DEFLOG - a logic of dialectical justification and defeat

Bart Verheij

Department of Metajuridica, Universiteit Maastricht
P.O. Box 616, 6200 MD Maastricht, The Netherlands
bart.verheij@metajur.unimaas.nl, http://www.metajur.unimaas.nl/~bart/

Draft finished: June 15, 2000
Last change: August 11, 2000

Contents

Abstract

The present paper is the result of a search for an analog for defeasible reasoning of valid consequence and proof in deductive reasoning. The abundance of research on nonmonotonic logics, and more specifically on defeasible reasoning, has shown the notoriety of the topic. One reason is that the received canon of views on logic and reasoning, as exemplified by standard logics, such as first-order predicate logic, is inappropriate in the context of defeasible reasoning. A goal of the present paper is to provide a revision of these views on logic and reasoning, by providing an abstract, formal theory of dialectical justification and defeat. Dialectical justification can be regarded as an analog of valid consequence.

The starting point is an analog of interpretation in the context of defeasible reasoning, viz. the notion of an extension of a theory, where a theory is regarded as a set of sentences. An extension of a theory can be thought of as an interpretation of the theory as a set of defeasible statements. In an extension, a theory's sentences can not only be justified, but also defeated. This is in contrast with the standard, non-defeasible interpretation of a theory in terms of models, where all sentences of the theory are assigned the same positive value, viz. true. In an extension of a theory, the justified part of the theory must provide an argument against the entire defeated part.

The search for an analog of valid consequence and proof started naïvely, in work on the graphical presentation of dialectical arguments in which statements can be supported by reasons and also attacked by counterarguments. The development of naïve dialectical arguments for the experimental argument assistance system ArguMed resulted in the discovery and investigation of the notion of *dialectical justification*: an argument is dialectically justifying if and only if the argument attacks all arguments that are incompatible with it.

Dialectical justification is analogous to valid consequence in the following two relevant ways. First, a dialectically justifying argument can be regarded as a set of premises justifying its conclusions, in the context of defeasible reasoning. The premises provide a basis justifying a conclusion, that is as solid as possible in the context of defeasible reasoning. Second, the investigation of the internal structure of a dialectically justifying argument leads to the notion of a justifying dialectical argument, that is a direct generalization of that of a proof, but incorporates counterarguments. A major difference between dialectical justification and valid consequence is of course that dialectical justification is nonmonotonic relative to a theory: when an argument is dialectically justifying with respect to a theory, it need not be dialectically justifying with respect to a larger theory. Another difference is the phenomenon of dialectical ambiguity: it can be the case that a statement is both dialectically justifiable and dialectically defeasible with respect to a theory. Dialectical ambiguity is analogous to inconsistency, but is not trivializing: the existence of a dialectically ambiguous statement with respect to a theory does not imply that any statement is dialectically justifiable.

The notion of dialectical justification plays a central role in an interesting necessary and sufficient condition for the existence of an extension of a theory. A characterization of the number of extensions (which is as usual zero, one or several) is given in terms of the notion of dialectical justification.

The notion of dialectical justification is closely related to the notion of admissibility that is currently regarded as state of the art: an argument is admissible if and only if it attacks all arguments that attack it. It is shown that the notion of dialectical justification is more satisfactory than the notion of admissibility, as a tool in the analysis of extensions. By a meta-analysis it is shown that three properties of dialectical justification are crucial: the union property, the localization property and the separation property. Admissibility lacks the latter, and as a result of that, does not allow a characterization of the existence of extensions analogous to that in terms of dialectical justification.

A useful instrument in the analysis of the dialectical interpretation of theories is the notion of a theory's *stages*. A stage of a theory is a partial dialectical interpretation of the theory, i.e., a dialectical interpretation of a subset of the theory. The stages of a theory correspond extensionally to the theory's satisfiable subsets (where satisfiability is used in the standard sense of having a model). There is an interesting intensional difference, which is relevant for the maximization of stages. Instead of maximizing the stage's justified part (which corresponds to maximizing a satisfiable subset), it is natural to maximize the stage's scope, i.e., the part of the theory that is interpreted in the stage, whether justified or defeated.

1 Introduction

Argumentation often has a dialectical character: it does not only involve arguments for a conclusion, but also arguments against. For instance, when the claim that Peter shot George is subject to debate, witness A's statement that Peter shot George could be adduced as a reason supporting the claim, while witness B's statement that Peter did not shoot George could be raised as a reason attacking the claim. The present paper attempts to show the nature of dialectical argumentation and how it should be modeled. It builds on and extends the work of many authors who have written about the subject. Among the most influential for my thinking about the subject are especially Reiter, Hage, Prakken, Pollock, Vreeswijk, Loui, Dung and Toulmin.[1]

In the present paper, a logic of dialectical justification and defeat, called DEFLOG, is presented. DEFLOG is about justification in the sense that it attempts to explain when conclusions are justified by a set of assumptions. DEFLOG deals with defeat in the sense that sentences cannot only be justified, but also defeated by a set of assumptions. The adjective 'dialectical' is used to suggest that in DEFLOG justification and defeat occurs in a context of juxtaposed opposing or contradictory claims (and not to suggest a dialogical setting, but see section 13.5). For instance, in DEFLOG, sets of assumptions can contain opposing claims and still be sensibly interpreted from the dialectical point of view. DEFLOG is a logic in the sense that it provides a formal specification of aspects of reasoning, viz. of dialectical argumentation. The logicality of DEFLOG is stressed by the fact that it contains analogues of several elements that often occur in logic, such as interpretations, valid consequence, proofs, satisfiability and inconsistency (cf. especially section 15).

DEFLOG uses a logical language with two connectives $\times$ and $\leadsto$. The first, the unary connective $\times$, is used to express the defeat of a statement. If φ is a sentence, then the sentence $\times\varphi$ expresses that the statement that φ is defeated. (I like to speak of the *dialectical negation* of a statement.) The second, the binary connective $\leadsto$, is used to express conditional justification. If φ and ψ are sentences, then the sentence $\varphi \leadsto \psi$ expresses that if the statement that φ is justified, then the statement that ψ is justified. A third connective $\rtimes$ is used to express attack. Attack is defined in terms of dialectical negation and conditional justification. That the statement that φ attacks the statement that ψ is considered to mean that if the statement that φ is justified, then the statement that $\times\psi$ is justified. As a result, that φ attacks ψ is expressed by the sentence $\varphi \leadsto \times\psi$, abbreviated as $\varphi \rtimes \psi$.[2] It is among the innovations of DEFLOG that its language is constructed using genuine sentential connectives, in the sense that nested expressions like $p \leadsto (q \rtimes (r \leadsto s))$ - that can be suggested by sensible examples ! - are allowed.

The central definition of DEFLOG is that of the dialectical interpretation of a theory in terms of extensions. There are two main differences between the idea of an interpretation of a theory in standard logic (often called a model of a theory) and that of a dialectical interpretation of a theory in DEFLOG. The first is that, in the interpretations of standard logic, all sentences in the theory are assigned the same positive status, in logic usually referred to as true. A model of a theory is then a logically possible world in which all sentences of the theory are true. In the dialectical interpretation of a theory in DEFLOG, however, not all sentences need to be given a positive evaluation: a sentence of the theory can be either positively evaluated, viz. as *justified*, negatively, viz. as *defeated*. The key idea is simple: in a dialectical interpretation of a theory, a sentence of the theory is defeated if and only if it is justified by the justified part of the theory that the statement is defeated.

The second main difference between standard interpretations and dialectical interpretations is that in the interpretations of standard logic, the whole language is interpreted, i.e., all sentences of the language are assigned a status (usually either true or false), while in dialectical interpretations, this need not be so: a dialectical interpretation has an extent, that consists of the sentences of the language that are assigned a status. The intuitive idea is that in a dialectical interpretation only those sentences are evaluated as justified by the theory. More precisely, a sentence φ (in the language) is evaluated as justified in a dialectical interpretation of a theory if and only if φ is supported by the justified part of the theory.

[1] The order in which the names appear only reflects the accidental chronology of my intellectual history. Some relevant sources are Reiter's (1980), Hage's (1996, 1997), Prakken's (1997), Pollock's (1995), Vreeswijk's (1997), Loui's (1998), Dung's (1995) and Toulmin's (1958).

[2] For convenience, here and in the following the phrase 'the statement that' - as in 'the statement that φ' - is often omitted. This is somewhat sloppy since it blurs the distinction between the sentence φ and the statement it expresses, but will hopefully not lead to confusion.

DefLog does not fall from the sky. The formal characterization of justification and defeat in a dialectical context has recently received much attention. A lot of research has been devoted to the formalization of these notions, which has resulted in a diversity of formalisms.[3]

Among the innovations of DefLog and the contributions of this paper are the following:

- The idea of considering extensions as interpretations of defeasible theories, contrasted with the standard notion of models as interpretations of strict theories.
- The discovery of the notion of dialectical justification and its role in the extension existence and extension multiplicity problems, and its subtle distinction from the notion of admissibility.
- The notion of naïve dialectical arguments as reason/attack-structures, and their evaluation, and an explication of the discovery of the extent to which naïve dialectical arguments can count as the counterpart in dialectical logic of proofs in standard logic.
- The notions of dialectical negation and conditional justification, and the discovery that attack and several other notions from dialectical logic (like rebutters and undercutters) can be expressed in terms of dialectical negation and conditional justification.
- The use of genuine sentential connectives $\times$, $\leadsto$ and $\rtimes$, allowing nested expressions, in the context of dialectical argumentation, thus normalizing and enhancing the expressiveness of logics for dialectical argumentation.
- The notion of stages as partial interpretations of defeasible theories, and the discovery of its extensional (but not intensional) equivalence to the maximal consistent subsets of the theory.
- The distinction of two fundamentally different ways of maximizing partial dialectical interpretations of theories, viz. the maximization of the theory's justified sentences, and the maximization of the theory's interpreted sentences.
- Discussion of the relations between several types of stages (or, better, of their non-relations).

Of course some of the above are not entirely new or original, but I claim that the ideas are here at least significantly extended or clarified, given suitable explicitness, or deservedly emphasized.

The paper is structured as follows. In the next section, the notion of a naïve dialectical argument is introduced. Naïve dialectical arguments are structured sets of statements, in which statements can be reasons for or counterarguments against other statements. By the graphical presentation of naïve dialectical arguments, several key ideas can be set out in an intuitive way.

In section 3, the 'standard' logical core of DefLog is explained, viz. its language, interpretations and models. Section 4 introduces the dialectical core, viz. the notion of extension as the dialectical interpretation of a theory. The sections 5 to 10 further elaborate on DefLog's dialectical core, in terms of among others stages, dialectical justification and dialectical arguments. Section 11 deals with some representational issues that arise in the context of dialectical argumentation, and section 12 with variations on DefLog. In section 12.4, a meta-analysis of some of the main properties of dialectical justification shows why it has been selected from among several alternatives. Related research is discussed in section 13. In section 14, two metaphors of dialectical argumentation, viz. the comparison and the attack metaphor, are discussed from DefLog's point of view. In section 15, DefLog is contrasted with standard propositional logic in an attempt to clarify the differences and the similarities between a dialectical and a deductive approach to logic.

2 Naïve dialectical arguments

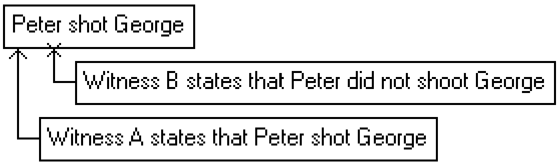
Part of the inspiration for the development of DefLog was my work on the graphical representation of arguments in defeasible argumentation (during the design of prototypical argument assistance systems; see, e.g., Verheij, 1998a, 1998b, 1999, *to appear*, and http://www.metajur.unimaas.nl/~bart/aaa/, where the systems can be downloaded). I introduced the term 'dialectical argument' for an argument possibly incorporating counterarguments against statements occurring in the argument. As a result of the possible occurrence of a counterargument, not all statements in a dialectical argument are to be considered justified. For instance, if a statement is attacked by a justified statement, the statement is defeated. It was my hunch that dialectical arguments would become the counterparts in defeasible reasoning of proofs in strict reasoning. Since it turned out that the dialectical arguments as they are studied here were too coarse

[3] Next to the sources mentioned in footnote 1, the reader might want to consult the work of Bondarenko *et al.* (1997), Prakken & Sartor (1996), Verheij (1996a, 1996b, 1999), and the overview by Prakken & Vreeswijk (*to appear*).

in structure for this purpose (see below, section 10), I now call them *naïve dialectical arguments*. The notion of naïve dialectical arguments is still intuitively attractive and provides a good illustration of some central ideas in DefLog. The discussion is rather informal and serves merely as an appetizer for the formalism to come.

2.1 The structure of naïve dialectical arguments

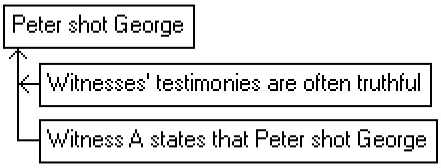
Naïve dialectical arguments consist of statements that can have two types of connections between them: a statement can *support* another, or a statement can *attack* another. The former is indicated by a pointed arrow between statements, the latter by an arrow ending in a cross. Here is an example:


Peter shot George
Witness B states that Peter did not shoot George
Witness A states that Peter shot George

The dialectical argument consists of three elementary statements, viz. that Peter shot George, that witness A states that Peter shot George, and that witness B states that Peter did not shoot George. As is indicated, the second is a reason supporting that Peter shot George, the second a reason attacking that Peter shot George.

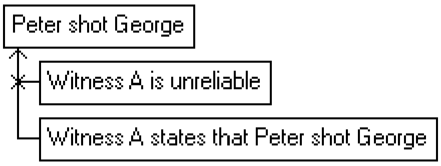
The expressiveness of naïve dialectical arguments is significantly enhanced by considering the connecting arrows (of both the supporting and the attacking type) as a kind of statements, that can as such be supported and attacked. The arrow of a supporting or attacking argument step is here called the warrant of the step (cf. also Toulmin's terminology, see also section 11.3 below).

For instance, one could ask why A's testimony supports that Peter shot George. In the following, the statement that witnesses' testimonies are often truthful is adduced as a reason:


Peter shot George
Witnesses' testimonies are often truthful
Witness A states that Peter shot George

The statement that witnesses' testimonies are often truthful serves as a backing of the supporting argument step (cf. also Toulmin's terminology, see also section 11.3 below). The same statement can back the attacking argument step of B's testimony attacking that Peter shot George.

That the connecting arrows can also be attacked can be seen in the following example:



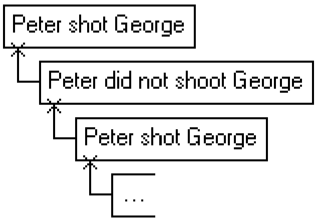
Here the unreliability of witness A is adduced as a counterargument against the supporting connection between the other two statements.

In general, naïve dialectical arguments are finite structures that result from a finite number of applications of three kinds of construction types:

1. Making a statement
2. Supporting a previously made statement by a reason for it
3. Attacking a previously made statement by a reason against it (also called a counterargument)

It should be borne in mind that the types two and three consist of making two statements: one an ordinary elementary statement, viz. the reason for or against a statement, the other the special statement that the reason and the supported or attacked statement are connected, as expressed by the warrant of the supporting or attacking argument step.

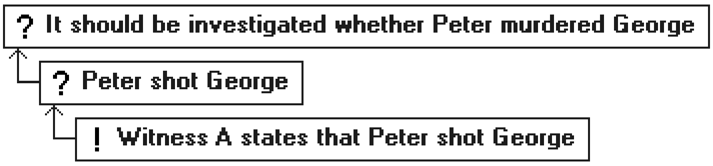
Though naïve dialectical arguments are here considered as the result of a finite construction, their corresponding tree structure can be virtually infinite. An example is suggested in the following picture:



The argument can be thought of as being the result of three construction steps. First the statement that Peter shot George is made, then that statement is attacked by the counterargument that Peter did not shoot George, and finally it is stated that the statement that Peter shot George is on its turn a counterargument to its attack. If the resulting (finite) looping structure is expanded as a tree (growing downward from the initial statement), the result is infinite.

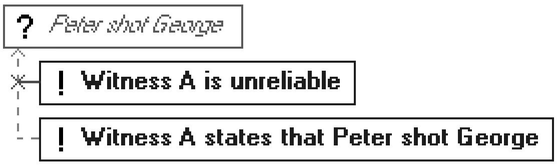
2.2 Evaluating naïve dialectical arguments

Naïve dialectical arguments can be evaluated with respect to a set of defeasible assumptions. An example of an evaluated naïve dialectical argument is the following:


? It should be investigated whether Peter murdered George
? Peter shot George
! Witness A states that Peter shot George

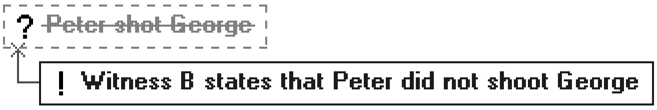
Defeasible assumptions are preceded by an exclamation mark, all others - called issues - by a question mark. Above the statement that witness A states that Peter shot George is a defeasible assumption. All three arguments that occur in the argument are evaluated as justified, as is indicated by the dark bold font. The statement about A's testimony is justified since it is an assumption that is not attacked, the statement that Peter shot George is justified since it is supported by a justifying reason (viz. A's testimony), and similarly for the statement about the investigation. (Here and in the following the warrants of argument steps are implicitly assumed to be defeasibly justified.)

The following example involves the attack of the support relation between two statements:


? Peter shot George
! Witness A is unreliable
! Witness A states that Peter shot George

The statements about A's testimony and unreliability are defeasible assumptions, while the statement that Peter shot George is an issue. The two assumptions are justified since they are not attacked. The statement that Peter shot George is unevaluated (as is indicated by the light italic font): it is not justified since it is an issue for which there is no justifying reason, nor is it defeated since there is no attacking counterargument.

An example of a naïve dialectical argument in which a statement is defeated is the following:



The issue that Peter shot George is defeated (as is indicated by the dark bold struck-through font) since it is attacked by the counterargument that witness B states that Peter did not shoot George.

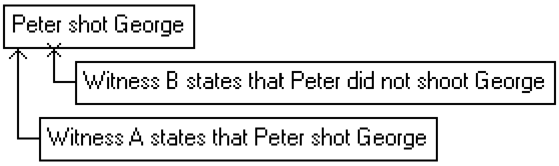
The evaluation of naïve dialectical arguments with respect to a set of defeasible assumptions is naturally constrained as follows:

1. A statement is *justified* if and only if
   a. it is an assumption, against which there is no defeating counterargument, or
   b. it is an issue, for which there is a justifying reason.

   A statement is *defeated* if and only if there is a defeating counterargument against it.
2. A reason is *justifying* if and only if the reason and the warrant of the corresponding supporting argument step are justified.
3. A counterargument is *defeating* if and only if the counterargument and the warrant of the corresponding attacking argument step are justified.

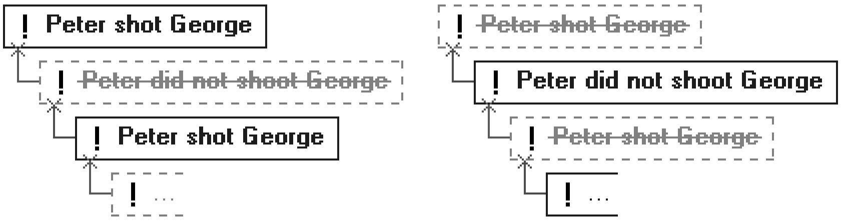
It is a fundamental complication of dialectical argumentation that a naïve dialectical argument can have any number of evaluations with respect to a set of defeasible assumptions: there can be no evaluation, or one, or several.

Assuming as we do that statements cannot be both justified and defeated, the following argument whether Peter shot George - already discussed above - has no evaluation with respect to the two testimonies as defeasible assumptions.


Peter shot George
Witness B states that Peter did not shoot George
Witness A states that Peter shot George

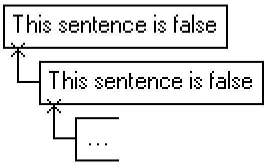
That the argument has no evaluation is seen as follows. Since both assumptions are not attacked they must be justified in any evaluation. But then A's testimony would require that it is justified that Peter shot George, while at the same time B's testimony would require that it is defeated that Peter shot George. This is impossible.

An example of a naïve dialectical argument with two evaluations is the looping argument discussed above:



The argument has two defeasible assumptions, viz. that Peter shot George and that Peter did not shoot George. The assumptions attack each other. In one evaluation, it is justified that Peter shot George, thus making it defeated that Peter did not shoot George, while in the other evaluation it is the other way around.

Note that the existence of the two evaluations is possible because the loop of attacks consists of an even number of statements. An odd length loop of attacks can cause that there is no evaluation. An example is the following:



If its only elementary statement (viz. that this sentence is false) is a defeasible assumption, the argument has no evaluation obeying the constraints.

3 DefLog's language, interpretations and models

The informally presented notion of naïve dialectical arguments leads the way to the formalization of dialectical justification and defeat in the formalism DefLog. As a first step, a suitable logical language and its interpretation is introduced, and some of its elementary - and relatively standard - properties are studied.

The first consideration towards DefLog's logical language is the recognition of the warrants of argument steps as logically compound sentences. Since warrants connect two statements, they can be expressed in a logical style using binary connectives. In DefLog, the warrant of a supporting step in which the statement that φ is a reason for the statement that ψ, is denoted as φ ↝ ψ using the binary connective ↝. The warrant of an attacking step in which the statement that φ is a counterargument to the statement that ψ is denoted as φ ⋊ ψ using the binary connective ⋊.

Though a logical language with the two binary connectives ↝ and ⋊ can successfully be used as the basis of a formalization of justification and defeat (see, e.g., Verheij, *to appear*), a second consideration leads to considerably simpler definitions and deeper understanding. This second consideration is that it is useful to express the defeat of a statement in the logical language. In DefLog, the defeat of a statement is expressed using the unary connective ×. A sentence ×φ expresses that the statement that φ is defeated.

As a result, it becomes possible to *define* attack in terms of conditional justification and defeat: the statement that φ attacks ψ can be defined as the statement that if φ is justified, then ψ is defeated. In other words, φ ⋊ ψ can be considered as shorthand for φ ↝ ×ψ.

DefLog's logical language is defined as follows.

Definition (3.1): the language

Given a set of elementary sentences, DefLog's *language* is the smallest set of sentences, such that if φ and ψ are sentences, then ×φ and (φ ↝ ψ) are sentences.

A sentence φ expresses that the statement that φ is justified.[4] For convenience, it is also said that a sentence φ expresses that φ is justified, or even simply that φ. A sentence ×φ expresses that the statement that φ is defeated, or that φ is defeated, for short. A sentence φ ↝ ψ expresses that if the statement that φ is justified, then the statement that ψ is justified, or that if φ, then ψ, for short.

If p, q and r are elementary sentences, then some examples of sentences are the following:

p, (p ↝ q), (p ↝ (q ↝ r)), ((p ↝ q) ↝ r), ×p, ××p, ×(p ↝ q), (p ↝ ×q), (×p ↝ q), (p ↝ p)

In the following, outer brackets are normally omitted, as for instance in (p ↝ q) ↝ r.

Convention (3.2)

If φ and ψ are sentences, then (φ ⋊ ψ) and (ψ ⋉ φ) are abbreviations of (φ ↝ ×ψ).

A sentence φ ⋊ ψ expresses that the statement that φ attacks the statement that ψ, or in other words, that the statement that φ is an argument against the statement that ψ. In DefLog, the statement expressed by φ ⋊ ψ is equivalent to the statement that if φ is justified, then ψ is defeated (or fully, that if the statement that φ is justified, then the statement that the statement that ψ is defeated, is defeated).[5]

Sentences of the form ×φ are *defeat sentences*, sentences of the form φ ↝ ψ *conditional sentences*, and sentences of the form φ ⋊ ψ *attack sentences*. (Note that attack sentences are also conditional sentences.) If φ ↝ ψ is a conditional sentence, φ is the conditional's antecedent and ψ its consequent. If φ is a sentence, then ×φ is its defeat sentence.

In DefLog, sets of sentences are interpreted by assigning each sentence in the set one of two justification statuses, either justified or defeated. The justification statuses are abbreviated as j and d,

[4] This reminds of Tarski's well-known scheme in standard logic, according to which a sentence φ expresses that the sentence φ is true.

[5] Note that in DefLog it is equivalent to say that φ is defeated, or that the statement that φ is defeated, is justified.

respectively. The assignment of the status j to a sentence φ, corresponds to the statement φ being justified, and the assignment of the status d to the statement being defeated.

In an interpretation, the assignment of justification statuses must obey two constraints, suggested by the intended meaning of the sentences of the language. First, the statement φ being defeated coincides with the statement $\times\varphi$ being justified. As a result, a sentence φ is assigned the status d if and only if the sentence $\times\varphi$ is assigned the value j.

Second, if it is justified that if φ then ψ, and it is justified that φ, then it should follow that it is justified that ψ. As a result, sentences $\varphi \rightsquigarrow \psi$ and φ being assigned the value j implies the sentence ψ being assigned the value j.

Definition (3.3): interpretations

An *interpretation* (or *world*) is a mapping W from a set of sentences S to the set {j, d}, such that the following two constraints obtain:

1. $W(\times\varphi)$ is equal to j if and only if $W(\varphi)$ is equal to d.
2. If $W(\varphi \rightsquigarrow \psi)$ and $W(\varphi)$ are both equal to j, then $W(\psi)$ is equal to j.

The set S is called the *extent* of the interpretation. If W is an interpretation, J(W) denotes the set of sentences that are assigned the value j under W, and D(W) the set of sentences assigned the value d. The elements of J(W) are said to be *justified* in W, those in D(W) *defeated*.

It should be noted that sentences of the language can be neither justified nor defeated in an interpretation, since an interpretation's extent is not necessarily equal to the whole language.[6] When an interpretation has the whole language as its extent, i.e., any sentence of the language is either justified or defeated in the interpretation, the interpretation is said to be *total*.

Note also that, while by the first constraint the justification status of a sentence $\times\varphi$ in an interpretation is a function of the status of φ, the second constraint shows that the justification status of a sentence $\varphi \rightsquigarrow \psi$ is *not* a function of the statuses of φ and ψ. Of the eight possible assignments of justification statuses to three sentences φ, ψ and $\varphi \rightsquigarrow \psi$, constraint 2 only excludes one, viz. that in which φ and $\varphi \rightsquigarrow \psi$ are justified, while ψ is defeated. The meaning of a conditional $\varphi \rightsquigarrow \psi$ is simply that its consequent follows if its antecedent applies. (This is in stark contrast with the truth functionality of the material implication of standard logic, of which the truth coincides with the antecedent's falsity or the consequent's truth.)

Example (3.4)

In each row of the table on the left, the justified and defeated sentences of an interpretation are listed. In the table on the right, examples of status assignments that are not interpretations are listed.

Examples of interpretations	
Justified sentences	*Defeated sentences*
$\times p$	p
$p, \times\times p$	$\times p$
$p, q, p \rightsquigarrow q$	-
$p, \times q, p \rtimes q$	q
$q, p \rightsquigarrow q$	$\times p$
$\times q, p \rightsquigarrow q$	q
$p, q, \times(p \rightsquigarrow q)$	$p \rightsquigarrow q$
$p, \times q, r, p \rtimes q, \times q \rightsquigarrow r$	q

Examples of non-interpretations	
Justified sentences	*Defeated sentences*
$p, \times p$	-
p	$\times p$
$p, \times p$	$\times p, \times\times p$
$p, p \rightsquigarrow q$	-
$p, p \rtimes q$	q
$\times p, q, \times p \rightsquigarrow q$	-

Notation (3.5)

For an interpretation W, $W \vDash \varphi$ denotes that the sentence φ is justified in the interpretation W.

There is no dedicated notation for a sentence being defeated in an interpretation, but note that by constraint 1 in definition (3.3), $W \vDash \times\varphi$ denotes that φ is defeated in W.

[6] This suggests that DEFLOG'S interpretations can be regarded as three-valued interpretations of the whole language, viz. 'justified', 'defeated' and 'uninterpreted'. Whether this is a fruitful way of looking at DEFLOG'S interpretations might be an interesting topic for further research.

Models can be regarded as interpretations of sets of sentences as *strict* theories: all sentences in the set of sentences are assumed to be justified. An interpretation is a model of a theory if all sentences in the theory are justified in the interpretation.

Definition (3.6): models of theories, satisfiability

If T is a set of sentences and W is an interpretation, such that $W \vDash \varphi$ for any sentence φ in T, then the interpretation W is a *model of the theory* T, which is denoted as $W \vDash T$. A set of sentences is *satisfiable* if it has a model.

Example (3.7)

The sets $T_1 = \{p, p \rtimes p\}$ and $T_2 = \{p, q, p \rightsquigarrow r, q \rtimes r\}$ are not satisfiable, while the sets $T_3 = \{p, p \rightsquigarrow q, q \rtimes r\}$ and $T_4 = \{p \rightsquigarrow q, (p \rightsquigarrow q) \rtimes r, \times r \rtimes \times(p \rightsquigarrow q)\}$ are satisfiable.

The consequences of a theory are as usual defined as those sentences that are justified in all models of the theory. Note that in determining the consequences of a theory it is considered as strict, i.e., non-defeasible. The interpretation of theories as defeasible is the topic of the next section.

Definition (3.8): consequences of theories

If T is a set of sentences and φ a sentence, then φ is a *consequence of the theory* T if, for any interpretation W, if $W \vDash T$, then $W \vDash \varphi$. That φ is a consequence of a theory T is denoted as $T \vDash \varphi$. The set of consequences of T is denoted Cn(T).

Example (3.9)

The sets T_1 and T_2 of example (3.7) are not satisfiable, and therefore have all sentences of the language as their consequences. The sets T_3 and T_4 have $T_3 \cup \{q, \times r\}$ and $T_4 \cup \{\times r, \times\times(p \rightsquigarrow q)\}$ as their sets of consequences.

The set Cn(T) of consequences of a theory T can be characterized by rules of inference. Cn(T) is the closure of T under the rules of inference $\varphi, \varphi \rightsquigarrow \psi \,/\, \psi$ ($\rightsquigarrow$-*Modus ponens*, or *Modus ponens,* for short) and $\varphi, \times\varphi \,/\, \psi$ (a variant of *Ex falso quodlibet*). The closure of T under ($\rightsquigarrow$-)*Modus ponens* alone is denoted as Mp(T). Mp(T) is the smallest set that contains T and that is closed under rule application. For satisfiable T, Mp(T) and Cn(T) coincide. Note that Cn(T) and Mp(T) are also closed under the rule of inference $\varphi, \varphi \rtimes \psi \,/\, \times\psi$ that might be called $\rtimes$-*Modus ponens*.

Definition (3.10)

A set of sentences S is *conflict-free* if there is no sentence φ in S, such that $\times\varphi$ is in S. A set of sentences S is *closed under Modus ponens* if whenever $\varphi \rightsquigarrow \psi$ and φ are in S, then ψ is in S.

The defeated sentences of an interpretation are 'encoded' in the justified sentences: a sentence φ in an interpretation is defeated when and only when the sentence $\times\varphi$ is justified. As a result, the defeated sentences of an interpretation are in a precise sense superfluous in the characterization of an interpretation: only the justified sentences suffice in order to characterize an interpretation, as in the following property.

Property (3.11)

A set of sentences C is conflict free and closed under *Modus ponens* if and only if there is an interpretation W such that C is equal to J(W), the set of justified sentences of W.

Proof: The 'if'-part follows by checking the definitions. The 'only if'-part is based on the following construction. If C is conflict free and closed under Modus ponens, then the mapping that assigns the value j to all sentences in C, and the value d to all sentences φ, for which $\times\varphi$ is in C, is an interpretation.

If C is conflict free and closed under *Modus ponens*, the interpretation that is constructed in the proof of the property above is denoted as W_C. Clearly, it follows that, for any such C, $J(W_C)$ is equal to C, and that, for any interpretation W, $W_{J(W)}$ is equal to W. This gives a convenient characterization of worlds in terms of sets of sentences that is often used throughout the paper.

Since the notion of a satisfiable set of sentences is important in DEFLOG, it is natural to look for a syntactic characterization. It is provided by the notion of argument.

Definition (3.12)
A set of sentences is an *argument* if its closure under Modus ponens is conflict free.

Property (3.13)
A set of sentences C is an argument if and only if C is satisfiable.

Proof: Use property (3.11).

Definition (3.14): arguments for and against, attack, incompatibility

(i) An argument C *supports* or is an *argument for* a sentence φ if $C \vDash \varphi$. An argument C *attacks* or is an *argument against* φ if $C \vDash \times\varphi$. The sentences in an argument C are also called its *premises*, the sentences φ such that $C \vDash \varphi$, its *conclusions*.

(ii) An argument C *attacks* an argument C' if C attacks a sentence in C'.

(iii) Arguments C and C' are *compatible* if $C \cup C'$ is an argument, and otherwise *incompatible*. The arguments in a collection $\{C_i\}_{i \in I}$ are *compatible* if their union $\cup_{i \in I} C_i$ is an argument, otherwise *incompatible*.

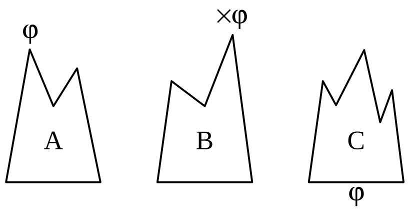
The set of sentences $\{p, p \leadsto q\}$ is an argument, the set $\{p, p \leadsto q, \times q\}$ is not. The argument $\{p, q, p \leadsto (q \rtimes r)\}$ has p, q and $p \leadsto (q \rtimes r)$ as premises, and p, q, $p \leadsto (q \rtimes r)$, $q \rtimes r$ and $\times r$ as conclusions.

If C attacks C', then C and C' are incompatible. If the arguments in a collection $\{C_i\}_{i \in I}$ are pairwise compatible, the collection is not necessarily compatible. For instance, the three arguments {p}, {q} and $\{p \rtimes q\}$ are pairwise compatible, but the collection containing all three arguments is not compatible. The incompatibility of two arguments does not imply that one of them attacks the other. E.g., the arguments {p, q} and $\{p \leadsto r, q \leadsto \times r\}$ are incompatible, but neither attacks the other.

*Property (*3.15*)*
If C is an argument for φ, then there is a *Modus ponens* derivation with premises in C and conclusion φ. If C is an argument against φ, then there is a *Modus ponens* derivation with premises in C and conclusion $\times\varphi$.

Proof: Use property (3.11).

In the following figure, three arguments are graphically suggested.



The bottoms of the alpine shapes consist of the premises of the argument; the tops are the conclusions. Argument A has conclusion φ, argument B conclusion $\times\varphi$ and argument C has premise φ. B attacks C, but not necessarily A (since φ might not be a premise of A). A and B are incompatible, and B and C too.

If C is an argument, then its closure under *Modus ponens* characterizes an interpretation, cf. (3.11). It is denoted as W_C. For arguments C, it does not in general hold that $J(W_C)$ is equal to C. It does hold that $J(W_C)$ is equal to Cn(C).

Definition (3.16)
Let C be an argument. Then W_C is the *interpretation specified by the argument* C.

The following monotonicity property obtains. Cf. the properties (4.6) and (6.6) below.

Property (3.17)

Let T and T' be theories, such that T is a subset of T'. If T has no model, T' does not have one. If T and T' both have a model, say W and W', respectively, and φ is justified in W, then φ is justified in W'. If φ is defeated in W, it is defeated in W'. If φ is in the extent of W, it is in the extent of W'.

Proof: Use properties (3.11) and (3.15).

4 Extensions as interpretations of defeasible theories

The models of a theory, defined in the previous section, can be regarded as interpretations of *strict* theories: a model of a theory is an interpretation in which all sentences of the theory are considered to express justified statements. In this section, the notion of *extensions* of a theory is introduced. Extensions are interpretations of theories as defeasible statements. The main idea is that an extension of a theory is an interpretation specified by a part of the theory that is an argument against the remainder of the theory. In other words, in an extension of a theory, the theory is split in a justified and a defeated part. The justified part is an argument against the defeated part and specifies the extension. In this way, many sets of sentences that are not satisfiable are given sensible interpretations as defeasible theories.

Before the formal definition is given, some examples are discussed. The following definition comes in handy.

Definition (4.1)

Let Δ be a set of sentences and C an argument. Then C is a Δ-*argument* if C is a subset of Δ.

Some simple but important examples are the following.

Example (4.2)

(i) Consider the set Δ = {p, q, q ⋊ p}. The theory Δ says that p, that q, and that q attacks p. Δ is clearly not satisfiable. It contains an argument however that attacks all sentences outside the argument: {q, q ⋊ p} is indeed an argument, and attacks p. In the interpretation specified by {q, q ⋊ p}, ×p, q and q ⋊ p are justified, and p is defeated. This interpretation is the theory's extension.

(ii) Consider the set Δ = {p, q, r, q ⋊ p, r ⋊ q}. The theory Δ says that p, that q, that r, and that q attacks p, while q is on its turn attacked by r. Again the theory is not satisfiable. Still there is a Δ-argument, viz. {p, r, q ⋊ p, r ⋊ q}, that attacks all sentences of the theory not in it, in this case only q. In the interpretation it specifies all sentences in Δ are justified, except q, which is defeated. This interpretation is the theory's extension.

(iii) Consider the set Δ = {p, ×p}. The theory Δ says that p, and that it is defeated that p. It is not satisfiable. However the interpretation specified by the argument {×p} in which ×p is justified and p is defeated is the theory's extension.

Here is the formal definition of extensions.

Definition (4.3): extensions

If Δ is a set of sentences and E an interpretation, then E is an *extension of the theory* Δ if and only if E is an interpretation that is specified by a Δ-argument J that attacks any sentence φ in Δ \ J. The set J(E) ∩ Δ is the *justified part* of the theory in the extension, the set D(E) ∩ Δ the *defeated part*.

If E is an extension of Δ and J is as in the definition, E = W_J. Since J is satisfiable, its set of consequences is equal to Mp(J). Any sentence φ in J is justified in E, i.e., E(φ) = j, and any sentence ψ in Δ \ J is defeated in E, i.e., E(ψ) = d.

In the table, the splitting of some theories (among them those of example (4.2)) into sets of justified and defeated sentences, as in the definition of extensions, is shown. Each splitting in the table corresponds to an extension, by taking the interpretation specified by the justified sentences of the theory.

Defeasible theory	*Justified part*	*Defeated part*
p, ×p	×p	p
p, p ⋉ q, q	p ⋉ q, q	p
p, p ⋉ q, q, q ⋉ r, r	p, p ⋉ q, q ⋉ r, r	q
p, p ⇝ q, ×p	p ⇝ q, ×p	p
p, p ⇝ q, ×(p ⇝ q)	p, ×(p ⇝ q)	p ⇝ q
p, p ⋉ q, q, q ⇝ r	p ⋉ q, q, q ⇝ r	p

Another fundamental characteristic for the interpretation of sets of sentences as defeasible theories is the following.

Property (4.4)

A theory can have zero, one or several extensions.

Proof: Cf. the following examples.

Example (4.5)

(i) The three theories {p, p ⋊ p}, {p, p ⇝ q, ×q}, {p_i | i is a natural number} ∪ {p_i ⋉ p_j | i and j are natural numbers, such that i < j} lack extensions. For the latter theory, this can be seen as follows. Assume that there is an extension E in which for some natural number n p_n is justified. Then all p_m with m > n must be defeated in E, for if such a p_m were justified, p_n could not be justified. But that is impossible, for the defeat of a p_m with m > n can only be implied by a justified $p_{m'}$ with m' > m. As a result, no p_i can be justified in E. But then all p_i must be defeated in E, which is impossible since the defeat of a p_i can only be implied by a justified p_j with j > i. (Note that any *finite* subset of the latter theory has an extension, while the whole theory does not. This shows a 'non-compactness' property[7] of extensions.) See also example (6.10) below.

(ii) The three theories {p, q, p ⋊ q, p ⋉ q}, {p_i, p_i ⋉ p_{i+1} | i is a natural number} and {$\times^i$p | i is a natural number} have two extensions. (Here $\times^i$p denotes, for any natural number i, the sentence composed of a length i sequence of the connective ×, followed by the constant p.)

(iii) The theory {p, ×p} has a unique extension, just as the other example theories in the table above.

It follows that, although a theory that is not satisfiable, can have an extension, not all theories have an extension: such theories are neither 'strictly satisfiable' nor 'defeasibly satisfiable'. Such sets of sentences can neither be interpreted as a strict theory nor as a defeasible theory.

The following nonmonotonicity property obtains. Cf. the properties (3.17) and (6.6).

Property (4.6)

Let Δ and Δ' be theories, such that Δ is a subset of Δ'. If Δ has an extension, Δ' need not have one. If Δ and Δ' both have an extension, say E and E', respectively, and φ is justified in E, then φ need not be justified in E'. If φ is defeated in E, it need not be defeated in E'. If φ is in the extent of E, it need not be in the extent of E'.

Proof: While {p} has an extension, {p, p ⋊ p} does not. While p is justified in the extension of {p}, it is not in that of {p, q, q ⋊ p}. While ×p is justified in the extension of {×p}, it is not in that of {×p, q, q ⋊ ×p}. While r is in the extent of the extension of {p, p ⇝ r}, it is not in that of {p, p ⇝ r, q, q ⋊ p}.

The following notational convention is sometimes useful.

Convention (4.7)

If S is a set of sentences, then ×S denotes the set {×φ | φ is an element of S} and $\times^{-1}$S the set {φ | ×φ is an element of S}.

For determining the extensions of a theory, the following simple property can be helpful.

[7] A property P of sets is called compact if a set S has property P whenever all its finite subsets have the property. Cf. the compactness of satisfiability in first-order predicate logic.

Property (4.8)

If E is an extension of a theory Δ, then $J(E) \subseteq Mp(\Delta)$ and $D(E) \subseteq \times^{-1}Mp(\Delta)$.

Proof: Let the set J be as in definition (4.3). Then $E = W_J$. Therefore $J(E) = Mp(J)$ and since $J \subseteq \Delta$ and Mp is monotonous, $J(E) \subseteq Mp(\Delta)$. Since $E = W_J$, it follows that $D(E) \subseteq \times^{-1}J(E)$. Then by $J(E) \subseteq Mp(\Delta)$, also $D(E) \subseteq \times^{-1}Mp(\Delta)$.

Obviously, it is not in general the case that $Mp(\Delta)$ is a subset of the extent of an extension of the theory Δ. An example is the theory $\Delta = \{p, p \rightsquigarrow q, \times p\}$ that has a unique extension specified by $\times p$ and $p \rightsquigarrow q$. In the extension, p is defeated and q not taken into account..
The following proposition gathers some alternative definitions of extensions.

Proposition (4.9)

Let E be an interpretation and Δ a set of sentences. Then the following are equivalent:

(i) E is an extension of the theory Δ.

(ii) There are sets of sentences J and D with $\Delta = J \cup D$, $J \cap D = \varnothing$, such that $J(E) = Mp(J)$ and $D(E) \supseteq D$.

(iii) There are sets of sentences J and D with $\Delta = J \cup D$, $J \cap D = \varnothing$, such that $J(E) = Mp(J)$ and $J(E) \supseteq \times D$.

(iv) $E = W_{\Delta \cap J(E)}$ and $E \vDash \times(\Delta \setminus J(E))$.

(v) E is an interpretation specified by a maximal satisfiable subset of Δ and with Δ in its extent.

Proof: (i) $\Rightarrow$ (ii): Let J be as in the definition of extensions and let D be $\Delta \setminus J$. Then J and D are as in (ii). (ii) $\Rightarrow$ (iii): For any interpretation E and any set of sentences D, it follows from $D(E) \supseteq D$ that $J(E) \supseteq \times D$. (iii) $\Rightarrow$ (iv): Note that $J = \Delta \cap J(E)$ and $D = \Delta \setminus J(E)$. (iv) $\Rightarrow$ (v): $J = \Delta \cap J(E)$ is a maximal satisfiable subset of Δ since for any φ in Δ not in J, it obtains that $J \vDash \times\varphi$. (v) $\Rightarrow$ (i): Let J be a maximal satisfiable subset of Δ specifying E. Any φ in $\Delta \setminus J$ is in E's extent. It cannot be in J(E) for then $J \cup (\varphi)$ would be satisfiable. Therefore it must be in D(E). But then $\times\varphi$ must be in J(E). Since J specifies E, $\times\varphi$ is a consequence of J.

By part (iv) of the proposition, the set J as it occurs in the definition of extensions can be extracted from a given extension of the theory: the set J is equal to $\Delta \cap J(E)$. Note however that an extension E of a theory Δ can be specified by other subsets of Δ than $\Delta \cap J(E)$, viz. subsets J' for which it holds that $\Delta \cap J(E) = Mp(J')$. For instance, the unique extension E of the theory $\{p, p \rightsquigarrow q, p \rtimes r, q\}$ is specified by $\{p, p \rightsquigarrow q\}$, which is a proper subset of $\Delta \cap J(E)$.

The following property characterizes the extensions of satisfiable theories: they are just the interpretations specified by the theory. In addition, it is stated that an extension of a theory is also an extension of certain other sets of sentences.

Property (4.10)

(i) If T is satisfiable, then T has a unique extension, viz. the interpretation W_T that is specified by T.

(ii) If E is an extension of Δ, then E is an extension of any set of sentences Δ', such that $J(E) \cap \Delta \subseteq \Delta' \subseteq J(E) \cup D(E)$.

Proof: Property (i) follows from the fact that if E is an extension of T and φ were a sentence in T that is defeated in E, then $T \cap J(E) \vDash \times\varphi$ and T would not be satisfiable. For property (ii), first note that $\Delta' \cap J(E) = \Delta \cap J(E)$ and $\varnothing \subseteq \Delta' \setminus J(E) \subseteq D(E)$, and then use part (iv) of proposition (4.9).

Note that, though according to property (ii) above an extension E of a theory Δ is also an extension of any set Δ', such that $J(E) \cap \Delta \subseteq \Delta' \subseteq J(E) \cup D(E)$, such a set Δ' can have an extension that is not an extension of Δ. The following is an example.

Example (4.11)

The theory $\Delta = \{p, p \rightsquigarrow q, q \rightsquigarrow r, r \rightsquigarrow s, q \rtimes (r \rightsquigarrow s), s \rtimes (p \rightsquigarrow q)\}$ has one extension, viz. the interpretation specified by $\Delta \setminus \{r \rightsquigarrow s\}$, in which $r \rightsquigarrow s$ is defeated. $J(E) \cup D(E)$ has a second extension, viz. the interpretation specified by $\{s\} \cup \Delta \setminus \{p \rightsquigarrow q\}$.

In the following sections, the notion of extensions as interpretations of defeasible theories is further investigated.

In the following, the notion of extensions is further investigated. One of the aims is to better understand the possibility that a theory has zero, one or several extensions. Let's call the possibility that a theory does not always have an extension the *extension existence problem*, and the possibility that a theory has more than one extension the *extension multiplicity problem*. One tool will be the notion of stages.

5 Stages

Even if a theory has no extension, its subsets can have extensions. The extensions of subsets of a theory are called the theory's *stages*.[8] The extensions of the subsets of a theory can be regarded as preliminary stages on the path towards an extension of the whole theory. One could say that at these preliminary stages less information as it is expressed in the theory, is taken into account than at an extension. Even if the theory as a whole lacks an extension, its stages can provide interesting information about the theory.

Definition (5.1): stages

An interpretation S is a *stage of the theory* Δ if and only if it is an extension of a subset of Δ. If S is a stage, the set Δ ∩ (J(S) ∪ D(S)) is the *scope* of the stage. If S is a stage, the sets J and D, where J := J(S) ∩ Δ and D := D(S) ∩ Δ, are the *j-scope* and the *d-scope* of the stage, respectively. A sentence φ in Δ that is in the scope of a stage S is *taken into account at the stage* S.

The scope of a stage a theory can be regarded as the subset of the theory that has been taken into account at the stage.[9] A stage's scope should be contrasted with the stage's extent, which is the whole set of sentences (not in general a subset of Δ) that are assigned a defeat status in it (cf. definition (3.3)). For instance, in the stage of the theory {p, ×p, p ⇝ q} specified by the set {p, p ⇝ q}, q is in the stage's extent, but not in its scope since it is not in the theory.

Example (5.2)

The sets ∅, {p}, {×p}, {p ⇝ q}, {p, p ⇝ q} and {×p, p ⇝ q} specify the stages of the theory {p, ×p, p ⇝ q}. Its unique extension is specified by {×p, p ⇝ q}. Note that the scopes of the stages specified by {×p} and {×p, p ⇝ q} include the sentence p.

Not all subsets of a theory occur as the scope of one of the theory's stages. There are two fundamentally different reasons for this. The first is that the subset does itself not have an extension. For instance, the subset {p, p ⋊ p} of the theory {p, p ⋊ p, q, q ⋊ p} does not occur as the scope of a stage. The second reason is that if S is an extension of a subset Δ' of a theory Δ, then the scope of the stage S of Δ is not necessarily equal to Δ'. The scope is then necessarily a *larger* subset of Δ. For instance, the stage (actually: the extension) of the (satisfiable) theory {p, q, p ⇝ q} specified by the set {p, p ⇝ q} has the whole theory as its scope.

The stages of a theory correspond exactly to the interpretations that are specified by the satisfiable subsets of Δ:

Property (5.3)

An interpretation S is a stage of the theory Δ if and only if S is specified by a satisfiable subset T of Δ, i.e., S = W_T.

Proof: If T is a satisfiable subset of Δ, then T has a unique extension, viz. W_T, by part (i) of property (4.10). As a result, it is a stage of Δ. If S is a stage of the theory Δ, it is by definition an extension of a subset Δ' of Δ. By part (iv) of proposition (4.9) it is specified by Δ' ∩ J(S), which is a satisfiable subset of Δ.

[8] For the development of my ideas on stages, see also Verheij, 1996a and 1996b.

[9] As a result, the stages of a theory can be regarded as three-valued interpretations of the theory, viz. the values 'justified', 'defeated' and 'not taken into account'. Together with the additional value 'uninterpreted' suggested in note 6, one can look at stages as four-valued interpretations of the whole language. Again, whether this is a fruitful view, is left for further research.

As a result, the stages of a theory are exactly the interpretations specified by the theory's arguments (cf. definition (3.14)).

Note that *extensionally* stages coincide with satisfiable subsets, but not *intensionally*. One notion associated with the stages of a theory is their scope, which is not as readily suggested by the theory's satisfiable subsets.

A characteristic phenomenon of argument defeat that is made explicit in the stages of a theory, is the possibility that the defeat status of a sentence changes on the basis of additional information.

Example (5.4)

(i) The sets {p, p ⋉ q} and {×p, q, p ⋉ q} specify stages of the theory {p, q, p ⋉ q}. The latter specifies the theory's extension. In the former, p is justified since the attack q is not yet taken into account. In the latter, p has become defeated since it is attacked by q.

(ii) The stages specified by the sets {p, p ⋉ q, q ⋉ r}, {×p, q, p ⋉ q, q ⋉ r} and {p, ×q, r, p ⋉ q, q ⋉ r} of the theory {p, q, r, p ⋉ q, q ⋉ r} (the latter of which is the theory's unique extension) show the *reinstatement* of a sentence: p is consecutively justified, defeated, and then again justified.

It is a natural step to accentuate the stages in which a maximal subset of the theory, is taken into account. Such stages that have maximal scope, are called maximal stages.

Definition (5.5): maximal stages

An interpretation E is a *maximal stage of the theory* Δ if and only if it has maximal scope among the stages of Δ.

Note that a maximal stage of a theory, i.e., a stage with maximal scope, also has maximal extent among the theory's stages, but that not all stages with maximal extent are maximal stages. The stage of the theory {p, p → q, ×p} that is specified by the set {p, p → q} has maximal extent (it is the stage that has maximal extent among the stages with q in their extent), but does not have maximal scope: its scope does not contain ×p, while the theory has a full-scope stage, namely its extension specified by {p → q, ×p}.

Property (5.6)

Extensions are maximal stages, but not in general vice versa.

Proof: The scope of an extension of a theory is maximal since it is equal to the whole theory. Example (5.7) below shows that maximal stages are not in general extensions.

Example (5.7)

The theory {p, p ⋊ p} has the interpretations specified by {p} and {p ⋊ p} as maximal stages, but no extension. The theory {p, p → q, ×q} has the interpretations specified by {p, p → q}, {p, ×q} and {p → q, ×q} as maximal stages, but has no extension. Cf. part (i) of example (4.5).

Property (5.6) implies that the number of maximal stages is equal to or larger than the number of extensions. Example (5.7) shows that the number of maximal stages can indeed be larger than the number of extensions. This can however only be the case if a theory lacks an extension, as the following property shows. It says that for theories with an extension the notion of extension coincides with the notion of maximal stage.

Property (5.8)

If a theory has an extension, then any maximal stage of the theory is an extension.

Proof: If E is an extension of a theory Δ, then its scope is equal to Δ. As a result, any maximal stage must have Δ as its scope, and is therefore also an extension.

As a result of this property, if a theory has an extension, the number of maximal stages of the theory equals the number of extensions.

While the number of maximal stages is equal to or larger than that of extensions, the question arises whether there is an analog for maximal stages of the extension existence problem: are there theories lacking a maximal stage? Two of the three sample theories lacking an extension (discussed in example

(4.5) above) were shown to have maximal stages (example (5.7) above). The third sample theory without an extension also lacks a maximal stage.

Example (5.9): a theory without maximal stage

The theory $\Delta = \{p_i \mid i$ is a natural number$\} \cup \{p_i \ltimes p_j \mid i$ and j are natural numbers, such that $i < j\}$ has no maximal stage. This can be seen as follows. Among its stages are the interpretations S_n specified by the sets $\{p_n\} \cup \{p_i \ltimes p_j \mid i$ and j are natural numbers, such that $i < j\}$, where n is a natural number. In a stage S_n, p_n is justified and any p_i with $i < n$ is defeated. The extents of the stages S_n exhaust the whole theory, so a maximal stage must have the whole theory as its extent, i.e., must be an extension. However, the theory does not have an extension, cf. example (4.5), part (i).

Note that in the example, the non-existence of a maximal stage proves a 'non-compactness' property: the sample theory Δ has the property that for any *finite* subset Δ' of the theory there is a stage the scope of which contains Δ', while for the whole theory there is not. (It is even the case that any finite subset of Δ has an extension.)

The analog of property (4.4) for maximal stages is the following.

Property (5.10)

A theory can have zero, one or several maximal stages.

Proof: The property follows from property (5.8), example (4.5) and example (5.9).

Maximal stages are the result of maximizing the scope of the stages of a theory. Another way to maximize stages is by maximizing only the justified sentences in the scope of the stages. Stages that are maximal in this second way are called the satisfiability classes of a theory, since they turn out to correspond exactly to the maximal satisfiable subsets of the theory.

Definition (5.11): satisfiability classes

A stage S is a *satisfiability class of the theory* Δ if S is a stage of Δ such that the j-scope of S is maximal among the stages of Δ.

Property (5.12)

Let S be an interpretation and Δ a set of sentences. Then the following are equivalent:

(i) A stage S is a satisfiability class of the theory Δ.

(ii) S is specified by a maximal satisfiable subset of Δ.

(iii) S is a stage such that J(S) is maximal among the stages of Δ.

Proof: Assume that S is a satisfiability class with j-scope J. J is satisfiable. If T with $\Delta \supseteq T \supseteq J$ is satisfiable, it specifies a stage with j-scope J. By the maximality of J, it follows that T = J. Assume that S is specified by a maximal satisfiable subset T of Δ. Then $S = W_T$ and J(S) = Mp(T). If S' is a stage of Δ with $J(S') \supseteq J(S)$, then S' is specified by the satisfiable set $J(S') \cap \Delta \supseteq T$. Then the maximality of T implies that $J(S') \cap \Delta = T$, and therefore S' = S. Assume that S is a stage with j-scope J such that J(S) is maximal among the stages of Δ. If S' is a stage with j-scope J' with $\Delta \supseteq J' \supseteq J$, then $J(S') = Mp(J') \supseteq Mp(J) = J(S)$ by the monotonicity of Mp. By the maximality of J(S), it follows that J(S) = J(S'). But then $J = J(S) \cap \Delta = J(S') \cap \Delta = J'$.

Example (5.13)

Consider the theory $\{p, q, r, p \ltimes q, q \ltimes r\}$ that was already discussed in part (ii) of example (5.4) above. Its satisfiability classes that have all the attack sentences of the theory in their extent, are specified by the sets $\{q, p \ltimes q, q \ltimes r\}$ (in which p is defeated, q justified and r not taken into account) and $\{p, r, p \ltimes q, q \ltimes r\}$ (in which p and r are justified and q defeated). The latter is the theory's extension.

Note that the example shows that satisfiability classes are 'insensitive' to the possibility of counterattack and reinstatement: the stage in which p comes out as defeated and q as justified is from the point of view of satisfiability classes as good as the stage in which the outcome is the other way around.

Satisfiability classes and maximal stages are the maxima of two different partial orderings on the set of stages of a theory. The partial orderings are defined as follows.

Definition (5.14): a stage extending another stage

S *extends* S', denoted as $S \preccurlyeq S'$, if the scope of S is a subset of that of S'.

S *compatibly extends* S', denoted as $S \sqsubseteq S'$, if the j-scope of S is a subset of the j-scope of S' and the d-scope of S is a subset of the d-scope of S'.

Note that $\preccurlyeq$ and $\sqsubseteq$ are not defined in terms of set inclusion of the *extents* of stages (which might be confusing given the terminology), but of their *scopes*.

Property (5.15)

Let S and S' be stages of a theory. Then, if $S \sqsubseteq S'$, $S \preccurlyeq S'$.

Proof: If $J(S) \cap \Delta \subseteq J(S) \cap \Delta$, then $J(S) \subseteq J(S)$. If $J(S) \subseteq J(S)$, then $D(S) \subseteq D(S)$.

The satisfiability classes of a theory are the $\sqsubseteq$-maxima among the theory's stages and its maximal stages the $\preccurlyeq$-maxima.

Property (5.16)

Any maximal stage of a theory is a satisfiability class, but not in general vice versa.

Proof: Any $\preccurlyeq$-maximum is also $\sqsubseteq$-maximal. The satisfiability class of the theory $\{p, q, p \rtimes q\}$ specified by $\{q\}$ is not a maximal stage since the stage (actually: extension) specified by $\{p, p \rtimes q\}$, in which q is defeated, has larger scope.

Corollary (5.17)

Any extension of a theory is a satisfiability class, but not in general vice versa.

Proof: Combine the properties (5.6) and (5.16).

The number of satisfiability classes is larger than or equal to the numbers of maximal stages and extensions. Indeed, in contrast with the situation for maximal stages and extensions, any theory has one or more satisfiability classes.

Theorem (5.18)

(i) Any theory Δ has one or more satisfiability classes.

(ii) If S is a stage of Δ, then Δ has a satisfiability class that is compatible with S.

Proof: For part (i), consider the partial ordering $\sqsubseteq$ on the set of stages. Apply Zorn's lemma (or, if you prefer, one of its weakenings) after observing that the stage specified by the empty set of sentences is a stage and that totally ordered chains of stages $(S_i)_i$ have a supremum, viz. W_J, where J is the union of all sets $J(S_i)$. For part (ii), consider the partial ordering $\sqsubseteq$ on the set of stages compatible with S.

Property (5.19)

A theory has a maximal stage if and only if the partial ordering $\preccurlyeq$ on the theory's satisfiability classes has a maximum.

Proof: The property follows directly from the definitions.

Corollary (5.20)

Any finite theory has a maximal stage.

Proof: The number of satisfiability classes of a finite theory is finite (e.g., since the number of partial justification status assignments is) and finite partial orderings have a maximum.

Recall that finite theories do not always have an extension. Cf. part (i) of example (4.5).

Definition (5.21): compatibility of stages

Stages S and S' of a theory Δ are *compatible* if there is a stage S'', such that $S \sqsubseteq S''$ and $S' \sqsubseteq S''$. Stages are *incompatible* if they are not compatible. A collection of stages $\{S_i\}_{i \in I}$ is compatible if there is a stage S, such that $S_i \sqsubseteq S$ for all i in I.

The definition matches the notion of compatibility of arguments, cf. the correspondence between stages and arguments following from (5.3).

Example (5.22)

The stages specified by the sets {p, q} and $\{p \leadsto r, q \leadsto \times r\}$ are not compatible. Note that the example shows that if two stages are incompatible, there need not be a sentence that is justified in one and defeated in the other (in contrast with a property of dialectically justifying stages, defined and discussed below, property (6.11)).

Compatibility in pairs of the stages in a collection of stages $\{S_i\}_{i \in I}$ does not imply compatibility. See the corresponding example for arguments below definition (3.14).

Property (5.23)

(i) Stages S and S' of a theory Δ are compatible if and only if $J(S) \cup J(S')$ is satisfiable.

(ii) If S and S' are compatible stages, then $S \sqsubseteq W_{J(S) \cup J(S')}$ and $S' \sqsubseteq W_{J(S) \cup J(S')}$.

(ii) If S and S' are compatible stages, and S'' is a stage such that $S \sqsubseteq S''$ and $S' \sqsubseteq S''$, then $W_{J(S) \cup J(S')} \sqsubseteq S''$.

Proof: Part (i) can be seen as follows. Let S'' be a stage such that $S \sqsubseteq S''$ and $S' \sqsubseteq S''$. Then $J(S) \cup J(S') \subseteq J(S'')$, which is satisfiable. If $J(S) \cup J(S')$ is satisfiable, then it specifies a stage $S'' = W_{J(S) \cup J(S')}$, such that $S \sqsubseteq S''$ and $S' \sqsubseteq S''$. Part (ii) follows immediately. For part (iii), note that from $S \sqsubseteq S''$ and $S' \sqsubseteq S''$, it follows that $J(S) \cup J(S') \subseteq J(S'')$ and therefore $W_{J(S) \cup J(S')} \sqsubseteq W_{J(S'')} = S''$.

Notation (5.24)

If S and S' are compatible stages, then $S \sqcup S'$ denotes the stage $W_{J(S) \cup J(S')}$. $S \sqcup S'$ is the *union* of the stages S and S'.

Property (5.25)

If S_1 and S_2 are different satisfiability classes of the theory Δ, then S_1 and S_2 are incompatible.

Proof: If S_1 and S_2 were compatible satisfiability classes, their union $S_1 \sqcup S_2$ would exist and would be a satisfiability class. If S_1 and S_2 are different, this would contradict their maximality with respect to the partial ordering $\sqsubseteq$.

Corollary (5.26)

If S_1 and S_2 are different maximal stages or extensions of a theory Δ, then S_1 and S_2 are incompatible.

Proof: Maximal stages and extensions are satisfiability classes.

6 Dialectical justification

An important question to ask is whether it is possible to find a criterion that determines whether a *particular* sentence is dialectically interpretable with respect to a theory, either as justified or as defeated. That is the topic of this section. The result is the notion of dialectical justification that can be regarded as an analog in defeasible reasoning of valid consequence in deductive reasoning.

A relevant property of extensions is expressed in the following proposition.

Proposition (6.1)

Let E be an extension of a theory Δ. Then J(E) is a Δ-argument that attacks any Δ-argument C that is incompatible with J(E).

Proof: Since E is an extension, J(E) is satisfiable. Hence a Δ-argument C that is incompatible with J(E) cannot be a subset of J(E) since J(E) is not incompatible with any of its subsets. Therefore there is a sentence φ in C that is not in J(E). Since E is an extension, it is in D(E). But for any sentence φ in D(E) it holds by the definition of extensions that J(E) ⊨ ×φ, i.e., J(E) attacks C.

The following corollary is a 'localized' version of proposition (6.1), that will be the starting point of section 10, where the internal structure of dialectical justification is investigated.

Corollary (6.2)

Let E be an extension of a theory Δ, φ a sentence that is justified in E, and C a J(E)-argument for φ. Then for any Δ-argument C' that is incompatible with C, there is a Δ-argument C" that is compatible with C (in fact a J(E)-argument), such that C" attacks C'.

Proof: Take J(E) in the role of C". Since C is a subset of J(E), C' is incompatible with J(E). Hence J(E) attacks C' according to the proposition.

The property of the argument J(E) in proposition (6.1) above is sufficiently important to deserve a name of its own.

Definition (6.3): dialectically justifying arguments

A Δ-argument C is *dialectically justifying* with respect to Δ if and only if C attacks any Δ-argument C' that is incompatible with C.

Definition (6.4): dialectically justifiable and defeasible sentences

A sentence φ is *dialectically justifiable* with respect to Δ if and only if there is a Δ-argument C for φ that is dialectically justifying with respect to Δ. Such an argument C is then called a *dialectical justification of* φ, and C *dialectically justifies* φ with respect to Δ. A sentence φ is *dialectically defeasible* with respect to Δ if and only if ×φ is dialectically justifiable with respect to Δ. If C is a dialectical justification of φ, then the argument C *dialectically defeats* φ with respect to Δ.

Example (6.5)

(i) The argument {p, r, q ⋉ r} dialectically justifies p with respect to the theory {p, q, r, p ⋉ q, q ⋉ r} (see part (ii) of example (5.4) and example (5.13)). The argument {p} does not dialectically justify p since the incompatible argument {q, p ⋉ q} is not attacked. The argument {r, q ⋉ r} dialectically defeats q with respect to the theory.

(ii) Sentences can be both dialectically justifiable and defeasible with respect to a theory. Consider the theory {p, q, p ⋊ q, p ⋉ q} (see part (ii) of example (4.5)). Then p and q are both dialectically justifiable and defeasible with respect to the theory. The argument {p, p ⋊ q} dialectically justifies p and dialectically defeats q, while the argument {q, p ⋉ q} dialectically defeats p and dialectically justifies q.

(iii) A sentence need not be dialectically justifiable or defeasible with respect to a theory. For instance, the sentence p is not dialectically justifiable and not dialectically defeasible with respect to the theory {p, p ⋊ p} (see part (i) of example (4.5)).

The following nonmonotonicity property obtains. Cf. the properties (3.17) and (4.6).

Property (6.6)

Let Δ and Δ' be theories, such that Δ is a subset of Δ'. If a sentence φ is dialectically justifiable with respect to Δ, it need not be dialectically justifiable with respect to Δ'. If a sentence φ is dialectically defeasible with respect to Δ, it need not be dialectically defeasible with respect to Δ'. If a sentence φ is dialectically justifiable or defeasible with respect to Δ, it can be neither dialectically justifiable nor dialectically defeasible with respect to Δ'.

Proof: While p is dialectically justifiable with respect to {p}, it is not with respect to {p, q, q ⋊ p}. While ×p is dialectically defeasible with respect to {×p}, it is not with respect to {×p, q, q ⋊ ×p}. While r is dialectically justifiable with respect to {p, p ⇝ r}, it is neither dialectically justifiable nor dialectically defeasible with respect to {p, p ⇝ r, q, q ⋊ p}.

Property (6.7)

(i) If C is a dialectical justification of φ with respect to Δ, then C is a dialectical justification for all its conclusions.

(ii) If C is a dialectically justifying argument with respect to Δ, then Cn(C) (which is equal to Mp(C)) is also dialectically justifying.

Proof: Part (i) of the property is a direct consequence of the definitions. Part (ii) follows by noting that any argument C' that is incompatible with Mp(C) is also incompatible with C.

The new terminology leads to the following rephrasing of proposition (6.1).

Corollary (6.8)

(i) The set of justified sentences of an extension of a theory is a dialectically justifying argument with respect to the theory.

(ii) If φ is a justified sentence in an extension of a theory Δ, then φ is dialectically justifiable with respect to Δ and J(E) is a dialectical justification of φ.

(iii) If φ is a defeated sentence in an extension of a theory Δ, then φ is dialectically defeasible with respect to Δ and J(E) is a dialectical justification of ×φ with respect to Δ.

Proof: The corollary is a reformulation of proposition (6.1) using the new terminology.

The following non-trivial sufficient condition for the non-existence of extensions is implied by the corollary.

Corollary (6.9)

A theory Δ has no extension if there is a sentence in Δ that is neither dialectically justifiable nor dialectically defeasible with respect to Δ.

Proof: Any sentence in Δ is either justified or defeated in an extension of Δ and therefore dialectically justifiable or defeasible with respect to Δ.

The corollary can explain all examples of theories without extensions that have been encountered above (part (i) of example (4.5)): in all there is a sentence that is neither dialectically justifiable nor dialectically defeasible. Nevertheless the condition in corollary (6.9) is *not* necessary for the non-existence of an extension. It does *not* obtain that a theory has an extension if any sentence in the theory is dialectically justifiable or defeasible, as the following example shows.

Example (6.10)

The theory Δ = {p, q, p ⋊ q, q ⋊ p, r, r ⋊ r, s, s ⋊ s, p ⋊ r, q ⋊ s} has no extension. Nevertheless all sentences in the theory are dialectically justifiable or defeasible with respect to Δ. The Δ-arguments {p, p ⋊ q}, {q, q ⋊ p}, {p, ×r, p ⋊ r} and {q, ×s, q ⋊ s} are dialectical justifications with respect to Δ of p, q, ×r and ×s, respectively.

The following proposition shows when the union of two dialectically justifying arguments is not dialectically justifying.

Proposition (6.11)

Let C and C' be dialectically justifying arguments with respect to a theory Δ. Then the following are equivalent:

(i) C ∪ C' is not a dialectically justifying argument with respect to Δ.

(ii) C ∪ C' is not an argument.

(iii) C and C' are not compatible.

(iv) There is a φ, such that C dialectically justifies φ while C' dialectically defeats φ.

(v) There is a φ in Δ, such that C dialectically justifies φ while C' dialectically defeats φ.

(vi) There is a φ in C ∪ C', such that C dialectically justifies φ while C' dialectically defeats φ.

(vii) C attacks C'.

(viii) C attacks C' and C' attacks C.

Proof: That (i) follows from (ii), which follows from (iii), which follows from (iv), which follows from (v), which follows from (vi), which follows from (vii) (use property (6.7)), which follows from (viii), is immediate. It is left to show that (viii) follows from (i). Assume first that $C \cup C'$ is not an argument. Then C and C' are incompatible, and since each is dialectically justifying it follows that C attacks C' and that C' attacks C. Assume second that $C \cup C'$ is an argument. The proof of the proposition is finished, when it is shown that $C \cup C'$ is then dialectically justifying. Let C" be an argument that is incompatible with $C \cup C'$. In the case that C" is incompatible with C', C' attacks C" since C' is dialectically justifying. In the case that C" is compatible with C', C is incompatible with the argument $C' \cup C''$, and therefore C attacks $C' \cup C''$ since C is dialectically justifying. Since C and C' are compatible it follows that C attacks C". In both cases $C \cup C'$ attacks C". It follows that $C \cup C'$ is dialectically justifying.

The result can be generalized to arbitrary collections of dialectically justifying arguments:

Proposition (6.12)

Let $\{C_i\}_{i \in I}$ be a collection of dialectically justifying arguments with respect to a theory Δ, and let C be its union $\cup_{i \in I} C_i$. Then the following are equivalent:

(i) C is not a dialectically justifying argument with respect to Δ.
(ii) C is not an argument.
(iii) $\{C_i\}_{i \in I}$ is not compatible.
(iv) There are i and j in I, such that C_i and C_j are incompatible.
(v) There is a φ and there are i and j in I, such that C_i dialectically justifies φ while C_j dialectically defeats φ.
(vi) There is a φ in Δ and there are i and j in I, such that C_i dialectically justifies φ while C_j dialectically defeats φ.
(vii) There is a φ in C and there are i and j in I, such that C_i dialectically justifies φ while C_j dialectically defeats φ.
(viii) There are i and j in I, such that C_i attacks C_j.
(ix) There is an attack loop among the C_i, i.e., there are no i(0), ..., i(n), such that $C_{i(k)}$ attacks $C_{i(k+1)}$ for k from 0 to n-1, and $C_{i(n)}$ attacks $C_{i(0)}$.
(x) There are i and j in I such that C_i attacks C_j and C_j attacks C_i.

Proof: Again the implications from bottom to top are immediate. It remains to show that (i) implies (x). Assume first that C is not an argument. Then there are i(0), ..., i(n) in I (with n > 0), such that $C_{i(0)} \cup C_{i(1)} \cup ... \cup C_{i(n)}$ is not an argument, while $C_{i(1)} \cup ... \cup C_{i(n)}$ is. Then $C_{i(0)}$ is incompatible with the argument $C_{i(1)} \cup ... \cup C_{i(n)}$. Since $C_{i(0)}$ is dialectically justifying, it therefore attacks $C_{i(1)} \cup ... \cup C_{i(n)}$. $C_{i(0)}$ then attacks one of $C_{i(1)}$, ..., $C_{i(n)}$, say $C_{i(1)}$. By applying the previous proposition it follows that $C_{i(1)}$ also attacks $C_{i(0)}$. Assume second that C is an argument. That C is then dialectically justifying can be seen as follows. Let C' be an argument that is incompatible with C. Then there are i(0), ..., i(n) in I (with n > 0), such that $C' \cup C_{i(0)} \cup C_{i(1)} \cup ... \cup C_{i(n)}$ is not an argument, while $C' \cup C_{i(1)} \cup ... \cup C_{i(n)}$ is. Therefore $C_{i(0)}$ is incompatible with the argument $C' \cup C_{i(1)} \cup ... \cup C_{i(n)}$. It follows that $C_{i(0)}$ attacks one of C', $C_{i(1)}$, ..., $C_{i(n)}$. Since C is an argument, $C_{i(0)}$ attacks C'. A fortiori, C attacks C'.

The propositions (6.11) and (6.12) have some important corollaries.

Corollary (6.13): reduction

Let Δ be a theory. Then the following are equivalent:

(i) There is an incompatible collection of dialectically justifying arguments.
(ii) There is an incompatible pair of dialectically justifying arguments.
(iii) There is a pair of dialectically justifying arguments that attack each other.

Corollary (6.14): union

If C and C' are compatible dialectically justifying arguments, then also $C \cup C'$ is dialectically justifying. (Similarly, for any compatible collection of dialectically justifying arguments: the union of a compatible collection of dialectically justifying arguments is again dialectically justifying.)

Corollary (6.15): separation

If C and C' are incompatible dialectically justifying arguments, then there are opposites φ and $\times\varphi$, such that $C \vDash \varphi$ and $C' \vDash \times\varphi$, or such that $C \vDash \times\varphi$ and $C' \vDash \varphi$. (Similarly, for any incompatible collection of dialectically justifying arguments: given an incompatible collection of dialectically justifying arguments, there are opposites that are the consequence of the unions of compatible subcollections.)

The union and separation properties are central in the treatment of the extension existence and multiplicity problems in section 9 (cf. section 12.4).

A stronger version of separation follows immediately from the definition of dialectical justification:

Corollary (6.16): separation at the base

If C and C' are incompatible dialectically justifying arguments, then there is a sentence φ in $C \cup C'$, such that $C \vDash \times\varphi$ or $C' \vDash \times\varphi$. (Similarly, for any incompatible collection of dialectically justifying arguments: given an incompatible collection of dialectically justifying arguments, there is a sentence in the union of the collection that is attacked by the union of a compatible subcollection.)

In section 12.4, some variants of dialectical justification are treated. Each lacks at least one of these properties.

7 Dialectically justified stages

Each justified sentence of an extension is dialectically justified by the justified sentences of the extension, as was established in proposition (6.1). The analog for the justified sentences of stages, maximal stages or satisfiability classes does not obtain. For instance, the justified sentences in a stage of a theory are not necessarily dialectically justified by the justified sentences of the stage. This leads to the notion of dialectically justified stages.

Definition (7.1): dialectically justified stages

A stage S is a *dialectically justified stage of the theory* Δ if and only if S is a stage of Δ, for which it obtains that J(S) dialectically justifies any sentence in J(S) and dialectically defeats any sentence in D(S).

Property (7.2)

Let S be a stage of the theory Δ. The following are equivalent:

(i) S is a dialectically justified stage.
(ii) J(S) dialectically justifies any sentence in J(S).
(iii) For any φ in J(S), there is a subset C of J(S) that dialectically justifies φ with respect to Δ.
(iv) S is specified by the union of a compatible collection of dialectically justifying arguments.
(v) S is specified by a dialectically justifying argument.

Proof: (ii) follows from (i) by the definition of dialectically justified stages. (iii) follows trivially from (ii). Assume (iii). For any φ in J(S), pick a $C_\varphi \subseteq J(S)$ that dialectically justifies φ. Then by the union property (6.14) the union C of the C_φ is a dialectically justifying argument. But obviously C is equal to J(S) and therefore specifies S. Assume (iv). Then by the union property (6.14) the union of the collection of justifying arguments is a dialectically justifying argument. It also specifies S. Assume (v) and let C be a dialectically justifying argument specifying S. Then Mp(C) = J(S). Now by part (ii) of property (6.7) J(S) is dialectically justifying and (i) follows.

In analogy with the case of stages in general, there are two different ways to 'maximize' dialectically justified stages. The first possibility is to maximize the j-scope, i.e., the set of sentences in the theory that are justified in the dialectically justified stage. The second possibility is to consider dialectically justified stages in which the scope is maximal, i.e., in which the interpreted part of the theory is as large as possible. Maximal dialectically justified stages of the first type are *dialectically preferred stages*, those of the second type *maximal dialectically preferred stages*. Dialectically preferred stages are the analog among dialectically justified stages of the satisfiability classes among stages in general, maximal dialectically preferred stages that of maximal stages.

Definition (7.3): dialectically preferred stages and maximal dialectically preferred stages

(i) A stage S is a *dialectically preferred stage of the theory* Δ if and only if S is a dialectically justified stage of Δ such that $J(S) \cap \Delta$ is maximal among the dialectically justified stages of Δ.

(ii) A stage S is a *maximal dialectically preferred stage of the theory* Δ if and only if S is a dialectically justified stage of Δ such that the scope of S is maximal among the dialectically justified stages of Δ.

The dialectically preferred stages of a theory are the $\sqsubseteq$-maxima among the theory's dialectically justified stages. A theory's maximal dialectically preferred stages the $\preccurlyeq$-maxima.

It will not be surprising that maximal dialectically preferred stages are dialectically preferred stages and that extensions are maximal dialectically preferred stages. However, there exist theories that have dialectically preferred stages that are not maximal dialectically preferred stages, and theories that have maximal dialectically preferred stages that are not extensions.

Property (7.4)

(i) Any maximal dialectically preferred stage of a theory is a dialectically preferred stage, but not in general vice versa.

(ii) Any extension of a theory is a maximal dialectically preferred stage, but not in general vice versa.

Proof: (i) Any $\preccurlyeq$-maximum is also $\sqsubseteq$-maximal. Example (7.5) below shows that there is a theory with a dialectically preferred stage that is not maximal dialectically preferred. (ii) Proposition (6.1) and its paraphrase (6.8) show that extensions are dialectically preferred. Extensions are maximal dialectically preferred stages since their scope is the whole theory. The theory $\{p, p \rtimes p\}$ has no extension, but the empty stage specified by the empty set as maximal dialectically preferred stage.

Example (7.5)

The theory $\{p, q, r, p \rtimes q, q \rtimes p, q \rtimes r, r \rtimes r\}$ has the stage specified by $\{q, q \rtimes p, q \rtimes r\}$, in which p and r are defeated and q is justified, as extension, and therefore as maximal dialectically preferred stage. The stage specified by $\{p, p \rtimes q\}$, in which p is justified, q is defeated and r is not taken into account, is a dialectically preferred stage, that is not maximal dialectically preferred.

Dialectically preferred stages are not necessarily satisfiability classes and satisfiability classes are not necessarily dialectically preferred stages, as the following example shows. The further investigation of the relation between satisfiability classes and dialectically preferred stages is postponed to the next section.

Example (7.6)

The theory $\{p, p \rtimes p\}$ has satisfiability classes $\{p\}$ and $\{p \rtimes p\}$, neither of which is a dialectically preferred stage. Its only dialectically preferred stage (which is therefore also maximal dialectically preferred) is the empty stage specified by the empty set of sentences.

Nevertheless an analog of property (5.25) and corollary (5.26) is easily proven.

Theorem (7.7)

If P_1 and P_2 are different dialectically preferred stages or maximal dialectically preferred stages of the theory Δ, then P_1 and P_2 are incompatible.

Proof: If P_1 and P_2 were compatible, their union $P_1 \sqcup P_2$ would be a dialectically justified stage by the properties (6.14) and (7.2). Since P_1 and P_2 are different, $P_1 \sqcup P_2$ would compatibly extend P_1 and P_2, while $P_1 \sqcup P_2$ would not be equal to one of P_1 and P_2, contradicting the $\sqsubseteq$-maximality of P_1 and P_2.

Since by property (7.4) there can be more dialectically preferred and maximal dialectically preferred stages than there are extensions, the question again arises whether all theories have a dialectically preferred stage and a maximal dialectically preferred stage. It turns out that indeed all theories have a dialectically preferred stage, but that not all theories have a maximal dialectically preferred stage.

Theorem (7.8)

(i) Any theory Δ has at least one dialectically preferred stage.

(ii) If S is a dialectically justified stage of Δ, then Δ has a dialectically preferred stage that is compatible with S.

(iii) Not all theories have a maximal dialectically preferred stage.

Proof: (i) Consider the partial ordering $\sqsubseteq$ of the set of dialectically justified stages. Observe that the empty stage (i.e., the stage with empty extent) is a dialectically justified stage for any theory. Unions of totally ordered chains $(S_i)_i$ of dialectically justified stages have a dialectically justified stage as supremum S, viz. the stage specified by the union of all sets $J(S_i)$. (Here the properties (6.14) and (7.2) are used.) By Zorn's lemma (or one of its weakenings) $\sqsubseteq$ has a maximum. (ii) Apply Zorn's lemma (or one of its weakenings) to the partial ordering $\sqsubseteq$ of the set of dialectically justified stages that are compatible with S. (iii) Example (7.12) below shows that there is a theory without maximal dialectically preferred stage.

The construction of an example of a theory without a maximal dialectically preferred stage, discussed below as example (7.12), is rather involved. For instance, example (5.9) of a theory without a maximal stage has a maximal dialectically preferred stage, viz. the empty stage. One reason that the construction of a counterexample is not simple, is that analogs of property (5.19) and corollary (5.20) about maximal stages obtain for maximal dialectically preferred stages.

Property (7.9)

A theory has a maximal dialectically preferred stage if and only if the partial ordering $\preccurlyeq$ on the theory's dialectically preferred stages has a maximum.

Proof: The property follows directly from the definitions.

Corollary (7.10)

Any finite theory has a maximal dialectically preferred stage.

Proof: The number of dialectically preferred stages of a finite theory is finite and finite partial orderings have a maximum.

In the discussion of example (7.12) of a theory without a maximal dialectically preferred stage, the following lemma is useful.

Lemma (7.11)

In a dialectically preferred stage S of a theory Δ, any sentence φ in Δ, for which any Δ-argument C that is incompatible with {φ} contains a sentence ψ that is defeated in S, is justified in S.

Proof: Let φ be a sentence, for which the condition of the lemma obtains. If φ is not justified in S, it is defeated or not taken into account. Assume first that φ is defeated in S. Then J(S) supports ×φ, and therefore J(S) is an argument that is incompatible with {φ}. The condition of the lemma then says that J(S) contains a sentence that is defeated in S, which is impossible. Assume second that φ is not taken into account in S. Then $J(S) \cup \{\varphi\}$ is dialectically justifying, which can be seen as follows. Let C be a Δ-argument incompatible with $J(S) \cup \{\varphi\}$. If C is incompatible with J(S), then J(S) attacks C since J(S) is dialectically justifying. If C is compatible with J(S), then $C \cup J(S)$ is incompatible with {φ}, so by the condition of the lemma $C \cup J(S)$ contains a sentence ψ that is defeated in S. The sentence ψ must be in C, and since it is defeated J(S) attacks ψ. It follows that $J(S) \cup \{\varphi\}$ is dialectically justifying. As a result, the stage specified by $J(S) \cup \{\varphi\}$ is dialectically justified (cf. property (7.2)), while it compatibly extends S, which implies that it is equal to S (since S is dialectically preferred). This contradicts that φ is not taken into account in S.

Example (7.12): a theory without maximal dialectically preferred stage

Consider the theory Δ consisting of the following sentences:

p_i, q_i, r_i, for any natural number i
$p_i \rtimes p_j$ for all i and j with $i < j$
$p_i \rtimes q_i$ and $p_i \ltimes q_i$ for all i
$p_i \rtimes r_k$ for all i and k with $k \leq i$
$r_k \rtimes r_k$ for all k

Then the following are the 'initials' of some of Δ's stages:

S_0:	p_0 (q_0) (r_0)	(p_1) q_1 -	(p_2) q_2 -	(p_3) q_3 -	(p_4) q_4 -	...
S_1:	(p_0) q_0 (r_0)	p_1 (q_1) (r_1)	(p_2) q_2 -	(p_3) q_3 -	(p_4) q_4 -	...
S_2:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	p_2 (q_2) (r_2)	(p_3) q_3 -	(p_4) q_4 -	...
S_3:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	(p_2) q_2 (r_2)	p_3 (q_3) (r_3)	(p_4) q_4 -	...
S_4:	(p_0) q_0 (r_0)	(p_1) q_1 (r_1)	(p_2) q_2 (r_2)	(p_3) q_3 (r_3)	p_4 (q_4) (r_4)	...

... ...

The sentences in brackets ( ) are defeated at the stage. The other listed sentences are justified. The hyphens - indicate sentences that are not taken into account. For instance, at S_0, p_0 is justified, q_0 is defeated and r_1 is not taken into account. For any natural number i, S_i is defined as the stage at which

(i) p_i is justified and, for any j such that $i \neq j$, p_j is defeated, and
(ii) q_i is defeated and, for any j such that $i \neq j$, q_j is justified, and
(iii) for any j such that $i \geq j$, r_i is defeated and, for any j such that $i < j$, r_j is not taken into account, and
(iv) any sentence in Δ of the form $\varphi \rtimes \psi$ is justified.

The following properties obtain, as is proven below:

a. Each stage S_i is dialectically preferred.
b. S_i and S_j are incompatible if $i \neq j$.
c. If $i < j$, then the scope of S_i is a proper subset of the scope of S_j.
d. If a stage S is dialectically preferred, such that, for some i, p_i is justified in S, then S is equal to S_i.
e. If a stage S is dialectically preferred, such that no p_i is justified, then all p_i are defeated, all q_i are justified and no r_i is taken into account in S. The scope of this stage is properly contained in the scope of any of the stages S_i.
f. Δ has no maximal dialectically preferred stage.

Proof: The properties a, b and c follow from the definitions. Property d is shown as follows. Assume that p_i is justified in a dialectically preferred stage S. Then q_i and all p_j with $j > i$ are defeated. If, for some j, p_j is defeated, then by lemma (7.11) q_j is justified since S is dialectically preferred. So any q_j with $j > i$ is justified. No p_k with $k < i$ is justified since then p_i would be defeated. Assume now that, for some $k < i$, p_k is not taken into account at S. Then also q_k is not taken into account, for otherwise q_k would be justified or defeated, making p_k defeated or justified, respectively, which would be impossible. But if p_k and q_k are not taken into account, then the stage specified by $J(S) \cup \{q_k\}$ would be dialectically justified, contradicting the $\sqsubseteq$-maximality of S among the dialectically justified stages. Therefore no p_k with $k < i$ is not taken into account. As a result, all p_k with $k < i$ are defeated, and then by the lemma all q_k with $k < i$ are justified. Property e follows from the lemma applied to the sentences q_i. Property f follows from the other properties as follows. Assume that S is a maximal dialectically preferred stage of Δ. Then either there is a p_i that is justified in S or there is no such p_i. The former is impossible since by d S would have to be equal to a stage S_i, but no stage S_i is maximal dialectically preferred by b and c. The latter is impossible since S would be equal to the dialectically preferred stage in property e, which has a scope that is properly contained in the scope of each of the dialectically preferred stages S_i, contradicting that S is maximal dialectically preferred.

Note that the example proves a non-compactness property: though any finite stage of the sample theory has a maximal dialectically preferred stage (by corollary (7.10)), the whole theory does not. Cf. also example (5.9) and one of the sample theories in part (i) of example (4.5).

8 The relations between the types of stages

Among the stages of a theory, the following special types have been distinguished: extensions, maximal stages, satisfiability classes, dialectically justified stages, dialectically preferred stages and maximal dialectically preferred stages. Several relations between types of stages have already been encountered. In this section, the previously found relations are recapitulated and a number of other relations are established.[10]

Satisfiability and dialectical justifiability divides the types in two main groups. The 'satisfiability group' consists of the extensions, the maximal stages, the satisfiability classes and the stages. The 'dialectical justification group' consists of the extensions, the dialectically justified stages, the dialectically preferred stages and the maximal dialectically preferred stages. Note that the type of extensions belong to both groups.

The relations between the types of stages within a group have already been investigated. If $\underline{E}$, $\underline{M}$, $\underline{SC}$, $\underline{S}$, $\underline{DJ}$, $\underline{P}$ and $\underline{MP}$ denote the sets of extensions, maximal stages, satisfiability classes, stages, dialectically justified stages, dialectically preferred stages and maximal dialectically preferred stages of a theory,

[10] This section extends my earlier work on the relations between types of stages (Verheij, 1996a).

respectively, the relations between the types within the same group can be summarized as in the following figure.

$$\begin{tikzcd} \textit{satisfiability types} & \underline{\mathrm{E}} \arrow[r, hook] & \underline{\mathrm{M}} \arrow[r, hook] & \underline{\mathrm{SC}} \arrow[r, hook] & \underline{\mathrm{S}} \\ \textit{dialectical justification types} & \underline{\mathrm{E}} \arrow[r, hook] & \underline{\mathrm{MP}} \arrow[r, hook] & \underline{\mathrm{P}} \arrow[r, hook] & \underline{\mathrm{DJ}} \end{tikzcd}$$

The arrows indicate inclusion maps between the sets of stages. All inclusions have been proven earlier (in the properties (5.6) and (7.4)). They were also shown to be proper inclusions, in the sense that for each inclusion there exists a theory for which the inclusion is proper.

In this section, the relations between the stage types in different groups are investigated. The main tool is the dialectically justified restriction of stages. The dialectically justified restriction of a stage is the largest substage of a stage that is dialectically justified. For any stage, the dialectically justified restriction exists, as the following proposition shows.

Proposition (8.1)

Let S be a stage of a theory Δ. Consider the union $J(S)|_{dj}$ of all dialectically justifying subsets of J(S). Then the stage $S|_{dj}$ specified by $J(S)|_{dj}$ is a justified stage.

Proof: The proposition follows immediately from the properties (6.14) and (7.2).

Definition (8.2): dialectically justified restrictions

If S is a stage of a theory Δ, then the stage $S|_{dj}$ occurring in proposition (8.1) is the *dialectically justified restriction* of S.

By proposition (8.1), there is a (surjective) map from the set of stages to the set of dialectically justified stages, that maps a stage S to its justified restriction $S|_{dj}$. In the following, the properties of this map are investigated.

The dialectically justified restriction of a stage is a dialectically justified stage. One might hope that, by dialectically justified restriction, stages of one of the other satisfiability types E, M and SC map nicely to stages of a dialectical justification type DJ, P or MP. For instance, it could be that the dialectically justified restriction of a maximal stage is always maximal dialectically preferred, or that any maximal dialectically preferred stage is the dialectically justified restriction of a maximal stage. One such relation is trivial: since any extension is its own dialectically justified restriction, the restriction of any extension is an extension and any extension is the restriction of an extension.

Surprisingly, as will be shown below, *no other relation of this kind obtains*. More precisely, the images of the sets M and SC under the restriction map are not in general included in MP or P, and the originals of MP and P do not in general include M or SC. The following four new types of stages are found in this way.

SCDJ: The set of dialectically justified stages that are the dialectically justified restriction of a satisfiability class.

MDJ: The set of dialectically justified stages that are the dialectically justified restriction of a maximal stage.

PSC: The set of satisfiability classes that have a dialectically preferred stage as dialectically justified restriction.

MPSC: The set of satisfiability classes that have a maximal dialectically preferred stage as dialectically justified restriction.

The SCDJ and MDJ types belong to the group of dialectical justification types. The PSC and MPSC types belong to the group of satisfiability types. Note that except for their existence as independent types of stages, these four new classes do not seem to be very interesting. Their existence stresses that satisfiability and dialectical justification are very different notions.

In the following figure, the inclusion and dialectically justified restriction maps between the stage types are summarized. The more interesting 'old' stage types have been highlighted by the use of a bold font. The vertical arrows indicate the dialectically justified restriction maps, all of which are surjective. The arrow from $\underline{\mathrm{E}}$ to $\underline{\mathrm{E}}$ indicates the identity map. All other arrows indicate inclusion maps.

$$\begin{tikzcd}
 & & & \underline{\mathbf{M}} \arrow[d, rr, hook] \arrow[ddd, two heads] & & & \\
\textit{satisfiability types} & \underline{\mathbf{E}} \arrow[u, rr, hook] \arrow[dr, hook] \arrow[ddd, equal] & & & & \underline{\mathbf{SC}} \arrow[r, hook] \arrow[ddd, two heads] & \underline{\mathbf{S}} \arrow[ddd, two heads] \\
 & & \underline{\mathrm{MPSC}} \arrow[rr, hook] \arrow[ddd, two heads] & & \underline{\mathrm{PSC}} \arrow[ur, hook] \arrow[ddd, two heads] & & \\
 & & & \underline{\mathrm{MDJ}} \arrow[d, rr, hook] & & & \\
\textit{dialectical justification types} & \underline{\mathbf{E}} \arrow[u, rr, hook] \arrow[dr, hook] & & & & \underline{\mathrm{SCDJ}} \arrow[r, hook] & \underline{\mathbf{DJ}} \\
 & & \underline{\mathbf{MP}} \arrow[rr, hook] & & \underline{\mathbf{P}} \arrow[ur, hook] & &
\end{tikzcd}$$

That the maps in the figure exist is easy to check using what has been discussed before. The surjectivity of the dialectically justified restriction maps follows from the definitions of the new stage types. For particular theories some or even all of the maps can collapse into identities. An extreme example is the theory {p} in which all sets in the figure except $\underline{\mathrm{DJ}}$ are equal to the singleton set consisting of the theory's extension (specified by p). For this theory, $\underline{\mathrm{DJ}}$ consists of two stages, viz. the empty stage and the theory's extension.

In general, however, any inclusion map is proper, in the sense that for any inclusion map there is a theory for which the inclusion of the sets is proper. For most inclusion maps, this is easy to check using examples encountered earlier. For instance, it follows from the existence of dialectically preferred stages that are not maximal dialectically preferred, that there are theories Δ, such that $\underline{\mathrm{PSC}}_{\Delta}$ is a proper subset of $\underline{\mathrm{MPSC}}_{\Delta}$. Showing the properness of the following inclusion maps requires new examples:

$\underline{\mathrm{P}} \subseteq \underline{\mathrm{SCDJ}}$, but $\underline{\mathrm{P}} \neq \underline{\mathrm{SCDJ}}$:	example (8.3), part (i)
$\underline{\mathrm{SCDJ}} \subseteq \underline{\mathrm{DJ}}$, but $\underline{\mathrm{SCDJ}} \neq \underline{\mathrm{DJ}}$:	example (8.3), part (ii)
$\underline{\mathrm{PSC}} \subseteq \underline{\mathrm{SC}}$, but $\underline{\mathrm{PSC}} \neq \underline{\mathrm{SC}}$:	example (8.3), part (i)

All obtaining inclusions are shown in the figure. Example (8.3), part (iii), shows that $\underline{\mathrm{PSC}} \not\supseteq \underline{\mathrm{M}}$ and $\underline{\mathrm{MDJ}} \not\subseteq \underline{\mathrm{P}}$. Example (8.3), part (iv), shows that $\underline{\mathrm{MPSC}} \not\subseteq \underline{\mathrm{M}}$, $\underline{\mathrm{MPSC}} \not\supseteq \underline{\mathrm{M}}$, $\underline{\mathrm{PSC}} \not\subseteq \underline{\mathrm{M}}$, $\underline{\mathrm{MDJ}} \not\supseteq \underline{\mathrm{MP}}$, $\underline{\mathrm{MDJ}} \not\supseteq \underline{\mathrm{MP}}$ and $\underline{\mathrm{MDJ}} \not\supseteq \underline{\mathrm{P}}$.

Example (8.3)

(i) Let's again look at the theory {p, q, r, $p \ltimes q$, $q \ltimes r$}. Its satisfiability class specified by {q, $p \ltimes q$, $q \ltimes r$} has the empty stage as dialectically justified restriction. The empty stage is not dialectically preferred. The only dialectically preferred stage is the theory's extension specified by {p, r, $p \ltimes q$, $q \ltimes r$}. The only satisfiability class of which it is the restriction is the extension itself.

(ii) The stage S specified by {p, $p \ltimes q$, $q \ltimes r$} of the theory in part (i) is dialectically justified. It has one satisfiability class compatibly extending it, viz., the theory's extension. The dialectically justified restriction of the extension is the extension itself and is not equal to S. As a result, the stage S is in $\underline{\mathrm{DJ}}$, but not in $\underline{\mathrm{SCDJ}}$.

(iii) The theory {p, q, r, $p \ltimes q$, $q \ltimes r$, $r \rtimes r$} has two maximal stages (but no extension). The first, M_1, is specified by {p, $p \ltimes q$, $q \ltimes r$, $r \rtimes r$}: p is justified, q is defeated and r is not taken into account. M_1 is the theory's maximal dialectically preferred stage, and therefore equal to its dialectically justified restriction. The second, M_2, is specified by {q, $p \ltimes q$, $q \ltimes r$, $r \rtimes r$}: p is not taken into account, q is justified and r is defeated. The dialectically justified restriction of M_2 is the empty stage, which is not dialectically preferred. M_2 is a maximal stage, that is not the dialectically justified restriction of a dialectically preferred stage.

(iv) The theory {p, q, r_1, r_2, r_3, $p \rtimes q$, $q \rtimes p$, $q \rtimes r_1$, $r_1 \rtimes r_2$, $r_2 \rtimes r_3$, $r_3 \rtimes r_1$, $r_2 \rtimes r_2$, $r_3 \rtimes r_3$} is an example of a theory with a maximal dialectically preferred stage, for which no compatible maximal stage with larger or equal extent exists. The theory has one maximal stage and one maximal dialectically preferred stage, but they are not compatible. The theory's maximal stage M is specified by p, r_1 and the attack sentences of the theory: in M, p is justified, q defeated, r_1 justified, r_2 defeated and r_3 not taken into account. Its dialectically justified restriction is the dialectically preferred stage P specified by p and the attack sentences (in which r_1 and r_2 are not taken into account). The theory's maximal dialectically preferred stage MP, is specified by q and the attack sentences of the theory: in MP, p is defeated, q justified, r_1 defeated, and r_2 and r_3 not taken into account. It is its own dialectically justified restriction. M has larger scope than MP, but M's dialectically justified restriction has smaller scope than that of MP.

Corollary (8.4)

Let $\#\underline{E}$, $\#\underline{M}$, $\#\underline{SC}$, $\#\underline{DJ}$, $\#\underline{P}$ and $\#\underline{MP}$ denote the number (or cardinality) of extensions, maximal stages, satisfiability classes, dialectically justified stages, dialectically preferred stages and maximal dialectically preferred stages of a theory, respectively. Then the following inequalities hold:

(i) $\#\underline{E} \leq \#\underline{M} \leq \#\underline{SC}$

(ii) $\#\underline{E} \leq \#\underline{MP} \leq \#\underline{P} \leq \#\underline{DJ}$

(ii) $1 \leq \#\underline{P} \leq \#\underline{SC}$

The inequalities are sharp, in the sense that all equalities can occur. No other inequalities hold in general.

Proof: The parts (i) and (ii) follow from the inclusions discussed above. Part (iii) follows from part (i) of theorem (7.8) above, and from the inclusion of $\underline{P}$ in $\underline{SCDJ}$ and the surjection of $\underline{SC}$ is onto $\underline{SCDJ}$. Example (8.3) provides counterexamples to several of the missing inequalities. The theory $\{p, q, r, p \rtimes q, q \rtimes r, r \rtimes q\}$ is a counterexample to the inequality $\#\underline{SC} \geq \#\underline{DJ}$. Its only dialectically justified stage is the empty stage, while there are several satisfiability classes. The theory $\{p, q, p \rtimes q\}$ is a counterexample to $\#\underline{SC} \leq \#\underline{DJ}$. It has four dialectically justified stages, viz. those specified by the subsets of $\{p, p \rtimes q\}$, and three satisfiability classes, specified by the three two-element subsets of $\{p, q, p \rtimes q\}$.

9 The extension existence problem and the extension multiplicity problem

As noted in property (4.4), a theory can have zero, one or several extensions. The possibility that a theory does not always have an extension was called the *extension existence problem*, and the possibility that a theory has more than one extension the *extension multiplicity problem*. In this section these problems are investigated.

According to corollary (6.8), all justified sentences of a theory's extension are dialectically justifiable with respect to the theory, and all defeated sentences dialectically defeasible. As a result, a theory has no extension if there is a sentence in the theory that is neither dialectically justifiable nor dialectically defeasible with respect to the theory (corollary (6.9)). The opposite of the latter does not hold: example (6.10) shows a theory for which all sentences are dialectically justifiable or defeasible with respect to the theory, while the theory lacks an extension.

It turns out that the dialectical justifiability (or defeasibility) of a sentence does not guarantee that there is an extension in which the sentence is justified (or defeated, respectively). A weaker conclusion does follow however: for any dialectically justifiable sentence there is a *dialectically preferred stage* in which the sentence is justified (and similarly for a dialectically defeasible sentence).

Proposition (9.1)

A sentence φ is dialectically justifiable with respect to a theory Δ if and only if there is a dialectically preferred stage of Δ in which φ is justified. A sentence φ is dialectically defeasible with respect to a theory Δ if and only if there is a dialectically preferred stage of Δ in which φ is defeated.

Proof: The proposition follows from the properties (6.14) and (7.2) and from part (ii) of theorem (7.8).

The proposition leads to the following important characterization. It solves the 'dialectically preferred stages multiplicity problem', i.e., the analog of the extension multiplicity problem for dialectically preferred stages. (Note that the 'dialectically preferred stage existence problem' has already been solved: any theory has one or more dialectically preferred stages.)

Theorem (9.2)

A theory Δ has two or more dialectically preferred stages if and only if there is a sentence φ that is both dialectically justifiable and defeasible with respect to Δ. Equivalently, a theory Δ has a unique dialectically preferred stage if and only if there is no sentence φ that is both dialectically justifiable and defeasible with respect to Δ.

Proof: The 'if'-part (of the first equivalence) follows from proposition (9.1) and theorem (7.7). The 'only if'-part is seen as follows. Two dialectically preferred stages P_1 and P_2 must be incompatible by theorem (7.7). Therefore $J(P_1) \cap \Delta$ and $J(P_2) \cap \Delta$ are incompatible. Since $J(P_1) \cap \Delta$ is a dialectically justifying Δ-argument, $J(P_1) \cap \Delta$ attacks $J(P_2) \cap \Delta$. Therefore there is a φ, such that $J(P_1) \cap \Delta \vDash \times\varphi$, while φ is in $J(P_2) \cap \Delta$. Choose such a φ. Since $J(P_1) = Cn(J(P_1) \cap \Delta)$, φ is in $J(P_1)$. As a result, φ is justified in P_1 and defeated in P_2. Since the stages are dialectically preferred, φ must then be dialectically justifiable and defeasible with respect to Δ.

Theorem (9.3)

Let *n* be a natural (or cardinal) number. A theory Δ has exactly *n* dialectically preferred stages if and only if *n* is equal to the maximal number of mutually incompatible dialectically justifying Δ-arguments C.

Proof: Combine proposition (9.1) and theorem (7.7).

Dealing with the extension existence problem requires the notions of dialectical justifiability in a context and of disambiguating arguments.

Definition (9.4): dialectical justifiability in a context

A sentence φ is *dialectically justifiable in the context* C *with respect to* Δ if and only if there is an argument C' that contains C and that dialectically justifies φ with respect to Δ. A sentence φ is *dialectically defeasible in the context* C *with respect to* Δ if and only if $\times\varphi$ is dialectically justifiable in the context C with respect to Δ.

Definition (9.5): disambiguating arguments

A Δ-argument C is *disambiguating* if there is no sentence that is both dialectically justifiable and defeasible in the context C with respect to Δ.

As a result, if a disambiguating argument C is dialectically justifying, there is only one dialectically preferred stage compatibly extending the stage specified by C.

Theorem (9.6)

A theory Δ has no extension if and only if, for any disambiguating Δ-argument C, there is a sentence in Δ that is neither dialectically justifiable nor dialectically defeasible in the context C with respect to Δ. Equivalently, a theory Δ has one or more extensions if and only if there is a disambiguating Δ-argument C, in the context of which any sentence in Δ is dialectically justifiable or dialectically defeasible with respect to Δ.

Proof: Consider a disambiguating argument C, such that any sentence φ in Δ is dialectically justifiable or defeasible in the context C with respect to Δ, say by a dialectically justifying argument C_φ containing C. The collection $\{C_\varphi\}_{\varphi \text{ in } \Delta}$ is compatible, since otherwise C would not be disambiguating (property (6.12)). The union of the collection specifies an extension of Δ. Consider now a theory Δ with an extension E. Then J(E) is a disambiguating dialectically justifying Δ-argument, such that any sentence in Δ is dialectically justifiable or defeasible in the context J(E) with respect to Δ.

The importance of the theorem is that it gives necessary and sufficient conditions for the (non-)existence of an extension in terms of the notion of dialectical justification. It shows that if all sentences of a theory are dialectically interpretable with respect to the theory (i.e., dialectically justifiable or dialectically defeasible), while the theory still lacks an extension, it must be the case that the dialectical justification or defeat of one sentence is incompatible with that of another. The sentences of the theory must be dialectically justifiable in different, incompatible 'choices' of context. Of course such incompatible dialectical justifications can be extended to different and therefore incompatible dialectically preferred stages. In a context that is not disambiguating, this cause for the non-existence of an extension can be obscured. In a disambiguating context, no incompatible choices of justifications can be made. The following provides an example.

The theorem shows that dialectical justification is the 'right' tool to investigate the local structure of the extensions of a theory.

Example (9.7)

Let's reconsider the theory Δ = {p, q, p ⋊ q, q ⋊ p, r, r ⋊ r, s, s ⋊ s, p ⋊ r, q ⋊ s} of example (6.10). Although all sentences are dialectically justifiable or defeasible with respect to Δ, there is no disambiguating dialectically justifying argument C, such that all are dialectically justifiable or defeasible in the context C with respect to Δ. The dialectically justifying Δ-arguments {p, p ⋊ q} for p and {p, ×r, p ⋊ r} for ×r are incompatible with the dialectically justifying Δ-arguments {q, q ⋊ p} for q and {q, ×s, q ⋊ s} for ×s.

Corollary (9.8)

A dialectically preferred stage P of a theory Δ is an extension if and only if any sentence in Δ is dialectically justifiable or defeasible in the context J(P) with respect to Δ. Equivalently, a dialectically preferred stage P of a theory Δ is not an extension if and only if there is a sentence that is neither dialectically justifiable nor defeasible in the context J(P) with respect to Δ.

Proof: Note that if P is a dialectically preferred stage, J(P) is disambiguating, and apply theorem (9.6).

Theorem (9.6) gives criteria for the case that no extension of a theory exists, and for the case that at least one extension exists. The following corollary gives a criterion for the case that at least two extensions exist.

Corollary (9.9)

A theory Δ has two or more extensions if and only if there are two or more incompatible disambiguating dialectically justifying Δ-arguments C and C', in the context of which any sentence in Δ is dialectically justifiable or defeasible with respect to Δ.

Proof: The corollary follows by the combination of theorems (9.6) and (9.2).

Corollary (9.10)

Let *n* be a natural (or cardinal) number. A theory Δ has exactly *n* extensions if and only if *n* is equal to the maximal number of mutually incompatible disambiguating Δ-arguments C, in the context of which any sentence in Δ is dialectically justifiable or defeasible with respect to Δ.

10 Dialectical arguments and the internal structure of dialectical justification

In the present section, the internal structure of arguments dialectically justifying a sentence φ is investigated. At the core of such an argument there is an argument for φ. In general, such an argument is not dialectically justifying, namely in case it does not attack all arguments incompatible with it. A dialectical justification of φ is thus in general *larger* than an argument for φ. In general, it not only contains an argument for φ, but also arguments attacking the arguments incompatible with the argument for φ (cf. corollary (6.2)). It is the goal of this section to investigate how an argument must be extended in order to become dialectically justifying.

In order to probe as deeply as possible into the internal structure of dialectical justification, the investigation will be in terms of elementary arguments and elementary incompatibility, defined as follows.

Definition (10.1): elementary arguments and elementary incompatibility

(i) An argument C is an *elementary argument* for a sentence φ if C is the only argument for φ that is contained in C.

(ii) Let C be an argument. An argument C' is *elementarily incompatible* with C if there is a minimal unsatisfiable subset C'' of C ∪ C', such that C' = C'' \ C. C' *elementarily attacks* C if C' and C' is elementarily incompatible with C and C' attacks C.

For instance, the argument {p, p ⇝ q, q} is an argument for q, but not an elementary argument. Also the argument {p, p ⇝ q, q ⇝ r, r ⇝ q} for q is not elementary. The argument {p, q} is incompatible with {×p}, but not elementarily incompatible. The set {×p} is elementarily incompatible with {p, q} though. This shows that the incompatibility relation is symmetric, while the elementary incompatibility relation is not. Also {p} and {p ⇝ q, p ⋊ q} are incompatible, showing that there need not be a sentence φ, such that φ is a consequence of one argument and ×φ of the other. If C elementarily attacks C' at φ, then C is an elementary argument for φ.

Another natural minimality definition for incompatible arguments would be minimal incompatibility. In that case some causes of incompatibility cannot be distinguished however. It can for instance be the case that an argument attacks another argument, while there is no corresponding minimally incompatible attacking argument. E.g., the argument {p, p ⇝ ×q} (minimally) attacks the argument {q, q ⇝ ×p}, but is not minimally incompatible with it. The former is elementarily incompatible with the latter though since {p, p ⇝ ×q, q} is minimally unsatisfiable.

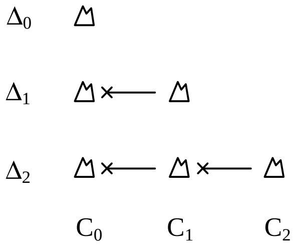
Cf. also definition (3.14) and the remarks following it.

In the following, some examples of increasing complexity are discussed as a preliminary to the systematic investigation of the internal structure of dialectical justification in terms of elementary arguments and elementary incompatibility.

Example (10.2)

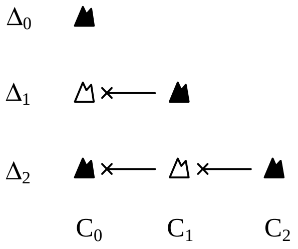
Consider the theories Δ_0 = {p}, Δ_1 = {p, ×p} and Δ_2 = {p, ×p, ××p}. Their respective unique extensions are specified by the arguments {p}, {×p} and {p, ××p}. As a result, p is dialectically justifiable with respect to Δ_0 and Δ_2, but dialectically defeasible with respect to Δ_1. Let's now try to explain this in terms of the elementary arguments and elementarily incompatible arguments of the theories. The only minimal argument for p is {p}. Let's call it C_0. In Δ_1 and Δ_2, there is one argument elementarily incompatible with it, viz. C_1 = {×p}. On its turn, there is one argument elementarily incompatible with C_1 in Δ_2, viz. C_2 = {××p}. Moreover, C_1 attacks C_0 and C_2 attacks C_1.

The situation is summarized in the figure below. Here and in the following, only elementary arguments and arguments elementarily incompatible with another argument are shown. The alpine shapes indicate the arguments. That an argument attacks another argument is indicated by a cross-headed arrow. Each row corresponds to one of the theories. The alpine shapes in a column indicate the same argument.



The three systems of arguments {C_0}, {C_0, C_1} and {C_0, C_1, C_2} contain all information that explains the status of p. For Δ_0 the situation is simple: C_0 is a Δ-argument for p, and there are no arguments incompatible with it. For Δ_1, the situation is thus: though C_0 is a Δ-argument for p, there is an argument attacking it, viz. C_1. Note that while C_0 is also incompatible with C_1, the situation is not symmetric, since C_0 does not attack C_1. Since from Δ_1, there is no argument attacking C_1, the argument C_0 cannot be extended to an argument dialectically justifying p. In Δ_2, this is remedied by the argument C_2: it is an argument attacking C_1, thereby making it possible to extend C_0 to the dialectically justifying argument $C_0 \cup C_2$. In Δ_1, there is one argument that is incompatible with C_1, viz. C_0. Since C_0 is itself attacked by C_1, all arguments incompatible with C_1 are attacked.

The figure below summarizes the situation. The black arguments are the arguments that are only incompatible with arguments that are themselves attacked.



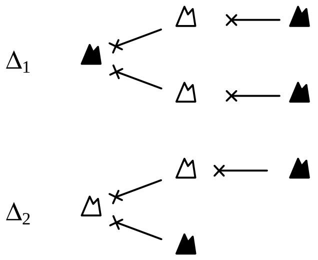
The three theories illustrate the following slogan: an argument can only be extended to a dialectically justifying argument if each argument incompatible with it is attacked.

Example (10.3)

The theory $\Delta = \{p, p \rightsquigarrow q, r, r \rightsquigarrow \times p\}$ allows non-trivial derivations (in contrast with the previous example). The theory's extension E is specified by $p \rightsquigarrow q$, r and $r \rightsquigarrow \times p$. The sentence q is not justified in E (q is not taken into account) since $\{r, r \rightsquigarrow \times p\}$ is an argument attacking the only argument $\{p, p \rightsquigarrow q\}$ for q, and there is no attack against it.

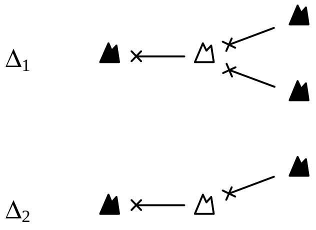
Example (10.4)

An argument can be incompatible with or attacked by more than one argument. The theories $\Delta_1 = \{p, q_1, q_2, r_1, r_2, q_1 \rightsquigarrow \times p, q_2 \rightsquigarrow \times p, r_1 \rightsquigarrow \times q_1, r_2 \rightsquigarrow \times q_2\}$ and $\Delta_2 = \Delta_1 \setminus \{r_2\}$ show that each attacking argument must be attacked. In the extension of Δ_1, p is justified, in that of Δ_2, defeated. There are two arguments attacking the (trivial) argument {p} for p, viz. the arguments $\{q_1, q_1 \rightsquigarrow \times p\}$ and $\{q_2, q_2 \rightsquigarrow \times p\}$. In Δ_1, there are arguments attacking each, viz. $\{r_1, r_1 \rightsquigarrow \times q_1\}$ and $\{r_2, r_2 \rightsquigarrow \times q_2\}$. In Δ_2, the latter is missing. The relations between the arguments concerning p are summarized in the following figure.



Example (10.5)

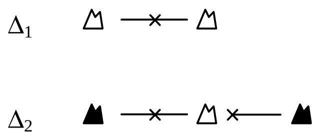
The theories $\Delta_1 = \{p, q, r_1, r_2, q \rightsquigarrow \times p, r_1 \rightsquigarrow \times q, r_2 \rightsquigarrow \times q\}$ and $\Delta_2 = \Delta_1 \setminus \{r_2\}$ show that one argument suffices as an attack against incompatible or attacking arguments, even if there are several. In the extension of Δ_1, p is justified. The argument $\{q, q \rightsquigarrow \times p\}$ attacks the argument {p} for p. In Δ_1, the argument is attacked by two elementary arguments, viz. $\{r_1, r_1 \rightsquigarrow \times q\}$ and $\{r_2, r_2 \rightsquigarrow \times q\}$. From Δ_2, the latter is missing, but still p is justified in the extension of Δ_2. Cf. the following figure.



Example (10.6)

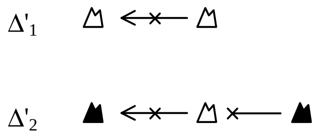
For an argument to be dialectically justifying, it does not suffice that there are arguments attacking against only the arguments *attacking* it (see also the sections 12.4 and 13.3 on the notion of admissibility). The theory $\Delta_1 = \{p, p \rightsquigarrow r, q, q \rightsquigarrow \times r\}$ is an example. The arguments $\{p, p \rightsquigarrow r\}$ and $\{q, q \rightsquigarrow \times r\}$ are incompatible with each other, but neither attacks the other. As a result, the arguments attack all arguments attacking them, since there are none. Still, Δ_1 has no extension. The theory $\Delta_2 = \Delta_1 \cup \{\times q\}$ does have an extension since the argument $\{\times q\}$ attacks $\{q, q \rightsquigarrow \times r\}$. Δ_2's extension is the interpretation specified by $\{p, p \rightsquigarrow r, q \rightsquigarrow \times r, \times q\}$, in which p and r are justified, and q is defeated.

The figure summarizes the situation. The crossed line indicates elementary incompatibility in both directions.



Example (10.7)

A related theory illustrating that a dialectically justifying argument must attack all arguments incompatible with it, is the theory $\Delta'_1 = \{p, p \rightsquigarrow q, p \rtimes q\}$. The elementary argument $\{p\}$ for p is elementarily incompatible with the argument $\{p \rightsquigarrow q, p \rtimes q\}$. Neither argument is attacked by an argument. The theory has no extension, while $\Delta'_2 = \Delta'_1 \cup \{\times(p \rtimes q)\}$ has one. Cf. the following figure, in which the crossed arrow indicates (one-directional) elementary incompatibility.



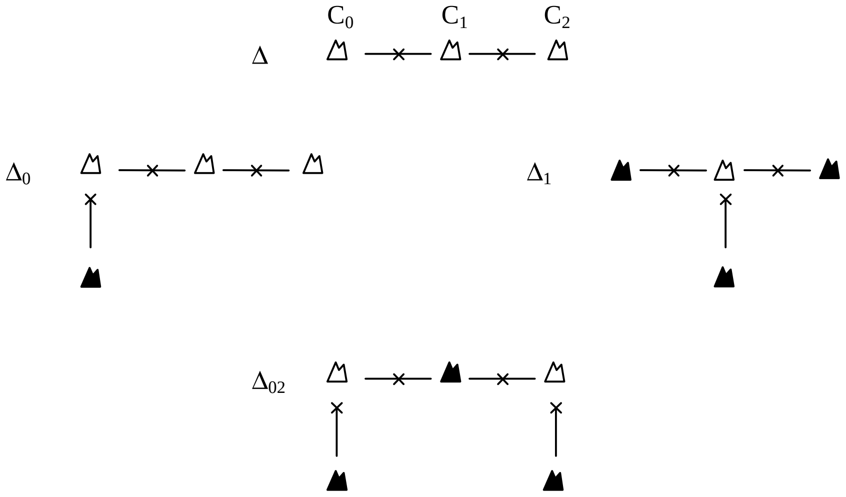
Definition (10.8): opposites

Sentences φ and ψ are *opposites* if φ is equal to $\times\psi$ or ψ is equal to $\times\varphi$.

A sentence φ that is not of the form $\times\psi$, has one opposite, viz. $\times\varphi$. A sentence φ of the form $\times\psi$ has two opposites, viz. $\times\varphi$ (which is equal to $\times\times\psi$) and ψ. For a sentence φ of the form $\times\psi$, ψ is conveniently denoted as $\times^{-1}\varphi$.

Example (10.9)

The sentence $\times p$ has two opposites, viz. p and $\times\times p$. Arguments for either of p's opposites are incompatible with arguments for p, and must be attacked if p is to be justified in an extension. Consider the theory $\Delta = \{q_0, q_1, q_2, q_0 \rightsquigarrow p, q_1 \rightsquigarrow \times p, q_2 \rightsquigarrow \times\times p\}$. There are three non-trivial arguments from Δ: the first $\{q_0, q_0 \rightsquigarrow p\}$, the second $\{q_1, q_1 \rightsquigarrow \times p\}$, the third $\{q_2, q_2 \rightsquigarrow \times\times p\}$. They are denoted C_0, C_1 and C_2, respectively. C_0 and C_1 are (elementarily) incompatible with each other, just as C_1 and C_2. (Cf. the figure below.) The theory Δ has no extension since none of the incompatible arguments is attacked. From each of the theories $\Delta_0 = \Delta \cup \{r_0, r_0 \rtimes q_0\}$, $\Delta_1 = \Delta \cup \{r_1, r_1 \rtimes q_1\}$ and $\Delta_2 = \Delta \cup \{r_2, r_2 \rtimes q_2\}$, a new argument $\{r_i, r_i \rtimes q_i\}$ (for i = 0, 1 or 2) can be constructed, denoted C'_i. C'_i attacks C_i. Only Δ_1 has a (unique) extension, viz. the interpretation specified by $\{q_0, q_2, q_0 \rightsquigarrow p, q_1 \rightsquigarrow \times p, q_2 \rightsquigarrow \times\times p, r_1, r_1 \rtimes q_1\}$, in which p and $\times\times p$ are justified, q_1 is defeated (by r_1) and $\times p$ is not taken into account (since the argument C_1 of $\times p$ is attacked). In Δ_1, C_1 is attacked by C'_1, so that all arguments (elementarily) incompatible with C_0 and C_2 (there is only one: C_1) are attacked. Δ_0 and Δ_2 have no extension. In Δ_0, for instance, only one of the two arguments C_0 and C_2 incompatible with C_1 is attacked, viz. the argument C_0, which is attacked by C'_0. From Δ_0, there is no argument attacking the argument C_2 incompatible with C_1. For Δ_2 similar remarks apply. From the union Δ_{02} of the theories Δ_0 and Δ_2, arguments attacking either argument incompatible with C_1 can be constructed, viz. C'_0 and C'_2. As a result, Δ_{02} has a (unique) extension, viz. the interpretation specified by $\{q_1, q_0 \rightsquigarrow p, q_1 \rightsquigarrow \times p, q_2 \rightsquigarrow \times\times p, r_0, r_0 \rtimes q_0, r_2, r_2 \rtimes q_2\}$, in which $\times p$ is justified, q_0 and q_2 are defeated (by r_0 and r_2), and p and $\times\times p$ are not taken into account (since the arguments C_0 and C_2 of p and $\times\times p$ are attacked).



In the previous examples, for determining whether a sentence φ is justified in an extension of a theory Δ, it sufficed to consider the *structure* of the graph of arguments formed by all elementary arguments for φ, all arguments elementarily incompatible with those arguments, the arguments elementarily incompatible with those, etc. and the *type* of incompatibility (viz., attacking or not). If all arguments incompatible with an argument were attacked, the argument would be effective.

In general, considering the *dialectical graph* concerning a sentence φ does not suffice to determine whether it is dialectically justifying. A complication occurs when the sentences in one argument occur in another. The following two examples show first a theory in which such a problem arises, and then a similar example that is not problematic. The relevant notions are *indirect support* are *indirect incompatibility*.

Definition (10.10)

Let Δ be a theory and C an elementary Δ-argument for a sentence φ. Then C *(directly) supports* φ. A Δ-argument C *indirectly supports* φ if there is a series of Δ-arguments C_0, C_1, ..., C_n with n an even natural number larger than 2, such that the following obtain:

(i) C_0 is an elementary argument for φ, and C_n is equal to C.

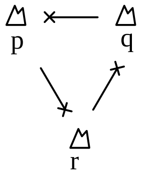
(ii) For all odd i from 0 to n, C_i is elementarily incompatible with C_{i-1}.

(iii) For all even i from 2 to n, C_i elementarily attacks C_{i-1}.

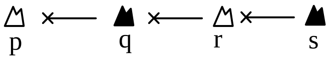
A Δ-argument C is *directly incompatible with* φ if there is such a series with n equal to 1, and *indirectly incompatible* if there is a such a series with n an odd natural number larger than 1.

Example (10.11)

A difficulty arises in the theory $\Delta = \{p, q, r, p \ltimes q, q \ltimes r, r \ltimes p\}$. A part of the graph of arguments relevant for justifying p is indicated in the following figure. The three arguments depicted are $\{p, r \ltimes p\}$, $\{q, p \ltimes q\}$ and $\{r, q \ltimes r\}$.



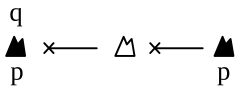
Note that any argument that attacks (or is incompatible with) another argument is itself attacked. Still, Δ has no extension. The problem arises by the fact that p occurs both at a (directly or indirectly) incompatible and at a (directly or indirectly) supporting place in the dialectical graph. The cause of the difficulty is that p plays opposing roles with respect to the justification of q. It can be seen that this is the cause of the missing extension if one considers the theory $\Delta' = (\Delta \setminus \{r \ltimes p\}) \cup \{s, r \ltimes s\}$, in which the role of attacking r is taken over from p by the sentence s. The theory Δ' has an extension, viz. the interpretation specified by {q, s, p ⋉ q, q ⋉ r, r ⋉ s}, in which p and r are defeated. The situation is depicted in the following figure.



Example (10.12)

The theory Δ = {p, p → q, r, r ⋊ (p → q), p ⋊ r} has the interpretation specified by {p, p → q, r ⋊ (p → q), p ⋊ r}, in which p and q are justified and r is defeated, as its extension. The role of p in justifying q is noteworthy: p occurs not only in the only elementary argument {p, p → q} for q, but also in the argument {p} that attacks the argument {r, r ⋊ (p → q)} against p → q that attacks the argument for q.

The sentence p occurs both in a (directly or indirectly) supporting and in a (directly or indirectly) incompatible position. A part of the graph of arguments relevant for justifying q is indicated in the following figure. It is suggested that q is a conclusion of the left-most argument and that p is a premise in two of the indicated arguments. The double role of p does not lead to a problem.



As a result, sentences should not occur in arguments with different roles, viz. both in a (directly or indirectly) supporting argument and in a (directly or indirectly) incompatible one.

The examples above lead to the following definitions of elementary dialectical graphs and justifying dialectical arguments. The dialectical graph concerning a sentence φ with respect to a theory Δ is the graph of elementary Δ-arguments for φ, the arguments elementarily incompatible with them, those elementarily incompatible with the latter, etc.

Definition (10.13): elementary dialectical graphs

The *elementary dialectical graph concerning* φ *with respect to* Δ is the smallest collection $\{C_i\}_{i \in I}$ of Δ-arguments, such that the following obtain:

(i) Any elementary Δ-argument for φ is in the collection.

(ii) If C is an argument in the collection and C' a Δ-argument that is elementarily incompatible with C, then C' is in the collection.

Note that the elementary dialectical graph concerning φ consists of the arguments that (directly or indirectly) support φ and those that are (directly or indirectly) incompatible with it.

Below it will be shown that the dialectical justifiability of a sentence φ with respect to Δ coincides with the existence of a subgraph of the elementary dialectical graph concerning φ that has properties as illustrated by the examples above. Such special subtrees are *dialectical arguments justifying* φ.

Definition (10.14): justifying dialectical arguments

A *dialectical argument justifying* φ *with respect to* Δ is a (non-empty) subcollection $\{C_i\}_{i \in I'}$ of the elementary dialectical graph $\{C_i\}_{i \in I}$ concerning φ with respect to Δ, such that the following obtain:

(i) There is an elementary Δ-argument for φ is in the subcollection.

(ii) No argument in the collection both (directly or indirectly) supports φ and is (directly or indirectly) incompatible with φ.

(iii) If C is an argument in the subcollection that is directly or indirectly incompatible with an argument for φ in the subcollection, then there is an argument in the subcollection that attacks it.

(iv) If C is an argument in the subcollection that directly or indirectly supports φ, then all Δ-arguments in the elementary dialectical graph that are incompatible with C are in the subcollection.

The union of the arguments in a dialectical argument justifying φ that directly or indirectly support φ, are the *justified premises* of the dialectical argument. The union of the arguments in a dialectical argument justifying φ that are directly or indirectly incompatible with φ, are the *defeated premises* of the dialectical argument. The sentences supported by the set of justified premises of a dialectical argument justifying φ are the *justified conclusions* of the dialectical argument. The sentences attacked by the set of justified premises of a dialectical argument justifying φ are the *defeated conclusions* of the dialectical argument.

Note that there is a slight abuse of terminology here, since dialectical arguments are not a special kind of arguments (in the sense of definition (3.14)), but graphs of arguments. Normally, no confusion is to be expected. If required, one could speak of *monolectical arguments* when referring to arguments in the sense of satisfiable sets of sentences, as in definition (3.14).

Note the difference between dialectically justifying arguments, which are satisfiable sets of sentences of a special kind, and justifying dialectical arguments, which are certain subcollections of elementary dialectical graphs. The closely related terminology is not coincidental, as the following proposition and theorem show.

Proposition (10.15)

Let $\{C_i\}_{i \in I'}$ be a dialectical argument justifying φ with respect to a theory Δ. Then the set C of justified premises of the dialectical argument is an argument dialectically justifying φ with respect to Δ.

Proof: First it must be shown that C is an argument, i.e., that it is satisfiable. Assume to the contrary that C is not satisfiable. Then there is a minimal subset C* of C that is not satisfiable. C* is finite, say $C^* = \{\varphi_0, ..., \varphi_n\}$. For each φ_t there is a C_t in the dialectical argument that contains φ_t and directly or indirectly supports φ. For any t from 0 to n, let C^*_t be the set $C^* \setminus C_t$. Then C^*_t is satisfiable and therefore C^*_t is elementarily incompatible with C_t. As a result, C^*_t is directly or indirectly incompatible with φ and occurs in the dialectical argument. Subsequently, there is a C^{**}_t in the dialectical argument that attacks C^*_t. So, there is a sentence ψ in C^*_t, attacked by C^{**}_t. But since C^*_t is a subset of C*, there is a t' with 0 ≤ t' ≤ n, such that $\varphi = \varphi_{t'}$. So C^{**}_t attacks also $C_{t'}$. A fortiori, C^{**}_t is incompatible with $C_{t'}$. But then C^{**}_t indirectly supports φ, since it attacks C^*_t, and is indirectly incompatible with φ, since it is incompatible with $C_{t'}$. This contradicts the definition of a justifying dialectical argument. So C is satisfiable after all.

Second it must be shown that C is dialectically justifying with respect to Δ. Consider an argument C' that is incompatible with C. Let ψ be a sentence, such that $C \cup C' \vDash \psi$ and $C \cup C' \vDash \times\psi$. By the compactness of the consequence notion ⊨, there is a finite subset C* of C, such that $C^* \cup C' \vDash \psi$ and $C^* \cup C' \vDash \times\psi$. Let $C_0, ..., C_n$ be arguments directly or indirectly supporting φ in the dialectical argument, such that $C^* \subseteq C_0 \cup ... \cup C_n$. Pick a maximal number of indices i(0), ..., i(m) from among the indices 0, ..., n, such that $C^{**} = C' \cup C_{i(0)} \cup ... \cup C_{i(m)}$ is satisfiable. Then there is an index i from among the indices 0, ..., n, such that C** is incompatible with C_i. Let C*** be a subset of C** that is elementarily incompatible with C_i. But C_i directly or indirectly supports φ, so C*** occurs in the dialectical argument, as an argument directly or indirectly incompatible with φ. As a result, there is a C**** in the dialectical argument that attacks a sentence χ in C***. C**** is a subset of C since it directly or indirectly supports φ. Therefore also C is an argument against χ. The sentence χ is an element of $C^{**} \setminus C_{i(0)} \cup ... \cup C_{i(m)} \subseteq C'$ since if χ were an element of $C_{i(0)} \cup ... \cup C_{i(m)}$, χ would be an element of C and C would not be satisfiable. This implies that C attacks C', as required.

Finally, it must be checked that φ is a conclusion of C, which follows from the fact that any dialectical argument justifying φ contains an elementary Δ-argument for φ.

Theorem (10.16)

There is an argument C dialectically justifying a sentence φ with respect to Δ if and only if there is a dialectical argument justifying φ with respect to Δ.

Proof: The 'if'-part of the theorem follows from the proposition. For the 'only if'-part of the theorem, a dialectical argument justifying φ must be constructed given an argument C dialectically justifying φ. The construction goes by induction on *n*, as follows. At *n* = 0, start with the collection of elementary C-arguments for φ. At an odd level *n*+1, add all arguments that are elementarily incompatible with the arguments added at level *n*. At an even level *n*+1, add all elementary C-arguments that attack an argument added at level *n*. That at least one such argument exists, is shown as follows. Pick an argument C_n added at the odd level *n*. C_n is added as an argument that is elementarily incompatible with an argument C_{n-1}, added at level *n*-1. C_{n-1} is a subset of C, since *n*-1 is even. Therefore C_n and C are incompatible. Since C is dialectically justifying, C attacks C_n. But then there is also an elementary C-argument that attacks C_n. It remains to be checked that no argument added in the construction both (directly or indirectly) supports φ and is (directly or indirectly) incompatible with φ. Assume to the contrary that such an argument C* is added in the construction. Then there would be sequences of arguments $C_0, ..., C_{2i+1}$ and

$C'_0, ..., C'_{2j}$, as in the definition of indirect support and incompatibility, all added in the construction, with C_0 and C'_0 elementary arguments for φ and C_{2i+1} and C'_{2j} equal to C*. But then there is an elementary C-argument C** added in the construction that attacks C_{2i+1} = C*. But both C* and C** are subsets of C, contradicting the satisfiability of C.

This theorem shows that a justifying dialectical argument reveals the internal structure of a dialectically justifying argument.

Note that the elementarity conditions on the arguments that form a justifying dialectical argument serve to enforce that justifying dialectical arguments expose the internal structure of dialectical justification to the highest possible degree.

How do the justifying dialectical arguments as defined here relate to the naïve dialectical arguments of section 2? Apart from the difference in presentation between the two, where justifying dialectical arguments are formally elaborated, while naïve dialectical arguments were only informally presented, there is also a conceptual difference. In justifying dialectical arguments, a stronger notion of defense against counterarguments is used, viz. dialectical justification, according to which any incompatible argument must be attacked. Naïve dialectical arguments can only model what might be called 'naïve dialectical justification', according to which any argument that attacks a consequence of an argument must be attacked. Naïve dialectical justification is a weaker notion of dialectical justification (just like admissibility) that does not suffice in the analysis of the internal structure of dialectical justification and dialectical interpretation as extensions. Cf. also section 12.4.

An interesting topic of research is whether the extension existence and multiplicity problems are simplified or even disappear for theories with well-behaved elementary dialectical graphs. A simple and often powerful property is the well-foundedness of the tree expansion of a graph. A tree is well-founded if it contains no infinite branches. In the literature, well-foundedness has been fruitfully used in the context of argument defeat. For instance, Dung (1995) proves that his argumentation frameworks have a unique stable extension when they are well-founded.

Unfortunately, the tree expansion of the elementary dialectical graph of an unsatisfiable theory Δ is never well-founded. This can be seen as follows. Let C be any minimal unsatisfiable subset of Δ. Then C can be split into two non-empty disjoint subsets C' and C" that are elementarily incompatible with each other. By choosing C', such that it is a minimal argument for a sentence φ, it follows that the tree expansion of the dialectical graph concerning φ with respect to Δ contains C', C", C', C", ... as an infinite branch. As a result, considering only theories with well-founded dialectical trees would exclude all unsatisfiable theories.

It is then natural to consider the *restricted* tree expansion of the dialectical graph, which does not include all arguments that are elementarily incompatible with other arguments, but only those that attack arguments in the graph. If one restricts the tree expansion of the dialectical graph to attacks only, the notion of well-foundedness does no longer exclude all unsatisfiable theories. A simple example is the theory $\{p, q, p \rtimes q\}$ that has a unique extension. Its 'attack tree' is clearly well-founded, though its full dialectical tree is not. It might be hoped that it holds in general that theories with well-founded attack have a unique extension. This is however not the case. An example is the theory $\{p, p \leadsto q, p \rtimes q\}$ that has well-founded attack, but no extension. Examining the example, it can be conjectured that the problem is that the theory does not only have *attack-type counterarguments*: the argument $\{p, p \rtimes q\}$ is a counterargument to $\{p, p \leadsto q\}$, but does not attack it.

It turns out that theories with well-founded attack and only attack-type counterarguments indeed have a unique extension. The following definition formalizes the relevant notions.

Definition (10.17): well-founded attack and only attack-type counterarguments

(i) The set of sentences Δ is a *theory with well-founded attack* if there is no infinite sequence of arguments $C_0, C_1, C_2, ...,$ where each C_i is attacked by its successor C_{i+1}.

(ii) The set of sentences Δ is a *theory with only attack-type counterarguments* if for all Δ-arguments C and C' with $C \vDash \times\varphi$ and $C' \vDash \varphi$, it holds that φ is an element of C'.

The following theorem can now be proven. It is a generalization of Dung's (1995) result that well-founded argumentation frameworks have a unique stable extension (cf. section 13.2 below).

Theorem (10.18)

If Δ is a theory with well-founded attack and only attack-type counterarguments, then Δ has a unique extension. The extension is also Δ's unique dialectically preferred stage.

Proof: For any Δ-argument C, define $F_\Delta(C)$ as the union of all Δ-arguments that are only attacked by Δ-arguments that are attacked by C. Claim: the union of the sets $\emptyset$, $F_\Delta(\emptyset)$, $F_\Delta(F_\Delta(\emptyset))$, ..., denoted $G(\Delta)$, specifies an extension of Δ. The claim is shown in two steps. First, it is shown that $G(\Delta)$ is satisfiable. Since F_Δ is monotonic and minimal incompatible subsets of $G(\Delta)$ would be finite, it suffices to show that $F_\Delta(C)$ is satisfiable for any argument C. Assume that $F_\Delta(C)$ is not satisfiable, i.e., there are $F_\Delta(C)$-arguments C' and C" for sentences $\times\varphi$ and φ, respectively. Since Δ has only attack-type counterarguments, φ is an element of C". Since C" is a subset of $C_1 \cup ... \cup C_n$, for certain arguments $C_1, ..., C_n$ that are only attacked by Δ-arguments that are attacked by C, φ is an element of one of the C_i, say $C_{i(0)}$. So C' attacks $C_{i(0)}$ and therefore C attacks C'. But then it follows that C attacks itself, contradicting that C is satisfiable.

Second, it is shown that if $G(\Delta)$ is not an argument against all sentences in $\Delta \setminus Cn(G(\Delta))$, it is not well-founded. Assume that φ_0 is in $\Delta \setminus Cn(G(\Delta))$, while $G(\Delta)$ is not an argument against φ_0. Since $G(\Delta)$ is a subset of $F_\Delta(G(\Delta))$ (even equal to it), $G(\Delta)$ does not attack all Δ-arguments attacking the argument $C_0 = \{\varphi_0\}$, for otherwise φ_0 would be in $G(\Delta)$. So there is a Δ-argument C_1 with $C_1 \vDash \times\varphi_0$ that is not attacked by $G(\Delta)$. C_1 attacks C_0. C_1 cannot be a subset of $Cn(G(\Delta))$, for then $G(\Delta)$ would be an argument against φ_0. So there is a sentence φ_1 in C_1 that is not in $Cn(G(\Delta))$. As a result, φ_1 is in $\Delta \setminus Cn(G(\Delta))$, while $G(\Delta)$ is not an argument against φ_1. Repeating the above with φ_1 in the place of φ_0, an argument C_2 attacking C_1 and a sentence φ_2 in $\Delta \setminus Cn(G(\Delta))$ that is not attacked by $G(\Delta)$, are found. Continuing inductively, one finds a sequence of arguments C_0, C_1, C_2, ..., each attacked by its successor.

11 Representational issues

In this section, a number of representational issues is discussed. How expressive is DefLog?[11]

11.1 Non-defeasible and defeasible assumptions

In several logical models for defeasible reasoning, theories are divided into two parts. One part of a theory consists of the non-defeasible assumptions, the other of the defeasible assumptions. Above, no such distinction has been made. The set of assumptions was encoded as an unstructured set of sentences. As a result, DefLog's definitions and proofs are simpler since they do not need to keep track of two distinct parts of a theory. The question arises whether the lack of this distinction in DefLog is a limitation. It is not, in the sense that it is easy to define the extensions of 'mixed theories', consisting of a non-defeasible and a defeasible part, in terms of the definition of DefLog's extensions of 'completely defeasible' theories, as follows.

Let (T, Δ) be a pair of sets of sentences. Then the following definition of extension has the effect that the sentences in T are interpreted non-defeasibly, while the sentences in Δ are interpreted defeasibly in an extension:

> An extension of the theory (T, Δ) is an extension E of $T \cup \Delta$, such that the sentences in T are justified in E.

Below, this definition of extensions of mixed theories is occasionally useful.

11.2 Defeasible vs. inconclusive conditionals

DefLog's conditionals are defeasible in the sense that a conditional $\varphi \leadsto \psi$ in a theory or following from a theory can be defeated in an extension of the theory. In this sense, conditionals are treated on a par with the non-conditional sentences. Any sentence, whether it is a conditional or not, can be defeated in an extension of a theory, even though it is in the theory's *Modus ponens* closure.

For conditionals, there is however a second way in which they can be considered to be defeasible, that is typical for conditionals only: it can be that under exceptional circumstances the conditional is not followed, while it is not itself defeated. Under such circumstances the conditional does not imply its consequent even though its antecedent does. In order to distinguish this second type of defeasibility for

[11] Section 13 on related research shows aspects of DefLog's expressiveness. For instance, the treatment of Reiter's logic for default reasoning (section 13.1) shows how defeasible rules of inference can be modeled in DefLog, while the treatment of Vreeswijk's abstract argumentation systems (section 13.2) shows the modeling of defeasible arguments (in the sense of derivations).

conditionals, such conditionals are categorized as *inconclusive* here. In contrast, when a conditional is said to be *defeasible*, the standard DefLog type of defeasibility is meant that is not specific for conditionals.

An example might clarify the distinction between defeasible and inconclusive conditionals. Let $p \Rightarrow q$ be a conditional and let r express an exceptional circumstance. Then if $p \Rightarrow q$ is a defeasible conditional, r will make $p \Rightarrow q$ itself defeated. The result is that q does not follow if p obtains. In a sense, the conditional itself has disappeared. If $p \Rightarrow q$ is an inconclusive conditional, however, r will not make the conditional $p \Rightarrow q$ defeated, but only its effect, viz. that the conditional makes that q follows from p. By r, $p \Rightarrow q$ does not have the effect that q follows from p, even though the conditional itself is not defeated. Only the conditional's effect has disappeared.

DefLog's conditionals are defeasible (just like any sentence), but not inconclusive since in any interpretation when a conditional and its antecedent are justified the conditional's consequent is also justified. Again the question arises whether this is a limitation.

It is not, since there is a simple way to incorporate inconclusive conditionals in DefLog. Let $\rightarrowtail$ be a new connective that will be used to express an inconclusive conditional. The following scheme of (defeasible) assumptions suffices to make $\rightarrowtail$ function as an inconclusive conditional:

$$(\varphi \rightarrowtail \psi) \leadsto (\varphi \leadsto \psi)$$

By the scheme, $\rightarrowtail$ is turned into a connective such that $\varphi \rightarrowtail \psi$ normally implies $\varphi \leadsto \psi$. As a result, it normally follows from $\varphi \rightarrowtail \psi$ and φ via $\varphi \leadsto \psi$ that ψ. But since the scheme is to be interpreted defeasibly, the effect of the conditional $\varphi \rightarrowtail \psi$ can be blocked. That there are exceptional circumstances in which the conditional $\varphi \rightarrowtail \psi$ does not have its normal effect, is straightforwardly expressed as $\times((\varphi \rightarrowtail \psi) \leadsto (\varphi \leadsto \psi))$. If required, this expression can be abbreviated as $\sim(\varphi \rightarrowtail \psi)$. As a result, ~ is a dedicated kind of negation for conditionals. Note that ~ is not an ordinary connective, since it cannot be attached to any sentence, but only to conditionals: while $\sim(p \rightarrowtail q)$ and $\sim(p \rightarrowtail (q \rightarrowtail r))$ are sentences, $\sim p$ is not.

As an example of the mixed theories discussed in section 11.1, it is shown how a system is arrived at in which all sentences are interpreted non-defeasibly, while the system's conditionals are inconclusive.

Let T be any theory consisting of sentences using only the connectives $\rightarrowtail$ and ~ (with the restriction that ~ only occurs in front of a $\rightarrowtail$-conditional sentence). Note that in T, DefLog's connectives $\leadsto$ and × do not occur, except 'hidden' in sentences of the form $\sim(\varphi \rightarrowtail \psi)$ that abbreviate $\times((\varphi \rightarrowtail \psi) \leadsto (\varphi \leadsto \psi))$. Let Δ consist of all sentences of the scheme $(\varphi \rightarrowtail \psi) \leadsto (\varphi \leadsto \psi)$. Consider the extensions of the mixed theory (T, Δ), as defined at the end of section 11.1, as the interpretations of a theory T about inconclusive conditionals. Such interpretations are referred to as the $\{\rightarrowtail, \sim\}$-extensions of T.

Note that in $\{\rightarrowtail, \sim\}$-extensions of a theory T no sentence of T is defeated. In particular, no sentence $\varphi \rightarrowtail \psi$ or $\sim(\varphi \rightarrowtail \psi)$ can be defeated. The only sentences that can be defeated are of the form $(\varphi \rightarrowtail \psi) \leadsto (\varphi \leadsto \psi)$. The conditional $\rightarrowtail$ is inconclusive as planned: it can be the case that sentences $\varphi \rightarrowtail \psi$ and φ are both justified while ψ is not. A simple example is provided by the theory T consisting of the following four sentences:

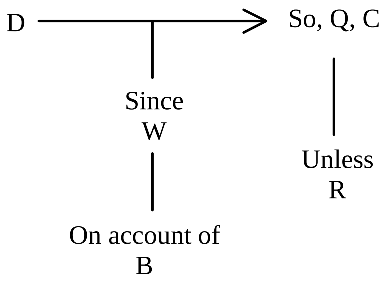
$$p \rightarrowtail q,\ r \rightarrowtail \sim(p \rightarrowtail q),\ p,\ r$$

In its unique extension, $p \rightarrowtail q$ and $\sim(p \rightarrowtail q)$ are justified and q is not interpreted. Note that in this example $p \rightarrowtail q$ cannot be reinstated (i.e., made effective again) by adding $\{\rightarrowtail, \sim\}$-sentences to T, since r is a strict assumption. If r would itself have been the conclusion of an inconclusive conditional (as e.g. in the theory consisting of $p \rightarrowtail q$, $r \rightarrowtail \sim(p \rightarrowtail q)$, p, r', $r' \rightarrowtail r$), the conditional $p \rightarrowtail q$ could be reinstated by blocking that conditional (in the example $r' \rightarrowtail r$, that is blocked by adding $\sim(r' \rightarrowtail r)$).

The system of $\{\rightarrowtail, \sim\}$-extensions shows that DefLog's use of defeasible conditionals does not preclude the modeling of inconclusive conditionals.

11.3 Toulmin's argument scheme

The following figure is adapted from Toulmin's *The Uses of Argument* (1958, p. 104).



D for *Datum*
Q for *Qualifier*
C for *Claim*

W for *Warrant*
B for *Backing*
R for *Rebuttal*

The Datum consists of certain facts that support the Claim. The Warrant justifies that the Datum supports the claim, while the Backing provides on its turn support for the Warrant. A Rebuttal provides conditions of exception, that weaken the Warrant, and the Qualifier can express a degree of force that the Datum gives to the Claim by the Warrant.

Toulmin's so-called argument scheme has had a continuous influence on argumentation researchers (cf., e.g., Bench-Capon, 1995, Van Eemeren *et al*., 1996). In the discussion of naïve dialectical arguments (section 2), some connections between Toulmin's scheme and DefLog have been mentioned.

As examples of the qualifier Q, Toulmin mentions modal qualifiers, such as 'probably' and 'presumably'. Since DefLog's language has no modal operators, the qualifier Q of the scheme has no counterpart in DefLog. In the following, the qualifier and the claim are therefore taken together in the 'qualified claim' QC.

A straightforward way to model the relation between datum D, claim C and warrant W in DefLog is to think of the warrant W as the conditional D ⇝ QC: it is the formal expression of the conditional connection between the two statements D and QC. A difference with Toulmin is that in his examples warrants often express a more general connection between statements, viz. one between patterns of statements. One way to deal with this is to extend DefLog's language with variables.

The relation between the backing B and the warrant W is expressible by the conditional B ⇝ W, which is then - using the conception of warrants above - equal to B ⇝ (D ⇝ QC). An example is the following:

Thieves should be punished ⇝ (John is a thief ⇝ John should be punished)

This conditional can be regarded as an instance of the following scheme:

Thieves should be punished ⇝ (*Person* is a thief ⇝ *Person* should be punished)

This scheme represents the connection between the 'unconditional' form of a rule statement (Thieves should be punished) with its 'conditional' form (*Person* is a thief ⇝ *Person* should be punished).

Note that in this conception of data, claims, warrants and backings both the connection between datum and claim and that between backing and warrant are expressed as a conditional, the former as D ⇝ QC, the latter as B ⇝ (D ⇝ QC). This suggests a slight generalization of Toulmin's scheme: there could be a statement supporting the conditional B ⇝ (D ⇝ QC). In other words, the backing B and the warrant W can themselves be considered as datum and claim of a Toulmin scheme. This would involve a backing B' for which B' ⇝ (B ⇝ W), i.e., B' ⇝ (B ⇝ (D ⇝ QC)). The following is an example:

The rule that thieves should be punished applies ⇝ (Thieves should be punished ⇝ (John is a thief ⇝ John should be punished))

Note by the way that the conception of data, claims, warrants and backings as presented here interprets them in a pure *Modus ponens* context, as in the following figure. On the left, the warrant W is replaced by the conditional D ⇝ QC with which it is identified.

$$\frac{D \qquad \dfrac{B \qquad B \leadsto (D \leadsto QC)}{D \leadsto QC}}{QC} \qquad\qquad \frac{D \qquad \dfrac{B \qquad B \leadsto W}{W}}{QC}$$

Each statement in the scheme can of course itself again occur in another scheme, possibly in a different role. The most obvious example of this is that the claim of one scheme can be the datum of the next.

What about Toulmin's notion of rebuttal R? A first obvious conception of the role of Toulmin's rebuttal R is as an attack of the qualified claim QC. This conception is in fact suggested by Toulmin's graphical representation. The connection would then be represented as $R \rtimes QC$. An example would be the following:

John died last year $\rtimes$ John should be punished

In a second conception of the role of Toulmin's rebuttal R, it attacks the connection between data and claim (just as Pollock's undercutting defeaters, to be discussed in the next section). The connection would then be represented as $R \rtimes (D \leadsto QC)$. Here is an example:

John is a minor first offender $\rtimes$ (John is a thief $\leadsto$ John should be punished)

In a third conception of the role of a rebuttal R, it is considered to attack the connection between backing and warrant, which can be represented as $R \rtimes (B \leadsto (D \leadsto QC))$. The following is an example:

John acted under *force majeure* $\rtimes$ (Thieves should be punished $\leadsto$ (*Person* is a thief $\leadsto$ *Person* should be punished))

The interpretation of Toulmin's scheme within DefLog as discussed here (with the three variants for the role of rebuttals) adds all of DefLog's machinery to it: the notions of extensions of theories and of dialectical justification become relevant. In this way, the DefLog interpretation of the scheme is more specific than Toulmin's original description.

Let's briefly consider extensions in the context of the DefLog interpretation of Toulmin's scheme. The following three theories correspond to the three conceptions of Toulmin's rebutters. In each, a datum, backing and rebutter are defeasibly assumed:

$D, B, R, B \leadsto (D \leadsto QC), R \rtimes QC$
$D, B, R, B \leadsto (D \leadsto QC), R \rtimes (D \leadsto QC)$
$D, B, R, B \leadsto (D \leadsto QC), R \rtimes (B \leadsto (D \leadsto QC))$

Of the three theories, only the latter has an extension. For instance, in the first theory the derivation $R, R \rtimes QC / \times QC$ of $\times QC$ does not suffice to block the derivation

$$\frac{D \qquad \dfrac{B \qquad B \leadsto (D \leadsto QC)}{D \leadsto QC}}{QC}$$

of QC. For blocking the latter, attacking one of its premises is required.

Arguably, it is then better to interpret Toulmin's scheme in terms of the inconclusive conditional $\rightarrowtail$ discussed in section 11.2. The warrant W is then interpreted as $D \rightarrowtail QC$. The backing is connected to the warrant by assuming $B \rightarrowtail (D \rightarrowtail QC)$, and the rebuttal for instance by assuming $R \rightarrowtail \sim(D \rightarrowtail QC)$ or $R \rightarrowtail \sim(B \rightarrowtail (D \rightarrowtail QC))$. The theories

$D, B, R, B \rightarrowtail (D \rightarrowtail QC), R \rightarrowtail \sim(D \rightarrowtail QC)$
$D, B, R, B \rightarrowtail (D \rightarrowtail QC), R \rightarrowtail \sim(B \rightarrowtail (D \rightarrowtail QC))$

both have a $\{\rightarrowtail, \sim\}$-extension.

11.4 Pollock's undercutting and rebutting defeaters

Pollock has distinguished two types of reasons leading to defeat (e.g., Pollock, 1987, 1995, p. 40-41, p. 85-86). He speaks of χ as a *rebutting defeater* when φ is a reason for ψ and χ is a reason for denying ψ. According to Pollock, rebutting defeaters are to be contrasted with *undercutting defeaters*. In Pollock's words, undercutting defeaters attack the connection between the reason and a conclusion rather than attacking the conclusion directly. In DefLog, no corresponding distinction between types of defeaters is made since in DefLog's language both types are expressible.

An undercutting defeater presupposes a reason φ for a conclusion ψ. In DefLog, this is represented by the pair of sentences φ and $\varphi \leadsto \psi$. If now χ is an undercutting defeater attacking the connection between φ and ψ, this can be represented by the sentence $\chi \rtimes (\varphi \leadsto \psi)$. Assume now that φ, $\varphi \leadsto \psi$, χ and $\chi \rtimes (\varphi \leadsto \psi)$ are part of a defeasible theory Δ. Then in any extension of Δ in which φ, χ and $\chi \rtimes (\varphi \leadsto \psi)$ are all three justified, it must be the case that $\varphi \leadsto \psi$ is defeated: it follows from χ and $\chi \rtimes (\varphi \leadsto \psi)$ that $\times(\varphi \leadsto \psi)$. As a result, it does not follow from φ that ψ. This is exactly as required: *defeasibly* it is the case that φ justifies ψ. By the undercutting exception χ the connection between the two is lost. Note that in DefLog the sentences φ, $\varphi \leadsto \psi$, χ and $\chi \rtimes (\varphi \leadsto \psi)$ do not have to be part of the defeasible theory Δ itself; it can be sufficient that they follow from the theory.

A rebutting defeater also presupposes a reason φ for a conclusion ψ. If now χ is a rebutting defeater (with respect to φ as a reason for ψ), χ must be a reason denying ψ. Let's interpret such a reason as a reason for not-ψ. Here the nature of the negation not-ψ of ψ is left implicit. It could be the classical or intuitionistic negation of ψ, but in the present context also the defeat of ψ (i.e., $\times\psi$) is a reasonable option. A first attempt to model the rebutting defeater χ in DefLog would include the five sentences φ, $\varphi \leadsto \psi$, χ, $\chi \leadsto$ not-ψ and $\chi \rtimes (\varphi \leadsto \psi)$ in a theory (or would make them follow from a theory). But then a rebutting defeater would be nothing more than an undercutting defeater - as represented by the four sentences φ, $\varphi \leadsto \psi$, χ and $\chi \rtimes (\varphi \leadsto \psi)$ - that is a reason for denying ψ - represented by the fifth sentence $\chi \leadsto$ not-ψ. Indeed this representation does not correctly capture the idea of a rebutter. For in this representation, χ would still attack the connection between φ and ψ if *it were defeated that* $\chi \leadsto$ not-ψ. So even if the reason χ *prima facie* denying ψ is not *actually* denying ψ, it would imply that $\varphi \leadsto \psi$ is defeated.

In the correct way to represent the rebutting defeater χ the statement $\chi \rtimes (\varphi \leadsto \psi)$ is replaced by $(\chi \leadsto$ not-$\psi) \leadsto (\chi \rtimes (\varphi \leadsto \psi))$. Only if the statements $(\chi \leadsto$ not-$\psi) \leadsto (\chi \rtimes (\varphi \leadsto \psi))$, $\chi \leadsto$ not-ψ and χ are all three justified, it follows that $\varphi \leadsto \psi$ is defeated. As a result, χ only has its rebutting effect if it is actually denying ψ.

This account of Pollock's rebutting defeaters can immediately be generalized to what might be called 'priority defeaters'. A *priority defeater* exists in the situation that the occurrence of one reason blocks the occurrence of another reason. When χ is a reason for ω that when it occurs blocks a reason φ for ψ, then χ as a reason for ω is a priority defeater for φ as a reason for ψ. Priority defeaters are analogous to rebutting defeaters with the difference that a priority defeater does not need to deny the conclusion of the reason it attacks, but can attack any reason. Priority defeaters occur frequently in the law (cf., e.g., Hage, 1997, and Prakken, 1997). It can for instance be the case that the application of one legal rule is excluded in case another rule is applied. An example is the *Lex superior derogat legi inferiori* principle, according to which of two rules with conflicting conclusions only the one made by the highest authority applies. When χ as a reason for ω is a priority defeater for φ as a reason for ψ, this can be expressed as $(\chi \leadsto \omega) \leadsto (\chi \rtimes (\varphi \leadsto \psi))$. When $\chi \leadsto \omega$ and χ are both justified, then $\varphi \leadsto \psi$ is defeated. As a result, χ only has its defeating effect when it is actually justifying ω.

Another kind of generalization of Pollock's rebutting defeaters is what might be called 'outweighing defeaters'. A set of reasons for a conclusion are an *outweighing defeater* when they block a set of reasons against the conclusion. The idea is that then the pros outweigh the cons. The idea of outweighing has for instance been studied in the context of legal reasoning by Hage (1997) and Verheij (1996b). Outweighing defeaters can be regarded as 'multi-reason' rebutters. An interesting difference with 'single-reason' is the intuition that it can be the case that reasons that are individually outweighed by another reason, might together be stronger than the opposing reason. Assume that φ_1 and φ_2 are reasons for χ, and that ψ is a reason for not-χ that individually rebuts both φ_1 and φ_2. We then have the following:

$\varphi_1, \varphi_2, \psi$
$\varphi_1 \leadsto \chi$
$\varphi_2 \leadsto \chi$

$\psi \leadsto \text{not-}\chi$
$(\psi \leadsto \text{not-}\chi) \leadsto (\psi \leadsto \times(\varphi_1 \leadsto \chi))$
$(\psi \leadsto \text{not-}\chi) \leadsto (\psi \leadsto \times(\varphi_2 \leadsto \chi))$

That φ_1 and φ_2 together outweigh ψ can be represented as follows:

$$(\varphi_1 \leadsto \chi) \leadsto ((\varphi_2 \leadsto \chi) \leadsto (\varphi_1 \leadsto (\varphi_2 \leadsto \times(\psi \leadsto \text{not-}\chi))))$$

By this sentence, ψ cannot be a reason for not-χ when φ_1 and φ_2 are reasons for χ. On its own, the representation of outweighing is not sufficient. The set of nine sentences above has two extensions, one in which χ is justified by φ_1 and φ_2 and the other in which not-χ is justified by ψ. The cause is that ψ can still rebut φ_1 and φ_2 individually, even when they are both justified.

There are two ways of completing the representation of outweighing to resolve this. The first uses the following two sentences:

$\varphi_1 \leadsto (\varphi_2 \leadsto \times((\psi \leadsto \text{not-}\chi) \leadsto (\psi \leadsto \times(\varphi_1 \leadsto \chi))))$
$\varphi_1 \leadsto (\varphi_2 \leadsto \times((\psi \leadsto \text{not-}\chi) \leadsto (\psi \leadsto \times(\varphi_2 \leadsto \chi))))$

These sentences have the effect that ψ does no longer rebut φ_1 and φ_2 individually when φ_1 and φ_2 are both justified. A second way of completing the representation of outweighing uses the idea of *accrual of reasons*. It requires the use of the inconclusive conditional $\rightarrowtail$ discussed in section 11.2 as a representation of *prima facie* reasons. So $\varphi_1 \leadsto \chi$, $\varphi_2 \leadsto \chi$ and $\psi \leadsto \text{not-}\chi$ are replaced by $\varphi_1 \rightarrowtail \chi$, $\varphi_2 \rightarrowtail \chi$ and $\psi \rightarrowtail \text{not-}\chi$. Then the accrual of φ_1 and φ_2 as reasons for χ is represented thus:

$$(\varphi_1 \rightarrowtail \chi) \leadsto ((\varphi_2 \rightarrowtail \chi) \leadsto ((\varphi_1 \rightarrowtail (\varphi_2 \rightarrowtail \chi))))$$

The sentence expresses that if φ_1 and φ_2 are each a reason for χ, then they also form together a reason for χ. (Note that $\varphi_1 \rightarrowtail (\varphi_2 \rightarrowtail \chi)$ can be regarded as a conditional with the conjunction of φ_1 and φ_2 as antecedent.) When each individual reason is defeated, their combination need not be. The idea is that combined reasons can be stronger than the individual reasons.

The $\leadsto$-analog of the accrual sentence, viz. $(\varphi_1 \leadsto \chi) \leadsto ((\varphi_2 \leadsto \chi) \leadsto ((\varphi_1 \leadsto (\varphi_2 \leadsto \chi))))$, does not work since it only has its accruing effect when both $\varphi_1 \leadsto \chi$ and $\varphi_2 \leadsto \chi$ are not defeated, while accruing is only interesting in case $\varphi_1 \leadsto \chi$ or $\varphi_2 \leadsto \chi$ is defeated.

The idea of accrual has been adopted by Hage (1997, e.g., p. 203-204) and Verheij (1996b, e.g., p. 161-162) and contested by Pollock (1995, p. 101-102) and Prakken (1997, p. 198-200).

A remark similar to the one ending the discussion of Toulmin's scheme is in place. Pollock's defeaters (and priority and outweighing defeaters) might well be represented somewhat better in terms of the inconclusive conditional $\rightarrowtail$ discussed in section 11.2, instead of in terms of DEFLOG's defeasible conditional $\leadsto$. We saw that in the case of outweighing defeaters the idea of accrual of reasons requires $\rightarrowtail$ instead of $\leadsto$. A $\rightarrowtail$-representation is also slightly closer to Pollock's account of defeaters since Pollock's reason statements seem to be intended as inconclusive rather than as defeasible conditionals (with obtaining antecedent).

11.5 Collective and indeterministic defeat

In much work on dialectical argumentation, some general principle to preserve consistency is modeled. For instance, it can be regarded as unwanted that the consequents of a set of conditionals of which the consequents are inconsistent, all follow from the antecedents. Two straightforward principles to preserve consistency in situations like this might be called the *collective* and *indeterministic defeat* (cf. Verheij, 1996b, p. 124-5). In collective defeat, none of the consequents of the conditionals follows, while in indeterministic defeat one of the consequents does not follow. Both in collective and indeterministic defeat, the inconsistency is resolved. In indeterministic defeat, each choice of blocked consequent is allowed, each leading to a different resolution of the inconsistency.

Collective defeat is for instance built into Pollock's (1995) OSCAR and in Reason-Based Logic (Hage, 1996, 1997, Verheij, 1996b). In both cases, collective defeat is an ultimate remedy: only when other means of conflict resolution (by explicit information, like in OSCAR on the basis of rebutting or undercutting defeaters, and in Reason-Based Logic by exclusionary reasons or weighing reasons) have failed, the remaining conflict is resolved by a form of collective defeat. Pollock handles collective defeat by an additional evaluation status that he calls provisional defeat (Pollock, 1995, p. 112-4). In Reason-Based Logic the conclusions of pros and cons are blocked when neither outweigh the other (e.g., Hage, 1997, p. 163-4).

Indeterministic defeat occurs for instance in the systems of Lin (1993) and Vreeswijk (1997). In Lin's system, only one of two arguments (in the sense of derivations) with opposing conclusions can be an element of an 'argument structure', that is Lin's counterpart of extensions. Vreeswijk's system refines indeterministic defeat by the use of a conclusive force relation on arguments (also in the sense of derivations). If there is no information about the conclusive force relations between the arguments involved in a conflict (i.e., a set of arguments with inconsistent conclusions), any choice of a single argument in the conflict can resolve it by being left out of an extension. However when an argument in the conflict has stronger conclusive force than another, it cannot be chosen to resolve the conflict by being left out of an extension. In Vreeswijk's system, indeterministic defeat is the primary resort for resolving conflict, but it can be influenced by the conclusive force relation. It can for instance be the case that by the conclusive force relation only one choice is left. (See also section 13.2.)

Modeling collective and indeterministic defeat in DEFLOG requires a method that is similar to the modeling of inconclusive conditionals in section 11.2. Let $\Rightarrow$ be a connective for which the principle of collective or indeterministic defeat should apply.

The first axiom scheme that is needed expresses that a $\Rightarrow$-conditional normally implies its $\leadsto$-counterpart:

Φ: $(\varphi \Rightarrow \psi) \leadsto (\varphi \leadsto \psi)$

The second axiom scheme expresses the defeat of the first axiom scheme in case of a conflict of conditionals:

Ψ_i: $(\varphi_1 \leadsto (\varphi_2 \leadsto (\ldots \leadsto (\varphi_n \leadsto ((\varphi_1 \Rightarrow \psi_1) \leadsto ((\varphi_2 \Rightarrow \psi_2) \leadsto (\ldots \leadsto ((\varphi_n \Rightarrow \psi_n) \leadsto \times(\Phi_i))\ldots))))\ldots)))$,
where $\psi_1, \psi_2, \ldots, \psi_n$ are inconsistent (with respect to some logical standard), and Φ_i is $(\varphi_i \Rightarrow \psi_i) \leadsto (\varphi_i \leadsto \psi_i)$ for some i with $1 \leq i \leq n$.

Ψ_1, Ψ_2, ... and Ψ_n together express that the conditionals $\varphi_1 \Rightarrow \psi_1$, $\varphi_2 \Rightarrow \psi_2$, ... and $\varphi_n \Rightarrow \psi_n$ are collectively defeated in case of a conflict, in the sense that then none of the conditionals $\varphi_1 \leadsto \psi_1$, $\varphi_2 \leadsto \psi_2$, ... or $\varphi_n \leadsto \psi_n$ follows, and therefore no consequent ψ_1, ψ_2, ... or ψ_n.

Indeterministic defeat can be modeled by blocking the Ψ_j for $j \neq i$ when Ψ_i is active. This is expressed by a third axiom scheme as follows:

$\times\Phi_i \leadsto \times\Psi_j$, for i and j with $1 \leq i, j \leq n$ and $j \neq i$.

The effect of this scheme is that when Ψ_i is active (i.e., when Φ_i obtains) the other Ψ_j should be blocked. In other words, when $\varphi_i \Rightarrow \psi_i$ does not lead to $\varphi_i \leadsto \psi_i$ all others do.

Note that collective defeat leads to one extension, while indeterministic defeat leads to many (in fact one for each choice of conflict resolution).

12 Variations

In the present section, some variations of the definitions of DEFLOG are discussed.

12.1 Standard logical connectives

No attention has been paid to the standard connectives, expressing conjunction, negation and material implication. DEFLOG's connectives × and $\leadsto$ differ from the standard connectives in two important respects. First the connective $\leadsto$ is not 'truth-functional'. It is a bit awkward to speak of truth functionality here since in DEFLOG justification statuses play the role of truth values.[12] Put more accurately, the justification value of a composite sentence $\varphi \leadsto \psi$ in an interpretation is not in general a function of the

[12] It should be noted however that speaking of truth values or of justification statuses is merely a matter of the use of different labels, especially in order to avoid the multitude of connotations that are related to the label 'truth values'.

justification values of the sentences φ and ψ in the interpretation. For instance, if φ and ψ are both justified in an interpretation, then $\varphi \leadsto \psi$ can be justified, defeated or unevaluated. Second the semantics of the connectives $\times$ and $\leadsto$ is partial. To be precise, sentences of the form $\times\varphi$ or $\varphi \leadsto \psi$ are not necessarily evaluated in an interpretation, not even if the sentences φ and ψ are.

This raises the question whether DefLog can incorporate the standard connectives. The answer is yes, and there are at least three interesting ways to do this.

A first, conservative way to incorporate the standard connectives is to add new connectives to DefLog's language, and axiomatize their intended meaning in terms of a set of assumptions. Let $\neg$ and $\rightarrow$ be connectives that are intended to express standard negation and material implication. (Recall that these two are functionally complete for the truth-functional connectives.) Consider the set of sentences T_{prop} consisting of all sentences of the following forms:

$\varphi \rightarrow (\psi \rightarrow \varphi)$
$(\varphi \rightarrow (\psi \rightarrow \chi)) \rightarrow ((\varphi \rightarrow \psi) \rightarrow (\varphi \rightarrow \chi))$
$(\neg\varphi \rightarrow \neg\psi) \rightarrow (\psi \rightarrow \varphi)$

$(\varphi \rightarrow \psi) \leadsto (\varphi \leadsto \psi)$

The first three schemes are familiar from Hilbert-style versions of standard proof theory, in which there are some axiom schemes and only one rule of inference, viz. *Modus ponens*. The fourth scheme links the standard connective $\rightarrow$ with DefLog's $\leadsto$. Its role is to validate *Modus ponens* for the connective $\rightarrow$ using the fact that it is already valid for $\leadsto$.[13]

The following property holds.

Property (12.1)

If $S \cup \{\varphi\}$ is a set of sentences that only contain the connectives $\neg$ and $\rightarrow$, then $S \cup T_{prop} \vDash_{\text{DefLog}} \varphi$ if and only if $S \vdash_{\text{Hilbert}} \varphi$.

In the second way to incorporate the standard connectives, DefLog's connectives $\times$ and $\leadsto$ are used to express standard negation and material implication, respectively. Consider the set of sentences T^*_{prop} consisting of all sentences of the following forms:

$\varphi \leadsto (\psi \leadsto \varphi)$
$(\varphi \leadsto (\psi \leadsto \chi)) \leadsto ((\varphi \leadsto \psi) \leadsto (\varphi \leadsto \chi))$
$(\times\varphi \leadsto \times\psi) \leadsto (\psi \leadsto \varphi)$

If one reads $\times$ as standard negation and $\leadsto$ as material implication, the three schemes are again those familiar from Hilbert-style versions of proof theory. The following property follows from the soundness and completeness (for standard logic) of the Hilbert-style proof theory, and from the fact that in DefLog $\leadsto$ validates *Modus ponens* (i.e., in any interpretation in which $\varphi \leadsto \psi$ and φ are justified, also ψ is justified). The property is stronger than property (12.1) above.

Let W be a DefLog interpretation. Then the following are equivalent:

1. W is total and a model of T_{prop}.
2. W is an interpretation with the following two properties:
   a. For any sentence φ, $W(\times\varphi) = j$ if and only if $W(\varphi) \neq j$.
   b. For all sentences φ and ψ, $W(\varphi \leadsto \psi) = j$ if and only if $W(\varphi) \neq j$ or $W(\psi) = j$.

Proof: That part of property 2 follows from property 1 is seen by noting that the following are all equivalent. (i) $W(\times\varphi) = j$. (ii) $\times\varphi \in J(W)$. (iii) $J(W) \vDash_{\text{DefLog}} \times\varphi$. (iv) $J(W) \vdash_{\text{Hilbert}} \times\varphi$. (v) It does not hold that $J(W) \vdash_{\text{Hilbert}} \varphi$. (vi) It does not hold that $J(W) \vDash_{\text{DefLog}} \varphi$. (vii) $\neg\varphi \notin J(W)$. (viii) $W(\varphi) \neq j$. The totality of the interpretation is used in

[13] The notions of validating and validity are here used in the standard sense: *Modus ponens* is valid for a conditional $\Rightarrow$ if the truth (or justifiedness or other positive evaluation) of sentences φ and $\varphi \Rightarrow \psi$ in some interpretation (or possible world or other semantic whole) implies the truth (or the justifiedness or the positive evaluation, respectively) of the sentence ψ in that interpretation (or that possible world or that semantic whole, respectively).

the equivalence of (iv) and (v). Part b follows similarly. An interpretation W that obeys part a of property 2 is total[14] since then if φ is not justified in W, it follows that $\times\varphi$ is justified, and therefore that φ is defeated in W.

The constraints a and b on interpretations are the analogues of the standard constraints on interpretations for standard negation and material implication.

Let's call interpretations that obey the equivalent conditions above *standard interpretations*. For standard interpretations, the following 'triviality result' obtains.

Property (12.2)

Let Δ be a set of sentences and W a standard interpretation. Then W is an extension of Δ if and only if W is a satisfiability class of Δ.

This triviality result is perhaps the primary reason why in standard logic the notions of extension and dialectical justification do not arise. It collapses into the notion of maximal consistency (the standard counterpart of DefLog's satisfiability classes).

Note that the result also gives a connection between DefLog and the maximal consistent set approach to defeasible reasoning, as it has been proposed by, e.g., Rescher (1964) and Poole (1988).

The triviality result follows from the following slightly more general theorem (cf. a similar result by Bondarenko *et al.*, 1997).

Theorem (12.3)

Let Δ be a set of sentences, such that, for any sentence φ and any subset S of Δ, it obtains that $S \cup \{\varphi\}$ is not satisfiable if and only if $S \vDash \times\varphi$. Then the extensions of Δ coincide with the satisfiability classes of Δ.

Proof: It suffices to show that each satisfiability class of Δ is an extension. Let SC be one of Δ's satisfiability classes, and let C be the maximal subset of Δ specifying SC. Consider a sentence φ in Δ that is not justified in SC, i.e., φ is not in Mp(C). Then $C \cup \{\varphi\}$ is not satisfiable, since C is a maximal satisfiable subset of Δ (property (5.12)). Hence, by the assumption of the theory, $C \vDash \times\varphi$.

The third way to incorporate the standard connectives is by extending DefLog's language and adding standard constraints that must obey in an interpretation. If the connectives ¬ and → are added to the language in order to express standard negation and material implication, the following constraints would have to be added to definition (3.3), in which DefLog's interpretations are defined:

3. For any sentence φ, $W(\neg\varphi) = j$ if and only if $W(\varphi) \neq j$.
4. For all sentences φ and ψ, $W(\varphi \to \psi) = j$ if and only if $W(\varphi) \neq j$ or $W(\psi) = j$.

Note that interpretations obeying these constraints are total with respect to standard negation ¬: for any sentence φ, either φ or $\neg\varphi$ is justified in an interpretation. (Here it is assumed that the background negation is standard: either φ is justified in an interpretation or it is not.) Interpretations are still not total with respect to 'dialectical negation' ×, since it can be the case that φ and $\times\varphi$ are both not justified.

Two new schemes of tautologies that obtain in interpretations obeying the additional constraints 3 and 4 above are the following:

$$\times\varphi \to \neg\varphi$$
$$(\varphi \rightsquigarrow \psi) \to (\varphi \to \psi)$$

Note that in these tautology schemes the occurrences of → cannot all be replaced by $\rightsquigarrow$, without making the scheme contingent. For example, the instances of the schemes $\times\varphi \rightsquigarrow \neg\varphi$ and $(\varphi \rightsquigarrow \psi) \rightsquigarrow (\varphi \rightsquigarrow \psi)$ are not justified in all interpretations obeying 3 and 4. Note that the scheme $(\varphi \to \psi) \rightsquigarrow (\varphi \rightsquigarrow \psi)$ that was used in the first way of incorporating the standard connectives is not tautologous.

[14] Recall that a total interpretation is an interpretation with the whole language as its extent, i.e., an interpretation in which any sentence of the language is either justified or defeated.

12.2 Symmetric defeat and symmetric dialectical justification

An obvious variation of the definitions of DEFLOG is *symmetric DEFLOG*. In symmetric DEFLOG, it does not only follow from the justifiedness of a sentence $\times\varphi$ that the sentence φ is defeated as in 'ordinary' DEFLOG, but it also follows from the justifiedness of φ that the sentence $\times\varphi$ is defeated. The latter does not obtain in ordinary DEFLOG.

An additional constraint on interpretations (definition (3.3)) is all that is needed. For any sentence φ it must hold that:

3. $W(\varphi) = j$ if and only if $W(\times\varphi) = d$.

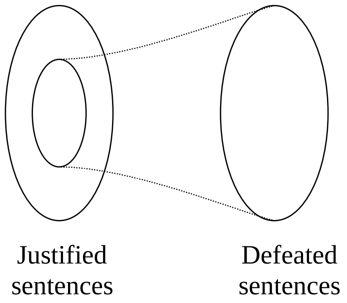
Note the subtle difference with the other constraint about defeat sentences:

1. $W(\times\varphi)$ is equal to j if and only if $W(\varphi)$ is equal to d.

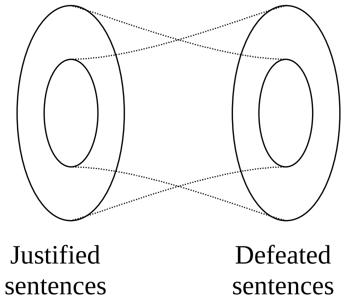
By constraint 1, the justifiedness of $\times\varphi$ expresses the defeat of φ. By constraint 3, the justifiedness of φ expresses the defeat of $\times\varphi$. With only constraint 1, the set of defeated sentences of an interpretation is faithfully mirrored in the set of justified sentences: each defeated sentence φ corresponds to the justified sentence $\times\varphi$. After adding constraint 3, the set of defeated sentences of an interpretation also contains a faithful mirror image of the set of justified sentences: each justified sentence φ corresponds to a defeated sentence $\times\varphi$. From constraints 1 and 3 it follows that a sentence φ is justified in an interpretation if and only if the sentence $\times\times\varphi$ is justified, and that a sentence φ is defeated in an interpretation if and only if the sentence $\times\times\varphi$ is defeated.

The figure below illustrates the situation. The large ovals represent the sets of justified and defeated sentences of an interpretation. The small ovals are the corresponding subsets of sentences of the form $\times\varphi$.

Ordinary DEFLOG



Symmetric DEFLOG



Let J and D denote the set of justified and defeated sentences in an interpretation, respectively. Then in ordinary DEFLOG it holds that $\times D$ is a subset of J. In symmetric DEFLOG, it also holds that $\times J$ is a subset of D. This leads to two chains of inclusions holding in symmetric DEFLOG:

$J \supseteq \times D \supseteq \times\times J \supseteq \times\times\times D \supseteq \ldots$
$D \supseteq \times J \supseteq \times\times D \supseteq \times\times\times J \supseteq \ldots$

The theory of ordinary DEFLOG developed in the previous sections can be naturally adapted for symmetric DEFLOG. Again satisfiable sets of sentences T have a unique model W_T with a minimal set of justified sentences. As a result satisfiable sets of sentences specify an interpretation W_T, and definition (4.3) of extensions still makes sense in the context of symmetric DEFLOG. The sets of justified sentences of interpretations can - in analogy with property (3.11) - be characterized as the sets that are conflict-free, closed under *Modus ponens* and closed with respect to double dialectical negation, i.e., φ is in the set if and only if $\times\times\varphi$ is in it. The set of consequences Cn(T) of a set of sentences T is the closure of T under the rules of inference $\varphi, \varphi \leadsto \psi \,/\, \psi$ (*Modus ponens*), $\varphi, \times\varphi \,/\, \psi$ (a variant of *Ex falso quodlibet*), $\times\times\varphi \,/\, \varphi$ and $\varphi \,/\, \times\times\varphi$.

A natural alternative to definition (4.3) of extensions for symmetric DEFLOG is the following:

Definition (12.4): symmetric extensions in symmetric DefLog

If Δ is a set of sentences and E an interpretation, then E is a *symmetric extension of the theory* Δ if and only if E is an interpretation that is specified by a Δ-argument J, such that, for any φ in $\Delta \setminus J$, an opposite of φ is a consequence of J.

Recall definition (10.8) of the opposites of a sentence: the opposites of a sentence φ are $\times\varphi$ and $\times^{-1}\varphi$, if it exists.

There are more theories with symmetric extensions in symmetric DefLog than in ordinary DefLog. For instance, the theory $\{p, p \leadsto q, \times q\}$ does not have an ordinary DefLog extension, but has a symmetric DefLog extension, viz. the (symmetric) interpretation in which q is justified and $\times q$ defeated. However, also in symmetric DefLog there are theories without extensions, e.g., the theory $\{p, p \rtimes p\}$.

That there are more extensions in symmetric DefLog is the result of the fact that more arguments serve as attacks of other arguments, in the sense that they can resolve incompatibilities. In ordinary DefLog, an argument C attacks another argument C' if the defeat of a sentence in C' follows from C. In symmetric DefLog, it makes sense to define that an argument C attacks another argument C' if the opposite of a sentence in C' follows from C.

The central theorems of ordinary DefLog seem to have analogues for symmetric DefLog. I have not discovered interesting properties holding for symmetric DefLog, but not for ordinary DefLog.

12.3 Deep and shallow attack

Assume that the sentence φ is justified in an extension of a theory Δ and that there is a non-trivial *Modus ponens* derivation of φ with premises in Δ, i.e., a derivation with at least one instance of *Modus ponens*. Then, in order to defeat φ, it does in general not suffice to add an attack χ to the theory. Often the theory $\Delta \cup \{\chi, \chi \rtimes \varphi\}$ (where χ is a sentence that is not an element of Δ, nor is a subsentence of an element of Δ) has no extension. A simple example shows this. The theory $\{p, q, q \leadsto p\}$ has a unique extension, while the theory $\{p, q, q \leadsto p, r, r \rtimes p\}$ has no extension. The reason why this happens is that blocking the derivation $q, q \leadsto p \,/\, p$ of p requires that one of its premises is defeated, viz. q or $q \leadsto p$. If only its conclusion p is attacked, an incompatibility remains.

In order to defeat φ, no reason ψ for φ can be justifying. In other words, if there is a justified reason ψ, the conditional $\psi \leadsto \varphi$ expressing its connection with φ should be attacked as well. However, if there is a non-trivial derivation of $\psi \leadsto \varphi$, the argument above can be repeated: it can be the case that the theory $\Delta \cup \{\chi, \chi \rtimes \varphi\} \cup \{\chi \rtimes (\psi \leadsto \varphi) \mid \Delta \vDash \psi \leadsto \varphi\}$ has no extension. An example is the theory $\{p, q_0, q_1, q_1 \leadsto (q_0 \leadsto p), r, r \rtimes p, r \rtimes (q_0 \leadsto p)\}$ that has no extension since the derivation $q_1, q_1 \leadsto (q_0 \leadsto p) \,/\, q_0 \leadsto p$ is not blocked.

The point can be repeated for conditionals with φ as its deep consequent. Here the *deep consequent* of a conditional is defined as follows. If $\varphi \leadsto \psi$ is a conditional for which ψ is not a conditional then the deep consequent of $\varphi \leadsto \psi$ is ψ. If $\varphi \leadsto \psi$ is a conditional for which ψ is a conditional $\chi \leadsto \omega$ then the deep consequent of $\varphi \leadsto \psi$ is the deep consequent of $\chi \leadsto \omega$. For instance, the deep consequents of $p \leadsto q$, $p \leadsto (q \leadsto r)$, $(p \leadsto q) \leadsto (r \rtimes (s \leadsto t))$ and are q, r and $\times(s \leadsto t)$, respectively.

This suggests a distinction between deep and shallow attack of a statement. A *shallow attack* of a statement φ in a theory Δ is then a statement χ in Δ for which also $\chi \rtimes \varphi$ is in Δ. A *deep attack* of a statement φ in a theory Δ consists of an attack of the statement, but also of each conditional with φ as its consequent or as its deep consequent.

Consider the following derivation of p with premises in the theory $\{q_0, q_1, q_2, q_2 \leadsto (q_1 \leadsto (q_0 \leadsto p))\}$. The leftmost sentences of the derivation express the statements attacked by a deep attack of p.

$$\dfrac{\dfrac{\dfrac{q_2 \leadsto (q_1 \leadsto (q_0 \leadsto p)) \qquad q_2}{q_1 \leadsto (q_0 \leadsto p)} \qquad q_1}{q_0 \leadsto p} \qquad q_0}{p}$$

For a statement r to be a deep attack of p it is required that all of the following are included in the theory:

$r \rtimes p$

$r \rtimes (q_0 \leadsto p)$
$r \rtimes (q_1 \leadsto (q_0 \leadsto p))$
$r \rtimes (q_2 \leadsto (q_1 \leadsto (q_0 \leadsto p)))$

Note that for blocking the derivation shown r and $r \rtimes (q_2 \leadsto (q_1 \leadsto (q_0 \leadsto p)))$ would suffice. For any individual derivation, attacking a premise would suffice.

However deep attack of a statement abstracts from individual derivations. It is a stronger type of attack with its own 'semantics'. Deep attack of a statement involves not only the attack of the statement itself, but of any possible 'backbone' of a derivation of the statement. Here the backbone of a *Modus ponens* derivation (with all majors in *Modus ponens* instances on the left as in the figure above) consists of the string of leftmost sentences that directly leads to the derivations conclusion. Formally, the backbone of a *Modus ponens* derivation is defined as follows. A trivial derivation φ has φ as its backbone. If $\Pi(\varphi \leadsto \psi)$ and $\Pi(\varphi)$ are derivations, then the derivation $\Pi(\varphi \leadsto \psi)$, $\Pi(\varphi)$ / ψ has the backbone of $\Pi(\varphi \leadsto \psi)$ extended with $\varphi \leadsto \psi$ as its backbone.

Deep attack can be thought of as a long - actually infinite - conjunction. If deep attack is expressed using the connective $\rtimes\rtimes$, the following suggests a definition:

$\varphi \rtimes\rtimes \psi := \varphi \rtimes \psi$, plus for any χ, $\varphi \rtimes (\chi \leadsto \psi)$, plus for any χ, χ', $\varphi \rtimes (\chi' \leadsto (\chi \leadsto \psi))$, plus ...

Each of the following theories has an extension:

$p, q, q \leadsto p, r, r \rtimes\rtimes p$
$p, q_0, q_1, q_1 \leadsto (q_0 \leadsto p), r, r \rtimes\rtimes p$
$p, q_0, q_1, q_2, q_2 \leadsto (q_1 \leadsto (q_0 \leadsto p)), r, r \rtimes\rtimes p$

It should be noted that deep attack does not guarantee the existence of an extension, as adding the sentence $\times p \rtimes\rtimes \times p$ to any of the theories above shows.

Both DefLog's ordinary, 'shallow' attack, and deep attack as discussed here, have useful characteristics. DefLog's shallow attack shows that attacking a statement does not suffice in order to defeat the statement, but that any of its derivations need be blocked. Deep attack is a tool to make it possible to reach the defeat of a statement by adding one attack without bothering about all possible derivations. DefLog's shallow attack has a simpler semantics than deep attack, in the sense that the semantics of deep attack is expressible in that of DefLog's shallow attack (as the infinite conjunction above suggests).

12.4 Admissibility, naïve dialectical justification and dialectical justification: a meta-analysis

An argument is dialectically justifying if it attacks any argument incompatible with it (cf. definition (6.3)). In the literature, a variant of dialectical justification occurs that goes by the name of admissibility, according to which an argument is admissible if it attacks any argument attacking it (cf. Dung, 1995, Bondarenko *et al.*, 1997, see also section 13.3). Obviously, in DefLog, dialectically justifying arguments are also admissible, while there are admissible arguments that are not dialectically justifying.[15]

Another variant of dialectical justification (in fact one that I at first thought to be the central notion) is naïve dialectical justification: an argument is *naïvely dialectically justifying* if the argument attacks any argument attacking the argument itself or one of its consequences. (Cf. also the naïve dialectical arguments of section 2. See the discussion after theorem (10.16).) Clearly, dialectically justifying arguments are also naïvely dialectically justifying, and naïvely dialectically justifying arguments are admissible.

Why has in DefLog dialectical justification been chosen instead of admissibility and naïve dialectical justification, that to some may seem more natural and at least simpler? The reason is that dialectical

[15] Admissibility in DefLog depends of course on its notions of argument and attack. The results in section 13.3 show that DefLog's admissibility are indeed an extrapolation of Dung's admissibility. In fact DefLog's dialectical justification is too, since as will be seen admissibility and dialectical justification coincide on Dung's restricted language. Though extensionally the definitions coincide on Dung's restricted language, the intensional difference remains: in admissibility, only attacks must be countered by attacks, while in dialectical justification all incompatibles must be countered by attacks. The intensional difference can however only be extensionally appreciated on DefLog's richer language, as is especially shown in the present section.

justification has properties that make it especially suitable for the analysis of extensions. In this section, some of these properties are discussed. By a meta-analysis, it is shown how the properties operate in some of DefLog's central theorems.

Among the useful properties of dialectical justification are the following:

Union

If C and C' are compatible dialectically justifying arguments, then also $C \cup C'$ is dialectically justifying. (Similarly, for any compatible collection of dialectically justifying arguments: the union of a compatible collection of dialectically justifying arguments is again dialectically justifying.)

Localization

Let E be an extension of a theory Δ. Then there is a collection $\{C_i\}_{i \in I}$ of dialectically justifying arguments that covers J(E), i.e., J(E) is equal to $\cup_{i \in I} C_i$.

Separation

If C and C' are incompatible dialectically justifying arguments, then there are opposites φ and $\times\varphi$, such that $C \vDash \varphi$ and $C' \vDash \times\varphi$, or such that $C \vDash \times\varphi$ and $C' \vDash \varphi$. (Similarly, for any incompatible collection of dialectically justifying arguments: given an incompatible collection of dialectically justifying arguments, there are opposites that are the consequence of the unions of compatible subcollections.)

The union and separation properties were stated earlier as the corollaries (6.14) and (6.15). The localization property is an immediate consequence of part (i) of corollary (6.8).

For our meta-analysis, these properties are generalized from dialectical justification to a general property of arguments Φ. Let $\Phi_\Delta(C)$, abbreviated $\Phi(C)$, express that the argument C has the property Φ with respect to a theory Δ. An argument C is a Φ-*argument* of a theory Δ if $\Phi_\Delta(C)$. Then the properties can be thus paraphrased:

Union

If C and C' are compatible Φ-arguments, then $C \cup C'$ is also a Φ-argument. (Similarly, for any compatible collection of Φ-arguments: the union of a compatible collection of Φ-arguments is again a Φ-argument.)

Localization

Let E be an extension of a theory Δ. Then there is a collection $\{C_i\}_{i \in I}$ of Φ-arguments that covers J(E), i.e., J(E) is equal to $\cup_{i \in I} C_i$.

Separation

If C and C' are incompatible Φ-arguments, then there are opposites φ and $\times\varphi$, such that $C \vDash \varphi$ and $C' \vDash \times\varphi$, or such that $C \vDash \times\varphi$ and $C' \vDash \varphi$. (Similarly, for any incompatible collection of Φ-arguments: given an incompatible collection of Φ-arguments, there are opposites that are the consequence of the unions of compatible subcollections.)

It is not hard to see that admissibility has the localization and union properties, but not the separation property, while naïve dialectical justification has the localization property, but lacks both union and separation.

For instance, the theory $\{p_1, p_1 \leadsto q, p_2, p_2 \rtimes q\}$ shows that naïve dialectical justification does not have the union property. The arguments $\{p_1\}$ and $\{p_1 \leadsto q\}$ are two compatible, naïvely dialectically justifying arguments with respect to the theory, while their union $\{p_1, p_1 \leadsto q\}$ is not, since it does not attack the argument $\{p_2, p_2 \rtimes q\}$ that attacks its consequence q.

That neither for admissibility nor for naïve dialectical justification, the separation property obtains, can be seen by inspection of the theory $\{p_1, p_1 \leadsto q, p_2, p_2 \leadsto (q \rtimes q)\}$. With respect to the theory, there are four maximal admissible arguments, viz. each three-element subset of the theory. These are also the maximal naïvely dialectically justifying arguments. (Note that each argument of the theory is admissible and naïvely dialectically justifying since there are no attacks.) Any pair of these arguments is incompatible, yet there is no sentence that is defeated by an argument, let alone by an admissible or naïvely dialectically justifying argument, as is required by the separation property.

The proofs of the localization property for admissibility and naïve dialectical justification are straightforward. The proof of the union property for admissibility is almost trivial (in contrast with the proof of its dialectical justification analogue) since any attack of the union of a collection of arguments is also an attack of one of the arguments in the collection.

Note also that the 'empty' property of arguments, viz. satisfiability or 'argumenthood', that any argument satisfies, has the localization and union properties. Localization follows from the satisfiability of the set of justified sentences of an extension, and the union property is for satisfiability trivial. The incompatible arguments $\{p\}$ and $\{p \rtimes p\}$ show that satisfiability lacks the separation property.

In order to show the use of the properties, several of DEFLOG's definitions are generalized from dialectical justification to a general property of arguments Φ, as follows.

Definition (12.5)

(i) A stage S is a Φ-*stage* if it is specified by a Φ-argument.

(ii) A stage S is Φ-*preferred* with respect to a theory Δ if $J(S) \cap \Delta$ is maximal among the theory's Φ-arguments.

Definition (12.6)

(i) A sentence φ is Φ-*justifiable* with respect to a theory Δ if it is a consequence of a Φ-argument of the theory, and Φ-*defeasible* if $\times\varphi$ is a consequence of a Φ-argument. A sentence is Φ-*interpretable* with respect to a theory Δ if it is Φ-justifiable or Φ-defeasible with respect to the theory. A sentence is Φ-*ambiguous* with respect to a theory Δ if it is Φ-justifiable and Φ-defeasible with respect to the theory.

(ii) Let C be an argument. A sentence φ is Φ-*justifiable in the context* C with respect to a theory Δ if it is a consequence of a Φ-argument of the theory that contains C, and Φ-*defeasible in the context* C if $\times\varphi$ is a consequence of a Φ-argument that contains C. A sentence is Φ-*interpretable in the context* C with respect to a theory Δ if it is Φ-justifiable or Φ-defeasible in the context C with respect to the theory. A sentence is Φ-*ambiguous in the context* C with respect to a theory Δ if it is Φ-justifiable and Φ-defeasible in the context C with respect to the theory.

Definition (12.7)

An argument C is Φ-*disambiguating* with respect to a theory Δ if there is no sentence that is Φ-ambiguous in the context C with respect to the theory.

The following theorem holds.

Theorem (12.8)

(i) Let Δ be a set of sentences and let Φ have the union property. Then the following hold:
  a. Any pair of Φ-preferred stages of the theory Δ is incompatible.
  b. If there is a Φ-ambiguous sentence with respect to the theory Δ, then there are at least two Φ-preferred stages.

(ii) Let Δ be a set of sentences and let Φ have the union and the localization property. Then the following hold:
  a. If E is an extension, then J(E) is a Φ-argument.
  b. There is no extension of the theory Δ if for any Φ-disambiguating context C there is a sentence φ in Δ that is not Φ-interpretable in the context C with respect to Δ.

(iii) Let Δ be a set of sentences and let Φ have the union and the separation property. Then the following hold:
  a. There is a Φ-ambiguous sentence with respect to the theory Δ if there are at least two Φ-preferred stages.
  b. If there is no extension of the theory Δ, then for any Φ-disambiguating context C there is a sentence φ in Δ that is not Φ-interpretable in the context C with respect to Δ.

Proof: (i) a. If P_1 and P_2 are preferred and compatible, their union U is a Φ-stage with $J(U) \cap \Delta \supseteq (J(P_1) \cap \Delta) \cup (J(P_2) \cap \Delta)$. (i) b. Apply Zorn's lemma to the Φ-arguments that are compatible with a Φ-justification of the Φ-ambiguous sentence, and to those compatible with a Φ-justification of its opposite. (ii) a. By localization, J(E) is the union of Φ-arguments, and therefore by union itself a Φ-argument. (ii) b. Let E be an extension. Then, by (ii) a., J(E) is a Φ-argument. J(E) is disambiguating, and any sentence in Δ is Φ-interpretable in the context C. (iii) a. Let P_1 and P_2 be different preferred stages. By (i) a. (and the union property), they are incompatible. So by the separation property, there are opposites φ and ψ, such that $J(P_1) \cap \Delta \vDash \varphi$ and $J(P_2) \cap \Delta \vDash \psi$. As a result, φ or ψ is a Φ-ambiguous sentence with respect to the theory Δ. (iii) b. Let C be Φ-disambiguating and let for any sentence φ in Δ C_φ be a Φ-justification of φ or of $\times\varphi$ that contains C. The collection of the C_φ is compatible, since otherwise there would (by the separation property) be a Φ-ambiguous sentence in the context C. The union of the C_φ is a φ-argument (by the union property) and specifies an extension of the theory Δ.

Since dialectical justification has the union, the separation and the localization properties, all parts of the theorem can be instantiated for dialectical justification. The instantiations for dialectical justification of all parts of the theorem have been proven earlier. Part (i) a. is theorem (7.7) and (i) b. is the 'if'-part of theorem (9.2). Part (ii) a. occurs in corollary (6.8), and (ii) b. is the 'if'-part of theorem (9.6). Part (iii) a. is the 'only if'-part of theorem (9.2), and (iii) b. the 'only if'-part of theorem (9.6).

Since admissibility has the union and the localization property, the parts (i) and (ii) of the theorem hold for admissibility. Since naïve dialectical justification only has the localization property, no part of the theorem is relevant for naïve dialectical justification. It is not hard to find counterexamples against part (iii) for admissibility and against all parts for naïve dialectical justification.

For a property Φ with the union, the localization and the separation property, like dialectical justification, the theorem can be summarized as follows.

Corollary (12.9)

Let Δ be a set of sentences and let Φ have the union, the separation and the localization property. Then the following hold:

(i) Any pair of Φ-preferred stages of the theory Δ is incompatible.

(ii) There is a Φ-ambiguous sentence with respect to the theory Δ if and only if there are at least two Φ-preferred stages.

(iii) There is no extension of the theory Δ if and only if for any Φ-disambiguating context C there is a sentence φ in Δ that is not Φ-interpretable in the context C with respect to Δ.

(iv) Let *n* be a natural (or cardinal) number. A theory Δ has exactly *n* extensions if and only if *n* is equal to the maximal number of mutually incompatible Φ-disambiguating arguments C, in the context of which any sentence in Δ is Φ-interpretable with respect to Δ.

Except for part (i) of the corollary (that only depends on the union property), which obtains for admissibility, no part of the corollary obtains for admissibility or naïve dialectical justification. All parts of the corollary obtain for dialectical justification, as was shown earlier.

I know of one other property of arguments than dialectical justification that has the union, the separation and the localization property. It is *weak dialectical justification*. An argument is weakly dialectically justifying if it attacks a consequence of any argument incompatible with it. The difference with dialectical justification is that the attack of a consequence of an incompatible argument suffices instead of an attack of the argument itself. For weak dialectical justification, all results of the theorem and the corollary obtain. My preference for the notion of dialectical justification stems from its property of *separation at the base*:

Separation at the base

If C and C' are incompatible Φ-arguments, then there is a sentence φ in $C \cup C'$, such that $C \vDash \times\varphi$ or $C' \vDash \times\varphi$. (Similarly, for any incompatible collection of Φ-arguments: given an incompatible collection of Φ-arguments, there is a sentence in the union of the collection that is attacked by the union of a compatible subcollection.)

Dialectical justification has the property of separation at the base, while naïve dialectical justification does not have it. By the property, Φ-ambiguity becomes an ambiguous Φ-interpretability of a sentence in a theory, and not merely of one of its consequences. To me, the former seems to be most appropriate. Note that separation at the base follows directly from the definition of dialectical justification. I do not know of other properties of arguments than dialectical justification with the properties of union, separation at the base and localization. Still there does not seem to be a direct proof that dialectical justification is the only such property of arguments.

13 Related research

In the following, research related to DefLog is discussed. Since terminology is not at all standard throughout the literature, it will sometimes be the case that a term as it is used by another author has a different meaning than it has in DefLog.

13.1 Reiter's logic for default reasoning

An important and still influential logical model of defeasible reasoning is Reiter's (1980) logic for default reasoning. The following is a restatement of Reiter's definition of extension. For any set S of first-order sentences (i.e., closed first-order formulas), $Th_{fo}(S)$ denotes the set of sentences that are first-order provable from S. Let L_{fo} denote the set of first-order sentences

Definition (13.10): Reiter's logic for default reasoning

(i) A *default* is an expression of the form $\alpha : M\beta_1, ..., M\beta_m / \gamma$, where $\alpha, \beta_1, ..., \beta_m$, and γ are first-order sentences.

(ii) A *default theory* is a pair (D, W), where D is a set of defaults and W a set of first-order sentences.

(iii) Let (D, W) be a default theory. For any subset S of L_{fo}, define $\Gamma(S)$ as the smallest set Γ of first-order sentences satisfying the following three properties:

D1. $W \subseteq \Gamma$

D2. $Th_{fo}(\Gamma) = \Gamma$

D3. If $\alpha : M\beta_1, ..., M\beta_m / \gamma \in D$ and $\alpha \in \Gamma$, and $\neg\beta_1, ..., \neg\beta_m \notin S$ then $\gamma \in \Gamma$.

(iv) A set of first-order sentences E is an *extension* for (D, W) if and only if $\Gamma(E) = E$.

Let now L_{DefLog} denote the language of DefLog that uses the first-order sentences as sentence constants. Let the DefLog translation of a default $\alpha : M\beta_1, ..., M\beta_m / \gamma$ be the set of m+1 sentences $\alpha \rightsquigarrow \gamma$, $\neg\beta_1 \ltimes (\alpha \rightsquigarrow \gamma)$, ... and $\neg\beta_m \ltimes (\alpha \rightsquigarrow \gamma)$, and let the DefLog translation D* of a set of defaults D be equal to the union of all the translations of the defaults in D. Let T_{fo} be the set of DefLog sentences $\{\varphi_1 \rightsquigarrow (\varphi_2 \rightsquigarrow (...(\varphi_n \rightsquigarrow \psi)...)) \mid \varphi_1, \varphi_2, ..., \varphi_n \vdash_{fo} \psi$ with $\varphi_1, \varphi_2, ..., \varphi_n, \psi \in L_{fo}\}$, where $\vdash_{fo}$ denotes first-order consequence. The following proposition establishes a formal connection between Reiter's logic for default reasoning and DefLog.

Proposition (13.11)

E is a Reiter extension of (D, W) if and only if $E = Th_{fo}(J(E^*) \cap L_{fo})$ for some DefLog extension E* of the theory $T_{fo} \cup W \cup D^*$ with $T_{fo} \cup W \subseteq J(E^*)$.

Proof: Let E be a Reiter extension of (D, W). Let J* be equal to $T_{fo} \cup E \cup \{\alpha \rightsquigarrow \gamma \mid \alpha : M\beta_1, ..., M\beta_m / \gamma \in D, \neg\beta_1 \notin E, ..., \neg\beta_m \notin E\} \cup \{\neg\beta_i \ltimes (\alpha \rightsquigarrow \gamma) \mid \alpha : M\beta_1, ..., M\beta_m / \gamma \in D\}$. J* is DefLog-satisfiable since for no default $\alpha : M\beta_1, ..., M\beta_m / \gamma \in D$ both $\alpha \rightsquigarrow \gamma$ is in J* and there is an i such that $\neg\beta_i$ is in J*. J* contains all sentences in $T_{fo} \cup W \cup D^*$ except the $\alpha \rightsquigarrow \gamma$ for which there is an i for which $\neg\beta_i$ is in J*. But for such $\alpha \rightsquigarrow \gamma$, J* contains $\times(\alpha \rightsquigarrow \gamma)$ since J* then contains $\neg\beta_i$ and $\neg\beta_i \ltimes (\alpha \rightsquigarrow \gamma)$ for some i. So J* specifies a DefLog extension E* of the theory $T_{fo} \cup W \cup D^*$. Since $J(E^*) = Mp(J^*)$, it follows that $T_{fo} \cup W \subseteq J(E^*)$ and that $E = Th_{fo}(J(E^*) \cap L_{fo})$.

Let E* be a DefLog extension of the theory $T_{fo} \cup W \cup D^*$ with $T_{fo} \cup W \subseteq J(E^*)$. It needs to be shown that $\Gamma(E) = E$, where $E = Th_{fo}(J(E^*) \cap L_{fo})$. In order to show that $\Gamma(E) \subseteq E$, it suffices to check that E satisfies the properties D1, D2 and D3 (with E in the places of both S and Γ). For D3, note that if, for some default $\alpha : M\beta_1, ..., M\beta_m / \gamma \in D$, it holds that α is in E and $\neg\beta_1, ..., \neg\beta_m \notin E$, then α and $\alpha \rightsquigarrow \gamma$ are both justified in E*, and therefore γ is in J(E*). For $\Gamma(E) \supseteq E$, note that any φ in $E \subseteq J(E^*) \subseteq Mp(T_{fo} \cup W \cup D^*)$ is a DefLog consequence of a minimal argument $C \subseteq J(E^*)$ consisting of sentences from T_{fo}, W and D*. Since $C \cap D^*$ contains only sentences of the form $\alpha \rightsquigarrow \gamma$ for a default $\alpha : M\beta_1, ..., M\beta_m / \gamma \in D$, for which there is no $\neg\beta_i$ in $J(E^*) \supseteq E$, it then follows that φ is in $\Gamma(E)$.

A formal connection between Reiter's logic for default reasoning and argument defeat has also been shown by Dung (1995).

13.2 Vreeswijk's abstract argumentation systems

In Vreeswijk's (1993, 1997) abstract argumentation systems, the defeat of arguments as derivations is studied. Vreeswijk's arguments are constructed from given sets of strict and defeasible rules of inference. In case a contradiction can be derived, one of the arguments involved in the derivation is considered to be defeated. As a result, the conflict is resolved. The selection of the defeated argument among all arguments involved in the conflict is, guided by a given conclusive force relation between arguments. If no defeasible argument involved in the conflict has stronger conclusive force than any of the others, then each can be selected as defeated. If one has stronger conclusive force than another, it cannot be selected as the defeated argument. If more than one argument can be selected as the defeated argument, each choice gives rise to a separate extension. Assume for instance that the three defeasible arguments σ_1, σ_2 and σ_3 can be extended to a derivation of a contradiction. If none of the arguments has stronger conclusive force than another, there are three extensions, viz. the three sets of two arguments $\{\sigma_1, \sigma_2\}$, $\{\sigma_1, \sigma_3\}$ and $\{\sigma_2, \sigma_3\}$. If for instance σ_1 has stronger conclusive force than σ_2, $\{\sigma_2, \sigma_3\}$ is not an extension.

Below some of Vreeswijk's (1997) definitions are recounted, some of them slightly adapted. Since in Vreeswijk's formalism rules and arguments are always treated separately, no notational precautions were necessary to distinguish a rule of inference from its instance in a derivation, i.e., in an argument. Since below the distinction is necessary in order to prevent ambiguity, rules of inference are denoted using the symbols $\rightarrow$ and $\Rightarrow$, while their instances in arguments are denoted using $\underline{\rightarrow}$ and $\underline{\Rightarrow}$. Vreeswijk's (1997) numbering is followed.

Definition (13.12): Vreeswijk's abstract argumentation systems

2.2 A *language* is a set L containing a distinguished element $\perp$.

2.3 A *strict rule of inference* is a formula of the form $\varphi_1, ..., \varphi_n \rightarrow \varphi$, where $\varphi_1, ..., \varphi_n$ is a finite, possibly empty, sequence in L and φ is a member of L. A *defeasible rule of inference* is a formula of the form $\varphi_1, ..., \varphi_n \Rightarrow \varphi$, where $\varphi_1, ..., \varphi_n$ is a finite, possibly empty, sequence in L and φ is a member of L.

2.5 An *argument* σ is

a. a member p of L. Its conclusion and only premise is p.

b. a formula of the form $\sigma_1, ..., \sigma_n \underline{\rightarrow} \varphi$, where $\sigma_1, ..., \sigma_n$ is a finite, possibly empty, sequence of arguments, such that the conclusions of $\sigma_1, ..., \sigma_n$ are $\varphi_1, ..., \varphi_n$, respectively, for some rule $\varphi_1, ..., \varphi_n \rightarrow \varphi$. Its conclusion is φ, its set of premises is the union of the sets of premises of $\sigma_1, ..., \sigma_n$.

c. a formula of the form $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$, where $\sigma_1, ..., \sigma_n$ is a finite, possibly empty, sequence of arguments, such that the conclusions of $\sigma_1, ..., \sigma_n$ are $\varphi_1, ..., \varphi_n$, respectively, for some rule $\varphi_1, ..., \varphi_n \Rightarrow \varphi$. Its conclusion is φ, its set of premises is the union of the sets of premises of $\sigma_1, ..., \sigma_n$.

2.10 An argument is *strict* if it is built using strict rules only, otherwise *defeasible*.

2.1 An *abstract argumentation system* is a triple (L, R, $\leq$), where L is a language, R is a set of rules of inference and $\leq$ is a reflexive and transitive order on arguments. For arguments σ and τ, $\sigma < \tau$ denotes that $\sigma \leq \tau$ while not $\tau \leq \sigma$.

3.2 A subset P of L is *incompatible* if there exists a strict argument with conclusion $\perp$.

4.1 A *base set* is a finite compatible subset of L. If P is a base set, an argument is *based on* P if its premises are in P. A set of arguments is *incompatible* if the set of conclusions of the arguments is incompatible. A set of arguments is *incompatible with* an argument σ if $\Sigma \cup \{\sigma\}$ is incompatible.

2.15 An argument σ *undermines* a set of arguments Σ if there is a τ in Σ with $\tau < \sigma$.

4.2 Let P be a base set, and let σ be an argument. A set of arguments Σ is a *defeater* of σ if Σ is incompatible with σ and not undermined by it.

4.17 Let P be a base set. A relation |~ between P and arguments based on P is a *defeasible entailment relation* if, for every argument σ based on P, it holds that P |~ σ if and only if

a. the set P contains σ, or

b. for some arguments $\sigma_1, ..., \sigma_n$ and a sentence φ in L, P |~ $\sigma_1, ..., \sigma_n$ and $\sigma = \sigma_1, ..., \sigma_n \underline{\rightarrow} \varphi$, or

c. for some arguments $\sigma_1, ..., \sigma_n$ and a sentence φ in L, P |~ $\sigma_1, ..., \sigma_n$ and $\sigma = \sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ and no set of arguments Σ with P |~ Σ is a defeater of σ.

4.18 A set of arguments Σ is an *extension* of P if there exists a defeasible entailment relation $\vdash$ such that $\Sigma = \{\sigma \mid P$ |~ $\sigma\}$.

Note that some of Vreeswijk's terminology also occurs in DefLog, but in a different meaning. Examples are Vreeswijk's arguments, incompatibility and extensions which are differently defined in DefLog. When confusion is likely, we speak for instance of DefLog arguments and AAS arguments, where AAS abbreviates 'abstract argumentation systems'.

In the following, a formal connection between Vreeswijk's abstract argumentation systems and DefLog is established. Assume an argumentation framework (L, R, $\leq$) as given.

As sentential constants in DEFLOG, the union of four sets is used: the set L, the set of rules R, the set of arguments and the set of conclusive force statements $\sigma < \tau$. In this way, a large part of the formal apparatus of Vreeswijk's abstract argumentation systems can be translated into DEFLOG's language. Consider for instance the construction of an AAS argument $\sigma(p) \underline{\Rightarrow} q$ from an AAS argument $\sigma(p)$ with conclusion p and a defeasible rule of inference $p \Rightarrow q$. This construction can now be expressed within DEFLOG by the sentence $(p \Rightarrow q) \rightsquigarrow (\sigma(p) \rightsquigarrow (\sigma(p) \underline{\Rightarrow} q))$. (Note that here the notational distinction between rules of inference and their instances in arguments are needed in order to prevent ambiguity.)

Consider now the following four schemes of DEFLOG sentences. The first two model AAS argument construction, the second two AAS defeat.

(i) $(\varphi_1, ..., \varphi_n \rightarrow \varphi) \rightsquigarrow (\sigma_1 \rightsquigarrow (... (\sigma_n \rightsquigarrow (\sigma_1, ..., \sigma_n \underline{\rightarrow} \varphi)) ...)$
Here $\varphi_1, ..., \varphi_n \rightarrow \varphi$ is a rule and $\sigma_1, ..., \sigma_n$ are arguments with conclusions $\varphi_1, ..., \varphi_n$.

(ii) $(\varphi_1, ..., \varphi_n \Rightarrow \varphi) \rightsquigarrow (\sigma_1 \rightsquigarrow (... (\sigma_n \rightsquigarrow (\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi)) ...)$
Here $\varphi_1, ..., \varphi_n \Rightarrow \varphi$ is a rule and $\sigma_1, ..., \sigma_n$ are arguments with conclusions $\varphi_1, ..., \varphi_n$.

(iii) $\rho_1 \rightsquigarrow (... (\rho_r \rightsquigarrow (\tau_1 \rightsquigarrow (... (\tau_t \rightsquigarrow \times((\varphi_1, ..., \varphi_n \Rightarrow \varphi) \rightsquigarrow (\sigma_1 \rightsquigarrow (... (\sigma_n \rightsquigarrow (\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi)) ...))) ...))) ...)$
Here $\rho_1, ..., \rho_r$ are strict rules of inference that can be used to extend $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ and the arguments $\tau_1, ..., \tau_t$ to an argument with conclusion $\perp$. The other elements of the scheme are as in (ii).

(iv) $(\sigma_1 > \tau_1) \rightsquigarrow (\rho_1 \rightsquigarrow (... (\rho_r \rightsquigarrow (\tau_1 \rightsquigarrow (... (\tau_t \rightsquigarrow \times(\rho_1 \rightsquigarrow (... (\rho_r \rightsquigarrow (\tau_1 \rightsquigarrow (... (\tau_t \rightsquigarrow \times((\varphi_1, ..., \varphi_n \Rightarrow \varphi) \rightsquigarrow (\sigma_1 \rightsquigarrow (... (\sigma_n \rightsquigarrow (\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi)) ...))) ...))) ...))) ...))) ...))$
All elements of the scheme are as in (iii).

Before the explanation of the schemes, it can be noted that the second scheme occurs as a subscheme of the third scheme, which on its turn is a part of the fourth. Using convenient abbreviations, the structure of the third and fourth schemes stands out more clearly as follows:

(iii) $\rho_1 \rightsquigarrow (... (\rho_r \rightsquigarrow (\tau_1 \rightsquigarrow (... (\tau_t \rightsquigarrow \times(\text{ii})) ...))) ...)$
(iv) $(\sigma_1 > \tau_1) \rightsquigarrow (\rho_1 \rightsquigarrow (... (\rho_r \rightsquigarrow (\tau_1 \rightsquigarrow (... (\tau_t \rightsquigarrow \times(\text{iii})) ...))) ...))$

By the schemes (i) and (ii), arguments can be expanded by the application of strict and defeasible rules. By scheme (iii), argument expansion by application of a defeasible rule is blocked if the expansion could lead to an argument with conclusion $\perp$. This is achieved by an attack of scheme (ii) in case there are strict rules and additional arguments from which an unwanted argument for $\perp$ could be constructed. By scheme (iv), the application of the defeasible rule is reinstated in case the resulting argument has stronger conclusive force than one of the other arguments needed to construct the unwanted argument for $\perp$. Formally this is expressed by an attack of scheme (iii) in case an appropriate conclusive force statement obtains, in addition to the rules and arguments needed for the construction of an argument for $\perp$.

As an illustration, one example is worked out. Let the language L consist of the sentences $\perp$, p_1, p_2, q and $\neg q$. Consider the three rules of inference $p_1 \Rightarrow q$, $p_2 \Rightarrow \neg q$ and $q, \neg q \rightarrow \perp$. With respect to the abstract argumentation framework with these rules and an empty conclusive force relation, the base set $\{p_1, p_2\}$ has two AAS extensions, viz. the sets of arguments $\{p_1, p_2, p_1 \underline{\Rightarrow} q\}$ and $\{p_1, p_2, p_2 \underline{\Rightarrow} \neg q\}$. If the argument $p_1 \underline{\Rightarrow} q$ has stronger conclusive force than the argument $p_2 \underline{\Rightarrow} \neg q$ in the abstract argumentation framework, then the base set $\{p_1, p_2\}$ has only $\{p_1, p_2, p_1 \underline{\Rightarrow} q\}$ as an extension.

In the following the instances of the schemes (i) to (iv) are listed that can be used to mimic Vreeswijk's technical apparatus in DEFLOG. Assume first that the conclusive force relation is empty. Then the following are needed.

(i) $(q, \neg q \rightarrow \perp) \rightsquigarrow ((p_1 \underline{\Rightarrow} q) \rightsquigarrow ((p_2 \underline{\Rightarrow} \neg q) \rightsquigarrow ((p_1 \underline{\Rightarrow} q, p_2 \underline{\Rightarrow} \neg q) \underline{\rightarrow} \perp)))$
(ii.a) $(p_1 \Rightarrow q) \rightsquigarrow (p_1 \rightsquigarrow (p_1 \underline{\Rightarrow} q))$
(ii.b) $(p_2 \Rightarrow \neg q) \rightsquigarrow (p_2 \rightsquigarrow (p_2 \underline{\Rightarrow} \neg q))$
(iii.a) $(q, \neg q \rightarrow \perp) \rightsquigarrow ((p_1 \underline{\Rightarrow} q) \rightsquigarrow \times((p_2 \Rightarrow \neg q) \rightsquigarrow (p_2 \rightsquigarrow (p_2 \underline{\Rightarrow} \neg q))))$
(iii.b) $(q, \neg q \rightarrow \perp) \rightsquigarrow ((p_2 \underline{\Rightarrow} \neg q) \rightsquigarrow \times((p1 \Rightarrow q) \rightsquigarrow (p_1 \rightsquigarrow (p_1 \underline{\Rightarrow} q))))$

By (i), the arguments $p_1 \underline{\Rightarrow} q$ and $p_2 \underline{\Rightarrow} \neg q$ for q and $\neg q$ can be extended to the argument $(p_1 \underline{\Rightarrow} q, p_2 \underline{\Rightarrow} \neg q) \underline{\rightarrow} \perp$ if the rule $q, \neg q \rightarrow \perp$ and the arguments $p_1 \underline{\Rightarrow} q$ and $p_2 \underline{\Rightarrow} \neg q$ obtain. By (ii.a), the defeasible argument $p_1 \underline{\Rightarrow} q$ can be formed from the rule $p_1 \Rightarrow q$ and the argument p_1. By (ii.b), $p_2 \underline{\Rightarrow} \neg q$ can be formed from $p_2 \Rightarrow \neg q$ and p_2. The sentences (iii.a) and (iii.b) make (ii.a) and (ii.b) defeasible. For instance, (ii.a) expresses the situation that the rule $q, \neg q \to \bot$ and the argument $p_1 \underline{\Rightarrow} q$ make the construction sentence (ii.a) for the argument $p_2 \underline{\Rightarrow} \neg q$ defeated.

The DEFLOG theory consisting of p_1, p_2 and the above sentences (i), (ii.a), (ii.b), (iii.a) and (iii.b) has two extensions, one in which (ii.a) is defeated and one in which (ii.b) is. In the first DEFLOG extension, the three argument expressing sentences p_1, p_2 and $p_1 \underline{\Rightarrow} q$ are justified, while $p_2 \underline{\Rightarrow} \neg q$ is not taken into account, and in the second p_1, p_2 and $p_2 \underline{\Rightarrow} \neg q$ are justified, while $p_1 \underline{\Rightarrow} q$ is not taken into account. As a result, the two DEFLOG extensions correspond to the two AAS extensions of the base set $\{p_1, p_2\}$.

Assume now that $(p_1 \underline{\Rightarrow} q) > (p_2 \underline{\Rightarrow} \neg q)$, i.e., that the argument $p_1 \underline{\Rightarrow} q$ has stronger conclusive force than the argument $p_2 \underline{\Rightarrow} \neg q$. Then the following instance of scheme (iv) does the trick.

(iv) $((p_1 \underline{\Rightarrow} q) > (p_2 \underline{\Rightarrow} \neg q)) \to ((q, \neg q \to \bot) \to ((p_1 \underline{\Rightarrow} q) \to$
$\times((q, \neg q \to \bot) \to ((p_2 \underline{\Rightarrow} \neg q) \to \times((p_1 \Rightarrow q) \to (p_1 \to (p_1 \underline{\Rightarrow} q)))))))$

It says that the conclusive force comparison $(p_1 \underline{\Rightarrow} q) > (p_2 \underline{\Rightarrow} \neg q)$, the rule $(q, \neg q \to \bot)$ and the argument $(p_1 \underline{\Rightarrow} q)$ make the blocking sentence (iii.b) defeated. As a result, the DEFLOG theory consisting of p_1, p_2 and the above sentences (i), (ii.a), (ii.b), (iii.a), (iii.b) and (iv) has only one extension, viz. the one in which (ii.b) and (iii.b) are defeated. Only the DEFLOG extension in which p_1, p_2 and $p_1 \underline{\Rightarrow} q$ are justified remains, in correspondence with the only remaining AAS extension.

Let now Δ_{AAS} consist of all sentences of one of the schemes (i) through (iv). Note that only for the sentences of the forms (ii) and (iii) an attack is available (as expressed in (iii) and (iv), respectively).

Proposition (13.13)

A set of AAS arguments E is an AAS extension of a base set P with respect to an abstract argumentation system $(L, R, \leq)$ if and only if E is equal to the set of justified statements in a DEFLOG extension of the theory $P \cup R \cup \leq \cup\, \Delta_{AAS}$ that express an AAS argument.

Proof: Given an AAS extension E as in the proposition, it is possible to construct a DEFLOG extension of $P \cup R \cup \leq \cup\, \Delta_{AAS}$. The defeat of no sentence in $P \cup R \cup \leq$ is derivable from the theory since the defeat sentences do not even occur in a sentence in the theory. (Recall that the defeat sentence of a sentence φ is $\times\varphi$.) The only defeat sentences that occur in a sentence in the theory are the defeat sentences of sentences in Δ_{AAS} of the forms (ii) and (iii). E determines which sentences of the forms (ii) and (iii) actually are to be considered defeated. Let $D_{(ii)}$ consist of the form (ii) sentences in Δ_{AAS} that express the construction of an argument $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ that is not in E, while the arguments $\sigma_1, ..., \sigma_n$ are in E and the rule $\varphi_1, ..., \varphi_n \Rightarrow \varphi$ is in R. Let $D_{(iii)}$ consist of the form (iii) sentences that express that an argument $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ cannot be constructed, while it is in E, and while there are arguments $\tau_1, ..., \tau_t$ in E and strict rules $\rho_1, ..., \rho_r$ in R that can be used to expand it to an argument for $\bot$. Claim: $(P \cup R \cup \leq \cup\, \Delta_{AAS}) \setminus (D_{(ii)} \cup D_{(iii)})$ specifies a DEFLOG extension of the theory $P \cup R \cup \leq \cup\, \Delta_{AAS}$, in which exactly the sentences in $D_{(ii)} \cup D_{(iii)}$ are defeated. The claim follows from two observations. First, observe that for any sentence in $D_{(ii)}$ expressing the construction of an argument $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ there must be arguments $\tau_1, ..., \tau_t$ in E and strict rules $\rho_1, ..., \rho_r$ in R that can be used to construct an argument for $\bot$, while $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$ does not have stronger conclusive force than one of the arguments $\tau_1, ..., \tau_t$. As a result, a corresponding form (iii) sentence can be used to derive its defeat, while no form (iv) sentence can defend against that. Second, observe that for any sentence in $D_{(iii)}$ expressing the attack of a form (ii) argument construction sentence, there must be an argument among the $\tau_1, ..., \tau_t$ that has weaker conclusive force than $\sigma_1, ..., \sigma_n \underline{\Rightarrow} \varphi$. As a result, its defeat can be derived from an appropriate form (iv) sentence.

That the set of justified statements in a DEFLOG extension of the theory $P \cup R \cup \leq \cup\, \Delta_{AAS}$ is an AAS extension of P with respect to $(L, R, \leq)$ follows from similar observations.

A major difference between DEFLOG and Vreeswijk's abstract argumentation systems is that the former is sentence-based, while the latter is derivation-based, in the sense that in DEFLOG the statements expressed by sentences can be defeated, while in Vreeswijk's abstract argumentation systems derivations (in his system called arguments) are the object of defeat. Verheij's CUMULA (1996b) is in a similar way derivation-based. The reconstruction of Vreeswijk's abstract argumentation systems shows how DEFLOG can incorporate the derivation-based approach by including sentences that express derivations in the logical language.

Another important difference is that Vreeswijk's abstract argumentation systems defeat is non-deterministic, in the following special sense: when some AAS arguments are involved in a conflict, each can be *chosen* as defeated, merely by its being involved in the conflict. The theory does not prescribe which from among the arguments to choose as defeated. The choice can be restricted by the conclusive force relation: an argument that is stronger than another argument in the conflict cannot be chosen as defeated. DEFLOG is deterministic, in the special sense that when a statement is defeated in an extension of a theory this is an explicit consequence of the justified part of the theory. That it still can occur that there are several extensions is not the result of non-deterministic choice, but of dialectical ambiguity (i.e., the possibility that a sentence is both dialectically justifiable and defeasible).

The result is for instance that Vreeswijk's abstract argumentation systems are not very well suited for modeling Pollock's undercutters. If for instance p undercuts q as a reason for r, then Vreeswijk interprets this in terms of the conditional q > r (cf. Vreeswijk's notation, 1997, p. 277; the conditional > is not to be confused with the conclusive force relation on arguments). He first enforces a contradiction between q > r and ¬(q > r), where the latter is made to follow from p, and then adds that the argument for ¬(q > r) has stronger conclusive force than an argument for q > r. Vreeswijk would use a strict rule q > r, ¬(q > r) → ⊥, a defeasible rule p ⇒ ¬(q > r), and the stipulation that the AAS argument p $\underline{\Rightarrow}$ ¬(q > r) (or any other argument ending like this) has stronger conclusive force than the AAS argument q > r (and any other argument with this conclusion). As a result, that p undercuts q as a reason for p must partly be expressed in the fixed conclusive force relation on arguments, that is expressed outside the logical object language. To me it seems much more natural that undercutters (and for that matter all other defeat information) are expressible in the logical object language, like in DEFLOG.

13.3 Dung's admissible sets

Dung's (1995) notion of admissible sets of unstructured arguments turned out to be a fruitful abstraction of ideas from nonmonotonic reasoning and logic programming.[16] Dung's definitions provided inspiration for several of DEFLOG's definitions (and for some of my earlier work on dialectical argumentation, e.g., Verheij, 1996a and 1996b). The following recapitulates some of Dung's definitions.

Definition (13.14): Dung's admissible sets

(i) An *argumentation framework* is a pair <AR, *attacks*> where AR is a set of arguments and *attacks* is a binary relation on AR. If (A, B) ∈ *attacks*, then the argument A *attacks* the argument B. A set of arguments S *attacks* an argument A if there is an argument B in S that attacks A.

(ii) Given an argumentation framework <AR, *attacks*>, a set S ⊆ AR of arguments is *conflict-free* if there are no arguments A and B in S such that A attacks B.

(iii) An argument A ∈ AR is *acceptable* with respect to a set S of arguments if, for each argument B ∈ AR, if B attacks A then S attacks B.

(iv) A conflict-free set of arguments S is *admissible* if each argument in S is acceptable with respect to S.

(v) A *preferred extension* of an argumentation framework <AR, *attacks*> is an admissible set that is maximal with respect to set inclusion.

(vi) A conflict-free set of arguments S is a *stable extension* if S attacks any argument not in S.

The DEFLOG translation of an argumentation framework <AR, *attacks*> goes as follows. The arguments of AR are used as the elementary sentences of DEFLOG's language. If <AR, *attacks*> is an argumentation framework, then the theory AR ∪ {A ⋊ B | (A, B) ∈ *attacks*} is its DEFLOG translation. The following proposition establishes a formal connection between Dung's admissible sets and DEFLOG.

Proposition (13.15)

Let Δ be the DEFLOG translation of an argumentation framework <AR, *attacks*>. Then the following obtain:

(i) A set of arguments S ⊆ AR is conflict free (in Dung's sense) if and only if S is satisfiable (in DEFLOG's sense).

(ii) A conflict free set of arguments S ⊆ AR is admissible (in Dung's sense) if and only if S is dialectically justifying with respect to Δ (in DEFLOG's sense).

(iii) A set of arguments S ⊆ AR is a preferred extension of <AR, *attacks*> (in Dung's sense) if and only if S specifies a preferred stage of Δ (in DEFLOG's sense).

[16] Admissibility has also been discussed in section 12.4.

(iv) A set of arguments $S \subseteq AR$ is a stable extension of <AR, *attacks*> (in Dung's sense) if and only if S specifies an extension of Δ (in DefLog's sense).

Proof: All parts of the proposition follow by straightforward definition checking.

The major difference between DefLog and Dung's definitions is that DefLog has a logical language in which information concerning justification, attack and defeat can be expressed, whereas Dung uses an unstructured language and a fixed set of attack relations. As a result of its expressive language, DefLog allows very flexible representations. Sentences like $p_1 \rightarrow (q \rtimes r)$ and $p_2 \rtimes (q \rtimes r)$ expressing that q attacks r if p_1 and that p_2 attacks that q attacks r, have no counterpart in Dung's argumentation frameworks.

A natural way to define dialectical justification in Dung's framework is the following:

> A conflict free set of arguments C (in Dung's sense) is *dialectically justifying* (in Dung's framework) if C attacks any conflict free set of arguments C', such that $C \cup C'$ is not conflict free.

It follows by straightforward definition checking[17] that the notions of admissibility and dialectical justification that differ on DefLog's language (see especially section 12.4) coincide on Dung's restricted language.

Bondarenko, Dung, Kowalski and Toni (1997) have used admissibility in their discussion of an abstract, argumentation-theoretic approach to default reasoning. Their setting is just as Dung's (1995) related to DefLog's, yet they focus on deductive systems. Interestingly, whereas in DefLog dialectical negation × is treated as an ordinary connective, Bondarenko, Dung, Kowalski and Toni consider the question which sentences are the contraries of others as part of the domain theory (as the mapping from sentences to their contraries is explicitly represented in their assumption-based frameworks). It seems that the notion of dialectical justification can be directly transplanted to their system. For the reasons, discussed in section 12.4, it is probable that dialectical justification has better properties for analyzing assumption-based frameworks than admissibility.

By the technical closeness of DefLog and the approaches of Dung (1995) and Bondarenko, Dung, Kowalski and Toni (1997), several of DefLog's properties have direct analogues in the latter work. Using the proposition, many results on DefLog are immediately relevant for Dung's admissible sets (and vice versa, of course). For instance, the results on the extension existence problem and the extension multiplicity problem in section 9 and on the internal structure of dialectical justification in section 10 can be easily translated to Dung's framework. The former give amongst others necessary and sufficient conditions for the existence of Dung's stable extensions and for the multiplicity of Dung's stable extensions in terms of dialectical justification, and equivalently in terms of admissibility (by the equivalence of dialectical justification and admissibility on Dung's restricted language, discussed above). The results of section 8 on types of stages can also be transplanted to Dung's preferred extensions. For instance, it is not the case that Dung's preferred extensions are in general maximal stages, or can in general be extended or compatibly extended to maximal stages (cf. also Verheij, 1996a).[18]

13.4 Reason-Based Logic

Reason-Based Logic (Hage, 1996, 1997; Verheij, 1996b) is a formal model of rules and reasons, inspired by the use of these notions in the field of law. In Reason-Based Logic, rules are individuals that can have properties. Key properties of rules distinguished in Reason-Based Logic are their validity, applicability or exclusion. These properties are part of the core of Reason-Based Logic - they belong to its 'logical constants'. Rules can however also have other properties. For instance, they can be just, or effective.

In order to allow rules to have properties, they are not represented as sentences, but as terms. In Reason-Based Logic, the validity of a rule with antecedent φ and consequent ψ is for instance expressed by the sentence Valid(rule(φ, ψ)). Reason-Based Logic assumes a translation from sentences to terms.

[17] Let C be admissible (and therefore conflict free), and C' conflict free while $C \cup C'$ is not conflict free. Then C attacks C' or C' attacks C. If C' attacks C, there is an argument α in C' that attacks an argument β in C. But β is acceptable with respect to C, so C attacks α. So C is dialectically justifying. Let C be dialectically justifying, and let α be an argument in C. If β attacks α, then $\{\beta\}$ is a conflict free set attacking C, while $C \cup \{\beta\}$ is not conflict free. Therefore C attacks $\{\beta\}$, α is acceptable with respect to C and C is admissible.

[18] As a result, Prakken and Vreeswijk's (*to appear*, section 5.1) claim that preferred extensions correspond to maximal partial status assignments is mistaken.

The main relations between the core properties of rules in Reason-Based Logic are (in Verheij's (1996b) version) encoded in its semantic constraints. For instance, rules can only be excluded if they are valid. It should be stressed that only the main relations between Reason-Based Logic's core properties are encoded in the semantic constraints. For instance, the semantic constraints do not determine which rules are valid. It is a starting point of Reason-Based Logic that the question whether a rule is valid can be answered differently in different contexts. As a result, the representation of the facts about rules is to a large extent left to the domain theory.

The semantic constraints of Verheij's (1996b) version of Reason-Based Logic can be used to define a monotonic consequence notion. However, since in Reason-Based Logic rules can be excluded and the reasons for a conclusion can outweigh the reasons against it, it is also natural to study nonmonotonic consequence notions for Reason-Based Logic. For instance, fixed point definitions in the style of Reiter's logic for default reasoning have been applied to Reason-Based Logic.

Similarly, DefLog can be used to specify nonmonotonic aspects of Reason-Based Logic. Here an axiom system is presented that gives an idea how the representation of rules and their properties in Reason-Based Logic can be modeled in DefLog.

(i) $\varphi \leadsto (\text{Valid}(\text{rule}(\varphi, \psi)) \leadsto \text{Reason}(\varphi, \psi))$

(ii) $\text{Reason}(\varphi, \psi) \leadsto \psi$

(iii) $\text{Excluded}(\text{rule}(\varphi, \psi)) \leadsto \times(\varphi \leadsto (\text{Valid}(\text{rule}(\varphi, \psi)) \leadsto \text{Reason}(\varphi, \psi)))$

(iv) $\text{Outweighs}(\{\varphi_1, ..., \varphi_n\}, \{\psi_1, ..., \psi_m\}, \chi) \leadsto \times(\text{Reason}(\psi_j, \text{not-}\chi) \leadsto \text{not-}\chi)$

(v) $\text{Reason}(\psi_{m+1}, \text{not-}\chi) \leadsto \times(\text{Outweighs}(\{\varphi_1, ..., \varphi_n\}, \{\psi_1, ..., \psi_m\}, \chi) \leadsto \times(\text{Reason}(\psi_j, \text{not-}\chi) \leadsto \text{not-}\chi))$

(vi) $\text{Reason}(\varphi, \psi) \leadsto \text{Valid}(\text{rule}(\varphi, \psi))$

(vii) $\text{Excluded}(\text{rule}(\varphi, \psi)) \leadsto \text{Valid}(\text{rule}(\varphi, \psi))$

(viii) $\text{Outweighs}(\{\varphi_1, ..., \varphi_n\}, \{\psi_1, ..., \psi_m\}, \chi) \leadsto \text{Reason}(\varphi_i, \chi)$

(ix) $\text{Outweighs}(\{\varphi_1, ..., \varphi_n\}, \{\psi_1, ..., \psi_m\}, \chi) \leadsto \text{Reason}(\psi_j, \text{not-}\chi)$

Note that φ, ψ etc. are used as metavariables for corresponding sentences and terms in these axiom schemes. In schemes (iv), (v) and (ix), not-χ is a metavariable that stands for the negation of χ, where the type of negation is left implicit here. In scheme (v), it is assumed that ψ_{m+1} differs from $\psi_1, ..., \psi_m$.

By axiom scheme (i), if the antecedent of a valid rule is satisfied, it becomes a reason for the rule's consequent. According to axiom scheme (ii), if there is a reason for some conclusion, the conclusion follows. Both schemes are defeasible, though, as the schemes (iii) and (iv) show. If a rule is excluded, it does not give rise to a reason (scheme (iii)). If the reasons $\varphi_1, ..., \varphi_n$ for a conclusion χ outweigh the reasons $\psi_1, ..., \psi_m$ against it, the latter do not lead to their conclusion not-χ (scheme (iv)). By scheme (v), outweighing has no effect if there is an opposing reason that is not considered.

While the axiom schemes (i) to (v) express central properties of rules in Reason-Based Logic, the schemes (v) to (ix) are mostly auxiliary. Schemes (vi) and (vii) state that only valid rules give rise to reasons or can be excluded. By schemes (viii) and (ix), outweighing actually concerns reasons for and against a conclusion.

Details about Reason-Based Logic can be found in the work of Hage (1996, 1997) and Verheij (1996b).

13.5 Winning strategies in dialogue games

In the context of dialectical argumentation, i.e., argumentation with arguments and counterarguments, it is natural to consider dialogue games in which one of the game players tries to justify some statement, while the other tries to show that it is not justified or that it is defeated. For instance, Prakken and Sartor (1997) have used a dialogue game in order to characterize their category of justified arguments. The basic idea is that an argument is justified if there is a *winning strategy* for the player that starts a dialogue game by claiming the argument.

Here the definitions as given by Prakken (1997) are given. Not all notions occurring in the definitions are formally defined here. They are recounted as an illustration. The numbering is taken from Prakken (1997, p. 166-167).

Definition (13.16)

6.5.2 A *dialogue* based on a default theory Γ is a nonempty sequence of moves $move_i$ = ($Player_i$, Arg_i) (with $i > 0$), such that

1. $Arg_i \in Args_\Gamma$
2. $Player_i$ = P if and only if i is odd; and $Player_i$ = O if and only if i is even;
3. If $Player_i = Player_j$ = P and $i \neq j$, then $Arg_i \neq Arg_j$;
4. If $Player_i$ = P, then Arg_i strictly defeats Arg_{i-1};
5. if $Player_i$ = O, then Arg_i defeats Arg_{i-1}.

6.5.3 A *dialogue tree* is a tree of moves such that

1. Each branch is a dialogue;
2. If $Player_i$ = P, then the children of $move_i$ are all defeaters of Arg_i.

6.5.4 A player *wins a dialogue* if the other player cannot move. And a player *wins a dialogue tree* if and only if it wins all branches of the tree.

6.5.5 An argument A is *justified* if and only if there exists a dialogue tree with A as its root, and won by the proponent.

Player P is the proponent and player O the opponent of the argument with which the dialogue starts. The players exchange arguments as allowed by the default theory (conditions 1 and 2 under 6.5.2). By condition 3, the proponent is not allowed to repeat his moves. Conditions 4 and 5 state that each newly adduced argument must be a counterargument to its predecessor (in the sense of the system defined by Prakken and Sartor). Note the asymmetry between the proponent and the opponent: while a proponent's argument must strictly defeat its predecessor, the opponent only needs to provide a defeating argument.

Dialogue trees (6.5.3) are those collections of dialogues that show the proponent's reaction to any possible counterargument by the opponent. Winning a tree (6.5.4) means that the proponent has a winning strategy. Finally, in 6.5.5, an argument is defined to be justified when its proponent has a winning strategy.

The idea of winning strategies in dialogue games is closely related to that of a justifying dialectical argument as it was defined in section 10 on the internal structure of dialectical argumentation (definition (10.14)). In fact, a justifying dialectical argument corresponds exactly to a winning strategy for the first player in the following argumentation game:

Definition (13.17)

(i) An *argumentation game* concerning φ with respect to a theory Δ is a (finite or infinite) sequence of Δ-arguments C_1, C_2, ... (where the indices are natural numbers > 0) , such that

a. C_1 is an elementary argument for φ, and
b. C_{i+1} and C_i are elementarily incompatible if i is odd, and
c. C_{i+1} elementarily attacks C_i if i is even, and
d. if i and j have different parity, then C_i and C_j are not equal.

If the argument sequence is finite, the *length* of the game is the number of arguments in the sequence.

(ii) An argumentation game *has ended* if it is not a proper initial of another argumentation game, or if it is infinite. An ended argumentation game *is won by the second player* if there is a final argument with even index. Otherwise the game is *won by the first player*.

(iii) The *first player has a winning strategy* in the argumentation game concerning φ with respect to Δ if there is a map S from the set of argumentation games Γ concerning φ with respect to Δ that have an even index final argument C_Γ, to the set of Δ-arguments, such that S(Γ) is an argument elementarily attacking C_Γ.

Since ended argumentation games can be finite or infinite, there are two types of winning for the first player: the game is finite and the last argument has odd index, or the game is infinite. The intuition behind the second type of winning is that in that case the first player has a reply to any move by the second player. As a result, the first player succeeds in attacking all counterarguments by the second player. The map S in the definition of the first player having a winning strategy indicates which move the first player can make in reply to any previous move by the second player.

Proposition (13.18)

There exists a dialectical argument justifying φ with respect to a theory Δ if and only if the first player has a winning strategy in the argumentation game concerning φ with respect to Δ.

Proof: The proposition follows from the observations that argumentation games correspond exactly to the initial parts of branches in a justifying dialectical argument, and that ended argumentation games correspond exactly to the full branches of a justifying dialectical argument. The correspondence is such that the length of the indices of the arguments in a branch of a justifying dialectical argument is equal to the index of the corresponding argument in an argumentation game. Since by the definition of a justifying dialectical argument, it has no branches ending with an even length index, each even length game can be continued by the first player by playing the next argument in the branch corresponding to the game.

14 The comparison and the attack metaphor in dialectical argumentation

In the field of dialectical argumentation, two guiding metaphors can be distinguished, viz. the comparison and the attack metaphor.[19] Both can be regarded as attempts to adapt pure maximal consistency for modeling dialectical argumentation. The first is the comparison metaphor. In this metaphor, the defeat of a statement (or argument or rule of inference or derivation, or whatever is one's favorite object of defeat) is the result of comparing the statements (or ...) that are involved in a conflict. Usually, the comparison involves notions like strength or priority, that are usually taken as a primitive. In the comparison metaphor, one starts with a symmetric notion like conflict, that is then asymmetricized by a comparison relation. When two statements are conflicting and one has priority, the other is defeated. Pollock's (1987) rebutting defeaters are a paradigmatic example of defeat like in the comparison metaphor. The comparison metaphor is also at the heart of Vreeswijk's (1997) work.

The second is the attack metaphor. In this metaphor, defeat is the result of a battle between statements (or arguments or whatever) some of which attack others. Some of the statements (or ...) are defeated, viz. those that are attacked by an undefeated statement; others remain undefeated, e.g., those that are not at all attacked, or that are only attacked by arguments that are themselves defeated. The notion of attack is taken as a primitive. Attack is usually taken as an asymmetric relation. Pollock's (1987) undercutting defeaters are a paradigmatic example of defeat like in the attack metaphor. Dung's (1995) work and Verheij's (1996b) CUMULA are based on the attack metaphor.

The two metaphors are to some extent interchangeable. Assume that the statements φ and ψ are conflicting. Then the priority of φ over ψ or the attack of ψ by φ has the same result: φ is undefeated and ψ defeated.

For quite some time, I considered the attack metaphor more satisfactory than the comparison metaphor.[20] My main reasons were threefold. First, undercutters are harder to explain using the comparison metaphor (cf. the discussion of Vreeswijk's treatment of undercutters in section 13.2). Second, I considered defeat as the immediate result of an asymmetric relation like attack, while the comparison metaphor naturally starts with a symmetric relation like conflict. Third, in the case of the defeat of structured arguments, the comparison metaphor seemed to imply the unnatural separation of argument construction on the one hand and comparison and defeat on the other.

I have changed my opinion, since in DEFLOG, the comparison and the attack metaphor merge into one, while none of my three reasons against the comparison metaphor obtain. Any defeated statement φ corresponds to a justified statement $\times\varphi$, with which it is conflicting (in the sense that φ and $\times\varphi$ cannot both be justified in an interpretation). This seems a choice for the comparison metaphor. Another of DEFLOG's traits is however more like the attack metaphor, viz. the inherent asymmetry between φ and $\times\varphi$ that is built into the definition of extensions. By this asymmetry, the defeat of φ coincides with the justification of $\times\varphi$, but not the other way around. The defeat of $\times\varphi$ does not coincide with the justification of φ, but with that of $\times\times\varphi$. (Note that the asymmetry is taken away in symmetric DEFLOG, discussed in section 12.2, suggesting that symmetric DEFLOG is less appropriate for modeling dialectical argumentation than ordinary, asymmetric DEFLOG.)

The mixture of the comparison and the attack metaphor in DEFLOG explains why both have been fruitfully adopted, while as yet neither has successfully claimed its primacy.

[19] In my dissertation (Verheij, 1996, p. 164-5), I spoke in a similar vein of inconsistency-triggered and counterargument-triggered defeat.

[20] See e.g. my lecture notes on attack and defeat at http://www.metajur.unimaas.nl/~bart/teaching/defarg/. Amongst others, a discussion with Alejandro García made me doubt my position.

15 DEFLOG as a dialectical logic

DEFLOG can be regarded as what might be called a dialectical logic, viz. a logic in which not only attention is paid to justification, but also to defeat. Dialectical logics can be contrasted with deductive logics, in which only justification is addressed. Here justification might be interpreted in terms of rules of inference. A paradigmatic example of a deductive logic is standard propositional logic with its semantics in terms of valuations and its proof theory in terms of (e.g.) rules of inference. Here the differences between DEFLOG and standard propositional logic as logics are briefly addressed.

The difference between the logical languages of DEFLOG and propositional logic is only superficial. Whereas DEFLOG only uses two sentential connectives ⤳ and ×, propositional logic uses → and ¬, and usually several more, like ∨, ∧ and ↔. It is well-known however that a propositional language using only → and ¬ as connectives does not diminish its expressiveness. The reason why in DEFLOG only ⤳ and × are used is that they seem to suffice for a dialectical logic. Since the interpretation of ⤳ validates *Modus ponens*, it is possible to mimic many deductive logics (in fact all deductive logics that have a Hilbert-style proof theory, based on *Modus ponens* only). It has been shown (especially in the sections 11 and 13) that many notions in the field of defeasible reasoning can be modeled when DEFLOG's dialectical negation × is added.

A difference between DEFLOG and propositional logic already arises at the level of interpretations. Whereas DEFLOG's interpretations are valuations of the language's subsets, those of propositional logic are always total: any sentence of the language must be assigned a truth value. The reason for this difference is that DEFLOG's 'partial' interpretations can be thought of as worlds as they are specified by a theory about the world. Since such a theory does not necessarily provide complete information (in the sense of having for any sentence φ either φ or its negation as a consequence), it makes sense to consider the partial interpretations of DEFLOG. Note that in propositional logic there is formally an analog for DEFLOG's partial interpretations, viz. the sets of sentences that are closed under valid consequence[21].

The next difference between DEFLOG and propositional logic occurs when theories are interpreted. The interpretation of a theory in propositional logic are its models, which are the interpretations in which all sentences in the theory are positively evaluated, viz. as true. In contrast, in DEFLOG, a theory is interpreted in terms of its extensions, in which some sentences of the theory are evaluated positively and others negatively, viz. as justified and defeated, respectively. The reason is that in DEFLOG a theory is interpreted as a *defeasible specification of the world*. On the one hand, the sentences in the theory express statements about the world that are assumed to state truths. On the other hand, the theory can itself express that statements in the theory or following from it are not to be assumed to state truths. As a result, DEFLOG provides a distinct way of interpreting sets of sentences, next to their ordinary interpretation as strict theories. In the ordinary, strict interpretation of theories, the sentences in a theory are all assumed to state truths. The distinct way of interpretation is to consider a set of sentences as a *dialectical theory* that expresses defeasible assumptions about the world. The result is that there are two kinds of satisfiability. The first is standard, 'strict' satisfiability, when a theory has a model. The other is 'dialectical' satisfiability, when a theory has an extension. Note that an extension of a theory can be regarded as a specific kind of consistency maintenance (just like taking maximal satisfiable subsets) since any extension of a theory selects a satisfiable subset of the theory. However it has been shown above (especially in section 8) that there are many satisfiable subsets of a theory (all corresponding to the theory's stages) that do not correspond to extensions.

The notion of valid consequence in propositional logic, according to which a conclusion follows from a theory in case it is true in any model of the theory, has a counterpart in DEFLOG's notion of dialectical justification. Just as a theory has valid consequences, it has dialectically justifiable consequences. An important difference is that the notion of validity is monotonic, in the sense that a valid consequence of a theory is also a valid consequence of any larger theory. Dialectical justification is nonmonotonic: if a statement is dialectically justifiable with respect to a theory it need not be dialectically justifiable with respect to a larger theory. Dialectical justification does also not obey the inclusion property since it is not the case that any statement expressed by the theory is dialectically justifiable with respect to the theory. Valid consequence has the inclusion property.

Another difference arises with respect to inconsistency: whereas an inconsistency in propositional logic trivializes the theory, since any conclusion follows from a theory that has both a conclusion and its negation as a consequence, its analog in DEFLOG, viz. dialectical ambiguity, is not trivializing: the

[21] Such a set is confusingly often called a theory, in contrast with the use of that term in the present paper, where a theory is just any set of sentences.

existence of a statement that is both dialectically justifiable and dialectically defeasible with respect to a theory does not imply that any statement is dialectically justifiable. Note also that while inconsistency corresponds to non-satisfiability in propositional logic, the analogous correspondence between dialectical ambiguity and dialectical non-satisfiability (in the sense of not having an extension) does not hold: dialectical ambiguity is the first step towards ambiguous dialectical satisfiability, in the sense of having more than one extension. The precise correspondence between dialectical ambiguity and dialectical non-satisfiability is stated in theorem (9.6) and corollary (9.10). It is significantly more complex than the equivalence of inconsistency and non-satisfiability in propositional logic.

The role of valid proofs in propositional logic is similar to that of the justifying dialectical arguments of DEFLOG. Whereas a valid proof shows the internal structure of valid consequence, in the sense that it explicates the inference steps that lead from a theory to its conclusion, justifying dialectical arguments show the internal structure of dialectical justification. Justifying dialectical arguments do however not only explicate the inference steps that lead from a theory to its conclusion (in DEFLOG simply a sequence of applications of ⤳-*Modus ponens*), but also which attacks are required against incompatibilities. An important limitation follows: the set of justifying dialectical arguments with respect to a theory is not recursively defined in terms of the theory as is the case for the set of valid proofs. As a result, the justifying dialectical arguments are not in general readily computable.

These differences show that the dialectical logic DEFLOG differs significantly from deductive logic as exemplified by standard propositional logic. DEFLOG can in an important sense, however, be regarded, not as a modification, but as an expansion of deductive logic: its core is a deductive logic built around a *Modus ponens* validating conditional ⤳, to which dialectical negation × has been added.

16 Conclusion

When theories are interpreted dialectically, i.e., as sets of sentences expressing juxtaposed opposing and contradicting statements, some of which can be justified and others defeated, more theories are interpretable then when theories are interpreted 'monolectically', i.e., as sets of sentences assumed to be all true. In other words, there are more theories with extensions than theories with models.

A fundamental complication of dialectical interpretation of theories in terms of extensions is that theories can have zero, one or several extensions. The extension existence problem asks for a necessary and sufficient criterion for the existence of an extension of a theory. The extension multiplicity problem asks for a necessary and sufficient criterion for the existence of multiple extensions of a theory.

In the present paper, the notion of dialectical justification has been introduced: an argument is dialectically justifying when it attacks all arguments that are incompatible with it. The properties of dialectical justification, especially the union, localization and separation properties, make it particularly suitable for the analysis of extensions. It has been shown that the notion of dialectical justification gives rise to necessary and sufficient criteria that solve the extension existence and the extension multiplicity problems. The idea is that an extension exists if and only if there is a part of the theory in the context of which no sentence of the theory is dialectically ambiguous (i.e., both dialectically justifiable and dialectically defeasible), while all sentences of the theory are dialectically interpretable (i.e., either dialectically justifiable or dialectically defeasible) in the context of that part of the theory. Multiple extensions exist if and only if there are multiple incompatible parts with these properties.

It has been shown that a simple, dialectically interpreted logical language using ordinary connectives × and ⤳ is suitable as a language for the analysis of central topics of dialectical argumentation, such as Toulmin's argument scheme, Pollock's rebutting and undercutting defeaters, and priority and weighing defeaters. An important consequence of the choice of language is that in DEFLOG all information concerning justification and defeat is expressible in the logical object language as contingent information. There is no need for separate classes of defeasible rules of inference, priority information or pre-defined conclusive force relations between arguments. All these kinds of information can be expressed directly in DEFLOG's language, along with the other contingent information.

The internal structure of dialectical justification has been analyzed, in terms of justifying dialectical arguments (that differ subtly from the naïve dialectical arguments of section 2).

The idea of stages provides a different approach towards the investigation of the local properties of dialectical interpretation of theories in terms of extensions. A theory's stages are the dialectical interpretations of parts of the theory. Instead of maximizing only the justified sentences of a theory in a stage, it is also possible to maximize the whole set of interpreted sentences of a theory. It turns out that the types of maximization are perpendicular, in the sense that maximization in one sense does not imply maximality in the other sense. The result is a plethora of types of stages, with few interrelations. To me, this suggests that one should not consider each as a different type of dialectical interpretation of theories, as is for some types suggested in the work of Dung (1995) and Bondarenko *et al.* (1997) and also in Prakken & Vreeswijk's overview (*to appear*), but merely as partial interpretations with an interesting special property. In other words, to me, there is only one 'genuine' dialectical semantics, viz. dialectical interpretation as extensions. All other notions, such as satisfiability classes, dialectically preferred stages and maximal stages, are in the first place tools in the investigation of the properties of extensions. The use of the notion of dialectical justification in the extension existence and the extension multiplicity problems is an example of the application of such tools.

Acknowledgments

The author gladly acknowledges the financial support by the Dutch national Programme Information Technology and Law (ITeR) for the research reported in this paper (project number 01437112). The author thanks Jaap Hage, Peter McBurney and Bram Roth for comments and discussion.

References

Bench-Capon, T. (1995). Argument in Artificial Intelligence and Law. *Legal knowledge based systems. Telecommunication and AI & Law* (eds. J.C. Hage, T.J.M. Bench-Capon, M.J. Cohen and H.J. van den Herik), pp. 5-14. Koninklijke Vermande, Lelystad.

Bondarenko, A., Dung, P.M., Kowalski, R.A., and Toni, F. (1997). An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, Vol. 93, pp. 63-101.

Dung, P.M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, Vol. 77, pp. 321-357.

Eemeren, F.H. van, Grootendorst, R., and Snoeck Henkemans, F. (1996). *Fundamentals of Argumentation Theory. A Handbook of Historical Backgrounds and Contemporary Developments.* Lawrence Erlbaum Associates, Mahwah (New Jersey).

Hage, J.C. (1997). *Reasoning with Rules. An Essay on Legal Reasoning and Its Underlying Logic.* Kluwer Academic Publishers, Dordrecht.

Lin, F. (1993). An argument-based approach to nonmonotonic reasoning. *Computational Intelligence*, Vol. 9, No. 3, pp. 254-267.

Loui, R.P. (1998). Process and Policy: Resource-Bounded Non-Demonstrative Reasoning. *Computational Intelligence*, Vol. 14, No. 1.

Pollock, J.L. (1995). *Cognitive Carpentry: A Blueprint for How to Build a Person.* The MIT Press, Cambridge (Massachusetts).

Poole, D. (1988). A logical framework for default reasoning. *Artificial Intelligence*, Vol. 36, pp. 27-47.

Prakken, H. (1997). *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law.* Kluwer Academic Publishers, Dordrecht.

Prakken, H., and Sartor, G. (1996). A Dialectical Model of Assessing Conflicting Arguments in Legal Reasoning. *Artificial Intelligence and Law*, Vol. 4, pp. 331-368.

Prakken, H., and Vreeswijk, G. (*to appear*). Logics for Defeasible Argumentation. *Handbook of Philosophical Logic* (ed. D. Gabbay). Kluwer Academic Publishers, Dordrecht.

Reiter, R. (1980). A Logic for Default Reasoning. *Artificial Intelligence*, Vol. 13, pp. 81-132.

Rescher, N. (1964). *Hypothetical reasoning*. North-Holland Publishing Company, Amsterdam.

Toulmin, S.E. (1958). *The uses of argument.* University Press, Cambridge.

Verheij, B. (1996a). Two approaches to dialectical argumentation: admissible sets and argumentation stages. *NAIC'96. Proceedings of the Eighth Dutch Conference on Artificial Intelligence* (eds. J.-J.Ch. Meyer and L.C. van der Gaag), pp. 357-368. Also presented at the *Computational Dialectics Workshop* at FAPR-96. June 3-7, 1996, Bonn.

Verheij, B. (1996b). *Rules, Reasons, Arguments. Formal studies of argumentation and defeat*. Dissertation. Universiteit Maastricht, Maastricht.

Verheij, B. (1998a). Argue! - an implemented system for computer-mediated defeasible argumentation. *NAIC '98. Proceedings of the Tenth Netherlands/Belgium Conference on Artificial Intelligence* (eds. H. La Poutré and H.J. van den Herik), pp. 57-66. CWI, Amsterdam.

Verheij, B. (1998). ArguMed - A Template-Based Argument Mediation System for Lawyers. *Legal Knowledge Based Systems. JURIX: The Eleventh Conference* (eds. J.C. Hage, T.J.M. Bench-Capon, A.W. Koers, C.N.J. de Vey Mestdagh and C.A.F.M. Grütters), pp. 113-130. Gerard Noodt Instituut, Nijmegen.

Verheij, B. (1999). Automated Argument Assistance for Lawyers. *The Seventh International Conference on Artificial Intelligence and Law. Proceedings of the Conference*, pp. 43-52. ACM, New York (New York).

Verheij, B. (*to appear*). Dialectical Argumentation as a Heuristic for Courtroom Decision Making. See http://www.metajur.unimaas.nl/~bart/publications.htm.

Vreeswijk, G. (1997). Abstract argumentation systems. *Artificial Intelligence*, Vol. 90, pp. 225-279.