Annals of Mathematics and Artificial Intelligence 9(1993)93–132

An argumentation framework in default logic

Henry Prakken
Computer/Law Institute, Free University, Amsterdam, The Netherlands

Abstract

This article presents a formal theory about nontrivial reasoning with inconsistent information, applicable, among other things, to defeasible reasoning. The theory, which is inspired by a formal analysis of legal argument, is based on the idea that inconsistency tolerant reasoning is more than revising an unstructural set of premises; rather it should be regarded as constructing and comparing arguments for incompatible conclusions. This point of view gives rise to two important observations, both pointing at some flaws of other theories. The first is that arguments should be compared as they are constructed, viz. step-by-step, while the second observation is that a knowledge representation language is needed with a defeasible conditional, since the material implication gives rise to arguments which are not constructed in actual reasoning. Accordingly, a nonmonotonic logic, default logic, is chosen as the formalism underlying the argumentation framework. The general structure of the framework allows for any standard for comparing pairs of arguments; in this study two such standards are investigated, based on specificity and on orderings of the premises.

1 Introduction

This paper connects two recent developments in formal research on commonsense reasoning: the study of argumentation and the study of inconsistency tolerant reasoning. The first development concerns the argumentation aspect of commonsense reasoning, i.e. the process of constructing and comparing arguments for or against a certain conclusion. Particularly in the legal field this development has become prevalent, in attempts to model the adversarial aspect of legal reasoning. These attempts are based on the idea that lawyers do not argue for the legally correct solution, if it exists at all, but for the solution which best serves the client's interests; in doing so a lawyer has available a large body of conflicting opinions, rules, precedents, principles and so on, with which to construct a coherent argument; of course, the opponent in a law suit will do the same and therefore not only ways of constructing arguments but also ways of comparing them become important. While in the legal domain this development has mainly resulted in implemented programs (e.g. [2, 6]), in other areas of AI research mainly formal argumentation systems have been developed [9, 22, 23]. These systems are meant to be an alternative to earlier approaches to formalize so-called nonmonotonic reasoning, in which conclusions can be invalidated by adding new information to the premises. This kind of reasoning is motivated by the fact that in real life people do not always have sufficient information to make a safe step towards the conclusion; instead, they often jump to conclusions by applying general, defeasible rules, since it is often too costly or even impossible to find the information which would guarantee a safe landing. Of course, in unusual circumstances this jump might appear to be unjustified, which is the reason why new information can invalidate old conclusions.

The second development, the study of inconsistency tolerant reasoning, has mainly taken place in the study of nonmonotonic reasoning. Usually, formalizations of nonmonotonic reasoning start with a consistent set of premises and provide for ways of deriving plausible but deductively unsound conclusions from them. However, a different approach is allowing the premises to be inconsistent and to prefer some part of the premises if indeed an inconsistency occurs (cf. [3, 13, 14, 20]). In this view, defeasible rules express approximations to reality, which might need to be corrected in specific circumstances: people reason with defaults as if they were true, until they give rise to an inconsistency. The attractiveness of this idea is that if nonmonotonic reasoning is regarded as a kind of inconsistency handling in classical logic, no new logic needs to be developed. For legal applications this approach is interesting for yet another reason, since it might result in theories which can also be used for modelling the way lawyers reason with hierarchies of regulations. In fact the second approach which will be discussed partly originates from legal philosophy.

The present paper is the result of a logical investigation of reasoning with inconsistent and incomplete information in legal reasoning; it has appeared that in the legal domain the above sketched developments are closely related, and since the law is just one domain of common-sense reasoning, this observation is relevant for other areas as well. Accordingly, the aim of the present paper is to investigate argumentation and inconsistency tolerant reasoning from a unifying perspective, which is that, unlike standard logic, in which from a contradiction anything can be derived, both phenomena deal with inconsistencies in a nontrivial way: in argumentation systems conflicting arguments are compared in order to choose the best one, and in inconsistency tolerant reasoning inconsistent premises are revised in order to derive nontrivial information. Although it seems very natural to study these phenomena in relation to each other, this up to now has been done surprisingly little. Still, such a combination is able to overcome some serious flaws of existing approaches to inconsistency tolerant reasoning, caused by the failure to recognize the argumentation aspect of this kind of reasoning. The unified study of the two developments will also yield a second important result, which is that information which is subject to conflict resolution metaprinciples is irreducibly defeasible in nature, for which reason classical knowledge is inadequate even as a knowledge representation language.

The structure of this report is as follows. In section 2 some problems of recent theories about inconsistency tolerant reasoning are discussed. An alternative theory is given in section 3 in the form of an argumentation framework based on default logic. In section 4 two applications of this framework with different standards for comparing arguments are discussed, which in section 5 are illustrated by examples. Then section 6 discusses related research and section 7 contains some remarks on implementing the theory.

2. Existing approaches to inconsistency tolerant reasoning

In this section three existing approaches to nontrivial reasoning with inconsistent information will be discussed, two of which originate from the study of nonmonotonic reasoning. Among the researchers modelling nonmonotonic reasoning as inconsistency tolerant reasoning are Brewka [3], Poole [13, 14] and Roos [20]; earlier Alchourrón and Makinson [1] applied inconsistency handling techniques to hierarchies of legal regulations. The general idea goes back to Rescher [19]: if the set of premises turns out to be inconsistent, maximal consistent subsets (for which we will use Brewka's term "subtheories") are identified and ordered according to some preference relation: in applications to nonmonotonic reasoning generally the subtheories containing the most specific information are preferred, while in legal applications those subtheories are preferred that best obey the legal conflict resolution metaprinciples. The preferred subtheories can be used for defining two nonstandard consequence relations, corresponding to following from one of the preferred subtheories or from all of them.

I will investigate three approaches based on this idea: first Poole's framework for default reasoning, then belief-revision related approaches inspired by Alchourrón and Makinson, and finally, Brewka's preferred-subtheories approach. The last two approaches use orderings on formulas to determine the preferred subtheories, for which reason some notational conventions have to be explained: "$x \leq y$" stands for "y is preferred over x"; "$x \approx y$" is shorthand for "$x \leq y$ and $y \leq x$"; and "$x < y$" abbreviates "$x \leq y$ and $y \not\leq x$". Unless stated otherwise, $\leq$ is assumed to be a partial preorder, i.e. a relation which is transitive and reflexive.

2.1. POOLE'S THEORY COMPARATOR

As Brewka [3] has convincingly argued, Poole's framework for default reasoning may be regarded as an application of the "preferred subtheories" idea. Poole [13] presents a formalization of the so-called specificity principle against the background of a general view on default reasoning presented in detail in [14]. Essentially, this view is that if defaults are regarded as possible hypotheses with which theories can be constructed to explain certain facts, there is no need to change the logic but only the way the logic is used. Accordingly, the semantics and proof theory of Poole's " logical framework for default reasoning" are simply those of first-order predicate logic. The basis of this framework are the sets F and δ. F is a set of closed first-order formulas, the *facts*, assumed consistent, and δ is a set of possibly inconsistent first-order formulas, the *defaults* or possible hypotheses. A *scenario* of a pair (F, δ)

is a consistent set $F \cup D$, where D is a set of ground instances of defaults of δ. An *explanation* of a closed formula is a scenario implying it. Theory formation consists in constructing an explanation for a given formula. An *extension* is a maximal (with respect to set inclusion) scenario. Extensions can be regarded as the preferred subtheories of a default theory. These definitions tell us that in constructing an explanation the facts must be obeyed but that the use of any default is free, as long as its use is consistent with the facts and the other applied defaults.

Conflicting explanations can be compared with respect to any criterion, one of which is their specificity. Consider explanations $Ai = F \cup Di$ for ϕ and $Aj = F \cup Dj$ for $\neg\phi$. Informally, the idea is that Ai is more specific than Aj iff there is a *possible* situation in which only Aj applies. To make this precise, the facts F are divided into necessary facts Fn, regarded as holding in all possible situations, and contingent facts Fc, the "input facts" of a case at hand. What is important is that the contingent facts are only used to derive conclusions about the actual circumstances; in determining specificity they are replaced by any set of "possible facts", which need not even be a subset of Fc. Now Ai is more specific than Aj with respect to ϕ iff there is a possible fact Fp which makes Aj explain $\neg\phi$ without making Ai explain ϕ or $\neg\phi$ (this last "non-triviality" requirement for Ai and $\neg\phi$ is essential, since without it the possible fact $\neg\phi$ would always make Ai more specific than Aj). These ideas are captured in the following definition, which differs from Poole's in some inessential respects.

DEFINITION 2.1

Let $Ai = (Fn \cup Fc \cup Di)$ and $Aj = (Fn \cup Fc \cup Dj)$ be explanations for, respectively, ϕ and $\neg\phi$. Then *Ai is more specific than Aj with respect to* ϕ iff there is a fact Fp such that

- $Fn \cup \{Fp\} \cup Dj \models \neg\phi$
- $Fn \cup \{Fp\} \cup Di$ not $\models \neg\phi$ and not $\models \neg\phi$. ($\models$ denotes first-order entailment.)

If, in addition, Aj is not more specific than Ai with respect to ϕ, then *Ai is strictly more specific than Aj with respect to* ϕ.

This elegant definition gives the intuitive results in the standard examples. Consider the following two pairs of conflicting explanations.

EXAMPLE 2.2

$Fc1 = \{a, c\}$	$Fc2 = \{a\}$ $Fn2 = \{c \supset a\}$
A1: $Fc1 \cup \{a \supset b\}$	$A2 = Fc2 \cup Fn2 \cup \{a \supset b\}$
A1′: $Fc1 \cup \{(a \wedge c) \supset \neg b\}$	$A2' = Fc2 \cup Fn2 \cup \{c \supset \neg b\}$

In both cases Fp = a is a fact making the first scenarios explain b without making the second ones explain $\neg b$, while all possible facts making the second scenarios explain $\neg b$ will make the first ones explain b: in case of A1′ because all such possible facts imply $(c \wedge a)$, which implies a, and in case of A2′ because all such facts imply c, which together with Fn also implies a: in result, A1′ and A2′ are strictly more specific with respect to b than their counterarguments. Consider next

A3: $Fc1 \cup \{a \supset b\}$

A3′: $Fc1 \cup \{c \supset \neg b\}$

In this case a is a possible fact making A3 explain b without making A3′ explain $\neg b$, and c is a possible fact making A3′ explain $\neg b$ without making A3 explain b, for which reason both arguments are more specific than each other with respect to b and neither wins the conflict. Applications of Poole's specificity definition to legal reasoning are discussed in [15].

However, despite their intuitive attractiveness, Poole's ideas also have some drawbacks. One problem is that definition 2.1 ignores the possibility of multiple conflicts, as is shown by the next example.

EXAMPLE 2.3

$A3 = Fc \cup \{a \supset b, \quad b \supset c, \quad c \supset d\}$

$A4 = Fc \cup \{(a \wedge f) \supset \neg b, \quad \neg b \supset e, \quad e \supset \neg d\}$

$Fc = \{a, f\} \quad Fn = \{c \supset e\}$

Poole's definition prefers A3 for d, because e is a fact which makes A4 explain $\neg d$ without A3 explaining d, while all facts which make A3 explain d imply c and therefore, since Fn contains $(c \supset e)$, they all imply e, which makes A4 explain $\neg d$ (note again that Fc is ignored). However, this ignores the fact that A4 uses the fact b, for which the scenario $A3' = Fc \cup Fn \cup \{a \supset b\}$ is strictly less specific than $A4' = Fc \cup Fn \cup \{(a \wedge f) \supset \neg b\}$ for $\neg b$.

Of course, as Poole [13, p. 146] himself recognizes when discussing a similar example, for an argument to be preferred not only the final conclusion but also all intermediate conclusions must be preferred. The problem with definition 2.1, however, is that it does not recognize b as an intermediate conclusion of A3. To repair this shortcoming (which Poole does not do), the specificity definition should be embedded in a more general theory, defining when and to which conflicts the specificity comparator should be applied. What is important is that such a framework reflects the step-by-step nature of constructing and comparing arguments: rather than at once, it should be checked after each step whether the argument constructed thus far is better than all counterarguments, and if at some step there are counterarguments which are strictly more specific with respect to the conclusions of that step, the argument should be "cut off". In this way A3 in the present example will be cut off after its intermediate step leading to "b" and therefore A4 will for trivial reasons be the argument which wins the conflict about "d", since its only counterargument, which is A3, has not survived a comparison for an earlier step. One of the main goals of this article will be to formalize this step-by-step nature of constructing and comparing arguments.

A second problem of Poole's framework, which it shares with the other approaches, will be discussed in the next subsections.

2.2. BELIEF REVISION APPROACHES

Belief revision (cf. [7]) is about the dynamics of "belief sets": it studies the process of revising a set of propositions with respect to a certain proposition. Belief revision can be applied to several problems, for instance, to testing scientific hypotheses, counterfactual reasoning or updating databases; in this section I will investigate its application to deriving nontrivial conclusions from inconsistent information. The reason why belief revision is a possible candidate framework for this purpose is that one way of characterizing the nontrivial consequences of inconsistent premises is defining them as the standard consequences of the set resulting from contracting "False" from the premises, an idea suggested by Alchourrón and Makinson [1]. The belief revision method which seems best suited for these purposes consists of identifying all maximal consistent subsets, i.e. all subtheories, of the premises, comparing them with respect to some ordering relation, and taking the intersection of all subtheories which are maximal in this ordering; to the result of this operation simply standard logic can be applied. This is called "partial meet contraction" applied to arbitrary sets [7, pp. 59, 80]. In the literature on belief revision several ways of comparing subtheories have been proposed. One of the best is Sartor's [21], particularly, since it correctly deals with multiple conflicts of the type of example 2.3. Nevertheless, all definitions share the same core idea: they all compare sets of formulas according to their *minimal* elements; and the problems already manifest themselves in cases in which the definitions boil down to directly applying this criterion to the subtheories of the premises. For this reason I will confine myself to discussing such cases.

Consider first an example in which this method gives satisfactory results. Assume that the inconsistent set $\{p, \neg p, r\}$ is ordered as $r > p, r > \neg p, p \approx \neg p$. This set has two subtheories, $\{p, r\}$ and $\{\neg p, r\}$; their minimal elements, p and $\neg$p, are of equal level, for which reason both sets are maximal in the ordering; their intersection is $\{r\}$, for which reason r is a nontrivial consequence of the premises, while nothing of interest can be concluded about p or $\neg$p.

However, there are also cases in which the belief revision methods give less natural results. Consider the example of an incoherent university library regulation of which one section says that misbehaviour can lead to removal from the library, while another section says that professors cannot be forced to leave the library; and the library regulation of the faculty of law, which is lower than the university regulation, saying that snoring is a case of misbehaviour. Finally, the facts, which are given highest priority, say that Bob is a professor who snores in the library.

EXAMPLE 2.4.

(1) Misbehaves(x) $\supset$ May-be-removed(x)

(2) Professor(x) $\supset \neg$May-be-removed(x)

(3) Snores(x) $\supset$ Misbehaves(x)

(4) Professor(Bob) $\wedge$ Snores(Bob)

$1 < 4,\ 1 \approx 2, 3 < 1.$

If we assume (1), (2) and (3) to be instantiated for Bob, then this example has four subtheories, each missing a different element of the premise-set: $A1 = \{1,2,3\}$, $A2 = \{1,2,4\}$, $A3 = \{1,3,4\}$, and $A4 = \{2,3,4\}$. If ordered on the basis of their minimal elements, their ordering is $A1 < A2$; $A1 \approx A3 \approx A4$, which implies that the only maximal element of the set of subtheories is A2. What does it say about the consequences of Bob's snoring? Since (3) is not in A2, (1) cannot be used any more to derive May-be-removed(Bob): the conclusion, then, is that Bob cannot be removed from the library.

However, I want to argue for a different view on the example, a view which is more in line with my remarks on example 2.3. On the one hand, clearly there is a potential conflict between the rules (1) and (2), since when a professor is held to be misbehaving a choice must be made about which of them takes precedence. On the other hand it seems natural to say that there is no dispute that Bob is misbehaving: in the same way as in example 2.3 it can be said that Misbehaves(Bob) is merely an intermediate conclusion for making (1) applicable, for which reason (3) is irrelevant to the conflict about whether Bob should be removed from the library. Instead of preferring $\neg$May-be-removed(Bob) by rejecting the conclusion that Bob is misbehaving, it seems more natural to make a choice between the norms which are certainly in conflict with each other, (1) and (2): and since these norms are of equal level, the outcome should be that the conflict cannot be resolved.

For mathematicians a natural way of replying would be: "well, if (3) should stay, then the ordering should be changed: (3) should be higher than (1) and (2)". This, however, although it may be acceptable for mathematical purposes, is cognitively inadequate, since it does not capture the way hierarchies are used in legal reasoning: such a hierarchy does not depend on desired outcomes in individual cases but is, instead, assigned generally and used to solve individual conflicts. What is required in modelling this use is a modification of the formal definitions rather than a change in the assignment of a specific ordering: premises like (3) should not be regarded as higher than (1) and (2), but as irrelevant to the conflict.

2.3. BREWKA'S PREFERRED SUBTHEORIES FRAMEWORK

An alternative way of constructing the preferred maximal consistent subsets is the one of Brewka [3]. The idea is that on the basis of a strict partial order of the individual premises a consistent set of premises is constructed by first adding as many formulas of the highest level to the set as is consistently possible, then adding as many formulas of the subsequent level as is consistently possible, and so on. If comparable formulas are in conflict, the resulting set branches into alternative and mutually exclusive sets. The nontrivial consequences of the initial premises can be defined as the standard consequences of the intersection of the resulting sets.

Applied to the set $\{p, \neg p, r\}$ with the ordering $r > p$, $r > \neg p$, this has the following result. First r is added to the set and then a choice must be made between p and $\neg p$, since adding them both would make the set inconsistent. Since they are incomparable, the set branches into two alternative preferred subtheories, $\{p, r\}$ and $\{\neg p, r\}$. The intersection of these sets is $\{r\}$, which is the same outcome as in the belief revision methods.

Let us now see how this method deals with example 2.4. First the facts (4) are added to the set, and then *both* norms about removal from the library, i.e. both (1) and (2), since without the lower norm (3) on snoring (1) cannot be used to derive May-be-removed(Bob), for which reason no contradiction occurs. finally, (3) is considered: adding this implication to the set would cause an inconsistency, for which reason it is left out. Thus the method of Brewka again results in the same set as the belief revision methods, viz. $\{1,2,4\}$, which makes it subject to the same criticism: it, in my opinion mistakenly, regards (3) as relevant to the conflict.

2.4. DIAGNOSIS

As already indicated, the problem with the methods of 2.1 and 2.2 is that they regard too many premises as relevant to the conflict about whether Bob may be removed from the library: in both approaches it is *all* members of the (classically) minimal inconsistent set which are regarded as relevant, whereas example 2.4 has illustrated that often it is only a subset of this set which matters: informally, only conditional rules with conflicting consequents are relevant to the conflict. At first sight this point seems to be rather ad hoc, in that it only pertains to the specific form of example 2.4. Nevertheless, it can be generalized if a different attitude is employed towards reasoning with inconsistent information, viz. if it is regarded as choosing between conflicting *arguments* instead of revising inconsistent premises. As already explained in the discussion of example 2.3, this is a step-by-step process and for this reason premises which are only needed to provide *intermediate* conclusions of an argument should be regarded as irrelevant to conflicts about conclusions drawn in *further* steps of the argument. In example 2.4 this means that (3), which in the argument for May-be-removed(Bob) is only used to derive the intermediate conclusion Misbehaves(Bob), is irrelevant to the conflict about May-be-removed(Bob).

Now if we loosely define an argument as a consistent set implying a conclusion, then at first sight nothing seems to have changed, since formally {1–4} not only contains an argument for "Misbehaves(Bob)", viz. {3,4}, but also one for the opposite, viz. {1,2,4}. However, a more natural view seems to be that arguments such as the one for "¬Misbehaves(Bob)" are not constructible: the formal constructibility of this argument depends on the fact that the truth of "Misbehaves(Bob)" would lead to a conflict between two other norms, but the very idea of inconsistency handling is to resolve such conflicts when they occur, and then it seems strange to allow arguments which are based on the idea that such conflicts cannot occur. And if this observation is combined with the remarks on the step-by-step nature of constructing arguments, then we have a general reason for regarding (3) as irrelevant to the conflicts: if rules are made subject to conflict resolution meta-principles, they are defeasible: modus tollens and other contrapositive inferences are invalid for them, for which reason no argument can be set up against Misbehaves(Bob). Therefore the premises should be represented with a defeasible conditional instead of with the material implication, after which the undesired argument is invalidated.

If we again take a look at Poole's framework, it is easy to see that these considerations also provide a solution for Poole's problems with multiple conflicts in example 2.3, since our conclusion was that, rather than at once, it should be checked after each step whether the argument constructed thus far is strictly more specific than all counter-arguments. Moreover, also in Poole's framework the material implication gives rise to undesired counterarguments, as can easily be checked by considering an example of the form $\delta = \{a \supset b, (b \wedge c) \supset d, c \supset \neg d\}$, $Fc = \{a, c\}$: while c should intuitively be preferred, in the same way as in example 2.4 a non-defeated argument against b can be constructed. Obviously the just-proposed solution, formalizing defaults with a defeasible conditional, will also apply to this example.

It is important to realize that the above problems cannot be solved by only replacing standard logic with some nonmonotonic formalism, since, as shown in [17], examples like example 2.4 can also be constructed in, for example, Brewka's prioritized default logic [4], which is in fact the default logic version of his preferred subtheories approach, and in Konolige's hierarchical autoepistemic logic [8]: the reason is that also these formalities fail to capture the step-by-step nature of comparing arguments.

To summarize the results of this section, we have, firstly, concluded that theories of inconsistency tolerant reasoning should take into account the step-by-step nature of argumentation; and we have, secondly, seen that the approach of modelling nonmonotonic reasoning as inconsistency tolerant reasoning in classical logic is more problematic than is often realized. For these reasons I will in the next section define the process of comparing arguments in an inductive way, and I will use Reiter's default logic [18] as the underlying formalism.

3. Comparing arguments in default logic

As an attempt to solve the problems identified in the previous section, this section develops a general framework for constructing and comparing arguments, combining an inductive definition of a preferred argument with default logic as the underlying formalism. Following a brief overview of default logic in 3.1, the formal definitions will be given in 3.2, while some definitions and results concerning default logic can be found in appendix A. In fact, my choice of default logic is partly based on pragmatic grounds, in that it may not be the only logic with a suitable defeasible conditional. However, alternatives on which an argumentation framework can be based do not seem to be available yet: for example, in recent approaches based on conditional logic (e.g. [5]) specificity is not regarded as a conflict resolution metaprinciple, but as a principle of the *semantics* of defaults, for which reason the use of other standards to compare arguments is excluded.

3.1. DEFAULT LOGIC

Like Poole's framework, also default logic is based on a set F of facts, expressed in the language of first-order logic and assumed consistent, and a set δ of defaults. However, a crucial difference between Poole's framework and default logic is that Reiter's defaults are inference rules: $\phi:\psi/\chi$, in which ϕ is the prerequisite, ψ the justification, and χ the consequent, informally reads as "If ϕ holds and ψ may be consistently assumed, χ may be inferred". Because of this reading defaults are directional: neither modus tollens nor contraposition is valid for them. Unlike first-order logic is *nonmonotonic*. A logic is called "monotonic" if the set of theorems grows in case the set of premises grows: formally, if $A \subseteq B$ then $Th(A) \subseteq Th(B)$. Default logic does not have this property: if a default $\phi:\psi/\chi$ is used to infer χ, and after that $\neg\psi$ is added to the facts, then the derivation of χ becomes invalid.

Relative to a given default theory (F, δ) new beliefs can be derived by using ground instances of any default of δ one wishes, as long as consistency is preserved. If as many defaults as possible are thus used, i.e. if applying any new default would cause an inconsistency, sets result which are called *extensions* of (F, δ). These sets can be regarded as possible maximal sets of beliefs which may be held on the basis of the facts F and default assumptions δ. Since defaults can conflict, a default theory may have several, mutually inconsistent, extensions. Extensions are similar to the deductive closure of the preferred subtheories in the above approaches.

Now below the idea is to represent the facts of the case at hand as elements of F, together with necessary truths such as "a man is a person" or "a lease contract is a contract", and to formalize defeasible rules as normal defaults $\phi:\psi/\psi$, in this paper written as $\phi \Rightarrow \psi$. Unconditional defeasible rules will be represented as defaults of the form $\Rightarrow \phi$, which is shorthand for $\top \Rightarrow \phi$, where $\top$ stands for any valid formula.

3.2. CONSTRUCTING AND COMPARING ARGUMENTS

In this subsection the above ideas will be formalized in the form of a framework for constructing and comparing arguments. The framework will be of a general nature, in that it allows for any standard for comparing pairs of arguments. It will be developed in three stages. First the notion of an argument and some related notions will be defined, and then it will be defined what it means that an argument is preferred. To capture the intended generality of the framework the latter definition will assume the existence of some unspecified standard R for comparing pairs of arguments; the only assumptions which will be made about R is that it is an asymmetric and upwardly bounded relation (below we will sometimes call such a standard a "kind of defeat"). It is this part of the framework where, for example, Poole's specificity definition or a standard on the basis of default orderings can be applied; this will be discussed in detail in the next section. The final stage is the definition of a unique set of formulas which are the "defeasible" consequences of a triple (F, δ, R), where F and δ are the facts and defaults of a default theory, and R is a kind of defeat. As for the presentation of the framework, it should be noted that all definitions below are given at the background of a fixed default theory (F, δ).

To start with arguments, they will be formally defined as a set of facts and a set of defaults, i.e. as a default theory. The facts are the set F of the background default theory (F, δ), while the defaults are a subset of the set of ground instances of δ. A reasonable view on arguments is that they should be internally coherent, which view will be formalized by requiring that they have unique extensions; furthermore, all defaults of an argument are required to be applicable. Finally, if a formula ϕ is in the extension of a certain argument A, A is said to explain ϕ.

DEFINITION 3.1

a. $A = (F, D)$ (where D is a finite subset of ground instances of δ) is an *argument* iff it has a unique extension E(A) such that of all elements of D both the prerequisites and the consequents are in E(A).

b. A *explains* a formula ϕ iff ϕ is in E(A).

c. $A' = (F, D')$ is a *subargument* of an argument $A = (F, D)$ iff A′ is an argument and if $D' \subset D$.

d. A *minimally explains* ϕ iff A explains ϕ and no subargument of A explains ϕ.

e. ϕ is a final conclusion of an argument A iff A minimally explains ϕ; ϕ is an *intermediate conclusion* of A iff a subargument of A explains ϕ.

f. The *combination* of two arguments $A = (F, D)$ and $A' = (F, D')$ is the argument $(F, D \cup D')$ which, overloading the symbol "$\cup$", will be denoted by $A \cup A'$.

Consider by way of illustration of the notion of a subargument $D = \{a \Rightarrow b, b \Rightarrow c\}$, $Fc = \{a\}$, and $A = (Fc, D)$. Then $(Fc, \{a \Rightarrow b\}$ is a subargument of A, but $(Fc, \{b \Rightarrow c\})$ is not, since in this default theory $b \Rightarrow c$ is not applicable.

COROLLARY 3.2

(i) Every argument has only a finite number of subarguments.

(ii) If A is an argument explaining ϕ but not minimally, then A has a subargument minimally explaining ϕ.

REMARK 3.3

It may be instructive to remark that condition (b) of definition 3.1 is equivalent to "A explains a formula ϕ iff $F \cup CONS(D) \models \phi$" (cf. definition A.2 and the remark on proposition A.4 in appendix A; CONS(D) denotes the set of consequents of all defaults in D).

The second element of the framework is the definition of a preferred argument, i.e. of an argument which is better than any counterargument. As just noted, it will in this section simply be assumed that some standard for comparing arguments is defined. In order to reflect the step-by-step nature of argumentation the notion of a preferred argument is defined inductively: the idea is that in each inductive step arguments are only compared with respect to their final conclusions; the preferred status of intermediate conclusions should already have been established at earlier steps in the induction. Conflicts about final conclusions will be captured by a notion of "minimal interference", defined in the following definition.

DEFINITION 3.4

(1) An argument Ai *interferes* with an argument Aj iff Ai and Aj explain contradictory facts. Furthermore, iff Aj interferes with Ai, then Aj is *a counterargument of* Ai.

(2) Ai *minimally interferes* with Aj *with respect to* ϕ iff Ai minimally explains ϕ and Aj minimally explains $\neg\phi$.

Another aspect of arguments which should be dealt with is that they can have more than one final conclusion, even logically independent ones: consider $\delta = \{a \Rightarrow b\}$, $Fc = \{a, (b \supset c\}$: both "b" and "c" are final conclusions of this argument. For this reason an argument is preferred only if it passes the test of R for *all* its final conclusions. Finally, the definition should account for the possibility that an argument which is not itself better than a counterargument is still saved by another argument which *is* better than this counterargument. This will be called *reinstatement.*

I will now present the definition which formalizes these observations. Since it is the central definition of the framework we will present it with some care: first I give a more intuitive definition, which at first sight would seem to work well, but which will appear to be too simple. Recall that R is a variable for a specific kind of defeat, assumed to be an asymmetric relation on arguments; if there is no danger of confusion, it will be left implicit.

DEFINITION 3.5

(*preferred arguments: first attempt*). An argument A = (F, D) is a R-*preferred argument* iff

(1) All subarguments of A are R-preferred arguments;

(2) For all formulas ϕ and arguments A′ such that A minimally interferes with A′ with respect to ϕ and such that neither A′ nor one of its subarguments is R-defeated by another R-preferred argument: A R-defeats A′ with respect to ϕ.

Because of the condition that all subarguments of an argument are also preferred, condition (1) ensures that multiple conflicts are dealt with correctly. Condition (2) formalizes the requirement that for all final conclusions an argument must itself defeat all arguments for a contradicting fact, unless these are already defeated by another preferred argument, which thereby reinstates the initial argument.

The reason that this definition is not satisfactory is that it is circular, which manifests itself in examples of the following kind.

EXAMPLE 3.6

(1) $a \Rightarrow b$ (2) $(b \wedge c) \Rightarrow \neg d$

(3) $c \Rightarrow d$ (4) $(a \wedge d) \Rightarrow \neg b$

$Fc = \{a, c\}$, $\delta = \{1-4\}$

Let A1 = (Fc, {1,2}) and A2 = (Fc, {3,4}) and assume that the kind of defeat is specificity (S-defeat). In order to know whether A1 is S-preferred we must first determine whether A1 S-defeats A′ with respect to d, where A2′ = (F, {3}). Although S-defeat is not yet formally defined, this definition will obviously be such that A1 indeed S-defeats A2′. Furthermore, we must also know whether A1′ = (F, {1}) is S-preferred: A2 will indeed S-defeat A′ with respect to b, but there is a problem, since, as just explained, A1 S-defeats A2′ with respect to d, and A2′ is a subargument of A2. Does this mean that A2 is by condition (1) of definition 3.5 not an S-preferred argument? This would be the case if A1 were S-preferred, but this is what we are trying to find out! Here the definition turns out to be circular. For this reason I will rewrite the conditions (1) and (2) as conditions on *sets* of arguments, denoted by R-PA, and I will define the set of preferred arguments as the smallest set of arguments satisfying these conditions.

DEFINITION 3.7

(*preferred arguments*). Let R-PA be a set of arguments satisfying the following two requirements

(1) If A $\in$ R-PA, then for all subarguments A′ of A: A′ $\in$ R-PA;

(2) A is in R-PA iff for all arguments B and formulas ϕ such that A minimally interferes with respect to ϕ, B or one of its subarguments is R-defeated by another of R-PA, or A R-defeats B with respect to ϕ.

Then an argument is an *R-preferred argument* iff it is in the smallest set R-PA satisfying these requirements.

COROLLARY 3.8

For any set R-PA it holds that its elements do not interfere with each other.

The reason why this definition does not loop in example 3.6 is that we can construct a set of PA from which all arguments involved in the loop are removed. This set still satisfies the conditions (1) and (2) of definition 3.7: the fact that PA does not contain A1 and A2 satisfies condition (1) since not all of their subarguments are in PA, and the fact that PA does not contain A1′ and A2′ satisfies condition (2) since these arguments are S-defeated by arguments which do not interfere with any other element of PA.

It should of course be proven that the set of preferred arguments is indeed the smallest, i.e. a *unique* set satisfying (1) and (2).

PROPOSITION 3.9

For all R there is a smallest set R-PA satisfying conditions (1) and (2) of definition 3.7.

Because of this result definition 3.7 can without harm be read as its simpler version 3.5. This is what I will do in the rest of this paper.

It is often useful to speak of a "defeated (sub-) argument".

DEFINITION 3.10

A is an R-*defeated* (*sub-*) *argument* iff A explains a formula ϕ such that there is an R-preferred argument for $\neg\phi$.

COROLLARY 3.11

If (F, D) is an R-defeated argument, then for every superset $D' \supset D$ the argument (F, D′) is R-defeated.

An interesting aspect of definition 3.7 is that it leaves room for a non-empty class of arguments which are neither preferred, nor defeated. For instance, in example 3.6 all arguments are of this type. The significance of this class is that an argument need not itself be preferred in order to prevent another argument from being preferred; it need merely be defensible. Assume by way of illustration that A and B minimally interfere with each other with respect to ϕ and that they do not interfere with any other argument. Assume furthermore that neither of the arguments R-defeats the other. Then neither of them is preferred since they do not satisfy condition (2), but then also neither of them is defeated, since the only argument with which they interfere is not preferred. For this reason a third class of so-called defensible arguments can be defined.

DEFINITION 3.12

An argument is *R-defensible* iff it is neither an R-preferred nor an R-defeated argument.

At this stage the third part of the framework can be defined, the set of *defeasible knowledge* of a default theory plus a kind of defeat R: this set should contain the facts for which there is an argument which according to R is better than any competing argument. Accordingly, it is simply defined as the collection of all formulas explained by some preferred argument. The term "defeasible" refers to the nonmonotonic nature of the theory: if the set of facts F of a default theory (F, δ) is extended, then it might be that some consequences cease to be preferred.

DEFINITION 3.13

The set of *defeasible knowledge* $DK(F, \delta, R)$ is the set of all formulas explained by an R-preferred argument (F, D) such that $D \subseteq \delta$.

A natural requirement for a theory of preferred defeasible knowledge is that if a formula is deductively implied by preferred formulas, it is also itself preferred. In the present framework this depends on the kind of defeat which is used (cf. section 4).

This completes the design of a general framework for comparing arguments. It should again be stressed that the framework does not refer to any specific kind of defeat, for which reason it can serve as a general framework for any theory of comparing pairs of arguments. This will be further illustrated in the next section, where alternatively two kinds of defeat will be defined.

4. Kinds of defeat

In this section two ways of comparing pairs of arguments will be analyzed: Poole's specificity definition and using hierarchical orderings of the set of defaults.

4.1. SPECIFICITY

As we concluded in section 2.1, one problem with Poole's specificity definition is that it is not embedded in a general context for comparing arguments, for which reason Poole is not able to specify when and how his theory comparator should be used. As a solution to this problem I will now embed Poole's specificity definition in the framework developed in the previous section. The adaptation of Poole's original definition to default logic is straightforward, but first a problem motivating a small change of this definition should be discussed; the discovery of this problem is due to Loui and Stiefvater [10], who solve it in a different way than I will do. As the next example shows, the notion of a "possible situation in which an argument applies" should be defined carefully.

EXAMPLE 4.1

$A1 = Fc \cup \{e \Rightarrow c \quad d \Rightarrow b \quad b \wedge c \Rightarrow a\}$

$A2 = Fc \cup \{d \Rightarrow b \quad b \Rightarrow \neg a\}$

$Fc = \{d, e\}$

Intuitively, A1 should be strictly more specific with respect to a than A2. However, according to Poole's definition this is not the case, since there is a possible fact making A1 explain a without making A2 explain $\neg a$: this possible fact is $c \wedge (c \supset b)$. The problem is that this fact "sneaks" a new way of deriving an intermediate conclusion into the argument by introducing a new "link" $c \supset b$, thereby intuitively making it a different argument: with the possible fact the argument A1 uses another default to explain b than with the actual facts, viz. $e \Rightarrow c$ instead of $d \Rightarrow b$, and therefore it cannot be said that A1 *itself* applies in the possible situation.

The solution is to require that if an argument $A = (Fc \cup Fn, D)$ explains ϕ, a possible fact Fp can only be said to make A *itself* explain ϕ if $(\{Fp\} \cup Fn, D)$ does not use other defaults to explain ϕ than A does. Applied to example 4.1, what we do not want to have is that $e \wedge (c \supset b)$ is a possible fact making A1 explain b; only facts with which A1 uses the same defaults (viz. $d \Rightarrow b$) to explain b as with Fc should be "valid" possible facts making A1 explain b. In fact, it is even a bit more complicated, since the same should hold for eventual other conclusions made explainable by the possible fact: also this should be done via the same defaults as with Fc. Now the set of defaults "used" by an argument to explain a formula will be defined as a *minimal* set of defaults of which the consequents together with the facts deductively imply the formula.

DEFINITION 4.2

Let $A = (F, D)$ be a normal default theory. Then a *ϕ-implying set A^{ϕ}* is a minimal set $D' \subseteq D$ such that $F \cup CONS(D') \models \phi$. If ϕ is left unspecified, I will also call this a *conclusion implying set.*

Consider by way of example the argument $A = (\{a\}, \{a \Rightarrow b, b \Rightarrow c\})$. $A^c = \{b \Rightarrow c\}$, but $A^{b \wedge c} = \{a \Rightarrow b, b \Rightarrow c\}$. A further complication in solving the problems is that even if A minimally explains ϕ, A^{ϕ} need not be unique: consider $D = \{a \Rightarrow (b \wedge (c \supset d)),$

$b \Rightarrow (e \wedge (c \supset d)), e \Rightarrow c\}$ and $\phi = d$. Then $(\{a\},D)$ has two d-implying sets, viz. $\{a \Rightarrow (b \wedge (c \supset d)), e \Rightarrow c\}$ and $b \Rightarrow (e \wedge (c \supset d)), e \Rightarrow c\}$. For this reason definition 4.3, defining when a possible fact makes an argument explain a conclusion, has to quantify over ψ-implying sets. In this definition $\sqcup A^{\phi}$ denotes the set of all sets A^{ϕ} of an argument A explaining ϕ.

DEFINITION 4.3

If $A = (Fc \cup Fn, D)$ explains ϕ, then a possible fact Fp *makes A explain* ϕ iff

(1) $E(Fn \cup \{Fp\}, D) \models \phi$

(2) for all ψ explained by A such that $E(Fn \cup \{Fp\}, D) \models \psi$: $\sqcup(Fn \cup \{Fp\}, D)^{\psi} = \sqcup A^{\psi}$.

The idea of clause (2) is that for all formulas explained by A with the possible fact Fp it needs exactly the same defaults for doing so as it needs with the contingent facts Fc; this holds for both all its final conclusions and all its subconclusions. This in fact reduces the role of possible facts to "filling in" prerequisites of defaults, which seems to be their proper role.

I admit that thus the theory on specificity has become rather complex, but it should be stressed that this problem is not caused by the present framework: it is a problem of Poole's original specificity definition; my framework is designed in such a way that if a better way of formalizing specificity is found, it can also be incorporated in the framework. Anyway, the adaptation of Poole's specificity definition to default logic is now straightforward.

DEFINITION 4.4

(1) $A1 = (Fc \cup Fn, D1)$ is *more specific than* $A2 = (Fc \cup Fn, D2)$ *with respect to* ϕ (m.s.$_{\phi}$) iff: if A2 explains $\neg\phi$, then there is a possible fact Fp such that:

Fp makes A2 explain $\neg\phi$;

Fp does not make A1 explain ϕ;

Fp does not make A1 explain $\neg\phi$(*non-triviality*);

(2) A1 is *strictly more specific* (*s.m.s.*) *than* A2 *with respect to* ϕ iff A1 m.s.$_{\phi}$ A2 and not A2 m.s.$_{\phi}$ A1.

At first sight it would now seem to suffice to say "A1 S-defeats A2 with respect to ϕ iff A1 s.m.s.$_{\phi}$ A2". However, again a problem has to be discussed, which this time is due to the present framework. Consider the following example.

EXAMPLE 4.5

D1: $\{a \Rightarrow b,\ b \Rightarrow c,\ c \Rightarrow d\}$

D2: $\{a \Rightarrow b,\ b \Rightarrow c,\ c \Rightarrow e,\ e \Rightarrow \neg d\}$

Fc = $\{a\}$

It would seem that $A = (Fc, D1)$ is preferred, since A1 is strictly more specific with respect to d than $A2 = (Fc, D2)$: all possible facts which make A1 explain d imply c and therefore make A2 explain $\neg d$, while, on the other hand, e is a possible fact making A2 explain $\neg d$ without making A1 explain d. However, if the example is examined more closely, more formulas can be found about which A1 and A2 are in conflict: one of them is the conjunction of all intermediate conclusions of A2 and its final conclusion, $b \wedge c \wedge e \wedge \neg d$. Clearly, this formula is explained minimally by A2 and, moreover, its negation is explained by A1, since $d \models \neg(b \wedge c \wedge e \wedge \neg d)$. Now, with respect to this formula A1 is not strictly more specific than A2, since c is a possible fact which makes A1 explain $\neg(b \wedge c \wedge e \wedge \neg d)$ without making A2 explain the opposite. Clearly, this is counterintuitive, and therefore conflicts about conjunctions of intermediate and final conclusions should not count for S-defeat if there are also conflicts about "real" final conclusions. To meet this requirement a notion of "conflicting with" will be used, which is defined in definition 4.6.

DEFINITION 4.6

A *conflicts* with A′ *with respect to* ϕ iff

(1) A and A′ minimally interfere with respect to ϕ;

(2) If ϕ or $\neg\phi$ is deductively implied by a conjunction of a final conclusion ψ and an intermediate conclusion χ of A but not by ψ alone, then A and A′ do not minimally interfere with respect to ψ.

In words this definition says that if ϕ or $\neg\phi$ is a "conjunctive" conclusion of one of the arguments, then there should be no "conjoining" conclusion with respect to which they also minimally interfere. In example 4.5 this ensures that A1 and A2 do not conflict with respect to the conjunctive conclusion $b \wedge c \wedge e \wedge \neg d$, since this formula is the conjunction of the intermediate conclusions b, c and e and the final conclusion $\neg d$, it is not implied by $\neg d$, while finally, A1 and A2 also minimally interfere with respect to d.

Now S-defeat can be defined.

DEFINITION 4.7

A1 *S-defeats* A2 with respect to ϕ iff: for all formulas ψ such that A1 and A2 conflict with respect to ψ, A1 s.m.s.$_\psi$ A2.

REMARK 4.8

Because of definition 4.7 clause (2) of definition 3.7 can for S-defeat be read as: For all formulas ϕ and arguments A′ such that A conflicts with A′ with respect to ϕ and A′ does not interfere with another S-preferred argument: A s.m.s.$_\phi$ A′. The reason is that for every ϕ mentioned in definition 4.7 the set of all ψ mentioned in this definition is the same.

The next proposition ensures that the assumption of the previous section that R-defeat is asymmetric is justified for S-defeat.

PROPOSITION 4.9

S-defeat is asymmetric.

In the previous section it was said that the deductive closure of DK depends on the kind of defeat which is used. For S-defeat the deductive closure indeed holds.

THEOREM 4.10

For every normal default theory (F, δ), $DK(F, \delta, S)$ is deductively closed.

Furthermore, it can be shown that DK is the extension of the default theory consisting of F and the set of all defaults which are used in any S-preferred argument, which set is denoted by D^{pref}.

THEOREM 4.11

For every normal default theory (F, δ), $DK(F, \delta, S)$ is the unique extension of (F, D^{pref}).

4.2. ORDERINGS OF DEFAULTS

In this subsection I will formalize reasoning with hierarchically ordered premises. In order to apply the framework to this kind of reasoning I will replace definition 4.7 of S-defeat by a definition of "hierarchical defeat", which assumes the set of defaults δ to be ordered by a partial preorder $\leq_\delta$ (normally, the subscript will be omitted). To satisfy the conclusions of section 2, the definitions should take care of two things: they should identify the defaults which are relevant to a conflict, and they should tell how to compare the relevant sets of conflicting arguments. The relevant defaults are picked out by definition 4.14, according to the following idea. Informally, since intermediate conclusions should not be relevant to a conflict between arguments, we would like to identify exactly the defaults which are "at the end of the argument chain" for ϕ; intermediate conclusions are then dealt with by the inductive part of definition 3.7. Note that, since the test for H-defeat will only be applied to formulas which are explained minimally, it suffices to define the relevant set for such formulas. Now for minimally explained formulas there will be no rules at the end of the chain (which we will call "top rules") which are not needed for explaining it, otherwise they should be deleted from the argument, which would mean that the argument does not explain the formula minimally: therefore we can regard all top rules of the argument as relevant. This leads to the following definition, which is inspired by a similar definition in [10].

DEFINITION 4.12

Let ϕ be a formula minimally explained by $A = (F, D)$.

- $d \in D$ is a *top rule* of A iff $(F, D-\{d\})$ explains the prerequisite of d.
- The *ϕ-relevant set* of A, denoted by $[\phi]A$, is the set of all top rules of A.

Consider by way of illustration $F = \{a\}$, $d1 = a \Rightarrow b$, $d2 = b \Rightarrow c$. Now $[b](F, \{d1\}) = \{d1\}, [c](F, \{d1, d2\}) = [b \wedge c](F, \{d1, d2\}) = \{d2\}$. Note that d1 is not in the latter set, which is in fact the relevant set of a conjunction of an intermediate and a "real" final conclusion; this shows that the problems with defining S-defeat illustrated by example 4.5 do not occur with H-defeat, since defaults not needed for the "real" final conclusion of an argument will not be among the top rules of the argument.

The following fact ensures that the priority test will always be applied to unique relevant sets.

FACT 4.13

If an argument A minimally explains a formula ϕ then $[\phi]A$ is unique.

For comparing the relevant sets of conflicting arguments a reasonable standard seems to be the one according to which an argument A for ϕ hierarchically defeats an argument A′ for $\neg\phi$ iff all members of $[\phi]A$ are higher than *all* members of $[\neg\phi]A'$. Below $\leq_{\wp(\delta)}$ is an ordering on the power set of δ, i.e. on the set of all subsets of δ. The subscript will be omitted if there is no danger for confusion.

DEFINITION 4.14

1. $A = (F, D)$ is *higher than* $A' = (F, D')$ *with respect to* ϕ (notation: $A >_\phi A'$) iff for every $d \in [\phi]A$ and $d' \in [\neg\phi]A'$ holds: $d >_\delta d'$. (notation: $[\phi]A >_{\wp(\delta)} [\neg\phi]A'$)
2. An argument A *H-defeats* an argument A′ with respect to ϕ iff $A >_\phi A'$.

For H-defeat the same formal properties hold as for S-defeat.

PROPOSITION 4.15

H-defeat is asymmetric.

THEOREM 4.16

For every normal default theory (F, δ), $DK(F, \delta, H)$ is deductively closed.

THEOREM 4.17

For every normal default theory (F, δ), DK(F, δ, H) is the unique extension of (F, D^{pref}).

4.3. COMBINING KINDS OF DEFEAT

Once various individual kinds of defeat have been investigated, the question naturally arises as to whether they can or even must be combined. Often specificity is regarded as the only criterion for comparing contradicting defeasible conclusions, but philosophically this is inadequate; for example, in legal reasoning conflicts between arguments are solved by a combined application of several conflict resolution metarules, based on hierarchical relations between norms, on specificity and on the time of enactment of norm. The general nature of the present argumentation framework also allows the definition of a combined use of kinds of defeat. However, because of space limitations the reader is referred to [17] for an extensive treatment of this topic.

5. Some applications

In this section the formal theory of the previous sections is applied to some examples. I will alternatively use S-defeat and H-defeat.

EXAMPLE 5.1

First it is shown how the present framework copes with the snoring-professor.

(1) Misbehaves(x) ⇒ May-be-removed(x)

(2) Professeor(x) ⇒ ¬May-be-removed(x)

(3) Snores(x) ⇒ Misbehaves(x)

Fc = {Professor(Bob) ∧ Snores(Bob)}, δ = {1-3}; 1 ≈ 2; 3 < 1

A1 = (F, {1,3}) is an argument for May-be-removed(Bob), while A2 = (F, {2}) is an argument for the opposite. If we apply H-defeat, then the sets relevant to the conflict are [May-be-removed(Bob)]A1 = {1} and [¬May-be-removed(Bob)]A2 = {2}. Since the only elements of these sets are of equal level, neither A1, nor A2 is preferred: both are merely H-defensible arguments. Note that, as desired, {3} is not in any relevant set.

EXAMPLE 5.2

The next example shows that the definitions are capable of handling exceptions to exceptions. It consists of a rule stating that contracts bind only the parties involved, an exception to this rule saying that lease contracts of houses also bind new owners of the house and an exception to this exception in case the tenant has agreed by contract with the opposite. In default logic:

(4) Contract(x) $\Rightarrow$ Binds-only-parties(x)

(5) House-lease-contract(x) $\Rightarrow$ ($\neg$Binds-only-parties(x) $\wedge$ Binds-all-owners(x))

(6) House-lease-contract(x) $\wedge$ Tenant-agreed-by(x) $\Rightarrow$ Binds-only-parties(x)

Fn = {$\forall$x(House-lease-contract(x) $\supset$ Contract(x))}

Fc = {House-lease-contract(c) $\wedge$ Tenant-agreed-by(c)}, δ = {4-6}

A3 = (Fc $\cup$ Fn, {4}) explains Binds-only-parties(c), A4 = (Fc $\cup$ Fn, {5}) explains $\neg$Binds-only-parties(c), and A5 = (Fc $\cup$ Fn, {6}) again explains Binds-only-parties(c). Clearly A4 is strictly more specific with respect to Binds-only-parties(c) than A3. However, in order to be S-preferred, A4 must also defeat A5, but it is the other way around, since A5 $\text{s.m.s.}_{\text{Binds-only-parties(c)}}$ A4. Therefore A5 is an S-preferred argument and the consequent of (5) is not in DK. Note also that A5 *reinstates* A3.

EXAMPLE 5.3

This example illustrates the inductive part of the definitions.

(7) Likes-beer $\Rightarrow$ $\neg$Individualistic

(8) Young $\Rightarrow$ Ambitious

(9) Ambitious $\Rightarrow$ Individualistic

(10) Young $\wedge$ Unemployed $\Rightarrow$ $\neg$Ambitious

Fc = {Likes-beer, Young, Unemployed}, δ = {7-10}

A6 = (Fc, {7}) explains "$\neg$Individualistic" while A7 = (Fc, {8,9}) explains "Individualistic". Although not A6 $\text{s.m.s.}_{\text{Individualistic}}$ A7, A6 is still S-preferred, since A7 contains an S-defeated subargument, viz. (Fc, {8}), explaining "Ambitious", which is defeated by (Fc, {10}), explaining the opposite.

EXAMPLE 5.4

Assume example 5.1 is modified to the effect that the lower regulation itself tries to put a sanction on misbehaviour.

(1) Misbehaves(x) $\Rightarrow$ May-be-removed(x)

(2) Professor(x) $\Rightarrow$ $\neg$May-be-removed(x)

(3′) Snores(x) $\Rightarrow$ Misbehaves(x) $\wedge$ May-be-removed(x)

F = {Professor(Bob) $\wedge$ Snores(Bob)}, δ = {1,2,3′}; $1 \approx 2$; $3 < 1$

This time the outcome is that $\neg$May-be-removed(Bob) is H-preferred, since (F, {1,3′}), explaining the opposite, now has a subargument explaining May-be-removed(Bob), viz. (F, {3′}), and [May-be-removed(Bob)](F, {3′}) = {3′}, [¬May-be-removed(Bob)](F, {2}) = {2} and 3′ < 2. Hence, (F, {1,3′}) now has a defeated subargument and is therefore itself defeated, although there is no argument against Misbehaves(Bob). At first sight this seems unsatisfactory, but in my view the correct answer in this example depends on whether the two consequents of (3′) are regarded as connected or not, and since this is not a logical but a legal matter, a formal system should have ways of formalizing both possibilities. In the present system this can indeed be done: the alternative interpretation can be presented if 3′ is split into the next two defaults.

3″ Snores(x) ⇒ Misbehaves(x)

3‴ Snores(x) ⇒ May-be-removed(x)

Thus (F, {3″}) is a preferred argument for Misbehaves(Bob), for which reason the same outcome is obtained as in the original example.

EXAMPLE 5.5: INAPPLICABILITY OF DEFAULTS

Sometimes a default $\phi \Rightarrow \psi$ is regarded inapplicable in special circumstances, but the opposite conclusion $\neg\psi$ is not drawn. Pollock [12] calls such exceptions "undercutting defeaters". In the present framework this can be formalized with the well-known techniques of general exception predicates and naming defaults. The idea is to add to the prerequisite of a general rule a condition expressing "there is no exception to this rule", and to assert the truth of this condition by way of an unconditional default. Below (11) is such a general rule and (12) such an unconditional default, while (13) is an undercutting defeater. Note that (12) contains a variable for names of defaults, for which reason it can make any other default inapplicable.

(11) $Ax \wedge \neg Exc(11,x) \Rightarrow Bx$

(12) $\Rightarrow \neg Exc(n,x)$

(13) $Cx \Rightarrow Exc(11,x)$

Now if Fc = {Aa, Ca}, then the effect of (13) is that A8 = (Fc, {13}) s.m.s.$_{Exc(11,a)}$ A9 = (Fc, {12}), for which reason A10 = (Fc, {11, 12}), explaining Ba, contains an S-defeated subargument, for which reason Ba is not in DK. However, neither is ¬Ba, since it is not explained by any argument.

EXAMPLE 5.6

The last example illustrates an issue for further research.

(14) $a \Rightarrow (b \wedge d)$

(15) $a \Rightarrow (\neg b \wedge d)$

(16) $a \Rightarrow (b \wedge \neg d)$

$F = \{a\}$, $\delta = \{14\text{-}16\}$; $15 \approx 16$, $14 < 15$

Both if only (15) and if only (16) were in δ, $(F, \{14\})$ would be H-defeated; however, since (15) and (16) are both in δ, the irresolvable conflict between $(F, \{15\})$ and $(F, \{16\})$ secures $(F, \{14\})$ from being defeated, which intuitively seems strange: it seems that at least $(b \veebar d)$ should be in DK ($\veebar$ stands for "exclusive or").

6. Related research

In recent years, the idea of regarding defeasible reasoning as constructing and comparing alternative arguments has also been developed by some other researchers. The most close to the present investigations are Simari and Loui [22], who combine Poole's theory comparator with the ideas of Pollock [12] on the interaction of arguments. Their definition of arguments and subarguments is almost the same as mine and they, too, use a metalinguistic connective to represent defeasible statements; although they do not regard this connective as a Reiter-like default, their and my requirements are internally coherent (in my terms: that they have a unique extension) seems to make the two connectives equivalent with respect to the argumentation systems. Nevertheless, there are some significant differences with the present approach. The first is the definition of a preferred argument, for which Simari and Loui adapt the one of [12], which does not explicitly reflect the step-by-step nature of argumentation. Other differences are that Simari and Loui only consider specificity as a kind of defeat, have no proof of the deductive closure of the set of preferred conclusions and do not discuss the problem of conjunctions of intermediate and final conclusions. On the other hand, they prove more formal properties of arguments and they thoroughly discuss the prospects for implementation, with some interesting formal results.

Formally less close to the present framework but concerned with similar problems is Vreeswijk [23]. He uses an idea which is also used by Lin and Shoham [9]; the language in which defeasible information is formulated is left unspecified and defeasible inference rules are not domain specific, as in [22] and the present system, but general inference rules, assumed to receive their justification from the logical interpretation of the object language. For example, Vreeswijk copes with defeasible statements by assuming an object language with a defeasible connective $a > b$ and by stating a defeasible inference rule of the form $\{a > b, a\} \Rightarrow b$, assuming that this is justified by the logical interpretation of $>$. Unlike [9], which is only concerned with capturing existing nonmonotonic logics in its framework, Vreeswijk adds to this the possibility to compare arguments. Interestingly, while [22] and the present framework define arguments as deductively closed, Vreeswijk regards, like [9], deductive derivations as an explicit argument step.

7. Implementation

The aim of this paper has not been to give a procedure for determining which arguments are preferred, but to give a definition of what it means that an argument is preferred. As a consequence, the present theory is not very well suited for a straightforward implementation. Moreover, implementing the full theory is problematic for a number of reasons. Firstly, it uses the full expressive power of first-order predicate logic, for which as a whole to date no theorem provers exist which are both complete and efficient. Furthermore, default logic is known to be non-semidecideable, in the sense that there is no algorithm which guarantees that every formula which is in some extension is shown to be in an extension [18]. Finally, unlike theorem provers for standard logics, which can stop when a proof has been found, systems which try to find *best* argument will have to continue searching the whole space of possible counterarguments.

In practice, problems of efficiency may be dealt with by restricting the language to an efficiently computable fragment, for example, to clause logic, as in [11]. Moreover, efficiency may be increased by sacrificing completeness with respect to our theory. Nevertheless, however difficult the implementation of the theory developed in this paper may be, it does at least make it possible to formulate exactly in which respects practical applications are or have to be imperfect, which makes the present study valuable as a piece of fundamental research.

8. Conclusion

Inspired by a formal analysis of legal reasoning, this article has investigated nontrivial reasoning with inconsistent information. Two important conclusions have emerged, both revealing that the formalization of this kind of reasoning is more complicated than is generally acknowledged in AI research. The first conclusion is that reasoning with inconsistent information should be modelled as constructing and comparing incompatible arguments, in a way which reflects the step-by-step nature of argumentation. However, this is not sufficient, since the second conclusion is that a theory of argumentation still faces serious problems if standard logic is not abandoned as the knowledge representation language: the reason is that the strong logical properties of the material implication give rise to arguments which in actual reasoning are not constructed. For this reason approaches to formalizing nonmonotonic reasoning by changing the way logic is used rather than changing the logic itself are far less attractive than is often claimed. It has turned out that the argumentation framework in default logic developed in this study respects the two conclusions. A further attractive feature of the framework is that it allows for the definition of any particular standard for comparing pairs of arguments, including a combination of these standards.

Appendix A: Some definitions and theorems of Reiter [18]

DEFINITION A.1

(*defaults*) [18, p. 88]. A *default* is any expression of the form $\alpha(x):\beta_1(x),\ldots,\beta_m(x)/\omega(x)$, where $\alpha(x)$, $\beta_1(x),\ldots,\beta_m(x)$ and $\omega(x)$ are first-order predicate logic well-formed formulas of which the free variables are among those of $x = x_1,\ldots,x_n$. $\alpha(x)$ is called the *prerequisite*, $\beta_1(x),\ldots,\beta_m(x)$ the *justifications* and $\omega(x)$ the *consequent*. A *normal default* is a default of the form $\alpha(x):\omega(x)/\omega(x)$.

A *default theory* is a pair (F,δ) where δ is a set of defaults and F a set of closed first-order predicate logic well-formed formulas. A *normal default theory* is a default theory with only normal defaults.

DEFINITION A.2

If δ is any set of defaults, then

PRE(δ) is the set of prerequisites of all defaults of δ.

CONS(δ) is the set of consequents of all defaults of δ.

DEFINITION A.3

(*extension*) [18, def. 1]. Let (F,D) be a closed default theory. For any set of closed wff's let $\Gamma(S)$ be the smallest set satisfying the following three properties:

(i) $F \subseteq \Gamma(S)$

(ii) $Th(\Gamma(S)) = \Gamma(S)$

(iii) If $\alpha:\beta_1,\ldots,\beta_m/\omega \in D$ and $\alpha \in \Gamma(S)$, and $\neg\beta_1,\ldots,\neg\beta_m \notin S$, then $\omega \in \Gamma(S)$.

A set of closed wff's E is an *extension* of (F,D) iff $\Gamma(E) = E$, i.e. iff E is a fixed point of the operator Γ.

PROPOSITION A.4

[18, Th. 2.5]. Suppose E is an extension for a closed default theory (F,δ). Then $E = Th(F \cup CONS(GD(F,\delta)))$.

REMARK

GD(F,δ) is the set of members of δ which are applicable, i.e. of which the justifications are consistent with E and therefore in E. By definition 3.1 all defaults of an argument are applicable. Hence, proposition A.4 and definition 3.1 imply that a formula ϕ is explained by an argument (F,D) iff $F \cup CONS(D) \models \phi$.

PROPOSITION A.5

[18, corr. 3.4]. Suppose (F,δ) is a closed normal default theory such that $F \cup CONS(\delta)$ is consistent. Then (F,δ) has a unique extension.

PROPOSITION A.6

(*semi-monotonicity of normal default theories*) [18, th. 3.2]. Suppose D and D′ are sets of normal defaults with $D' \subseteq D$. Let E′ be an extension for the normal default theory $\delta' = (F, D')$ and let $\delta = (F, D)$. Then δ has an extension E such that

(1) $E' \subseteq E$ and

(2) $GD(E', \delta') \subseteq GD(E, \delta)$.

Appendix B: Proofs[1)]

COROLLARY 3.2

(i) Every argument has only a finite number of subarguments.

(ii) If A is an argument explaining ϕ but not minimally, then A has a subargument minimally explaining ϕ.

Proof of (i)

Every argument has a finite number of defaults and every argument of a given default theory (F, δ) contains F. □

Proof of (ii)

Observe first that for no $D' \subseteq D$ (F, D') has more than one extension, since otherwise by semimonotonicity (proposition A.6) also (F,D) would have more than one extension and not be an argument. Furthermore, since D is finite, there is a smallest subset $D' \subseteq D$ such that $E(F, D')$ contains ϕ. If (F, D') is not an argument, then for some $d \in D'$ its consequent is not in $E(F, D')$; but then d is not among the generating defaults of $E(F, D')$, for which reason by proposition A.4 it can be deleted from D′ without affecting the content of $E(F, D')$. If this is repeated for all inapplicable elements of D′, this results in a set $D'' \subseteq D'$ with only applicable defaults and such that $E(F, D'') = E(D, F')$, for which reason (F, D'') is an argument explaining ϕ. □

COROLLARY 3.8

For any set R-PA it holds that its elements do not interfere with each other.

Proof

By definition 3.4 every pair of arguments A1, A2 interfering with each other explain contradictory formulas ϕ and $\neg\phi$, respectively. Then by corollary 3.2 there is a subargument A1′ of A1 minimally explaining ϕ and a subargument A2′ of A2 minimally explaining $\neg\phi$. Then by clause (2) of definition 3.7 and

1) In this appendix the kind of defeat will be left implicit if there is no danger of confusion.

asymmetry of R at most one of A1′,A2′ R-defeats the other, for which reason for any set R-PA at most one A1′,A2′ and therefore of A1,A2 can be in R-PA. □

PROPOSITION 3.9

For all R there is a smallest set R-PA satisfying (1) and (2).

Proof

Below the fact that the notions "minimal interference" and "defeat" are relative to formulas will for ease of notation be left implicit. Now consider the set PA = $\cap_{i=1}^{n}$ PAi. which is the intersection of all sets PAi ($1 \leq i \leq n$) satisfying the conditions (1) and (2) of definition 3.7. It will be shown that PA satisfies (1) and (2) as well, because of which it is by its construction the smallest set doing so. The proof needs the following observation.

OBSERVATION B.1

Since all of PA1, . . . ,PAn satisfy corollary 3.8, by construction of PA no element of $\cup_{i=1}^{n}$ PAi has counterarguments in PA. □

I will abbreviate $\cup_{i=1}^{n}$ PAi as ∪PAi. Consider any argument C′ such that C ∉ PA. I will concentrate on the smallest subargument C of C′ for which this holds: such a smallest subargument exists by corollary 3.2. Then all subarguments of C are, since they are in PA, by construction of PA also in all of PA1, . . . , PAn. Now the proof has to distinguish two kinds of arguments not being in PA, depending on whether they are or are not in ∪PAi.

Assume first that C ∉ ∪PAi and assume that C has no counterarguments in ∪PAi: then it interferes with no element of ∪PAi. But since all its subarguments are in all of PA1, . . . , PAn, C is by definition 3.7 also in all of PA1, . . . , PAn, for which reason it is by construction of PA in PA. But this contradicts the assumption that it is not. Therefore C has a counterargument D in ∪PAi defeating C, and since by observation B.1 the argument D has no counterarguments in PA, D does not interfere with any number of PA. But then, since C does not defeat D, PA satisfies (2) of definition 3.7 in not containing C.

The second kind of argument which is not in PA is members of ∪PAi. Consider any such argument A ∈ ∪PAi: it will be shown that PA satisfies (1) and (2) in not containing A. Two cases have to be considered. The first is that some subargument A′ of A is not in PA. Then PA satisfies (1) in not containing A. The second situation is that all subarguments of A are in PA. Then there are sets PAi and PAj ($1 \leq i,j \leq n$) such that A ∈ PAi but A ∉ PAj. Then PAj contains a counterargument B of A defeating A, otherwise by the construction of PA and the assumption that all subarguments of A are in PA and therefore in PAj, PAj violates (2) in not containing A. But then, since by observation B.1, B does not interfere with any member of PA, PA satisfies (2) in not containing A.

In conclusion, PA satisfies definition 3.7 in not containing A. □

PROPOSITION 4.9

S-defeat is asymmetric.

The proof needs the following lemma.

LEMMA B.2

For all arguments A and A′ and formulas ϕ such that A and A′ minimally interfere with respect to ϕ there is a formula ψ such that A and A′ conflict with respect to ψ.

Proof of lemma B.2

Let $A1 = (F, D1)$ and $A2 = (F, D2)$ such that A1 and A2 minimally explain ϕ and $\neg\phi$, respectively. Then by finiteness of an argument there exists a minimal ϕ-implying set $A1^{\phi}$ such that $F \cup CONS(A1^{\phi}) \cup CONS(D2)$ is inconsistent; likewise there exists a $\neg\phi$-implying set $A2^{\neg\phi}$ such that $F \cup CONS(A2^{\neg\phi}) \cup CONS(D1)$ is inconsistent. If neither ϕ nor $\neg\phi$ is a conjunction of a final and an intermediate conclusion, then A1 and A2 conflict with respect to ϕ and we are done. If otherwise, then assume without loss of generality that it is otherwise for A1, that is, that ϕ is deductively implied by a final conclusion ψ and an intermediate conclusion χ of A1, but not by ψ alone. Then, there is a ψ-implying set $A1^{\psi}$ such that $A1^{\psi} \subset A1^{\phi}$ and $F \cup CONS(A1^{\psi}) \cup CONS(D2)$ is inconsistent. Now if A1 and A2 also minimally interfere with respect to χ, then this contradicts the assumed minimality of $A1^{\phi}$. Therefore A1 and A2 do not minimally interfere with respect to ψ, for which reason they conflict with respect to ϕ. □

Proof of proposition 4.9

By corollary 3.2 for every pair of arguments (Ai, Aj) interfering with each other there is a pair of subarguments (Ai′, Aj′) minimally interfering with each other with respect to some formula ϕ. Then by lemma B.2 for some formula ϕ Ai′ and Aj′ conflict with each other. Furthermore, the relation "s.m.s.$_{\phi}$" of definition 4.4 is obviously asymmetric, and since this relation must by definition 4.7 hold for all formulas ϕ with respect to which arguments conflict in order to have a relation of S-defeat between Ai′ and Aj′, S-defeat is an asymmetric relation as well. □

THEOREM 4.10

For every normal default theory (F, δ), $DK(F, \delta, S)$ is deductively closed.

Proof

Note that the proof is relative to a fixed (F,δ): hence all Di occurring below are sets of ground instances of elements of δ. First a pair of S-preferred arguments is considered, Ai = (F,Di) and Aj = (F,Dj). It will be proven that if these S-preferred arguments are combined, the result is also an S-preferred argument. According to definition 3.7 this boils down to proving the following proposition.

PROPOSITION B.3

For every Ai = (F,Di) and Aj = (F,Dj): if Ai and Aj are S-preferred arguments, then

(i) Ai ∪ Aj is an argument;

(ii) All subarguments of Ai ∪ Aj are S-preferred;

(iii) For all formulas ϕ and arguments Ak such that Ai ∪ Aj minimally interferes with Ak with respect to ϕ and neither Ak nor one of its subarguments is S-defeated by another S-preferred argument: Ai ∪ Aj S-defeats Ak.

Proof of proposition B.3, (i)

This clause is proven by the following lemma and the fact that preferred arguments do not interfere with each other. The lemma says that the combination of two non-interfering arguments is again an argument.

LEMMA B.4

If Ai and Aj are non-interfering arguments (cf. def. 3.4), then Ai ∪ Aj is an argument.

Proof of lemma B.4

By proposition A.5 Ai ∪ Aj has a unique extension if F ∪ CONS(Di ∪ Dj) is consistent. Furthermore, since all defaults of an argument are by definition 3.1 applicable, for any argument A = (F,D) it holds that GD(A) = D, for which reason by proposition A.4 the extension E(A) = Th(F ∪ CONS(D)). By assumption E(Ai) ∪ E(Aj) is consistent, which means that Th(F ∪ CONS(Di)) ∪ Th(F ∪ CONS(Dj)) is consistent. Then also Th(F ∪ CONS(Di) ∪ CONS(Dj)) = Th(F ∪ CONS(Di ∪ Dj)) is consistent. But this is equal to E(Ai ∪ Aj), which because of proposition A.5 proves that Ai ∪ Aj has a unique extension. Furthermore, since Ai ∪ Aj contains only defaults of Ai and Aj, which are arguments, all defaults of Ai ∪ Aj are applicable. Hence Ai ∪ Aj is an argument. □

The proof of proposition B.3, (i) now immediately follows from lemma B.4 and proposition 4.9, saying that preferred arguments do not interfere with each other. □

COROLLARY B.5

If Ai is a preferred argument and Aj a not defeated argument then Ai $\cup$ Aj is an argument.

Proof of corollary B.5

Ai and Aj are not interfering with each other, since otherwise Aj would be defeated. Then by lemma B.4 Ai $\cup$ Aj is an argument. □

The clauses (ii) and (iii) of proposition B.3 are proven with induction on the definition of a preferred argument. Informally, the initial step takes care of the smallest subargument of Ai $\cup$ Aj, which is (F,{ }). Every inductive step consists of adding one more default either to the subargument of Ai created so far, or to the subargument of Aj created so far, and this in every possible order. The choice which default to add next is restricted by the requirement of definition 3.1 that every default of an argument is applicable: only defaults can be added of which the prerequisite is explained by the argument to which it is added.

Proof of proposition B.3, (ii,iii)

Initial step: If Di′ = Dj′ = the empty set, then Ai $\cup$ Aj = F is trivially preferred because, since F is assumed consistent, there are no interfering arguments.

Notation B.6: For any argument A = (F,D), A-1 is a maximal subargument of A, which means that it is a subargument of A obtained by deleting one element of D.

Induction hypothesis: For every Ai′ = (F,Di′) and Aj′ = (F,Dj′) such that Di′ $\subseteq$ Di and Dj′ $\subseteq$ Dj, and for every Aj′-1 : Ai′ $\cup$ Aj′-1 is a preferred argument.

Induction step: Consider an arbitrary pair of preferred arguments A1 = (F,D1) and A2 = (F,D2) such that D1 $\subseteq$ Di and D2 $\subseteq$ Dj. Then by the induction hypothesis both A1 $\cup$ A2-1 and A1-1 $\cup$ A2 are preferred. What is left to prove is that for all formulas ϕ and arguments A3 such that A1 $\cup$ A2 and A3 minimally interfere with respect to ϕ, A3 is defeated, since then A1 $\cup$ A2 is a preferred argument. First the following lemma is needed.

LEMMA B.7

If A = (F,D) and A′ = (F,D′) are non-interfering arguments and A $\cup$ A′ minimally explains ϕ, then there is a formula p, minimally explained by A, and a formula p′, minimally explained by A′, such that {p,p′} $\models \phi$ and neither p $\models \phi$, nor p′ $\models \phi$.

Proof

Consider a $(A \cup A')^{\phi}$. It contains elements of both D and D′, since otherwise A or A′ would individually explain ϕ. Let q and p′ be the conjunction of all elements of D and D′, respectively, which are in $(A \cup A')^{\phi}$. Then $F \cup \{q, p'\} \models \phi$. By compactness of first-order predicate logic the same holds for a minimal finite subset F′ of F. Let f be the conjunction of all elements of F′ and let p be $(q \wedge f)$. Then A explains p and $\{p, p'\} \models \phi$. Furthermore, A and A′ minimally explain p and p′, because if a subargument Ai of, say, A would explain p, then there would be a subargument of $A \cup A'$ explaining ϕ, viz. $Ai \cup A'$. □

Now a crucial step in the proof of proposition B.3 is the following. By lemma B.7 the arguments A1 and A2 minimally explain formulas p1 and p2 such that $\{p1, p2\} \models \phi$. But then $\{\neg\phi, p1\} \models \neg p2$ and $\{\neg\phi, p2\} \models \neg p1$. Hence if there is an argument A3 for $\neg\phi$, then $A1 \cup A3$ interferes with A2 and $A2 \cup A3$ interferes with A1. What will be shown is that the existence of such an A3 leads to a contradiction if A3 is not S-defeated by another preferred argument.

ASSUMPTION B.8 (for contradiction)

For some formula ϕ minimally explained by $A1 \cup A2$ there is an argument A3 minimally interfering with $A1 \cup A2$ with respect to ϕ and not defeated by another preferred argument.

Two situations must be considered: first the one in which not both $A1 \cup A3$ and $A2 \cup A3$ are an argument and then the situation in which they both are an argument. The first situation causes no problems.

FACT B.9

If $A1 \cup A3$ or $A2 \cup A3$ is not an argument, then A3 is defeated by A1 or A2.

Proof

If $A1 \cup A3$ or $A2 \cup A3$ is not an argument, then by corollary B.5 A1 and A3 or A2 and A3 interfere; but then by corollary 3.2 these arguments also minimally interfere with respect to some formula, for which reason by definition 3.10 A3 is already defeated by A1 or A2 individually. □

The second situation which has to be considered is that both $A1 \cup A3$ and $A2 \cup A3$ are an argument. Now an important observation is that since both A1 and A2 are preferred, both $A1 \cup A3$ and $A2 \cup A3$ are defeated. It is this observation which will make it possible to derive a contradicting refuting assumption B.8. To summarize, the situation is as follows.

SITUATION 1

For some argument A3 and formula ϕ there are formulas p1 and p2 such that

$\{p1, p2\} \models \phi$

A1 minimally explains p1 and is a preferred argument;

A2 minimally explains p2 and is a preferred argument;

A3 minimally explains $\neg\phi$;

A2 $\cup$ A3 explains $\neg$p1 and is a defeated argument;

A1 $\cup$ A3 explains $\neg$p2 and is a defeated argument.

The rest of the proof will be devoted to showing that this situation leads to a contradiction refuting assumption B.8. First I shall show that A2 $\cup$ A3 and A1 $\cup$ A3 have no defeated subarguments. Assume for contradiction that this is otherwise. Then three situations have to be considered, which will without loss of generality be done for A2 $\cup$ A3.

Note first that if A2 $\cup$ A3 = A2 then this contradicts the assumption that A2 is a preferred argument. Otherwise if A2 $\cup$ A3 = A3 then A3 is defeated by A1, contrary to assumption B.8.

Consider next the situation that there is a subargument A2′ of A2 such that A2′ $\cup$ A3 is defeated. Then by the same line of reasoning as the one leading to situation 1 it holds that A3 interferes with A1 $\cup$ A2′, but this argument is preferred by the induction hypothesis, for which reason A3 is defeated by another preferred argument, which contradicts assumption B.8 that it is not.

Let there finally be a subargument A3′ of A3 such that A2 $\cup$ A3′ is defeated. Then A1 $\cup$ A2 interferes with A3′, which contradicts the assumption B.8 that A3 is minimal in interfering with A1 $\cup$ A2. In conclusion, A2 $\cup$ A3 contains no defeated subarguments.

Observe now that, since clause (2) of definition 3.7 leaves room for reinstatement (cf. example 5.2), A1 need not itself defeat A2 $\cup$ A3 in order to be preferred. Therefore I will now derive from situation 1 the basic situation which justifies the qualifications "preferred" and "defeated" in situation 1. Let A4 be a preferred argument which is possibly distinct from A1 and which S-defeats A2 $\cup$ A3. Then for all formulas ϕ with respect to which A2 $\cup$ A3 conflicts with A4 it holds that A4 s.m.s.$_\phi$ A2 $\cup$ A3. Since A2 $\cup$ A3 minimally explains $\neg\phi$, there is by lemma B.7 a formula ψ, minimally explained by A2, and a formula α, minimally explained by A3, such that $\{\psi, \alpha\} \models \neg\phi$. Then $\{\phi, \psi\} \models \neg\alpha$ but neither $\psi \models \neg\alpha$, nor $\phi \models \neg\alpha$, since by assumption B.8 that A3 is not defeated by a preferred argument other than A1 $\cup$ A2, A3 does not interfere with A4 or A2, which are both preferred. Furthermore, by lemma B.2 there is a formula χ with respect to which A3 not only minimally interferes but also conflicts with A4 $\cup$ A2 and for which further the same holds as for α. Observe finally that all this also holds for an argument A5 S-defeating A1 $\cup$ A3 with respect to ψ and thereby reinstating A2.

Now the idea is to establish direct relations of defeat in situation 1 by replacing A1 by A4 and A2 by A5. However, it should then first be shown that A4 s.m.s.$_\phi$ A5 $\cup$ A3. Consider therefore an even more specific argument A6 explaining ϕ which *does* S-defeat A5 $\cup$ A3 with respect to ϕ. Then A6 also defeats

A2 ∪ A3, for which reason we can make A4 equal to A6. Since the same line of reasoning holds for A5 explaining ψ we can, in sum, say that A5 ∪ A3 is S-defeated with respect to ϕ by A4 and A4 ∪ A3 is S-defeated with respect to ψ by A5. But then we can concentrate on the following situation. Note that it is structurally similar to situation 1, apart from the difference that all defeated arguments in the situation are directly defeated by other arguments in the situation.

FACT B.10

There are arguments A4, A5 and formulas ϕ, ψ, and χ such that

- $\{\phi, \psi\} \models \neg\chi$
- A4 minimally explains ϕ and is a preferred argument such that A4 and A5 ∪ A3 conflict with respect to ϕ and A4 s.m.s.$_\phi$ A5 ∪ A3;
- A5 minimally explains ψ and is a preferred argument such that A5 and A4 ∪ A3 conflict with respect to ψ and A5 s.m.s.$_\phi$ A4 ∪ A3;
- A3 minimally explains χ;
- A5 ∪ A3 minimally explains $\neg\phi$ and is a defeated argument;
- A4 ∪ A3 minimally explains $\neg\psi$ and is a defeated argument.

REMARK B.11

Because of the various relations of interference none of ϕ, ψ and χ and their negations is entailed by the set F of facts. Below this will be left implicit.

Recall that the situation of fact B.10 has been constructed as the basic case justifying all relations of defeat in situation 1, which situation was the result of assumption B.8 for contradiction. Now, if an inconsistency can be derived from this situation, assumption B.8 has been refuted. In this derivation the next lemma is a key element.

LEMMA B.12

If A = (F,D) and A′ = (F,D′) minimally explain ϕ and ϕ', $\{\phi, \phi'\} \models \psi$ and A ∪ A′ minimally explains ψ, then every possible fact Fp makes A ∪ A′ explain ψ iff Fp makes A explain ϕ and A′ explain ϕ'.

Proof

⇐ Trivial.

⇒ : Since A ∪ A′ minimally explains ψ, it needs all defaults of D and D′ to explain ψ. Therefore it explains both ϕ and ϕ', and therefore for some ψ-implying set D″ = $(A \cup A')^{\phi}$ there is a $A^{\phi} \subseteq D''$ and a $A'^{\phi'} \subseteq D''$. Furthermore, by definition 4.3 (Fn ∪ {Fp}, D ∪ D′) has the same set of ψ-implying sets as A ∪ A′, for which reason D″ is also a ψ-implying set of (Fn $\cup$ {Fp},D $\cup$ D′). Then Fp also makes A explain ϕ and A′ explain ϕ'. □

We are now in the position to derive from fact B.10 a contradiction refuting assumption B.8. Note that all conditions of minimal explanation and of conflict which are implicitly assumed below are satisfied by fact B.10.

(1) since A3 $\cup$ A4 explains $\neg\psi$ and A5 S-defeats A3 $\cup$ A4 it holds that A5 s.m.s.$_\psi$ A3 $\cup$ A4; the same holds for A4 and A3 $\cup$ A5 with respect to ϕ.

(2) Every possible fact making A4 explain ϕ, makes A3 $\cup$ A5 explain $\neg\phi$ (*fact B.10*).

(3) Every possible fact making A4 explain ϕ, makes A3 explain χ and A5 explain ψ (*lemma B.12,2*).

(4) Every possible fact making A4 explain ϕ and A3 explain χ, makes A5 explain ψ (*4*).

(5) Every possible fact making A3 $\cup$ A4 explain $\neg\psi$, makes A5 explain ψ (*5, lemma B.12*).

(5) contradicts (1), which says that A5 s.m.s.$_\psi$ A4 $\cup$ A3. The only assumption for contradiction which can be retracted to restore consistency is assumption B.8, which says that there is a counterargument A3 against A1 $\cup$ A2. Together with fact B.9 this proves (ii) and (iii) of proposition B.3. □ (end of proof of proposition B.3,(ii,iii)).

COROLLARY B.13

If two formulas ϕ and ϕ' are in DK, and $\{\phi,\phi'\} \models \alpha$, then α is in DK.

Proof

By definition 3.13 a formula is in DK iff it has a preferred argument. Assume ϕ has a preferred argument A and ϕ' has a preferred argument A′. Now, by proposition B.3 and definition 3.7 A $\cup$ A′ is preferred and by definition 3.13 α, which is explained by A $\cup$ A′, is in DK. □

Now the proof of the deductive closure of DK(F,δ,S) (theorem 4.10) can be completed. According to the compactness theorem for first-order predicate logic, if a formula is implied by an infinite DK, it is implied by a finite subset of DK. For this reason it suffices to show for any finite subset Ω of DK: if $\Omega \models \phi$, then $\phi \in$ DK. This is proven by induction on the number (n) of elements of Ω.

Initial step: If $n \leq 2$, then if $n = 0$, then $\models \phi$ and $\phi \in$ DK by the definition of an argument, otherwise $\phi \in$ DK by corollary B.13 (for $n = 1$ let ϕ' be ϕ).

Induction hypothesis: If Ω is a finite subset of Ω with n elements, then for all $\Omega' \subset \Omega$ with n-1 elements and all ϕ: if $\Omega' \models \phi$, then $\phi \in$ DK.

Induction step: Assume $\Omega = \{\alpha_1, \ldots, \alpha_n\}$ is a finite subset of DK and $\Omega \models \phi$. Then by the deduction theorem for first-order logic $\{\alpha_1, \ldots, \alpha_{n-1}\} \models (\alpha_n \supset \phi)$, and by the induction hypothesis $(\alpha_n \supset \phi) \in$ DK. Since by assumption also $\alpha_n \in$ DK, by corollary B.13 $\phi \in$ DK. □ (end of proof of theorem 4.10).

THEOREM 4.11

For every normal default theory (F,δ), DK(F,δ,S) is the unique extension of (F,D^{pref}).

Proof

It must be shown that DK is the smallest set satisfying the following properties (definition A.3):

(i) $F \subseteq DK$

(ii) $Th(DK) = DK$

(iii) If $\alpha : \beta / \beta \in D^{pref}$ and $\alpha \in DK$, and $\neg\beta$ not $\in DK$, then $\beta \in DK$.

Proof of (i)

Since by definition 3.1 F is explained by every argument, it is also explained by every preferred argument, for which reason all elements of F are in DK.

Proof of (ii)

This is theorem 4.10.

Proof of (iii)

$\alpha \Rightarrow \beta \in D^{pref}$ iff $\alpha \Rightarrow \beta$ is in some preferred argument S. Then, since a preferred argument is an argument, both α and β are explained by A and by definition of DK both α and β are in DK.

To prove the minimality of DK, assume there is a $DK' \subset DK$ satisfying the above three properties. Then there is a formula $\phi \in DK$ for which there is a preferred argument A = (Fi,Di), but which formula is not in DK′. Assume that of all defaults $\alpha \Rightarrow \beta$ of Di α is in DK′. Then, if all β are in DK′ as well, because of (ii), clause (b) of definition 3.1 and the fact that $F \cup CONS(Di) \models \phi$, also φ is in DK′, which contradicts our assumption. But if some β is not in DK′, condition (iii) is violated. Therefore, of one element of Di the prerequisite α is not in DK′. Consider the smallest subargument A′ of A such that of some of its defaults the prerequisite is not in DK′. Then there is a subargument A″ of A′ such that $CONS(A'') \cup F \models PRE(A')$, since otherwise some defaults of Di are not applicable. Since by condition (iii) all elements of CONS(A″) and by condition (i) all elements of F are in DK′, by condition (ii) also all elements of PRE(A′) are in DK′, which contradicts the observation that at least one of them is not and thereby the assumption that DK′ ⊂ DK. □

FACT 4.13

If an argument A minimally explains a formula ϕ then $[\phi]$A is unique.

Proof

Assume A has two distinct $[\phi]$A, T1 and T2. Then (F,D-T1) is a subargument of A explaining ϕ, which contradicts the minimality of A. □

PROPOSITION 4.15

H-defeat is asymmetric.

Proof

Trivial from the observation that the $>_\phi$-relation between arguments is asymmetric. □

THEOREM 4.16

For every normal default theory (F,δ), DK(F,δ,H) is deductively closed.

Proof

The proof is largely similar to that of theorem 4.10. Below the differences will be discussed. Up to and including "situation 1" both proofs are completely similar. Then by a similar line of reasoning as from situation 1 above a situation can be obtained in which A1 ∪ A3 and A2 ∪ A3 are directly H-defeated by another argument in the situation.

FACT B.14

There are arguments A4, A5 and formulas ϕ, ψ, χ such that

$\{\phi, \psi\} \models \neg\chi$

A4 minimally explains ϕ and is a preferred argument such that A4 $>_\phi$ A5 ∪ A3;

A5 minimally explains ψ and is a preferred argument such that A5 $>_\psi$ A4 ∪ A3;

A3 minimally explains χ;

A5 ∪ A3 minimally explains $\neg\phi$ and is a defeated argument;

A4 ∪ A3 minimally explains $\neg\psi$ and is a defeated argument.

Now again the aim is to derive from this situation a contradiction refuting assumption B.8. For this derivation the following lemmas are needed.

LEMMA B.15

If $A = (F,D)$ minimally explains ϕ, $A' = (F,D')$ minimally explains ϕ', $\{\phi,\phi'\} \models \psi$ and $A \cup A'$ is an argument minimally explaining ψ, then $[\psi](A \cup A') \subseteq [\phi]A \cup [\phi']A'$.

Proof

Consider any $d \in [\psi](A \cup A')$. Then for all $d' \in D \cup D'$ it holds that $(F,(D \cup D')-\{d\})$ explains PRE(d'). Consider for any such d' a minimal subset $D'' \subseteq (D \cup D')-\{d\}$ such that (F,D'') explains PRE(d') (since by compactness of first-order logic there is a finite set for which this holds, there is also such a minimal subset). Now observe that by definition 3.1(a) all defaults of both A and A' are applicable, for which reason no element of D needs an element of $D'-D$ to be applicable and vice versa. Then, since $d' \in D$ or $d' \in D'$, and since d' is applicable, it holds that for at least one D'' mentioned above $D'' \subseteq D$ or $D'' \subseteq D'$. Furthermore, since $d \notin D''$ also $D'' \subseteq D-\{d\}$ or $D'' \subseteq D'-\{d\}$. Then since (F,D'') explains PRE(d') and normal default logic is semimonotonic (proposition A.6) an extension of $(F,D-\{d\})$ or one of $(F,D'-\{d\})$ contains PRE(d'); and since these default theories have a unique extension since they are subarguments of the argument $A \cup A'$, we can say that $(F,D-\{d\})$ or $(F,D'-\{d\})$ explains PRE(d'); but then by definition 4.12 of a top rule $d \in [\phi]A$ or $d \in [\phi']A'$. Hence $[\psi](A \cup A') \subseteq [\phi]A \cup [\phi']A'$. □

LEMMA B.16

Let $A = (F,D)$ minimally explain ϕ, $A' = (F,D')$ minimally explains ϕ', and $\{\phi,\phi'\} \models \psi$. If $[\psi](A \cup A')$ does not contain elements of $[\phi]A$, then A is a subargument of A'.

Proof

Assume that no element of $[\phi]A$ is a top default of $A \cup A'$, that is, for every $d \in [\phi]A$ there is a $d' \in D'$ such that $(F,D \cup D'-\{d\})$ does not explain PRE(d'). Then, since by definition of an argument d' is applicable, d is needed to make d' applicable in A', for which reason d is in D'. But then also all $d'' \in D$ which are needed to make d applicable in A are in D'; if not, then for some such d'' $(F,D \cup D' - \{d''\})$ would be a subargument of $A \cup A'$ explaining ψ. Finally, since this holds for all $d \in [\phi]A$ and since D contains by minimality of A no defaults irrelevant for the explanation of ϕ, every $d \in D$ is in D'. □

We are now again in the position to derive a contradiction refuting assumption B.8. In informal terms the idea is that if A4 H-defeats $A5 \cup A3$ and A5 H-defeats

A4 ∪ A3, some "relevant" defaults of A4 and A5 are higher than each other, which is a contradiction. Observe first that, because of the relations of defeat stated in fact B.14, according to the definition of being higher D3, D4 and D5 are totally ordered. Furthermore,

(1) Since A4 $>_\phi$ A5 ∪ A3 (*fact B.14*), it holds that $[\phi]A4 > [\neg\phi](A5 \cup A3)$ (*definition 4.14*); For the same reasons, since A5 $>_\psi$ A4 ∪ A3, it holds that $[\psi]A5 > [\neg\psi](A4 \cup A3)$.

(2) $[\neg\phi](A5 \cup A3) \subseteq [\psi]A5 \cup [\chi]A3$ and $[\neg\psi](A4 \cup A3) \subseteq [\phi]A4 \cup [\chi]A3$ (*fact B.14, lemma B.15*).

(3) $[\neg\phi](A5 \cup A3) \not\subseteq [\chi]A3$, since otherwise by *lemma B.16* A5 is a subargument of A3 and A3 is defeated by A4, contrary to assumption B.8. Then because of (2), $[\neg\phi](A5 \cup A3)$ contains elements of $[\psi]A5$. These remarks also hold for $[\phi]A4$ and $[\neg\psi](A4 \cup A3)$.

(4) There is a $d1 \in [\psi]A5$ such that $d1 \in [\neg\phi](A5 \cup A3)$ (*3*) and for all $d' \in [\phi]A4$ $d1 < d'$ (*1, definition 4.14*); likewise is there a $d2 \in [\phi]A4$ such that $d2 \in [\neg\psi](A4 \cup A3)$ and for all $d' \in [\psi]A5$ $d2 < d'$.

(5) Both $d1 < d2$ and $d2 < d1$ (*4*), which is a contradiction.

(6) The only assumption which can be retracted to restore consistency is assumption B.8, saying that there is a counterargument A3 of A1 ∪ A2 not defeated by another preferred argument. The proof of theorem 4.16 is then completed in the same way as for theorem 4.10. □ (end of proof of theorem 4.16).

Finally, the proof of theorem 4.17 for H-defeat is identical to the proof of theorem 4.11 for S-defeat.

Acknowledgements

This article, which is a revised and extended version of [16], was supported by a grant from the Legal Research Foundation, which is part of the Netherlands Foundation for Scientific Research. I wish to thank John-Jules Meyer, Marek Sergot, Arend Soeteman, Frans Voorbraak and the referees for their valuable comments on earlier versions of this article.

References

[1] C.E. Alchourrón and D. Makinson, Hierarchies of regulations and their logic, in: *New Studies in Deontic Logic*, ed. R. Hilpinen (Reidel, Dordrecht, 1981) pp. 125–148.

[2] K.D. Ashley and E.L. Rissland, A case-based system for trade secrets law, *Proc. 1st Int. Conf. on Artificial Intelligence and Law*, Boston (1987) pp. 60–66.

[3] G. Brewka, Preferred subtheories: an extended logical framework for default reasoning, *Proc. IJCAI-1989*, pp. 1043–1048.

[4] G. Brewka, *Nonmonotonic Reasoning: Logical Foundations of Commonsense*, (Cambridge Univ. Press, 1991).
[5] J. Delgrande, An approach to default reasoning based on a first-order conditional logic: revised report, Artificial Intelligence 36(1988)63–90.
[6] A. v.d. Lieth Gardner, *An Artificial Intelligence Approach to Legal Reasoning* (MIT Press, 1987).
[7] P. Gärdenfors, *Knowledge in Flux. Modeling the Dynamics of Epistemic States* (MIT Press, 1988).
[8] K. Konolige, Hierarchic autoepistemic theories for nonmonotonic reasoning, *Proc. AAAI-88*, (1988).
[9] F. Lin and Y. Shoham, Argument systems. A uniform basis for nonmonotonic reasoning, *Proc. 1st Int. Conf. on Principles of Knowledge Representation and Reasoning* (Morgan Kaufmann, 1989) pp. 245–255.
[10] R. Loui and K. Stiefvater, Corrigenda to Poole's rules and a lemma of Simari–Loui, Department of Computer Science, Washington University, St. Louis (1992).
[11] D. Nute, Defeasible reasoning: a philosophical analysis in Prolog, in: *Aspects of Artificial Intelligence*, ed. J.H. Fetzer (Kluwer, 1988) pp. 251–288.
[12] J.L. Pollock, Defeasible reasoning, Cognitive Sci. 11(1987)481–518.
[13] D.L. Poole, On the comparison of theories: Preferring the most specific explanation, *Proc. IJCAI-1985*, pp. 144–147.
[14] D.L. Poole, A logical framework for default reasoning, Artificial Intelligence 36(1988)27–47.
[15] H, Prakken, A tool in modelling disagreement in law: preferring the most specific argument, *Proc. 3rd Int. Conf. on Artificial Intelligence and Law*, Oxford, 1991 (ACM Press, 1991) pp. 165–174.
[16] H. Prakken, Reasoning with normative hierarchies (extended abstract). *Proc. 1st Int. Workshop on Deontic Logic and Computer Science*, Amsterdam (1991) pp. 315–334.
[17] H. Prakken, Logical tools for modelling legal argument, Doctoral Dissertation, Free University Amsterdam (1993).
[18] R. Reiter, A logic for default reasoning, Artificial Intelligence 13(1980)81–132.
[19] N. Rescher, *Hypothetical Reasoning* (North-Holland, Amsterdam, 1964).
[20] N. Roos, What is on the machine's mind? Models for reasoning with incomplete and uncertain knowledge, Doctoral Dissertation, Technical University Delft (1991).
[21] G. Sartor, Reasoning with hierarchies of premises: derivation versus assumption based approaches, Draft, University of Bologna (1992).
[22] G.R. Simari and R.P Loui, A mathematical treatment of defeasible reasoning and its implementation, Artificial Intelligence 2(1992)125–157.
[23] G.A.W. Vreeswijk, Abstract argumentation systems, to appear in Studia Logica.