{H\cup\Pi^{\mathrm{G}}}\,\cup\,\mathcal{A}\,\cup\,\Pi^{\mathrm{G}}}$.

Activation sets that are not simplified differ from simplified ones by the admission of facts from Π^{F} (in addition to the general rules Π^{G}) after the defeasible part of the derivation is completed (cf. Example 6.6 for an occurrence of this difference).

Our introduction of activation sets that are not simplified is a conceptually important correction of Poole's approach: It must be admitted to use the facts besides the general rules in a purely strict derivation that is based on literals resulting from completed defeasible arguments, simply because the defeasible parts of a derivation (cf. Sect. 3) should not get more specific by the later use of additional facts that do not provide input to the defeasible parts.

5.1 Poole's Specificity Relations P1, P2, P3

In this section we will define the binary relations $\lesssim_{\mathrm{P1}}$, $\lesssim_{\mathrm{P2}}$, $\lesssim_{\mathrm{P3}}$ of "being more or equivalently specific according to David Poole" with implicit reference to our sets of facts and of general and defeasible rules (i.e. to Π^{F}, Π^{G}, and Δ, respectively).

The relation $\lesssim_{\mathrm{P1}}$ of the following definition is precisely Poole's original relation $\geq$ as defined at the bottom of the left column on page 145 of (Poole 1985). See Sect. 4 for our reasons to write "$\gtrsim$" instead of "$\geq$" as a first change. Moreover, as a second change required by mathematical standards, we have replaced the symbol "$\gtrsim$" with the symbol "$\lesssim$" (such that the smaller argument becomes the more specific one), so that the relevant well-foundedness becomes the one of its ordering $<$ instead of the reverse $>$.

Definition 5.2 $(\mathcal{A}_1, L_1) \lesssim_{\mathrm{P1}} (\mathcal{A}_2, L_2)$ if $(\mathcal{A}_1, L_1)$ and $(\mathcal{A}_2, L_2)$ are arguments, and if, for every $H \subseteq \mathfrak{T}_{\Pi\cup\Delta}$ that is a simplified activation set for $(\mathcal{A}_1, L_1)$ but not a simplified activation set for $(\mathcal{A}_2, L_1)$, H is also a simplified activation set for $(\mathcal{A}_2, L_2)$.

The relation $\lesssim_{\mathrm{P2}}$ of the following definition is the relation $\succeq$ of Definition 10 on page 94 of (Stolzenburg et al. 2003) (attributed to Poole 1985). Moreover, the relation $>_{\mathrm{spec}}$ of Definition 2.12 on page 132 of (Simari and Loui 1992) (attributed to (Poole 1985) as well) is the relation $<_{\mathrm{P2}} := \lesssim_{\mathrm{P2}} \setminus \gtrsim_{\mathrm{P2}}$.

Definition 5.3 $(\mathcal{A}_1, L_1) \lesssim_{\mathrm{P2}} (\mathcal{A}_2, L_2)$ if $(\mathcal{A}_1, L_1)$ and $(\mathcal{A}_2, L_2)$ are arguments, and if, for every $H \subseteq \mathfrak{T}_{\Pi\cup\Delta}$ that is a simplified activation set for $(\mathcal{A}_1, L_1)$ but not a simplified activation set for $(\emptyset, L_1)$, H is also a simplified activation set for $(\mathcal{A}_2, L_2)$.

The only change in Definition 5.3 as compared to Definition 5.2 is that "$(\mathcal{A}_2, L_1)$" is replaced with "$(\emptyset, L_1)$". We did not encounter any example yet where this most appropriate correction of the counter-intuitive variant "$(\mathcal{A}_2, L_1)$" of Definition 5.2 makes any difference to today's standard "$(\emptyset, L_1)$" in Definition 5.3.

The relations $\lesssim_{\mathrm{P1}}$ and $\lesssim_{\mathrm{P2}}$ were not meant to compare arguments for literals that do not need any defeasible rules — or at least they do not show an intuitive behavior on such arguments, as shown in Example 5.6 (right after the next definition and its corollary).

To overcome this minor flaw, we finally add an implication as an additional requirement in Definition 5.4. This implication guarantees that no argument that requires defeasible rules can be more specific than an argument not requiring any defeasible rules at all.

Definition 5.4 $(\mathcal{A}_1, L_1) \lesssim_{\mathrm{P3}} (\mathcal{A}_2, L_2)$ if $(\mathcal{A}_1, L_1)$ and $(\mathcal{A}_2, L_2)$ are arguments, $L_2 \in \mathfrak{T}_\Pi$ implies $L_1 \in \mathfrak{T}_\Pi$, and if, for every $H \subseteq \mathfrak{T}_{\Pi\cup\Delta}$ that is a [minimal] simplified activation set for $(\mathcal{A}_1, L_1)$ but not a simplified activation set for $(\emptyset, L_1)$, H is also a simplified activation set for $(\mathcal{A}_2, L_2)$.

Trivially, $\lesssim_{\mathrm{P3}} \subseteq \lesssim_{\mathrm{P2}}$. As every simplified activation set that passes the condition of Definition 5.2 passes the one of Definition 5.3, we have:

Corollary 5.5 $\lesssim_{\mathrm{P3}} \subseteq \lesssim_{\mathrm{P2}} \subseteq \lesssim_{\mathrm{P1}}$.

Example 5.6 (Minor Flaw of $\lesssim_{\mathrm{P1}}$ and $\lesssim_{\mathrm{P2}}$)
$\Pi^{\mathrm{F}}_{5.6} := \{\ \mathsf{thirst}\ \}$, $\Pi^{\mathrm{G}}_{5.6} := \{\ \mathsf{drink} \Leftarrow \mathsf{thirst}\ \}$,
$\Delta_{5.6} := \{\ \mathsf{beer} \leftarrow \mathsf{thirst}\ \}$, $\mathcal{A}_1 := \Delta_{5.6}$.
Let us compare the specificity of the arguments $(\mathcal{A}_1, \mathsf{beer})$ and $(\emptyset, \mathsf{drink})$. $\mathfrak{T}_{\Pi_{5.6}} = \{\mathsf{thirst}, \mathsf{drink}\}$, $\mathfrak{T}_{\Pi_{5.6}\cup\Delta_{5.6}} = \{\mathsf{beer}\} \cup \mathfrak{T}_{\Pi_{5.6}}$.
We have $(\mathcal{A}_1, \mathsf{beer}) \lesssim_{\mathrm{P2}} (\emptyset, \mathsf{drink})$ because for every $H \subseteq \mathfrak{T}_{\Pi_{5.6}\cup\Delta_{5.6}}$ that is a simplified activation set for $(\mathcal{A}_1, \mathsf{beer})$, but not a simplified activation set for $(\emptyset, \mathsf{beer})$, we have $H = \{\mathsf{thirst}\}$, which is a simplified activation set also for $(\emptyset, \mathsf{drink})$.
We have $(\emptyset, \mathsf{drink}) \lesssim_{\mathrm{P2}} (\mathcal{A}_1, \mathsf{beer})$ because there cannot be a simplified activation set for $(\emptyset, \mathsf{drink})$ that is not a simplified activation set for $(\emptyset, \mathsf{drink})$.
All in all, we get $(\mathcal{A}_1, \mathsf{beer}) \approx_{\mathrm{P2}} (\emptyset, \mathsf{drink})$, although $(\emptyset, \mathsf{drink}) <_{\mathrm{P3}} (\mathcal{A}_1, \mathsf{beer})$ should be given according to intuition, because, if beer produces a conflict with our drinking habits, there is no reason to prefer it to another drink. Finally note that by Corollary 5.5, we get $(\mathcal{A}_1, \mathsf{beer}) \approx_{\mathrm{P1}} (\emptyset, \mathsf{drink})$ as well.

Corollary 5.7 *If* $(\mathcal{A}_1, L_1)$, $(\mathcal{A}_2, L_2)$ *are arguments and we have* $\mathcal{A}_1 \subseteq \mathcal{A}_2$ *and* $L_1 = L_2$, *then we have* $(\mathcal{A}_1, L_1) \lesssim_{\mathrm{P3}} (\mathcal{A}_2, L_2)$.

By Corollaries 5.5 and 5.7, $\lesssim_{\mathrm{P1}}, \lesssim_{\mathrm{P2}}, \lesssim_{\mathrm{P3}}$ are reflexive relations, but — as we will show in Example 5.8 and state in Theorem 5.10 — not quasi-orderings in general.

Example 5.8 (Counterexample to Transitivities)
$\Pi^{\mathrm{F}}_{5.8} := \{\ \mathsf{alcohol},\ \mathsf{blessing},\ \mathsf{thirst}\ \}$,
$\Pi^{\mathrm{G}}_{5.8} := \{\ \mathsf{wine} \Leftarrow \mathsf{e}\ \}$, $\Delta_{5.8} := \mathcal{A}_1 \cup \mathcal{A}_2 \cup \mathcal{A}_3$,

$$\mathcal{A}_1 := \left\{ \begin{array}{l} \mathsf{e} \leftarrow \mathsf{alcohol} \wedge \mathsf{blessing} \wedge \mathsf{thirst}, \\ \mathsf{beer} \leftarrow \mathsf{e} \end{array} \right\},$$

$\mathcal{A}_2 := \{\ \mathsf{wine} \leftarrow \mathsf{alcohol} \wedge \mathsf{blessing}\ \}$,
$\mathcal{A}_3 := \{\ \mathsf{vodka} \leftarrow \mathsf{alcohol}\ \}$.
Compare the arguments $(\mathcal{A}_1, \mathsf{beer})$, $(\mathcal{A}_2, \mathsf{wine})$, and $(\mathcal{A}_3, \mathsf{vodka})$!

Lemma 5.9 *There are a specification* $(\Pi^{\mathrm{F}}, \Pi^{\mathrm{G}}, \Delta)$ *without any negative literals (i.e.* $\Pi^{\mathrm{F}} \cup \Pi^{\mathrm{G}} \cup \Delta$ *is non-contradictory), and arguments* $(\mathcal{A}_1, L_1)$, $(\mathcal{A}_2, L_2)$, $(\mathcal{A}_3, L_3)$ *with respective minimal sets* $\mathcal{A}_1$, $\mathcal{A}_2$, $\mathcal{A}_3$ *(i.e.* $(\mathcal{A}'_i, L_i)$ *is not an argument for any proper subset* $\mathcal{A}'_i \subsetneq \mathcal{A}_i$*), such that*
$(\mathcal{A}_1, L_1) \lesssim_{\mathrm{P3}} (\mathcal{A}_2, L_2) \lesssim_{\mathrm{P3}} (\mathcal{A}_3, L_3) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_1, L_1)$ *and*
$(\mathcal{A}_1, L_1) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_2, L_2) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_3, L_3)$.

Proof of Lemma 5.9 Looking at Example 5.8, we see that only the quasi-ordering properties in the last two lines of Lemma 5.9 are non-trivial.
$\mathfrak{T}_{\Pi_{5.8}} = \{\mathsf{alcohol}, \mathsf{blessing}, \mathsf{thirst}\}$, $\mathfrak{T}_{\Pi_{5.8}\cup\Delta_{5.8}} = \{\mathsf{e}, \mathsf{beer}, \mathsf{wine}, \mathsf{vodka}\} \cup \mathfrak{T}_{\Pi_{5.8}}$. Thus, regarding the arguments $(\mathcal{A}_1, \mathsf{beer})$, $(\mathcal{A}_2, \mathsf{wine})$, $(\mathcal{A}_3, \mathsf{vodka})$, the additional implication condition of Definition 5.4 as compared to Definitions 5.2 and 5.3 is always satisfied, simply because its condition is always false.
<u>$(\mathcal{A}_3, \mathsf{vodka}) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_1, \mathsf{beer}) \lesssim_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{wine})$:</u> The minimal simplified activation sets for $(\mathcal{A}_1, \mathsf{beer})$ that are subsets of $\mathfrak{T}_{\Pi_{5.8}\cup\Delta_{5.8}}$ and no simplified activation sets for $(\emptyset, \mathsf{beer})$ (or, without any difference, for $(\mathcal{A}_3, \mathsf{beer})$) are $\{\mathsf{alcohol}, \mathsf{blessing}, \mathsf{thirst}\}$ and $\{\mathsf{e}\}$, which are simplified activation sets for $(\mathcal{A}_2, \mathsf{wine})$ — but $\{\mathsf{e}\}$ is no simplified activation set for $(\mathcal{A}_3, \mathsf{vodka})$.
<u>$(\mathcal{A}_1, \mathsf{beer}) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_2, \mathsf{wine}) \lesssim_{\mathrm{P3}} (\mathcal{A}_3, \mathsf{vodka})$:</u> The only minimal simplified activation set for $(\mathcal{A}_2, \mathsf{wine})$ that is a subset of $\mathfrak{T}_{\Pi_{5.8}\cup\Delta_{5.8}}$ and no simplified activation set for $(\emptyset, \mathsf{wine})$ (such as $\{\mathsf{e}\}$) (or, without any difference, for $(\mathcal{A}_1, \mathsf{wine})$) is $\{\mathsf{alcohol}, \mathsf{blessing}\}$, which is a simplified activation set for $(\mathcal{A}_3, \mathsf{vodka})$, but not for $(\mathcal{A}_2, \mathsf{beer})$.
<u>$(\mathcal{A}_2, \mathsf{wine}) \not\gtrsim_{\mathrm{P1}} (\mathcal{A}_3, \mathsf{vodka})$:</u> The only minimal simplified activation set for $(\mathcal{A}_3, \mathsf{vodka})$ that is a subset of $\mathfrak{T}_{\Pi_{5.8}\cup\Delta_{5.8}}$ and no simplified activation set for $(\mathcal{A}_2, \mathsf{vodka})$ is $\{\mathsf{alcohol}\}$, which is not a simplified activation set for $(\mathcal{A}_2, \mathsf{wine})$. **Q.e.d. (Lemma 5.9)**

The relations stated in Lemma 5.9 hold not only for the given indices, but — by Corollary 5.5 — actually for all of P1, P2, P3; and so we immediately get:

Theorem 5.10 *There is a specification* $(\Pi^{\mathrm{F}}, \Pi^{\mathrm{G}}, \Delta)$*, such that* $\Pi^{\mathrm{F}} \cup \Pi^{\mathrm{G}} \cup \Delta$ *is non-contradictory, but none of* $\lesssim_{\mathrm{P1}}$, $\lesssim_{\mathrm{P2}}$, $\lesssim_{\mathrm{P3}}$, $<_{\mathrm{P1}}$, $<_{\mathrm{P2}}$, $<_{\mathrm{P3}}$ *is transitive.*

Note that in the given counterexample to transitivity (Example 5.8) all arguments have minimal sets of ground instances of defeasible rules.

As a consequence of Theorem 5.10, the respective relations in (Simari and Loui 1992) and (Stolzenburg et al. 2003) are not transitive. This means that these relations are not quasi-orderings, let alone reflexive orderings. See (Wirth and Stolzenburg 2013, Sect. 5.2) for details.

5.2 Our Novel Specificity Ordering CP1

In the previous section, we have seen that *minor corrections* of Poole's original specificity relation P1 (such as P2, P3) do not cure the (up to our finding of Example 5.8) hidden and even denied formal deficiency of these relations, namely their lack of transitivity. Therefore, in this section, we now define our major correction of Poole's specificity — the binary relation $\lesssim_{\mathrm{CP1}}$ — with implicit reference to our sets of facts and of general and defeasible rules (i.e. to Π^{F}, Π^{G}, and Δ, respectively) as follows.

Definition 5.11 (Our Version of Specificity $\lesssim_{\rm CP1}$)
$(\mathcal{A}_1, L_1) \lesssim_{\rm CP1} (\mathcal{A}_2, L_2)$ if $(\mathcal{A}_1, L_1)$ and $(\mathcal{A}_2, L_2)$ are arguments, and we have
1. $L_1 \in \mathfrak{T}_\Pi$ or
2. $L_2 \notin \mathfrak{T}_\Pi$ and every $H \subseteq \mathfrak{T}_\Pi$ that is an [minimal] activation set for $(\mathcal{A}_1, L_1)$ is also an activation set for $(\mathcal{A}_2, L_2)$.

The crucial change in Definition 5.11 as compared to Definition 5.4 is not the merely technical emphasis it puts on the case "$L_1 \in \mathfrak{T}_\Pi$", which has no effect on the extension of the relation as compared to $\lesssim_{\rm P3}$. The crucial changes actually are

- the replacement of "$H \subseteq \mathfrak{T}_{\Pi\cup\Delta}$" with "$H \subseteq \mathfrak{T}_\Pi$", and the thereby enabled
- omission of the previously technically required, but unintuitive negative condition on derivability ("but not a simplified activation set for $(\emptyset, L_1)$").

An additional minor change, which we have already discussed in Sect. 5, is the one from simplified to (non-simplified) activation sets.

Corollary 5.12 *If $(\mathcal{A}_1, L_1)$, $(\mathcal{A}_2, L_2)$ are arguments and we have $\mathcal{A}_1 \subseteq \mathcal{A}_2$ and $L_1 = L_2$, then we have $(\mathcal{A}_1, L_1) \lesssim_{\rm CP1} (\mathcal{A}_2, L_2)$.*

Theorem 5.13
$\lesssim_{\rm CP1}$ is a quasi-ordering on arguments.

The proof of Theorem 5.13 is straightforward and can be found in (Wirth and Stolzenburg 2013, Sect. 5.3, Proof of Theorem 5.13).

Example 5.14 (continuing Example 5.8)
The following holds for our specification of Example 5.8 by Lemma 5.9 and Corollary 5.5:
$(\mathcal{A}_1, \mathsf{beer}) <_{\rm P3} (\mathcal{A}_2, \mathsf{wine}) <_{\rm P3} (\mathcal{A}_3, \mathsf{vodka}) \not\gtrsim_{\rm P3} (\mathcal{A}_1, \mathsf{beer})$. We have now:
$(\mathcal{A}_1, \mathsf{beer}) <_{\rm CP1} (\mathcal{A}_2, \mathsf{wine}) <_{\rm CP1} (\mathcal{A}_3, \mathsf{vodka}) >_{\rm CP1} (\mathcal{A}_1, \mathsf{beer})$, simply because the trouble-making set $\{\mathsf{e}\}$ is not to be considered: it is not a subset of $\mathfrak{T}_{\Pi_{5.8}}$!

Obviously, an argument is ranked by $\lesssim_{\rm CP1}$ firstly on whether its literal is in $\mathfrak{T}_\Pi$, and, if not, secondly on the set of its activation sets, which is an element of the power set of the power set of $\mathfrak{T}_\Pi$. So we get:

Corollary 5.15
If $\mathfrak{T}_\Pi$ is finite, then $<_{\rm CP1}$ is well-founded.

Theorem 5.16 *$\lesssim_{\rm P3} \subseteq \lesssim_{\rm CP1}$ holds if, for each instance $L \Leftarrow L'_0 \wedge \ldots \wedge L'_n$ with $n \geq 1$ of each rule in $\Pi^{\rm G}$, we have $L'_j \notin \mathfrak{T}_\Pi$ for all $j \in \{0, \ldots, n\}$.*

Please find the lengthy proof of Theorem 5.16 in (Wirth and Stolzenburg 2013, Sect. 5.4, Theorem 5.15). Here, however, is a typical application of it:

Example 5.17 (continuing Example 2.2)
We have $(\mathcal{A}_2, \mathsf{flies(edna)}) \not\lesssim_{\rm CP1} (\emptyset, \neg\mathsf{flies(edna)})$ because $\mathsf{flies(edna)} \notin \mathfrak{T}_{\Pi_{2.2}}$ and $\neg\mathsf{flies(edna)} \in \mathfrak{T}_{\Pi_{2.2}}$.
We have $(\emptyset, \neg\mathsf{flies(edna)}) \lesssim_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$, because $\neg\mathsf{flies(edna)} \in \mathfrak{T}_{\Pi_{2.2}}$ and because the premise of the last condition in Definition 5.4 is contradictory for $\mathcal{A}_1 := \emptyset$, and cannot be satisfied by any set $H \subseteq \mathfrak{T}_{\Pi_{2.2}\cup\Delta_{2.2}}$. All in all, by Theorem 5.16
we get $(\emptyset, \neg\mathsf{flies(edna)}) <_{\rm CP1} (\mathcal{A}_2, \mathsf{flies(edna)})$
and $(\emptyset, \neg\mathsf{flies(edna)}) <_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$.

6 Specificity and Human Intuition

Let us now put the two notions of specificity — as formalized in the two binary relations $\lesssim_{\rm P3}$ and $\lesssim_{\rm CP1}$ — to test w.r.t. our changed phase 1 of Sect. 5 in a series of classical examples.

6.1 Preference of the "More Concise"

Example 6.1 *(continuing Example 3.1)*
Let us compare the arguments $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$ and $(\mathcal{A}_2, \mathsf{flies(edna)})$. $\mathfrak{T}_{\Pi_{3.1}} = \{\mathsf{bird(tweety)}, \mathsf{emu(edna)}, \mathsf{bird(edna)}\}$, $\mathfrak{T}_{\Pi_{3.1}\cup\Delta_{3.1}} = \{\neg\mathsf{flies(edna)}, \mathsf{flies(edna)}, \mathsf{flies(tweety)}\} \cup \mathfrak{T}_{\Pi_{3.1}}$. We have
$(\mathcal{A}_2, \mathsf{flies(edna)}) \not\lesssim_{\rm CP1} (\mathcal{A}_1, \neg\mathsf{flies(edna)})$
because $\mathsf{flies(edna)} \notin \mathfrak{T}_{\Pi_{3.1}}$ and because $\{\mathsf{bird(edna)}\} \subseteq \mathfrak{T}_{\Pi_{3.1}}$ is an activation set for $(\mathcal{A}_2, \mathsf{flies(edna)})$, but not for $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$. We have
$(\mathcal{A}_1, \neg\mathsf{flies(edna)}) \lesssim_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$,
because $\mathsf{flies(edna)} \notin \mathfrak{T}_{\Pi_{3.1}}$ and, if $H \subseteq \mathfrak{T}_{\Pi_{3.1}\cup\Delta_{3.1}}$ is a simplified activation set for $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$, but not for $(\emptyset, \neg\mathsf{flies(edna)})$, then we have $\mathsf{emu(edna)} \in H$, and thus H is a simplified activation set also for $(\mathcal{A}_2, \mathsf{flies(edna)})$. By Theorem 5.16, we get $(\mathcal{A}_1, \neg\mathsf{flies(edna)}) <_{\rm CP1} (\mathcal{A}_2, \mathsf{flies(edna)})$
and $(\mathcal{A}_1, \neg\mathsf{flies(edna)}) <_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$.

Example 6.2 *(continuing Example 3.2)*
Let us compare the arguments $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$ and $(\mathcal{A}_2, \mathsf{flies(edna)})$. $\mathfrak{T}_{\Pi_{3.2}} = \{\mathsf{emu(edna)}\}$, $\mathfrak{T}_{\Pi_{3.2}\cup\Delta_{3.2}} = \{\mathsf{bird(edna)}, \mathsf{flies(edna)}, \neg\mathsf{flies(edna)}\} \cup \mathfrak{T}_{\Pi_{3.2}}$.
We have $(\mathcal{A}_2, \mathsf{flies(edna)}) \lesssim_{\rm CP1} (\mathcal{A}_1, \neg\mathsf{flies(edna)})$ because $\neg\mathsf{flies(edna)} \notin \mathfrak{T}_{\Pi_{3.2}}$ and, for every activation set $H \subseteq \mathfrak{T}_{\Pi_{3.2}}$ for $(\mathcal{A}_2, \mathsf{flies(edna)})$, we get $\mathsf{emu(edna)} \in H$, so H is an activation set also for $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$.
$(\mathcal{A}_2, \mathsf{flies(edna)}) \not\lesssim_{\rm P3} (\mathcal{A}_1, \neg\mathsf{flies(edna)})$ is still the case because $\{\mathsf{bird(edna)}\} \subseteq \mathfrak{T}_{\Pi_{3.2}\cup\Delta_{3.2}}$ is a simplified activation set for $(\mathcal{A}_2, \mathsf{flies(edna)})$, but neither for $(\emptyset, \mathsf{flies(edna)})$, nor for $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$.
We have $(\mathcal{A}_1, \neg\mathsf{flies(edna)}) \lesssim_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$, because of $\mathsf{flies(edna)} \notin \mathfrak{T}_{\Pi_{3.2}}$ and because, if $H \subseteq \mathfrak{T}_{\Pi_{3.2}\cup\Delta_{3.2}}$ is a simplified activation set for $(\mathcal{A}_1, \neg\mathsf{flies(edna)})$, but not for $(\emptyset, \neg\mathsf{flies(edna)})$, then we have $\mathsf{emu(edna)} \in H$ and thus H is a simplified activation set also for $(\mathcal{A}_2, \mathsf{flies(edna)})$.
All in all, by Theorem 5.16, this time
we get $(\mathcal{A}_1, \neg\mathsf{flies(edna)}) \approx_{\rm CP1} (\mathcal{A}_2, \mathsf{flies(edna)})$
and $(\mathcal{A}_1, \neg\mathsf{flies(edna)}) <_{\rm P3} (\mathcal{A}_2, \mathsf{flies(edna)})$.
From a conceptual point of view, we have to ask ourselves, whether we would like a *defeasible* rule instance such as $\mathsf{bird(edna)} \leftarrow \mathsf{emu(edna)}$ to reduce the specificity of $\mathcal{A}_2$ as compared to a system that seems equivalent for the given argument for $\mathsf{flies(edna)}$, namely the argument
$(\{\mathsf{flies(edna)} \leftarrow \mathsf{emu(edna)}\}, \mathsf{flies(edna)})$?
Does the specificity of a defeasible reasoning step really reduce if we introduce intermediate literals?

According to human intuition, this question has a negative answer, as we have already explained at the end of Sect. 3.1. Moreover, Example 6.3 will exhibit a strong reason to deny it.

Finally, see Example 6.5 for another example that makes even clearer why defeasible rules should be considered for their global semantical effect instead of their syntactical fine structure.

6.2 Monotonicity w.r.t. Conjunction

Monotonicity w.r.t. conjunction means for a quasi-ordering $\lesssim_N$ that, in case of $(\mathcal{A}_1^i, L_1^i) \lesssim_N (\mathcal{A}_2^i, L_2^i)$ for $i \in \{1,2\}$, we always have $(\mathcal{A}_1^1 \cup \mathcal{A}_1^2, L_1') \lesssim_N (\mathcal{A}_2^i \cup \mathcal{A}_2^i, L_2')$ for fresh constant literals L_j' with additional general rules $L_j' \Leftarrow L_j^1 \wedge L_j^2 \in \Pi^{\mathrm{G}}$ $(j \in \{1,2\})$. For $\lesssim_{\mathrm{CP1}}$, this property is trivially given in case of $L_1^1, L_1^2 \in \mathfrak{T}_\Pi$, but cannot be expected in case of $L_1^i \notin \mathfrak{T}_\Pi \ni L_1^{3-i}$ (for some $i \in \{1,2\}$), simply because then we get $L_1' \notin \mathfrak{T}_\Pi$. Also for the only remaining and most interesting case of $L_1^1, L_1^2 \notin \mathfrak{T}_\Pi$, this property is obviously given. For $\lesssim_{\mathrm{P1}}$, however, monotonicity is not even given in this most interesting case, as already noted in (Poole 1985):

Example 6.3 (Poole 1985, Example 6)

$\Pi^{\mathrm{F}}_{6.3} := \left\{ \begin{array}{l} \mathsf{a}, \\ \mathsf{d} \end{array} \right\}, \quad \Pi^{\mathrm{G}}_{6.3} := \left\{ \begin{array}{l} \mathsf{g}_1 \Leftarrow \neg\mathsf{c} \wedge \neg\mathsf{f}, \\ \mathsf{g}_2 \Leftarrow \mathsf{c} \wedge \mathsf{f} \end{array} \right\},$

$\Delta_{6.3} := \mathcal{A}_1 \cup \mathcal{A}_2, \quad \mathcal{A}_1 := \{\ \neg\mathsf{c} \leftarrow \mathsf{a},\ \neg\mathsf{f} \leftarrow \mathsf{d}\ \},$

$\mathcal{A}_2 \ := \{\ \mathsf{b} \leftarrow \mathsf{a},\ \mathsf{c} \leftarrow \mathsf{b},\ \mathsf{e} \leftarrow \mathsf{d},\ \mathsf{f} \leftarrow \mathsf{e}\ \}$

Let us compare the specificity of the arguments $(\mathcal{A}_1, \mathsf{g}_1)$ and $(\mathcal{A}_2, \mathsf{g}_2)$. $\mathfrak{T}_{\Pi_{6.3}} = \{\mathsf{a}, \mathsf{d}\}$. $\mathfrak{T}_{\Pi_{6.3} \cup \Delta_{6.3}} = \{\mathsf{b}, \mathsf{c}, \mathsf{e}, \mathsf{f}, \mathsf{g}_1, \mathsf{g}_2, \neg\mathsf{c}, \neg\mathsf{f}\} \cup \mathfrak{T}_{\Pi_{6.3}}$. $(\mathcal{A}_1, \mathsf{g}_1) \approx_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{g}_2)$ because $H \subseteq \mathfrak{T}_{\Pi_{6.3}}$ is an activation set for $(\mathcal{A}_i, \mathsf{g}_i)$ if and only if $H = \{\mathsf{a}, \mathsf{d}\}$. We have $(\mathcal{A}_1, \mathsf{g}_1) \triangle_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{g}_2)$: $\{\mathsf{a}, \neg\mathsf{f}\} \subseteq \mathfrak{T}_{\Pi_{6.3} \cup \Delta_{6.3}}$ is a simplified activation set for $(\mathcal{A}_1, \mathsf{g}_1)$, but neither for $(\emptyset, \mathsf{g}_1)$, nor for $(\mathcal{A}_2, \mathsf{g}_2)$. $\{\mathsf{a}, \mathsf{f}\} \subseteq \mathfrak{T}_{\Pi_{6.3} \cup \Delta_{6.3}}$ is a simplified activation set for $(\mathcal{A}_2, \mathsf{g}_2)$, but neither for $(\emptyset, \mathsf{g}_2)$, nor for $(\mathcal{A}_1, \mathsf{g}_1)$.

In (Poole 1985), the same result for $\lesssim_{\mathrm{P1}}$ is described as "seemingly unintuitive", because, as we have seen in the isomorphic sub-specification of Example 3.2, we have both $(\mathcal{A}_1, \neg\mathsf{c}) <_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{c})$ and $(\mathcal{A}_1, \neg\mathsf{f}) <_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{f})$. Indeed, as already listed as an essential requirement in Sect. 4, the conjunction of two respectively more specific derivations should be more specific. On the other hand, considering $\lesssim_{\mathrm{CP1}}$ instead of $\lesssim_{\mathrm{P3}}$, the conjunction of two equivalently specific derivations results in an equivalently specific derivation — exactly as one intuitively expects.

Example 6.4 (1st Variation of Example 6.3)

$\Pi^{\mathrm{F}}_{6.4} := \Pi^{\mathrm{F}}_{6.3}, \qquad \mathcal{A}_1 := \{\ \neg\mathsf{c} \leftarrow \mathsf{a},\ \neg\mathsf{f} \leftarrow \mathsf{d}\ \},$

$\Pi^{\mathrm{G}}_{6.4} := \{\ \mathsf{g}_1 \Leftarrow \neg\mathsf{c} \wedge \neg\mathsf{f},\ \mathsf{g}_2 \Leftarrow \mathsf{c} \wedge \mathsf{f},\ \mathsf{b} \Leftarrow \mathsf{a}\ \},$

$\Delta_{6.4} := \mathcal{A}_1 \cup \mathcal{A}_2, \quad \mathcal{A}_2 := \{\mathsf{c} \leftarrow \mathsf{b},\ \mathsf{e} \leftarrow \mathsf{d},\ \mathsf{f} \leftarrow \mathsf{e}\}.$

Let us compare the specificity of the arguments $(\mathcal{A}_1, \mathsf{g}_1)$ and $(\mathcal{A}_2, \mathsf{g}_2)$. $\mathfrak{T}_{\Pi_{6.4}} = \{\mathsf{a}, \mathsf{b}, \mathsf{d}\}$. $\mathfrak{T}_{\Pi_{6.4} \cup \Delta_{6.4}} = \{\mathsf{c}, \mathsf{e}, \mathsf{f}, \mathsf{g}_1, \mathsf{g}_2, \neg\mathsf{c}, \neg\mathsf{f}\} \cup \mathfrak{T}_{\Pi_{6.4}}$. $(\mathcal{A}_2, \mathsf{g}_2) \not\lesssim_{\mathrm{CP1}} (\mathcal{A}_1, \mathsf{g}_1)$ because $\{\mathsf{b}, \mathsf{d}\} \subseteq \mathfrak{T}_{\Pi_{6.4}}$ is an activation set for $(\mathcal{A}_2, \mathsf{g}_2)$, but not for $(\mathcal{A}_1, \mathsf{g}_1)$. $(\mathcal{A}_1, \mathsf{g}_1) \lesssim_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{g}_2)$ because, for any activation set $H \subseteq \mathfrak{T}_{\Pi_{6.4}}$ for $(\mathcal{A}_1, \mathsf{g}_1)$, we have $\{\mathsf{a}, \mathsf{b}\} \subseteq H$; so H is also an activation set for $(\mathcal{A}_2, \mathsf{g}_2)$.

Again $(\mathcal{A}_1, \mathsf{g}_1) \triangle_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{g}_2)$, for the same reason as in Example 6.3. Thus, the situation for $\lesssim_{\mathrm{P3}}$ is just as in Example 6.3, and just as "seemingly unintuitive" for exactly the same reason.

We have $(\mathcal{A}_1, \mathsf{g}_1) <_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{g}_2)$, which is intuitive because the conjunction of a more specific and an equivalently specific element, respectively, should be more specific. Indeed, from the isomorphic sub-specifications in Examples 3.1 and 3.2, we know that $(\mathcal{A}_1, \neg\mathsf{c}) <_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{c})$ and $(\mathcal{A}_1, \neg\mathsf{f}) \approx_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{f})$, resp.

All in all, $\lesssim_{\mathrm{P3}}$ fails in this example again, whereas $\lesssim_{\mathrm{CP1}}$ satisfies the monotonicity w.r.t. conjunction required in Sect. 4.

6.3 Preference of the "More Precise"

As primary sources of differences in specificity, the previous examples illustrated the effect of a chain of implications. We now consider examples where the primary source is an essentially required condition that is a super-conjunction of the condition of another rule.

Example 6.5 (2nd Variation of Example 6.3)

$\Pi^{\mathrm{F}}_{6.5} := \Pi^{\mathrm{F}}_{6.3}, \quad \Pi^{\mathrm{G}}_{6.5} := \{\ \mathsf{g}_1 \Leftarrow \neg\mathsf{c},\ \mathsf{g}_2 \Leftarrow \mathsf{c} \wedge \mathsf{f}\ \},$

$\Delta_{6.5} := \mathcal{A}_1 \cup \mathcal{A}_2, \quad \mathcal{A}_1 := \{\ \neg\mathsf{c} \leftarrow \mathsf{a}\ \},$

$\mathcal{A}_2 \ := \{\ \mathsf{b} \leftarrow \mathsf{a},\ \mathsf{c} \leftarrow \mathsf{b},\ \mathsf{e} \leftarrow \mathsf{d},\ \mathsf{f} \leftarrow \mathsf{e}\ \}$

Let us compare the specificity of the arguments $(\mathcal{A}_1, \mathsf{g}_1)$ and $(\mathcal{A}_2, \mathsf{g}_2)$. $\mathfrak{T}_{\Pi_{6.5}} = \{\mathsf{a}, \mathsf{d}\}$. $\mathfrak{T}_{\Pi_{6.5} \cup \Delta_{6.5}} = \{\mathsf{b}, \mathsf{c}, \mathsf{e}, \mathsf{f}, \mathsf{g}_1, \mathsf{g}_2, \neg\mathsf{c}\} \cup \mathfrak{T}_{\Pi_{6.5}}$. We have $(\mathcal{A}_1, \mathsf{g}_1) \not\lesssim_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{g}_2)$ because $\{\mathsf{a}\} \subseteq \mathfrak{T}_{\Pi_{6.5}}$ is an activation set for $(\mathcal{A}_1, \mathsf{g}_1)$, but not for $(\mathcal{A}_2, \mathsf{g}_2)$. We have $(\mathcal{A}_2, \mathsf{g}_2) \lesssim_{\mathrm{CP1}} (\mathcal{A}_1, \mathsf{g}_1)$ because any activation set for $(\mathcal{A}_2, \mathsf{g}_2)$ that is a subset of $\mathfrak{T}_{\Pi_{6.5}}$ includes a, and so is also an activation set for $(\mathcal{A}_1, \mathsf{g}_1)$. Considering Theorem 5.16 and the the activation set $\{\mathsf{b}, \mathsf{d}\}$ for $(\mathcal{A}_2, \mathsf{g}_2)$, we see $(\mathcal{A}_1, \mathsf{g}_1) \triangle_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{g}_2)$.

All in all, $\lesssim_{\mathrm{CP1}}$ realizes the intuition that the super-conjunction $\mathsf{a} \wedge \mathsf{d}$ — which is essential to derive $\mathsf{c} \wedge \mathsf{f}$ according to $\mathcal{A}_2$ — is more specific than the "less precise" a.

Just like Example 3.2, this example shows again that $\lesssim_{\mathrm{P3}}$ does not really implement the intuition that defeasible rules should be considered for their global semantical effect instead of their syntactical fine structure.

Example 6.6 (Stolzenburg et al. 2003, p. 96)

$\Pi^{\mathrm{F}}_{6.6} := \{\ \mathsf{c}, \mathsf{d}, \mathsf{e}\ \}, \quad \Pi^{\mathrm{G}}_{6.6} := \{\ \mathsf{x} \Leftarrow \mathsf{a} \wedge \mathsf{f}\ \},$

$\Delta_{6.6} := \mathcal{A}^1 \cup \mathcal{A}^2 \cup \mathcal{A}^3 \cup \mathcal{A}^4 \cup \mathcal{A}^5,$

$\mathcal{A}^1 := \{\ \mathsf{x} \leftarrow \mathsf{a} \wedge \mathsf{b} \wedge \mathsf{c}\ \}, \quad \mathcal{A}^2 := \{\ \neg\mathsf{x} \leftarrow \mathsf{a} \wedge \mathsf{b}\ \},$

$\mathcal{A}^3 := \{\mathsf{f} \leftarrow \mathsf{e}\}, \ \mathcal{A}^4 := \{\mathsf{a} \leftarrow \mathsf{d}\}, \ \mathcal{A}^5 := \{\mathsf{b} \leftarrow \mathsf{e}\}.$

Let us compare the specificity of the arguments $(\mathcal{A}^1 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \mathsf{x})$, $(\mathcal{A}^2 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \neg\mathsf{x})$, $(\mathcal{A}^3 \cup \mathcal{A}^4, \mathsf{x})$. $\mathfrak{T}_{\Pi_{6.6}} = \{\mathsf{c}, \mathsf{d}, \mathsf{e}\}$. $\mathfrak{T}_{\Pi_{6.6} \cup \Delta_{6.6}} = \{\mathsf{a}, \mathsf{b}, \mathsf{f}, \mathsf{x}, \neg\mathsf{x}\} \cup \mathfrak{T}_{\Pi_{6.6}}$.

We have $(\mathcal{A}^1 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \mathsf{x}) <_{\mathrm{CP1}} (\mathcal{A}^2 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \neg\mathsf{x}) \approx_{\mathrm{CP1}} (\mathcal{A}^3 \cup \mathcal{A}^4, \mathsf{x})$, because of $\mathsf{x}, \neg\mathsf{x} \notin \mathfrak{T}_{\Pi_{6.6}}$, and because any activation set $H \subseteq \mathfrak{T}_{\Pi_{6.6}}$ for any of $(\mathcal{A}^1 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \mathsf{x})$, $(\mathcal{A}^2 \cup \mathcal{A}^4 \cup \mathcal{A}^5, \neg\mathsf{x})$, $(\mathcal{A}^3 \cup \mathcal{A}^4, \mathsf{x})$ contains $\{\mathsf{d}, \mathsf{e}\}$, which is an activation set only for the latter two. This matches our intuition well, because the first of these arguments essentially requires the "more precise" $\mathsf{c} \wedge \mathsf{d} \wedge \mathsf{e}$ instead of the less specific $\mathsf{d} \wedge \mathsf{e}$.

We have $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x}) \triangle_{\mathrm{P3}} (\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x}) \triangle_{\mathrm{P3}} (\mathcal{A}^3{\cup}\mathcal{A}^4, \mathsf{x}) \triangle_{\mathrm{P3}} (\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$, however. This means that $\lesssim_{\mathrm{P3}}$ cannot compare these counterarguments and cannot help us to pick the more specific argument. What is most interesting under the computational aspect is that, for realizing

$$(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x}) \not\lesssim_{\mathrm{P3}} (\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x}),$$

we have to consider the defeasible rule of $\mathcal{A}^3$ (implicitly via $\{\mathsf{d}, \mathsf{f}\} \subseteq \mathfrak{T}_{\Pi_{6.6} \cup \Delta_{6.6}}$), which is not part of any of the two arguments under comparison. Note that such considerations are not required, however, for realizing the properties of $\lesssim_{\mathrm{CP1}}$, because defeasible rules not in the given argument can be completely ignored when calculating the minimal activation sets as subsets of $\mathfrak{T}_{\Pi}$ instead of $\mathfrak{T}_{\Pi\cup\Delta}$. This means in particular that the complication of *pruning* — as discussed in detail by Stolzenburg et al. (2003, Sect. 3.3) — does not have to be considered for the operationalization of $\lesssim_{\mathrm{CP1}}$.

Example 6.7 (Variation of Example 6.6)
$\Pi^{\mathrm{F}}_{6.7} := \{\ \mathsf{c}, \mathsf{d}, \mathsf{e}\ \}$, $\Pi^{\mathrm{G}}_{6.7} := \{\ \mathsf{x} \Leftarrow \mathsf{a}{\wedge}\mathsf{f},\ \mathsf{f} \Leftarrow \mathsf{e}\ \}$,
$\Delta_{6.7} := \mathcal{A}^1 \cup \mathcal{A}^2 \cup \mathcal{A}^4 \cup \mathcal{A}^5$,
$\mathcal{A}^1 := \{\ \mathsf{x} \leftarrow \mathsf{a}{\wedge}\mathsf{b}{\wedge}\mathsf{c}\ \}$, $\mathcal{A}^2 := \{\ \neg\mathsf{x} \leftarrow \mathsf{a}{\wedge}\mathsf{b}\ \}$,
$\mathcal{A}^4 := \{\ \mathsf{a} \leftarrow \mathsf{d}\ \}$, $\mathcal{A}^5 := \{\ \mathsf{b} \leftarrow \mathsf{e}\ \}$.
Let us compare the specificity of the arguments $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$, $(\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x})$, $(\mathcal{A}^4, \mathsf{x})$. $\mathfrak{T}_{\Pi_{6.7}} = \{\mathsf{c}, \mathsf{d}, \mathsf{e}, \mathsf{f}\}$. $\mathfrak{T}_{\Pi_{6.7} \cup \Delta_{6.7}} = \{\mathsf{a}, \mathsf{b}, \mathsf{x}, \neg\mathsf{x}\} \cup \mathfrak{T}_{\Pi_{6.7}}$.
Obviously, $\mathsf{x}, \neg\mathsf{x} \notin \mathfrak{T}_{\Pi_{6.7}}$. Moreover, $\{\mathsf{d}\} \subseteq \mathfrak{T}_{\Pi_{6.7}}$ is an activation set for $(\mathcal{A}^4, \mathsf{x})$ (but not a simplified one) and, *a fortiori* (by Corollary 5.12), for $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$, but not for $(\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x})$. Furthermore, every activation set $H \subseteq \mathfrak{T}_{\Pi_{6.7}}$ for $(\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x})$ satisfies $\{\mathsf{d}, \mathsf{e}\} \subseteq H$, which is an activation set for $(\mathcal{A}^4, \mathsf{x})$ and $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$. Furthermore, every activation set $H \subseteq \mathfrak{T}_{\Pi_{6.7}}$ for $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$ satisfies $\{\mathsf{d}\} \subseteq H$ which is an activation set for $(\mathcal{A}^4, \mathsf{x})$. All in all, we have
$(\mathcal{A}^4, \mathsf{x}) \approx_{\mathrm{CP1}} (\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x}) >_{\mathrm{CP1}} (\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x})$.
This is intuitively sound because $(\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x})$ is activated only by the more specific $\mathsf{d}{\wedge}\mathsf{e}$, whereas $(\mathcal{A}^4, \mathsf{x})$ is activated also by the "less precise" d. Moreover, $\mathsf{c}{\wedge}\mathsf{d}{\wedge}\mathsf{e}$ is not essentially required for $(\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x})$, which thus is equivalent to $(\mathcal{A}^4, \mathsf{x})$.
We have $(\mathcal{A}^4, \mathsf{x}) <_{\mathrm{P3}} (\mathcal{A}^1{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \mathsf{x}) \triangle_{\mathrm{P3}} (\mathcal{A}^2{\cup}\mathcal{A}^4{\cup}\mathcal{A}^5, \neg\mathsf{x}) \triangle_{\mathrm{P3}} (\mathcal{A}^4, \mathsf{x})$, however. This means that $\lesssim_{\mathrm{P3}}$ fails here completely w.r.t. Poole's intuition.

Example 6.8 *(continuing Example 3.3)*
Let us compare the specificity of the arguments $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a}))$ and $(\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$. $\mathfrak{T}_{\Pi_{3.3}} = \{\mathsf{q}(\mathsf{a}), \mathsf{s}(\mathsf{a})\}$, $\mathfrak{T}_{\Pi_{3.3} \cup \Delta_{3.3}} = \{\mathsf{p}(\mathsf{a}), \neg\mathsf{p}(\mathsf{a})\} \cup \mathfrak{T}_{\Pi_{3.3}}$.
We have $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a})) \approx_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$, because of $\mathsf{p}(\mathsf{a}), \neg\mathsf{p}(\mathsf{a}) \notin \mathfrak{T}_{\Pi_{3.3}}$, and because, for $H \subseteq \mathfrak{T}_{\Pi_{3.3} \cup \Delta_{3.3}}$, $i \in \{1, 2\}$, $L_1 := \neg\mathsf{p}(\mathsf{a})$, and $L_2 := \mathsf{p}(\mathsf{a})$, we have the logical equivalence of $H = \{\mathsf{q}(\mathsf{a})\}$ on the one hand, and of H being a minimal simplified activation set for $(\mathcal{A}_i, L_i)$ but not for $(\emptyset, L_i)$, on the other hand. By Theorem 5.16, we also get $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a})) \approx_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$. This makes perfect sense because $\mathsf{q}(\mathsf{a}){\wedge}\mathsf{s}(\mathsf{a})$ is not at all strictly "more precise" than $\mathsf{q}(\mathsf{a})$ in the context of $\Pi_{3.3}$.
Note that nothing is changed here if $\mathsf{s}(x) \Leftarrow \mathsf{q}(x)$ is replaced by setting $\Pi^{\mathrm{G}}_{3.3} := \{\mathsf{s}(\mathsf{a})\}$. If $\mathsf{s}(x) \Leftarrow \mathsf{q}(x)$ is replaced, however, by setting $\Pi^{\mathrm{G}}_{3.3} := \emptyset$ and $\Pi^{\mathrm{F}}_{3.3} := \{\mathsf{q}(\mathsf{a}), \mathsf{s}(\mathsf{a})\}$, then we get both $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a})) <_{\mathrm{P3}} (\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$ and $(\mathcal{A}_1, \neg\mathsf{p}(\mathsf{a})) <_{\mathrm{CP1}} (\mathcal{A}_2, \mathsf{p}(\mathsf{a}))$.

7 Relation to Other Approaches

Our new notion of specificity $\lesssim_{\mathrm{CP1}}$ follows the lines of Poole (1985). It is a transitive relation that provides us a quasi-ordering on arguments, which is as monotonic w.r.t. conjunction as can be expected (cf. Sect. 6.2). In addition, the effort for computing $\lesssim_{\mathrm{CP1}}$ is lower than that of $\lesssim_{\mathrm{P3}}$ because of $\mathfrak{T}_{\Pi} \subseteq \mathfrak{T}_{\Pi\cup\Delta}$, though not w.r.t. asymptotic complexity: In both cases already the number of possible (simplified) activation sets is exponential in the number of literals in the respective sets $\mathfrak{T}_{\Pi}$ and $\mathfrak{T}_{\Pi\cup\Delta}$, because in principle each possible subset has to be tested.

Stolzenburg et al. (2003, Definition 12) introduce the concept of pruning derivation trees because, for the case of $\lesssim_{\mathrm{P2}}$, attention cannot be restricted to derivations which make use only of the instances of defeasible rules given in the arguments. The reason for this is that the specificity notions of (Poole 1985) and (Simari and Loui 1992) admit literals L in activation sets that cannot be derived solely by strict rules, i.e. $L \in \mathfrak{T}_{\Pi\cup\Delta} \setminus \mathfrak{T}_{\Pi}$. Since this is not possible with the relation $\lesssim_{\mathrm{CP1}}$, this problem vanishes with our new version of specificity. See also Example 6.6.

Yet still, the new relation $\lesssim_{\mathrm{CP1}}$ inherits several properties from $\lesssim_{\mathrm{P3}}$. For instance, in general the specificity criterion requires us to compare sets of derivations, in principle all possible derivations for a given argument. This is true for both versions of the specificity relation. The reason for this complication is that we consider a very general setting of defeasible reasoning here, because — in contrast to other approaches (Gelfond and Przymusinska 1990; Dung and Son 1996; Benferhat and Garcia 1997) — we admit more than one antecedent in rules, i.e. bodies containing more than one literal, and (possibly) non-empty sets of background knowledge, namely the general rules in Π^{G} in addition to the facts in Π^{F}.

All in all, there are numerous frameworks for argumentation in logic. The overall process usually includes a dialectical process used for answering queries. Different arguments are pro or contra a certain answer. By means of an attack relation conflicts between contradicting arguments can be determined in abstract argumentation frameworks (Dung 1995; Prakken and Vreeswijk 2002). A concrete specificity or similar relation helps then to decide among conflicting arguments. There are a lot of works on abstract argumentation frameworks. However, as the discussion in this paper demonstrates, it is not that easy to find an effective concrete specificity relation. The main problem is that the orderings tend to be either computationally highly complex (Kern-Isberner and Thimm 2012) or not really appropriate for specificity (Besnard and Hunter 2001).

8 Conclusion

We would need further discussions on our most surprising new findings — after all, defeasible reasoning with Poole's notion of specificity is being applied now for over a quarter of century, and it was not to be expected that our investigations could discover a flaw in its foundations.

One remedy for the discovered lack of transitivity of $\lesssim_{\rm P3}$ could be to consider the transitive closure of the non-transitive relation $\lesssim_{\rm P3}$. Only under the condition of Theorem 5.16, the transitive closure of $\lesssim_{\rm P3}$ is a subset of $\lesssim_{\rm CP1}$, and therefore a possible choice. Moreover, it will still have all the intuitive shortcomings made obvious in Sect. 6. We do not see how this transitive closure could be decided efficiently. Furthermore, this transitive closure lacks a direct intuitive motivation, and after the first extension step from $\lesssim_{\rm P3}$ to its transitive closure, we had better take the second extension step to the more intuitive $\lesssim_{\rm CP1}$ immediately.

Finally, contrary to the transitive closure of $\lesssim_{\rm P3}$, our novel relation $\lesssim_{\rm CP1}$ also solves the problem of non-monotonicity of specificity w.r.t. conjunction (cf. Sect. 6.2) insofar as it was realized as a problem of $\lesssim_{\rm P1}$ in (Poole 1985).

Further work is needed to improve efficiency considerably. As a first step, we have narrowed the concept of $\lesssim_{\rm CP1}$ further down according to Sect. 3, resulting in a similar (no difference in any of the examples presented in this paper!), but more tractable relation $\lesssim_{\rm CP2}$. We are currently developing and testing strong methods for efficient operationalization of $\lesssim_{\rm CP2}$, which can hardly be found for any of $\lesssim_{\rm P1}$, $\lesssim_{\rm P2}$, $\lesssim_{\rm P3}$, $\lesssim_{\rm CP1}$.

References

Benferhat, S., and Garcia, L. 1997. A coherence-based approach to default reasoning. In Gabbay, D.; Kruse, R.; Nonnengart, A.; and Ohlbach, H.-J., eds., *Proc. 1st Int. Joint Conf. on Qualitative and Quantitative Practical Reasoning, 1997, June 9–12, Bad Honnef (Germany)*, number 1244 in Lecture Notes in Computer Science, 43–57. Springer. http://dx.doi.org/10.1007/BFb0035611.

Besnard, P., and Hunter, A. 2001. A logic-based theory of deductive arguments. *Artif. Intell.* 128(1-2):203–235.

Chesñevar, C. I.; Dix, J.; Stolzenburg, F.; and Simari, G. R. 2003. Relating defeasible and normal logic programming through transformation properties. *Theoretical Computer Sci.* 290:499–529. Received Jan. 8, 2001; rev. Nov. 9, 2001. http://dx.doi.org/10.1016/S0304-3975(02)00033-6.

Dung, P. M., and Son, T. C. 1996. An argumentation-theoretic approach to reasoning with specificity. In Aiello, L. C.; Doyle, J.; and Shapiro, S. C., eds., *Proc. 5th Int. Conf. on Principles of Knowledge Representation and Reasoning, 1996, Nov. 5–8, Cambridge (MA)*, 506–517. Morgan Kaufmann, Los Altos (CA) (Elsevier).

Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* 77(2):321–358.

García, A. J., and Simari, G. R. 2004. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming, Cambridge Univ. Press* 4:95–138.

Gelfond, M., and Przymusinska, H. 1990. Formalization of inheritance reasoning in autoepistemic logic. *Fundamenta Informaticae* XIII:403–443.

Gillman, L. 1987. *Writing Mathematics Well*. The Mathematical Association of America.

Kern-Isberner, G., and Thimm, M. 2012. A ranking semantics for first-order conditionals. In Raedt, L. D.; Bessière, C.; Dubois, D.; Doherty, P.; Frasconi, P.; Heintz, F.; and Lucas, P. J. F., eds., *ECAI*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, 456–461. IOS Press.

Poole, D. L. 1985. On the comparison of theories: Preferring the most specific explanation. In Joshi, A., ed., *Proc. 9th Int. Joint Conf. on Artificial Intelligence (IJCAI), 1985, Aug. 18–25, Los Angeles (CA)*. Morgan Kaufmann, Los Altos (CA) (Elsevier). http://ijcai.org/Past%20Proceedings.

Prakken, H., and Vreeswijk, G. 2002. Logics for defeasible argumentation. In Gabbay, D., ed., *Handbook of Philosophical Logic*. Kluwer Academic Publisher, 2nd edition. 218–319.

Simari, G. R., and Loui, R. P. 1992. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence* 53:125–157. Received Feb. 1990, rev. April 1991.

Stolzenburg, F.; García, A. J.; Chesñevar, C. I.; and Simari, G. R. 2003. Computing generalized specificity. *J. Applied Non-Classical Logics* 13:87–113. http://www.tandfonline.com/doi/abs/10.3166/jancl.13.87-113.

Wirth, C.-P., and Stolzenburg, F. 2013. *David Poole's Specificity Revised*. SEKI-Report SR–2013–01 (ISSN 1437–4447). DFKI Bremen GmbH, Safe and Secure Cognitive Systems, Cartesium, Enrique Schmidt Str. 5, D–28359 Bremen, Germany: SEKI Publications. pp. ii+22, http://arxiv.org/abs/1308.4943.