DEPARTMENT OF INFORMATION AND COMPUTING SCIENCES

UTRECHT UNIVERSITY

AUGUST 2025

Course reader

Computational Argumentation

Author:

HENRY PRAKKEN

Contents

Preface

This reader[1] is meant for the course *Computational Argumentation*, which gives an introduction to the computational study of argumentation in AI, a currently popular subfield of symbolic AI. The course, which was previously named *Commonsense Reasoning and Argumentation*, especially focuses on formal models of argumentation and their application in areas like commonsense reasoning, legal reasoning and multi-agent interaction. The computational study of argumentation concerns two aspects: reasoning and dialogue. Argumentation as a form of **reasoning** makes explicit the reasons for the conclusions that are drawn and how conflicts between reasons are resolved. Systems for argumentation-based inference were orginally developed in the field of nonmonotonic logic, which formalises qualitative reasoning with incomplete, uncertain or inconsistent information. Argument-based systems have been very successful as nonmonotonic logics, since they are based on very natural concepts, such as argument, counterargument, rebuttal and defeat. In this reader the following formalisms are discussed:

- The theory of abstract argumentation frameworks (the generally accepted formal foundation of the field) and its extension to bipolar argumentation frameworks.
- The theory of structured argumentation frameworks, with a special focus on the *ASPIC*$^+$ approach.
- Formal accounts of change operations on argumentation frameworks.
- Formal models of legal case-based reasoning

Argumentation as a form of **dialogue** concerns the rational resolution of conflicts of opinion by verbal means. Intelligent agents may disagree, for instance, about the pros and cons of alternative proposals, or about the factual basis of such proposals. Dialogue systems for argumentation formally define protocols for argumentation dialogues and thus enable a formal study of the dynamics of argumentative agent interaction, including issues of strategic choice. In this course two examples of such dialogues systems will be discussed.

This reader partly contains original texts and partly reuses and adapts texts from existing publications. The abstract approach to argumentation is discussed in Chapters 2 and 3, based on and extending Prakken and Vreeswijk (2002) and Vreeswijk and Prakken (2000). The *ASPIC*$^+$ framework for structured argumentation is discussed in Chapter 4, based on Modgil and Prakken (2014, 2018). Recent work on preferences, support relations and gradual argument acceptability are discussed in Chapter 5, which

[1]Thanks are due to the students of earlier years and in particular to Bas van Gijzel, Marc van Zee, Elisa Friscione, Daphne Odekerken and Heleen Kaemingk, for their corrections to previous versions of this reader.

combines parts of Prakken (2012, 2020) and Prakken (2021b). Dynamic aspects of argumentation are discussed in Chapter 6, partly based on Modgil and Prakken (2012) and Prakken (2023). Dialogue systems for agent interaction with argumentation are the topic of Chapter 7, based on Prakken (2006). Finally, legal applications of argumentation formalisms are discussed in Chapter 8, which mixes texts written especially for this course with parts of Prakken and Sartor (2015); Prakken (2021a) and Prakken (2015). Exercises can be found at the end of the relevant chapters, while their answers are in Chapter 9.

In the 2025 version of this reader, two chapters on non-argumentation-based nonmonotonic logics have been removed. Moreover, Exercise 3.5.7 and several exercises in Chapter 4 have been simplified, Section 4.4.2 has been extended with formal definitions of *ASPIC** and the order of some subsections and exercises in Chapter 4 has been changed. Finally, in Chapter 7 some new exercises on Sections 1-3 have been added.

Chapter 1

Argumentation logics: introduction

This chapter introduces the idea of an argumentation logic. In such a logic, arguments for and against a certain claim are produced and evaluated, to test the tenability of the claim. In the present chapter some motivating examples will be presented and the main concepts will be informally introduced, while in Chapters $2-4$ the formal theory of argumentation systems will be developed.

1.1 Motivating examples

We shall illustrate the idea of argumentation-based inference with a dispute between two persons, A and B. They disagree on whether it is morally acceptable for a newspaper to publish a certain piece of information concerning a politician's private life. Let us assume that the two parties have reached agreement on the following points.

(1) The piece of information I concerns the health of person P;
(2) P does not agree with publication of I;
(3) Information concerning a person's health is information concerning that person's private life

A now states the moral principle that

(4) Information concerning a person's private life may not be published if that person does not agree with publication.

and A says "So the newspapers may not publish I" (Fig. 1.1, page 10). Although B accepts principle (4) and is therefore now committed to (1-4), B still refuses to accept the conclusion that the newspapers may not publish I. B motivates her refusal by replying that:

(5) P is a cabinet minister
(6) I is about a disease that might affect P's political functioning
(7) Information about things that might affect a cabinet minister's political functioning has public significance

Furthermore, B maintains that there is also the moral principle that

(8) Newspapers may publish any information that has public significance

B concludes by saying that therefore the newspapers may write about P's disease (Fig. 1.2, page 11). A agrees with (5–7) and even accepts (8) as a moral principle,

but A does not give up his initial claim. Instead he tries to defend it by arguing that he has the stronger argument: he does so by arguing that in this case

(9) The likelihood that the disease mentioned in I affects P's functioning is small.

(10) If the likelihood that the disease mentioned in I affects P's functioning is small, then principle (4) has priority over principle (8).

Thus it can be derived that the principle used in A's first argument is stronger than the principle used by B (Fig. 1.3, page 11), which makes A's first argument stronger than B's, so that it follows after all that the newspapers should be silent about P's disease.

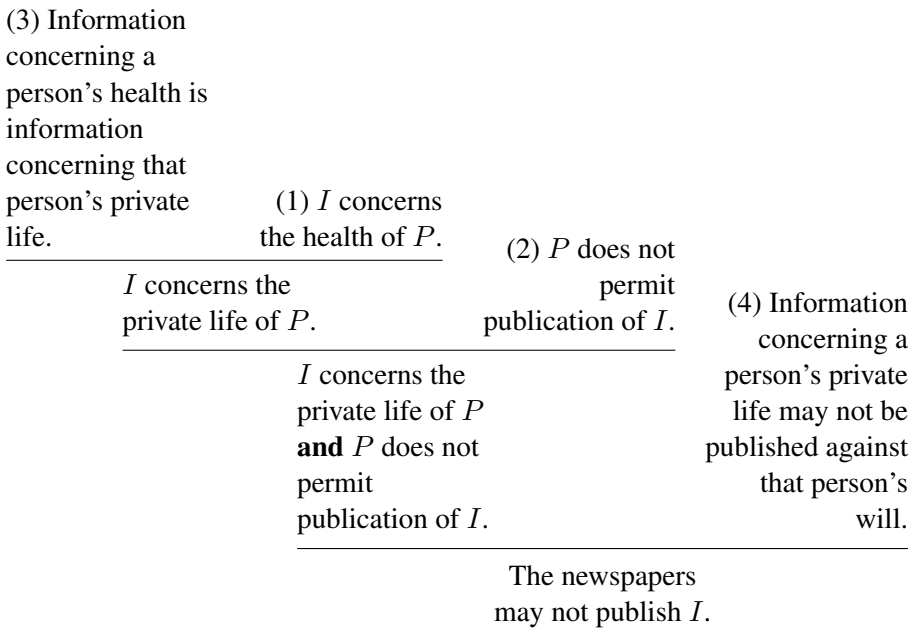


Figure 1.1: A's argument.

Let us examine the various stages of this dispute in some detail. Intuitively, it seems obvious that the accepted basis for discussion after A has stated (4) and B has accepted it, viz. (1,2,3,4), warrants the conclusion that the piece of information I may not be published. However, after B's counterargument and A's acceptance of its premises (5-8) things have changed. At this stage the joint basis for discussion is (1-8), which gives rise to two conflicting arguments. Moreover, (1-8) does not yield reasons to prefer one argument over the other: so at this point A's conclusion has ceased to be warranted. But then A's second argument, which states a preference between the two conflicting moral principles, tips the balance in favour of his first argument: so after the basis for discussion has been extended to (1-10), we must again accept A's moral claim as warranted.

Logical systems that formalise this kind of reasoning are called 'argumentation logics', or 'argumentation systems'. As the example shows, these systems lack the monotonicity property of 'standard', deductive logic (say, first-order predicate logic, FOL). According to FOL, if A's claim is implied by (1–4), it is surely also implied by (1–8). From the point of view of FOL it is pointless for B to accept (1–4) and yet

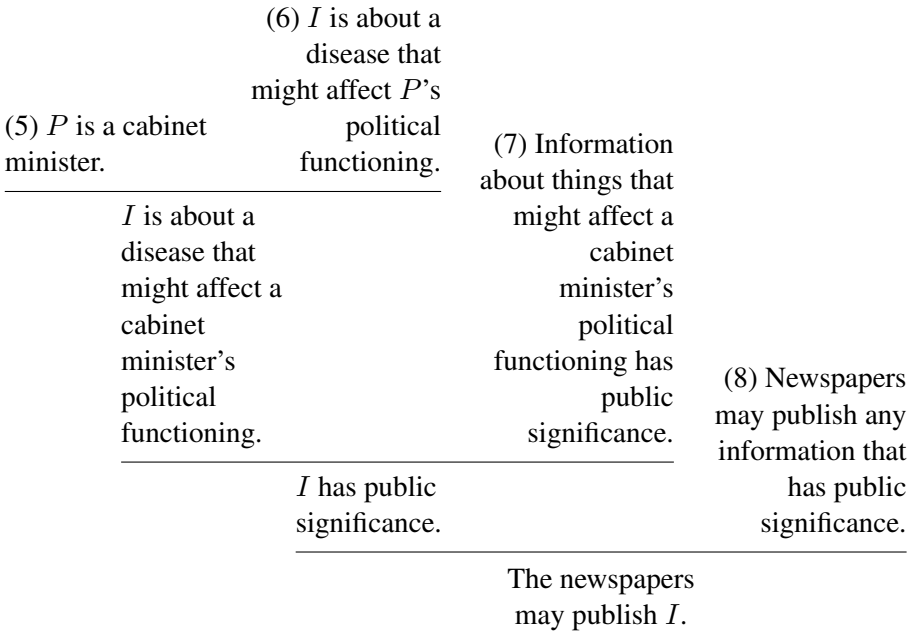


Figure 1.2: B's argument.

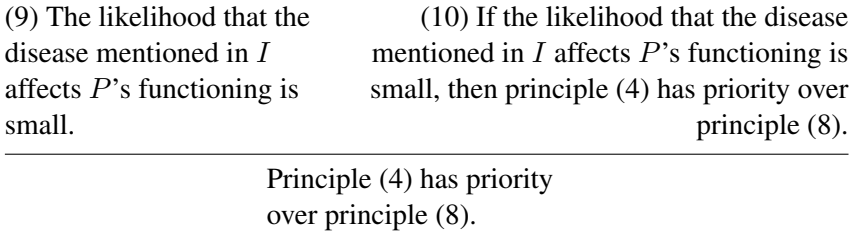


Figure 1.3: A's priority argument.

state a counterargument; B should also have refused to accept one of the premises, for instance, (4).

Does this mean that our informal account of the example is misleading, that it conceals a subtle change in the interpretation of, say, (4) as the dispute progresses? This is not so easy to answer in general. Although in some cases it might indeed be best to analyse an argument move like B's as a reinterpretation of a premise, in other cases this is different. In actual reasoning, rules are not always neatly labelled with an exhaustive list of possible exceptions; rather, people are often forced to apply 'rules of thumb' or 'default rules', in the absence of evidence to the contrary, and it seems natural to analyse an argument like B's as an attempt to provide such evidence to the contrary. When the example is thus analysed, the force of the conclusions drawn in it can only be captured by a consequence notion that is nonmonotonic: although A's claim is warranted on the basis of (1–4), it is not warranted on the basis of (1–8).

Argumentation logics are the most direct attempt to formalise examples like the above one, by defining notions like argument, counterargument, attack and defeat, and by defining nonmonotonic consequence in terms of the interaction of arguments for and against certain conclusions. This approach was initiated by the philosopher John Pol-

lock (Pollock; 1987), based on his earlier work in epistemology, e.g. (Pollock; 1974), and the AI researcher Ronald Loui (Loui; 1987).

One application of argumentation logics is to to formalise 'quick-and-dirty' commonsense reasoning with empirical generalisations. In everyday life people often reason with generalisations such as 'Birds fly', 'Italians usually like coffee', 'Chinese usually do not like coffee', 'Witnesses usually speak the truth' or 'When the streets are wet, it must have rained'. In commonsense reasoning, people apply such a generalisation if nothing is known about exceptions, but they are prepared to retract a conclusion if further knowledge tells us that there is an exception (for instance, a given bird is in fact a penguin, a witness has a reason to lie or the streets are wet because they are being cleaned).

However, argumentation systems have wider scope than just reasoning with such empirical generalisations. Firstly, argumentation systems can be applied to any form of reasoning with contradictory information, whether the contradictions have to do with generalisations and exceptions or not. For instance, the contradictions may arise from reasoning with several sources of information, or they may be caused by disagreement about beliefs or about moral, ethical or political claims. Moreover, it is important that several argumentation systems allow the construction and attack of arguments that are traditionally called 'ampliative', such as inductive, analogical and abductive arguments; these reasoning forms fall outside the scope of most other nonmonotonic logics.

One domain in which argumentation systems have become popular is legal reasoning. This is not surprising, since legal reasoning often takes place in an adversarial context, where notions like argument, counterargument, rebuttal and defeat are very common. Argumentation systems have also been applied in, for instance, the medical domain and in multi-agent models of negotiation and collaboration.

1.2 Argumentation systems: a conceptual sketch

In this section we give a conceptual sketch of the general ideas behind argumentation logics. First we sketch the general idea, and then we discuss the five main elements of such logics.

1.2.1 The general idea

Argumentation systems formalise nonmonotonic reasoning as the construction and comparison of arguments for and against certain conclusions. The idea is that the construction of arguments on the basis of a theory is monotonic, i.e., an argument stays an argument if the theory is enlarged with new information. Nonmonotonicity is explained in terms of the interactions between conflicting arguments: it arises from the fact that the new information may give rise to stronger counterarguments, which defeat the original argument. For instance, in case of Tweety the penguin we may construct one argument that Tweety flies because it is a bird, and another argument that Tweety does not fly because it is a penguin, and then we may prefer the latter argument because it is about a specific class of birds, and is therefore an exception to the general rule.

1.2.2 Five elements of argumentation systems

Argumentation systems contain the following five elements (although sometimes implicitly): an underlying logical language plus inference rules, definitions of an argument, of conflicts between arguments and of defeat between arguments and, finally, a definition of the dialectical status of arguments, which can be used to define a nonmonotonic notion of logical consequence.

A logical language plus inference rules

Argumentation systems are built around an underlying logical language and a set of inference rules defined over this language. Some systems assume a specific logical language and set of inferene rules, while other systems leave these things partly or wholly unspecified. The latter systems can thus be instantiated in alternative ways, which makes them frameworks rather than systems. An example of such a framework will be presented in Chapter 4.

Arguments

The notion of an argument corresponds to a tentative proof (or the existence of such a proof) in the 'logic' of the chosen logical language, where this 'logic' is expressed in the set of inference rules over the language. 'Logic' is here written between quotes because the logic does not need to be a standard deductive logic but can also contain defeasible inference rules (cf. the defaults of default logic). The nature of the inference rules of an argumentation system will be further discussed in Chapter 4. For now it suffices to say that the underlying logic of an argumentation system is still monotonic in the sense that new information cannot invalidate arguments as arguments but can only give rise to new counterarguments.

As for the layout of arguments, in the literature on argumentation systems three basic formats can be distinguished, all familiar from the logic literature. Sometimes arguments are defined as a tree of inferences grounded in the premises, and sometimes as a sequence of such inferences, i.e., as a deduction. Finally, some systems simply define an argument as a premises - conclusion pair, leaving implicit that the underlying logic validates a proof of the conclusion from the premises.

The notions of an underlying logic and an argument still fit with the standard picture of what a logical system is. The remaining three elements are what makes an argumentation system a framework for nonmonotonic reasoning.

Conflicts between arguments

The first is the notion of a *conflict* between arguments (also used are the terms 'attack' and 'counterargument'). In the literature, three types of conflicts are discussed. Firstly, arguments can be attacked on one of their premises, with an argument whose conclusion negates that premise. For example, an argument 'Tweety flies, because it is a bird' can be attacked by arguing that Tweety is not a bird. This kind of attack will in Chapter 4 be called *undermining* attack. The second type of attack is to negate the conclusion of an argument, as in 'Tweety flies, because it is a bird' and 'Tweety does not fly because it is a penguin' (cf. the left part of Fig. 1.4). Finally, when an argument uses a nondeductive, or *defeasible* inference rule, it can be attacked on its inference by arguing

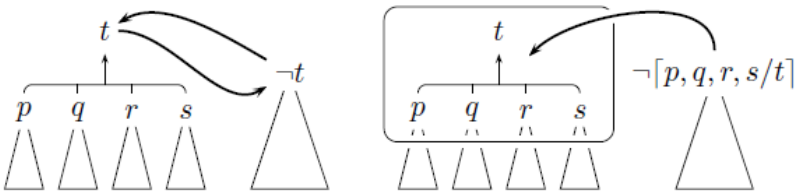


Figure 1.4: Rebutting attack (left) vs. undercutting attack (right).

that there is a special case to which the inference rule does not apply (cf. the right part of Fig. 1.4). After Pollock (1974, 1987), this is usually called *undercutting* attack. Unlike a rebutting attack, an undercutting attack does not negate the conclusion of its target but just says that its conclusion is not supported by its premises and can therefore not be drawn. In order to formalise this type of conflict, the rule of inference that is to be undercut (in Fig. 1.4: the rule that is enclosed in the dotted box, in flat text written as $p, q, r, s/t$) must be expressed in the object language: $\lceil p, q, r, s/t \rceil$) and denied: $\neg\lceil p, q, r, s/t \rceil$. [1] While all arguments can be attacked on their premises, only defeasible arguments can be attacked on their conclusion or inference. The reason why deductive arguments cannot be rebutted or undercut is that deductive inferences are by definition truth-preserving, i.e., the truth of their premises guarantees the truth of their conclusion, so the only way to disagree with the conclusion of a deductive argument is to deny one of its premsies. By contrast, the conclusion of a defeasible argument can be rejected even if all its premises are accepted. In Chapter 4 the difference between deductive and defeasible inference rules will be formalised and several examples of defeasible rules will be discussed. For now, consider the following example of a defeasible argument applying the principle of induction: the argument 'Raven $_{101}$ is black since the observed ravens raven$_1$ ...raven$_{100}$ were black' is undercut by an argument 'I saw raven$_{102}$, which was white'.

Note, finally, that all three kinds of attack have a direct and an indirect version; indirect attack is directed against a subconclusion or a substep of an argument, as illustrated by Figure 1.5 for indirect rebutting.

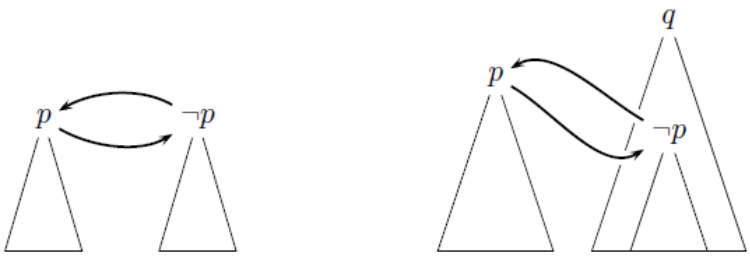


Figure 1.5: Direct attack (left) vs. indirect attack (right).

[1] Ceiling brackets around a meta-level formula denote a conversion of that formula to the object language, provided that the object language is expressive enough to enable such a conversion.

Defeat between arguments

The notion of conflicting, or attacking arguments does not embody any form of evaluation; evaluating conflicting pairs of arguments, or in other words, determining whether an attack is successful, is another element of argumentation systems. It has the form of a binary relation between arguments, standing for 'attacking and not weaker' (in a weak form) or 'attacking and stronger' (in a strong form). The terminology varies: some terms that have been used are 'defeat', 'attack' and 'interference'. Other systems do not explicitly name this notion but leave it implicit in the definitions. In this text we shall use 'defeat' for the weak notion and 'strict defeat' for the strong, asymmetric notion. Note that the several forms of attack, rebutting vs. assumption vs. undercutting and direct vs. indirect, have their counterparts for defeat.

Argumentation systems vary in their grounds for determining the defeat relations. Often only domain-specific criteria are available, which, moreover, are often defeasible. For this reason argumentation systems have been developed that allow for defeasible arguments on these criteria. To give some examples of domain-specific criteria, in domains where observations are important, defeat may depend on the reliability of tests, observers or sensors. In advice giving or consultancy, defeat may be determined by the level of expertise of the advisors or consultants. And in legal applications, defeat may depend on the legal hierarchy among statutes, on the court's level of authority, or on social or moral values. Our example in the introduction contains an argument on the criteria for defeat, viz. A's use of a priority rule (10) based on the expected consequences of certain events. This argument might, for instance, be attacked by an argument that in case of important officials even a small likelihood that the disease affects the official's functioning justifies publication, or by an argument that the negative consequences of publication for the official are small.

The dialectical status of arguments

The notion of defeat is a binary relation on the set of arguments. It is important to note that this relation does not yet tell us with what arguments a dispute can be won; it only tells us something about the relative strength of two individual conflicting arguments. The ultimate status of an argument depends on the interaction between all available arguments: it may very well be that argument B defeats argument A, but that B is itself defeated by a third argument C; in that case C 'reinstates' A (see Figure 1.6)[2]. Sup-

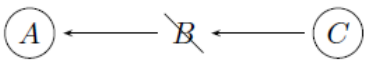


Figure 1.6: Argument C reinstates argument A.

pose, for instance, that the argument A that Tweety flies because it is a bird is regarded as being defeated by the argument B that Tweety does not fly because it is a penguin (for instance, because conflicting arguments are compared with respect to specificity).

[2]While in figures 1.4 and 1.5 the arrows stood for attack relations, from now on they will depict defeat relations.

And suppose that B is in turn defeated by an argument C, attacking B's intermediate conclusion that Tweety is a penguin. C might, for instance, say that the penguin observation was done with faulty instruments. In that case C reinstates argument A.

Therefore, what is also needed is a definition of the dialectical status of arguments on the basis of all the ways in which they interact. Besides reinstatement, this definition must also capture the principle that an argument cannot be justified unless all its subarguments are justified. There is a close relation between these two notions, since reinstatement often proceeds by indirect attack, i.e., attacking a subargument of the attacking argument (as illustrated by Figure 1.5). It is this definition of the status of arguments that produces the output of an argumentation system: it typically divides arguments in at least two classes: arguments with which a dispute can be 'won' and arguments with which a dispute should be 'lost'. Sometimes a third, intermediate category is also distinguished, of arguments that leave the dispute undecided. The terminology varies here also: terms that have been used are justified vs. defensible vs. defeated (or overruled), defeated vs. undefeated, in force vs. not in force, preferred vs. not preferred, etcetera. Unless indicated otherwise, we shall use the terms 'justified', 'defensible' and 'overruled' arguments.

These notions can be defined both in a 'declarative' and in a 'procedural' form. The declarative form, usually with fixed-point definitions, just declares certain sets of arguments as acceptable, (given a set of statements and evaluation criteria) without defining a procedure for testing whether an argument is a member of this set; the procedural form amounts to defining just such a procedure. Thus the declarative form of an argumentation system can be regarded as its (argumentation-theoretic) semantics, and the procedural form as its proof theory. Note that it is very well possible that, while an argumentation system has an argumentation-theoretic semantics, at the same time its underlying logic for constructing arguments has a model-theoretic semantics in the usual sense, for instance, the semantics of standard first-order logic, or a possible-worlds semantics of some modal logic.

EXERCISE 1.2.1 Reinstatement.

1. Extend Figure 1.6 (p. 15) with an argument D, such that D defeats C. Are there arguments that are justified? If so, which arguments? Are there arguments that are reinstated by D? If so, which?

2. Extend the figure just drawn with a fifth argument, E, such that E defeats D. Are there arguments that are justified? If so, which arguments? Are there arguments that are reinstated by D? If so, which? Are there arguments that are reinstated by E? If so, which?

The content of the remaining chapters on argumentation is as follows. Chapter 2 presents a fully abstract formal framework for the semantics of argumentation systems, which leaves the structure of arguments and the nature of the defeat relation unspecified. Chapter 3 discusses the proof-theory of these abstract argumentation systems in the form of so-called argument games. Chapter 4 then presents an instantiation of the abstract framework with structured arguments and two kinds of inference rules, deductive and defeasible ones. This framework is still partly abstract in that it abstracts from the nature and origin of these rules and from the nature of the logical language.

Chapter 2

An abstract framework for argumentation

This chapter presents a fully abstract framework for the semantics of argumentation, which leaves the internal structure of arguments and the nature of the defeat relation completely unspecified. As input it assumes nothing else but a set (of arguments) ordered by a binary relation (of defeat) and then defines several 'semantics', that is, properties that subsets of the set of all arguments should satisfy to be justified or defensible. Note that such argumentation semantics are, unlike the semantics of, say, standard first-order logic, not based on the notion of truth: since argumentation systems formalise reasoning that is defeasible, they are not concerned with truth of propositions, but with justification of accepting a proposition as true. In particular, one is justified in accepting a proposition as true if there is an argument for the proposition that one is justified in accepting. Argument-based semantics specify the conditions for when this is the case.

The abstract framework was introduced by Dung (1995). Historically, it came after the development of a number of more concrete argumentation systems, such as the systems of Pollock (1987)−(1994) and Vreeswijk (1993a) (which are both predecessors of the framework to be discussed in Chapter 4). Nevertheless, Dung's article is by now widely regarded as seminal. It was a breakthrough in several ways. Firstly, it contains a general account of argumentation semantics, applicable to all systems that instantiate his framework. Secondly, it made a precise comparison possible between different systems by translating them into his abstract format. Third, it made a general study of formal properties of systems possible, which are inherited by all systems that instantiate his framework. Finally, all this applies not just to argumentation systems but also to other nonmonotonic logics, since Dung (1995) showed for several such logics how they can be translated into his abstract framework. In Section 2.6 we shall discuss his argument-based reconstruction of default logic.

2.1 The status of arguments: preliminary remarks

We now start the discussion of abstract argument-based semantics. As explained above, the task of argument-based semantics is to specify the conditions under which it is justified to accept an argument. These conditions assume an 'input' set of arguments,

ordered by a binary relation of 'defeat'.[1] The framework is as abstract as possible, leaving both the structure of arguments and the grounds for defeat unspecified.

With Dung (1995) we shall call the input of the framework an 'abstract argumentation framework (sometimes 'argumentation framework' for short), abbreviated as AF.

Definition 2.1.1 [Abstract argumentation frameworks.]

1. An *abstract argumentation framework* (AF) is a pair $\mathcal{A}, \mathcal{D}$, where $\mathcal{A}$ is a set of arguments, and $\mathcal{D}$ a binary relation of defeat on $\mathcal{A}$.
2. We say that a set S of arguments defeats an argument A iff some argument in S defeats A; and S defeats a set S' of arguments iff it defeats a member of S'.

As for applications of the framework, one might think of the set $\mathcal{A}$ as all arguments that can be constructed in a given logic from a given set of premises (although this is not always the case: the framework equally applies to cases where just some of the constructible arguments are constructed). Unless stated otherwise, we shall below implicitly assume an arbitrary but fixed argumentation framework. Recall that we read 'A defeats B' in the weak sense of 'A conflicts with B and is not weaker than B'; so in some cases it may happen that A defeats B and B defeats A. If A defeats B, then if B does not defeat A we say that A *strictly defeats* B, otherwise A *weakly defeats* B.

Let us now concentrate on the task of defining the notion of a justified argument. Which properties should such a definition have? A simple definition is the following.

Definition 2.1.2 Arguments are either justified or not justified.

1. An argument is *justified* iff all arguments defeating it (if any) are not justified.
2. An argument is *not justified* iff it is defeated by an argument that is justified.

This definition works well in simple cases, in which it is clear which arguments should emerge victorious, as in the following example.

Example 2.1.3 Consider three arguments A, B and C such that B defeats A and C defeats B:

$$A \longleftarrow B \longleftarrow C$$

A concrete version of this example is

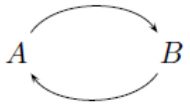
$A =$ 'Tweety flies because it is a bird'
$B =$ 'Tweety does not fly because it is a penguin'
$C =$ 'The observation that Tweety is a penguin is unreliable'

C is justified since it is not defeated by any other argument. This makes B not justified, since B is defeated by C. This in turn makes A justified: although A is defeated by B, A is reinstated by C, since C makes B not justified.

In other cases, however, Definition 2.1.2 is circular or ambiguous. In particular when arguments of equal strength interfere with each other, it is unclear which argument should remain undefeated.

[1] Dung (1995) uses the term 'attack', but to maintain uniformity throughout this text, we shall use 'defeat'.

Example 2.1.4 (Even cycle.) Consider the arguments A and B such that A defeats B and B defeats A.



A concrete example is

$A =$ 'Nixon was a pacifist because he was a quaker'
$B =$ 'Nixon was not a pacifist because he was a republican'

Can we regard A as justified? Yes, we can, if B is not justified. Can we regard B as not justified? Yes, we can, if A is justified. So, if we regard A as justified and B as not justified, Definition 2.1.2 is satisfied. However, it is obvious that by a symmetrical line of reasoning we can also regard B as justified and A as not justified. So there are two possible 'status assignments' to A and B that satisfy Definition 2.1.2: one in which A is justified at the expense of B, and one in which B is justified at the expense of A. Yet intuitively, we are not justified in accepting either of them.

In the literature, two approaches to the solution of this problem can be found. The first approach consists of changing Definition 2.1.2 in such a way that there is always precisely one possible way to assign a status to arguments, and which is such that with 'undecided conflicts' as in our example both of the conflicting arguments receive the status 'not justified'. The second approach instead regards the existence of multiple status assignments not as a problem but as a feature: it allows for multiple assignments and defines an argument as 'genuinely' justified if and only if it receives this status in all possible assignments. The following two sections discuss the details of both approaches.

First, however, another problem with Definition 2.1.2 must be explained, having to do with self-defeating arguments.

Example 2.1.5 (Self-defeat.) Consider an argument L, such that L defeats L (Figure 2.1). Suppose L is not justified. Then all arguments defeating L are not justified, so by clause 1 of Definition 2.1.2 L is justified. Contradiction. Suppose now L is justified. Then L is defeated by a justified argument, so by clause 2 of Definition 2.1.2 L is not justified. Contradiction.

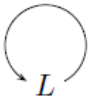


Figure 2.1: A self-defeating argument.

Thus, Definition 2.1.2 implies that there are no self-defeating arguments. Yet in ordinary discourse examples of self-defeating arguments can be found, as in the following example.

Example 2.1.6 (The Liar.) An elementary self-defeating argument can be fabricated on the basis of the so-called *paradox of the Liar*. There are many versions of this paradox. The one we use here, runs as follows:

> Dutch people can be divided into two classes: people who always tell the truth, and people who always lie. Hendrik is Dutch monk, and from Dutch monks we know that they tend to be consistent truth-tellers. Therefore, it is reasonable to assume that Hendrik is a consistent truth-teller. However, Hendrik *says* he is a liar. Is Hendrik a truth-teller or a liar?

The Liar-paradox is a paradox, because either answer leads to a contradiction.

1. Suppose that Hendrik tells the truth. Then what Hendrik says must be true. So, Hendrik is a lier. Contradiction.
2. Suppose that Hendrik lies. Then what Hendrik says must be false. So, Hendrik is not a lier. Because Dutch people are either consistent truth-tellers or consistent liers, it follows that Hendrik always tells the truth. Contradiction.

From this paradox, a self-defeating argument L can be made out of (1):

$$\dfrac{\dfrac{\text{Hendrik says: ``I lie''} \qquad \dfrac{\text{Dutch monks tend to be consistent truth-tellers} \qquad \text{Hendrik is a Dutch monk}}{\text{Hendrik is a consistent truth-teller}}}{\text{Hendrik lies}}}{\text{Hendrik is } \mathbf{not} \text{ a consistent truth-teller}}$$

If the argument for "Hendrik is *not* a consistent truth-teller" is as strong as its subargument for "Hendrik is a consistent truth-teller," then L defeats one of its own subarguments, and thus is a self-defeating argument.

In conclusion, the treatment of self-defeating arguments deserves special attention. Below we shall discuss for each particular semantics how it deals with self-defeat.

2.2 The unique-status-assignment approach

We now discuss an approach that changes Definition 2.1.2 in such a way that there is always precisely one possible way to assign a status to arguments. This 'unique-status-assignment' approach can best be explained by the way it formalises 'reinstatement' (see above, Section 1.2). It does so by combining a notion of *acceptability* with a fixed-point operator. Recall that an argument that is defeated by another argument can only be justified if it is reinstated by a third argument, viz. by a justified argument that defeats its defeater. Part of this idea is captured by the notion of *acceptability* (which, by the way, is also relevant for the multiple-status-assignments approach, as we shall see below in Section 2.3).

Definition 2.2.1 [Acceptability.] An argument A is *acceptable* with respect to a set S of arguments iff each argument defeating A is defeated by S. When A is acceptable with respect to S, we also say that S *defends* A.

The arguments in S can be seen as the arguments capable of reinstating A in case A is defeated. To illustrate acceptability, consider again Example 2.1.3: A is acceptable with respect to $\{C\}$, $\{A, C\}$, $\{B, C\}$ and $\{A, B, C\}$, but not with respect to $\varnothing$ and $\{B\}$.

The notion of acceptability is not yet sufficient. Consider in Example 2.1.4 the set $S = \{A\}$. It is easy to see that A is acceptable with respect to S, since all arguments defeating A (viz. B) are defeated by an argument in S, viz. A itself. Clearly, we do not want that an argument can reinstate itself, and this is the reason why, to obtain a unique status assignment, a fixed-point operator must be used.

Intermezzo: fixed point operators Below we need some basics on fixed-point operators. Let S be a set and $O : Pow(S) \longrightarrow Pow(S)$ be an operator which for any subset of S returns a subset of S. $T \subseteq S$ is a *fixed point* of O iff $O(T) = T$. It is known that if O satisfies certain properties, it has a *least fixed point*, i.e. a fixed point which is a subset of all other fixed points of O. The most important of these properties is monotonicity, which is that $O(T) \subseteq O(T')$ whenever $T \subseteq T'$.

Consider now the following operator, which for each set of arguments returns the set of all arguments that are acceptable to it.

Definition 2.2.2 [Grounded semantics.] Let AF be an abstract argumentation framework, and let $S \subseteq \mathcal{A}_{AF}$. Then the operator F^{AF} is defined as follows:

- $F^{AF}(S) = \{A \in \mathcal{A}_{AF} \mid A \text{ is acceptable with respect to } S\}$

The *grounded extension* of AF is defined as the least fixed point of F^{AF}.

It can be shown that the operator F has a least fixed point, so that the notion of a grounded extension is well-defined[2]. (The basic idea is that if an argument is acceptable with respect to S, it is also acceptable with respect to any superset of S, so that F is monotonic.) Self-reinstatement can then be avoided by defining the set of justified arguments as that least fixed point. Note that in Example 2.1.4 the set $\{A\}$ and $\{B\}$ are fixed points of F but not its least fixed point, which is the empty set. In general we have that if no argument is undefeated, then $F(\varnothing) = \varnothing$.

These observations allow the following definition of a justified argument.[3]

Definition 2.2.3 [Justified arguments in grounded semantics.] An argument is *justified* with respect to grounded semantics iff it is a member of the grounded extension.

In applying these definitions, it is useful to know that the least fixed point of F can be approximated, and under certain conditions even obtained, by iterative application of F to the empty set.

Proposition 2.2.4 Dung (1995) Consider the following sequence of arguments.

[2] Below the superscript of F will usually be omitted.

[3] Henceforth, the definitions in this and the next chapter will, unless specified otherwise, impicitly assume an arbitrary but fixed abstract argumentation framework.

- $F^0 = \varnothing$
- $F^{i+1} = \{A \in \mathcal{A} \mid A \text{ is acceptable with respect to } F^i\}$.

Let $F^\omega = \cup_{i=0}^{\infty}(F^i)$. The following observations hold.

1. All arguments in F^ω are justified.
2. If each argument is defeated by at most a finite number of arguments, then an argument is justified iff it is in F^ω.

Note that if the condition of (2) does not hold, it is possible that $F^\omega \subset F(F^\omega)$.

In the iterative construction of the set of justified arguments first all arguments that are not defeated by any argument are added, and at each further application of F all arguments that are reinstated by arguments that are already in the set are added. This is achieved through the notion of acceptability. To see this, suppose we apply F for the ith time: then for any argument A, if all arguments that defeat A are themselves defeated by an argument in F^{i-1}, then A is in F^i.

It is instructive to see how this works in Example 2.1.3. We have that

$F^1 = F(\varnothing) = \{C\}$
$F^2 = F(F^1) = \{A, C\}$
$F^3 = F(F^2) = F^2$

The following example, with an infinite chain of defeat relations, provides another illustration.

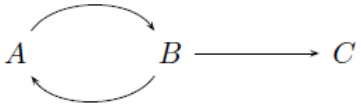
Example 2.2.5 Consider an infinite chain of arguments $A_1, \ldots, A_n, \ldots$ such that A_1 is defeated by A_2, A_2 is defeated by A_3, and so on.

The least fixed point of this chain is empty, since no argument is undefeated. Consequently, $F(\varnothing) = \varnothing$. Note that this example has two other fixed points, which also satisfy Definition 2.1.2, viz. the set of all A_i where i is odd, and the set of all A_i where i is even.

Defensible arguments

Definition 2.2.3 allows a distinction between two types of arguments that are not justified. Consider first again Example 2.1.3 and observe that, although B defeats A, A is still justified since it is reinstated by C. Consider next the following extension of Example 2.1.4.

Example 2.2.6 (Zombie arguments.) Consider three arguments A, B and C such that A defeats B, B defeats A, and B defeats C.



A concrete example is

$A =$ 'Dixon is no pacifist because he is a republican'
$B =$ 'Dixon is a pacifist because he is a quaker, and he has no gun because he is a pacifist'
$C =$ 'Dixon has a gun because he lives in Chicago'

According to Definition 2.2.3, neither of the three arguments are justified. For A and B this is since their relation is the same as in Example 2.1.4, and for C this is since it is defeated by B. Here a crucial distinction between the two examples becomes apparent: unlike in Example 2.1.3, B is, although not justified, not defeated by any justified argument and therefore B retains the potential to prevent C from becoming justified: there is no justified argument that reinstates C by defeating B. Sometimes arguments like B are called 'zombie arguments': B is not 'alive', (i.e., not justified) but it is not fully dead either; it has an intermediate status, in which it can still influence the status of other arguments.

We shall call the intermediate status of zombie arguments 'defensible'. In the unique-status-assignment approach it can be defined as follows.

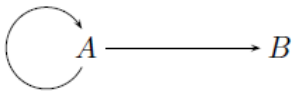
Definition 2.2.7 [Overruled and defensible arguments in grounded semantics.] With respect to grounded semantics, an argument is:

- *overruled* iff it is not justified, and defeated by a justified argument;
- *defensible* iff it is not justified and not overruled.

Self-defeating arguments

How does Definition 2.2.2 deal with self-defeating arguments? Consider the following extension of Example 2.1.5.

Example 2.2.8 Consider two arguments A and B such that A defeats A and A defeats B.

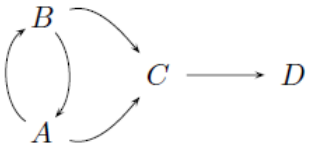


We have that $F(\varnothing) = \varnothing$, so neither A nor B are justified. Moreover, they are both defensible, since they are not defeated by any justified argument. At first sight, it might be thought that this is undesired since it would seem that self-defeating arguments should always be overruled. However, in Chapter 4 we will see that that things are more subtle and that a proper analysis of self-defeating arguments can only be given if the internal structure of arguments is made explicit.

Unique status assignments: problems

We have seen that the unique-assignment approach can be formalised in a mathematically elegant way, and that it produces intuitive results in many cases. However, there are also problems, in particular with examples of the following kind.

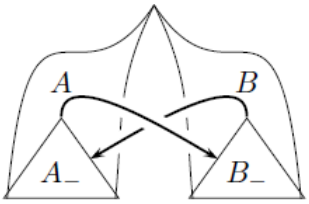
Example 2.2.9 (Floating arguments.) Consider the arguments A, B, C and D such that A defeats B, B defeats A, A defeats C, B defeats C and C defeats D.



Since no argument is undefeated, Definition 2.2.3 tells us that all of them are defensible. However, it might be argued that for C and D this should be otherwise: since C is defeated by both A and B, C should be overruled. The reason is that as far as the status of C is concerned, there is no need to resolve the conflict between A and B: the status of C 'floats' on that of A and B. And if C should be overruled, then D should be justified, since C is its only defeater.

A variant of this example is the following piece of default reasoning. To analyse this example, we must make two assumptions on the structure of arguments, viz. that they have a conclusion and that they have subarguments.

Example 2.2.10 (Floating conclusions.) Consider the arguments A^-, A, B^- and B such that A^- and B^- defeat each other and A and B have the same conclusion.



An intuitive reading is

A^-	=	Brigt Rykkje is Dutch because he was born in Holland
B^-	=	Brigt Rykkje is Norwegian because he has a Norwegian name
A	=	Brigt Rykkje likes ice skating because he is Dutch
B	=	Brigt Rykkje likes ice skating because he is Norwegian

The point is that whichever way the conflict between A^- and B^- is decided, we always end up with an argument for the conclusion that Brigt Rykkje likes ice skating, so it seems that it is justified to accept this conclusion as true, even though it is not supported by a justified argument. In other words, the status of this conclusion floats on the status of the arguments A^- and B^-.

While the unique-assignment approach is inherently unable to capture floating arguments and conclusions, there is a way to capture them, viz. by working with multiple status assignments. To this approach we now turn.

2.3 The multiple-status-assignments approach

A second way to deal with competing arguments of equal strength is to let them induce two alternative status assignments, in both of which one is justified at the expense of the other. In this approach, an argument is 'genuinely' justified iff it receives this status in all status assignments. This approach can be formalised in various ways, of which so-called stable and preferred semantics are the two best-known.

2.3.1 Stable semantics

The first way to allow for multiple status assignments, called stable semantics, is to take Definition 2.1.2 as the basis, and simply use the fact that it allows for multiple assignments. To this end, we turn this definition into one of a 'stable status assignment'.

Definition 2.3.1 [stable status assignments.]

Let $AF = (\mathcal{A}, \mathcal{D})$ be an abstract argumentation framework and In and Out two subsets of $\mathcal{A}$. Then (In, Out) is a *stable status assignment* on the basis of AF iff $In \cap Out = \varnothing$ and $In \cup Out = \mathcal{A}$ and for all $A \in \mathcal{A}$ it holds that:

1. A is *in* (that is, $A \in In$) iff all arguments defeating A (if any) are *out*.
2. A is *out* (that is, $A \in Out$) iff A is defeated by an argument that is *in*.

Note that the conditions 1 and 2 are just the conditions of Definition 2.1.2.

Definition 2.3.1 is said to define *stable* status assignments for the following reasons. Firstly, with each stable status assignment a so-called *stable argument extension* can be associated, containing all the arguments that are *in* in the status assignment.

Definition 2.3.2 [Stable argument extensions.] A set of arguments is a *stable argument extension* iff for some stable status assignment it is the set of all arguments that are assigned the status *in*.

Now stable argument extensions coincide with what Dung (1995) calls *stable extensions*. In fact, Dung gives another but equivalent definition, which uses the notion of a conflict-free set of arguments.

Definition 2.3.3 [Conflict-free sets.] A set S of arguments is *conflict-free* iff no argument in S defeats an argument in S.

Then Dung defines stable extensions as follows.

Definition 2.3.4 [Stable extensions.] A set S of arguments is a *stable extension* iff S is conflict-free and every argument that is not in S, is defeated by S.

Proposition 2.3.5 The stable argument extensions induced by Definition 2.3.1 are precisely the stable extensions defined by Definition 2.3.4.

Proof: $\Rightarrow$:

Suppose (In, Out) is a stable status assignment. To be proven:

1. In is conflict-free.

   Assume for contradiction that In contains arguments A and B such that A defeats B. Then by condition (2) of Definition 2.3.1 B is in Out. But since $In \cap Out = \varnothing$, we have that B is not in In. Contradiction. So there are no such A and B, so In is conflict-free.

2. In defeats every argument outside In.

   Since stable status assignments assign a status to all arguments in $\mathcal{A}$ and $In \cap Out = \varnothing$, every argument outside In is in Out. Then by condition (2) of Definition 2.3.1 every such argument is defeated by an argument in In.

$\Leftarrow$:

Suppose S is a stable extension. To be proven: $(S, \mathcal{A}/S)$ is a stable status assignment. Note first that by construction this is a partition of $\mathcal{A}$, so $In \cap Out = \varnothing$ and $In \cup Out = \mathcal{A}$. Then it must be verified that the two labelling conditions of Definition 2.3.1 are satisfied.

1. Condition (1) of Definition 2.3.1 is satisfied as follows. For the only if-part, if $A \in S$ then since S is conflict-free, no $B \in S$ defeats A, so all defeaters of A are in $\mathcal{A}$. For the if-part, if all defeaters of an argument A are in $\mathcal{A}$, then A cannot be in $\mathcal{A}/S$, since no defeater of A is in S. So A is in S.

2. Condition (2) of Definition 2.3.1 is satisfied as follows. For the only-if part, suppose $A \in \mathcal{A}/S$. Then since S defeats all arguments outside it, A is defeated by an argument in S. For the if-part, suppose A is defeated by an argument in S. Then since S is conflict-free, $A \in \mathcal{A}/S$. □

Below we shall use the term *stable extension* both for stable argument extensions and for Dung's stable extensions.

Example 2.1.3 has only one stable extension, viz. $\{A, C\}$, while Example 2.1.4 has two, induced by the following two status assignments:

Recall that an argumentation system is supposed to define when it is justified to accept an argument. What can we say in case of A and B in Example 2.1.4? Since both of them are *in* in one stable status assignment but *out* in the other, we must conclude that with respect to stable semantics neither of them is justified. This is captured by the following definition:

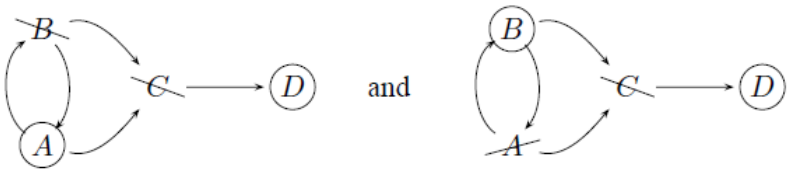
Definition 2.3.6 [Justified arguments in stable semantics.] With respect to stable semantics, an argument is *justified* iff it is *in* in all stable status assignments.

However, this is not all; just as in the unique-status-assignment approach, it is possible to distinguish between two different categories of arguments that are not justified. Some of those arguments are in no stable status assignment, but others are at least in some extensions. The first category can be called the *overruled*, and the latter category the *defensible* arguments.

Definition 2.3.7 [Overruled and defensible arguments in stable semantics.] With respect to stable semantics, an argument is:

- *overruled* iff it is *out* in all stable status assignments;
- *defensible* iff it is *in* in some but not in all stable status assignments.

It is easy to see that the unique-assignment and multiple-assignments approaches are not equivalent. Consider again Example 2.2.9. Argument A and B form an even defeat loop, thus, according to the multiple-assignments approach, either A and B can be assigned *in* but not both. So the above defeat relation induces stable two status assignments:



While in the unique-assignment approach all arguments are defensible, we now have that, while A and B are defensible, D is justified and C is overruled.

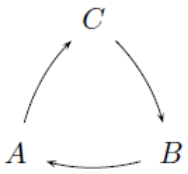
Multiple status assignments also make it possible to capture floating conclusions. Informally, this can be done by defining that a formula φ is justified as 'all extensions contain an argument for φ', rather than as 'there exists an argument for φ that is in all extensions'. In Chapter 4, in which the structure of arguments is formally defined, these alternative consequence notions for formulas will be fully formalised.

2.3.2 Preferred semantics

There is reason to discuss a second variant of the multiple-status-assignments approach. Since a stable extension is conflict-free, it reflects in some sense a coherent point of view. It is also a maximal point of view, in the sense that every possible argument is either accepted or rejected. In fact, stable semantics is the most 'aggressive' type of semantics, since a stable extension defeats every argument not belonging to it, whether or not that argument is hostile to the extension.

This feature is the reason why not all argumentation frameworks have stable extensions, as the following example shows. It contains an 'odd loop' of defeat relations.

Example 2.3.8 (Odd loop.) Let A, B and C be three arguments, represented in a triangle, such that A defeats C, B defeats A, and C defeats B.



In this situation, Definition 2.3.1 has some problems, since this example has no stable status assignments.

1. Assume that A is *in*. Then, since A defeats C, C is *out*. Since C is *out*, B is *in*, but then, since B defeats A, A is *out*. Contradiction.

2. Assume next that A is *out*. Then, since A is the only defeater of C, C is *in*. Then, since C defeats B, B is *out*. But then, since B is the only defeater of A, A is *in*. Contradiction.

Note that a self-defeating argument is a special case of Example 2.3.8, viz. the case where B and C are identical to A. This means that argumentation frameworks containing a self-defeating argument may have no stable status assignment.

To give such examples also a multiple-assignment semantics, we need allow for the possibility of *partial* status assignments.

Definition 2.3.9 [(Preferred) status assignments.] Let $AF = (\mathcal{A}, \mathcal{D})$ be an abstract argumentation framework and In and Out two subsets of $\mathcal{A}$. Then (In, Out) is a *status assignment* on the basis of AF iff $In \cap Out = \varnothing$ and for all $A \in \mathcal{A}$ it holds that:

1. A is *in* (that is, $A \in In$) iff all arguments defeating A (if any) are *out*.

2. A is *out* (that is, $A \in Out$) iff A is defeated by an argument that is *in*.

A status assignment (In, Out) is *preferred* iff it maximises the set of argument labelled *in*, that is, if there exists no status assignment (In', Out') such that $In \subset In'$.

To go back to Example 2.3.8, preferred semantics gives it a unique preferred status assignment, viz. $(\varnothing, \varnothing)$.

The notions of justified, overruled and defensible arguments defined in Definitions 2.3.6 and 2.3.7 can be easily defined also for preferred semantics, by uniformly replacing 'stable' by 'preferred'. However, in preferred semantics there are reasonable alternatives for the definitions of defensible and overruled arguments (and conclusions). This is because in each status assignment the status of an argument can be one of three kinds: *in*, *out* or undefined. Hence there are, unlike in stable semantics, situations where an argument is *in* in some but not in all assignments but yet not *out* in any assignment. Likewise, there are situations where an argument is *out* in some but not in all assignments but yet not *in* in any assignment. In the remainder of this reader we will for simplicity interpret the notions of defensible and overruled arguments as defined in Definitions 2.3.7.

To return to the notion of preferred extensions, Dung (1995) defines it not in terms of partial status assignments but with the notion of an admissible set, which in turn is defined in terms of acceptability.

Definition 2.3.10 [conflict-free and admissible sets.]

1. A set of arguments is *conflict-free* iff no argument in the set defeats an argument in the set.

2. A set of arguments S is *admissible* iff S is conflict-free and each argument in S is acceptable with respect to S.

Intuitively, an admissible set represents an admissible, or defendable, point of view. In Example 2.1.3 the sets $\varnothing$, $\{C\}$ and $\{A, C\}$ are admissible but all other subsets of $\{A, B, C\}$ are not admissible.

Definition 2.3.11 [Preferred extensions.] A conflict-free set of arguments is a *preferred extension* iff it is a maximal (with respect to set inclusion) admissible set.

There is a one-to-one correspondence between preferred status assignments and preferred extensions (cf. Caminada (2006).

Proposition 2.3.12

1. If (In, Out) is a status assignment, then In is an admissible set;

2. Let $Out(E)$ be the set of all arguments defeated by E. If E is a preferred extension, then $(E, Out(E))$ is a status assignment;

3. (In, Out) is a preferred status assignment iff In is a preferred extension.

It follows from Definition 2.3.11 that:

Proposition 2.3.13 (Dung; 1995) Every abstract argumentation framework has at least one preferred extension.

Grounded status assignments It turns out that grounded semantics can also be formulated in terms of status assignments, namely, as those assignments that minimise the set of arguments that is labelled *in*.

Definition 2.3.14 [Grounded status assignments.] A status assignment $S = (In, Out)$ is *grounded* iff there is no status assignment $S' = (In', Out')$ such that $In' \subset In$.

Proposition 2.3.15 (Caminada; 2006) S is the grounded extension of AF if and only if (S, Out) is a grounded status assignment of AF.

Self-defeat in preferred semantics Finally, how does preferred semantics deal with self-defeating arguments? It turns out that, just as in grounded semantics, self-defeating arguments can prevent other arguments from being justified. This can be illustrated with Example 2.2.8 (two arguments A and B such that A defeats A and A defeats B). The set $\{B\}$ is not admissible, so the only preferred extension is the empty set. As said above, a full analysis of self-defeat requires that the internal structure of arguments is made explicit; this will be further discussed in Chapter 4, Section 4.4.2.

2.4 Formal relations between grounded, stable and preferred semantics

We now give some results on the relation between the various semantics proven by Dung (1995).

Proposition 2.4.1 Every stable extension is preferred, but not vice versa.

Proof: It is clear that each stable extension is a preferred extension. And Example 2.2.8 shows that the reverse does not hold: the empty set is a preferred extension of this argumentation framework, but it is not stable. □

The following results are listed without proofs.

1. The grounded extension is contained in the intersection of all preferred extensions (Example 2.2.9 is a counterexample against 'equal to').

2. If an abstract argumentation framework does not give rise to infinite paths $A_1, \ldots, A_n, \ldots$ through the defeat graph such that each A_{i+1} defeats A_i then it has exactly one stable extension, which is also grounded and preferred. (Note that the even loop of Example 2.1.4 and the odd loop of Example 2.3.8 give rise to such an infinite defeat path.)

3. Finally, Dung (1995) identifies conditions under which preferred and stable semantics coincide. A necessary condition is that an abstract argumentation framework does not contain odd defeat loops.

2.5 Comparing the two approaches

How do the unique- and multiple-assignment approaches compare to each other? It is sometimes said that their difference reflects a difference between a 'skeptical' and 'credulous' attitude towards drawing defeasible conclusions: when faced with an unresolvable conflict between two arguments, a skeptic would refrain from drawing any conclusion, while a credulous reasoner would choose one conclusion at random (or both alternatively) and further explore its consequences. However, the distinction skeptical-credulous is independent of the distinction between the unique- and multiple-status-assignment approach. When deciding what to accept as a justified belief, what is important is not whether one or more possible status assignments are considered, but how the arguments are ultimately evaluated given these assignments. And this evaluation is captured by the qualifications 'justified' and 'defensible', which thus capture the distinction between 'skeptical' and 'credulous' reasoning. And since, as we have seen, the distinction justified vs. defensible arguments can be made in both the unique-assignment and the multiple-assignments approach, these approaches are independent of the distinction 'skeptical' vs. 'credulous' reasoning.

The use of skeptical reasoning (in whatever way it is formalised) is often defended by saying that since in an unresolvable conflict no argument is stronger than the other, neither of them can be accepted as justified, while the use of credulous reasoning has sometimes been defended by saying that the practical circumstances often require a person to act, whether or not s/he has conclusive reasons to decide which act to perform. In our opinion the notions of skeptical and credulous reasoning do not exclude but completement each other: whether it is better to reason skeptically or credulously may depend on the application context. For example, for a judge in a law court the reasoning about whether the suspect is guilty must clearly be skeptical, while for an intelligent software agent faced with two conflicting goals it makes sense to reason credulously, to achieve at least one of the goals.

As for their outcomes, the unique- and multiple-assignment approaches mainly differ in their treatment of floating arguments and conclusions. With respect to these examples, the question easily arises whether one approach is the right one. However, we prefer a different attitude: instead of speaking about the 'right' or 'wrong' definition, we prefer to speak of 'senses' in which an argument or conclusion can be justified. For instance, the sense in which the conclusion that Brigt Rykkje likes ice skating in Example 2.2.10 is justified is different from the sense in which, for instance, the conclusion that Tweety flies in Example 2.1.3 is justified: only in the second case is the conclusion supported by a justified argument. And the status of D in Example 2.2.9 is not quite the same as the status of, for instance, A in Example 2.1.3. Although both arguments need the help of other arguments to be justified, the argument helping A is itself justified, while the arguments helping D are merely defensible. Again it may depend on the application context which sense of justification is the best.

2.6 Argument-based reconstruction of other nonmonotonic logics

The application of Dung's abstract argumentation framework is not restricted to argument-based systems; it can also be used to reformulate other nonmonotonic logics in argument-based terms. The advantage of this is that these logics can thus be compared in terms

of a general theory: it can be systematically investigated in which respects they differ, and what the consequences are of these differences. Moreover, it becomes easier to formulate alternative versions of these logics. For instance, it is very easy to switch from one type of semantics to another.

We shall illustrate this for one of the best-known nonmonotonic logics, default logic. Our reconstruction is based on the one of Dung (1995), but somewhat deviates from it: while Dung bases his reconstruction on Reiter's original version of default logic, we base it on Antoniou's (1999) reformulation in terms of processes.

One way to reconstruct default logic in argument-based terms is by defining an argument as a finite *process* in the sense of Antoniou (1999). Recall that (informally) a process is a sequence of defaults without multiple occurrences such that the prerequisite of each default is logically implied by the union of the 'hard' knowledge W and the consequents of all preceding defaults in the sequence. A process is *closed* iff no more defaults can be appended to the sequence, and it is *successful* iff each of its assumptions is consistent with what is derived during the process. Clearly, processes as arguments do not have to be closed, since arguments are typically constructed to prove a particular conclusion. Moreover, they do not have to be successful, since unsuccessful processes correspond to self-defeating arguments.

A default theory can now be interpreted as an abstract argumentation framework as follows.

Definition 2.6.1 For any default theory $\Delta = (W, D)$, the abstract argumentation framework $AF(\Delta) = (\mathcal{A}_\Delta, \mathcal{C}_\Delta)$ is defined as follows.

- $\mathcal{A}_\Delta = \{\Pi \mid \Pi \text{ is a finite process of } \Delta\}$;
- Π *defeats*$_\Delta$ Π' iff $\varphi \in In(\Pi)$ for some $\varphi \in Out(\Pi')$.

A formula φ is a *conclusion* of an argument Π iff $\varphi \in In(\Pi)$.

Thus an argument can be defeated by deriving the negation of one of its assumptions.

Under this translation of default logic into an argumentation system, a correspondence can be proven between default logic and stable semantics. More precisely, let Δ be a default theory, and

- for any set E of formulas, let $Args(E)$ be the set of all $\Pi \in Args_\Delta$ such that for all $k \in Out(\Pi) : \{\neg k\} \cup E$ is consistent,
- for any set $S \subseteq Args_\Delta$, let $Concs(S)$ be the union of all sets $In(\Pi_i)$ such that $\Pi_i \in S$.

Then the following holds:

Proposition 2.6.2 For any default theory Δ:

1. If S is a stable extension of $AF(\Delta)$, then $Concs(S)$ is a Reiter-extension of Δ;
2. If E is a Reiter-extension of Δ, then $Args(E)$ is a stable extension of $AF(\Delta)$.

The proof of this proposition is not mandatory for this course but since it is not reported elsewhere in the literature, it is still included here. The proof uses some notation: if d is a default, then $Pre(d)$, $Jus(d)$ and $Cons(d)$ respectively denote d's prerequisite, justifications and consequent. We first prove the following lemma, which in effect says that violating the consistency check in testing applicability of a default gives rise to a defeating counterargument.

Lemma 2.6.3 If S is a stable extension of $AF(\Delta)$ and $\Pi \in S$, then:

1. all subsequences Π' of Π that are arguments are in S;
2. all arguments in S are processes.

Proof: For (1), observe that any defeater of Π' also is a defeater of Π, so is outside S; but then $\Pi' \in S$ by definition of a stable extension.

For (2), suppose $\Pi \in S$ and Π is not a process. Then for some subsequence $\Pi[i]$ of Π and $d_i \in \Pi$ the negation of some $j \in Jus(d_i)$ is in $In(\Pi[i])$. So $\Pi[i]$ defeats Π. Also, $\Pi[i] \in S$ by (1); but then S is not conflict-free. Contradiction. □

Proof: To prove (1) of Proposition 2.6.2, we first append all arguments in S into a sequence of defaults Π and delete each repeated occurrence of every default. Clearly, by Lemma 2.6.3 and conflict-freeness of S we have that Π is a process. We claim that Π is a closed and successful process.

Firstly, since S is conflict-free, it follows by definition of defeat that $In(\Pi) \cap Out(\Pi) = \varnothing$, so Π is successful.

Next, consider any default d not in Π and suppose that $Pre(d) \in In(\Pi)$. We claim that $In(\Pi) \vdash \neg k$ for some $k \in Jus(d)$. By compactness[4] of first-order logic, $Pre(d)$ is implied by some finite subset of $In(\Pi)$. With this subset a finite subprocess $\Pi[i]$ of Π can be associated. Since d is not an element of Π, we have that $\Pi[i], d$ is not a subprocess of Π. So by construction of Π we have that $\Pi[i], d \notin S$. But then since S is stable, S defeats $\Pi[i], d$ so $In(\Pi) \vdash \neg k$ for some $k \in Jus(d)$. Hence Π is closed.

Next, to prove (2), consider a closed process Π generating E and let $Args(\Pi)$ be the set of all finite processes that only use defaults from Π. Since Π is closed, we have that $Args(\Pi) = Args(E)$.

We next show that $Args(\Pi)$ is a stable extension. Conflict-freeness of $Args(\Pi)$ follows immediately from successfulness of Π. To show that $Args(\Pi)$ defeats any argument outside it, consider any such argument $A = d_1, \ldots, d_n$ and let d_i be the first default in A that is not in Π. Then since Π is closed, we have that $In(\Pi) \vdash \neg k$ for some $k \in Jus(d)$. But then by compactness of first-order logic, some argument in $Args(\Pi)$ defeats A. □

Example 2.6.4 Consider the following default theory $\Delta_1 = (W, D)$ where $W = \{p\}$ and

$$D = \left\{ d_1 : \frac{p : q \wedge r}{q}, \quad d_2 : \frac{q : s}{t}, \quad d_3 : \frac{p : u \wedge \neg t}{\neg t} \right\}$$

The argumentation framework $AF(\Delta_1)$ consists of the following arguments.

$A = \varnothing$
$B = d_1$
$C = d_1, d_2$
$D = d_3$

[4]Compactness means that if a sentence follows from an infinite set of premises, it also follows from a finite subset of these premises.

$E = d_1, d_3$
$F = d_3, d_1$
$G = d_1, d_3, d_2$
$H = d_3, d_1, d_2$

And the defeat relations are depicted in figure 2.2. This figure leaves implicit that G and H also defeat all other arguments except argument A.

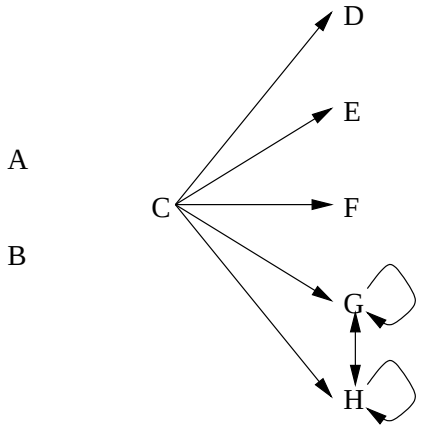


Figure 2.2: $AF(\Delta_1)$

It is easy to verify that the default theory Δ_1 has one default logic extension, viz. $Th(\{p, q, t\})$, generated by the process d_1, d_2. Correspondingly, $AF(\Delta_1)$ has a unique stable extension, viz. $\{A, B, C\}$. Note that this stable extension contains the process that generates the default logic extension of Δ_1, as well as all its subprocesses.

Example 2.6.5 Consider next a default theory $\Delta_2 = (\varnothing, \{\frac{:p}{\neg p}\})$. We know from Antoniou (1999) that this default theory has no extensions. We have that $AF(\Delta_2)$ contains two arguments, viz. $\varnothing$ and $\frac{:p}{\neg p}$. The only defeat relation is that the latter argument defeats itself. Then it is easy to see that this argumentation framework has no stable extensions.

2.7 Final remarks

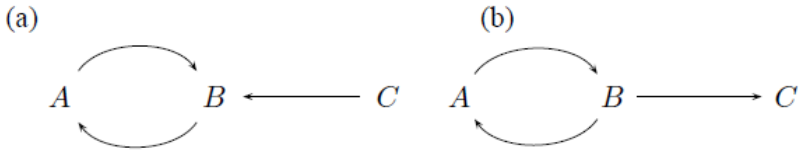
As remarked above, Dung's fully abstract approach was a major innovation in the study of defeasible argumentation, in that it provided an elegant general framework for investigating the various argumentation systems. Moreover, the framework also applies to other nonmonotonic logics, since Dung showed how several of these logics can be translated into argumentation systems. Thus it becomes very easy to formulate alternative semantics for nonmonotonic logics. For instance, default logic, which above was shown to have a stable semantics, can very easily be given an alternative semantics in which extensions are guaranteed to exist, like preferred or grounded semantics. Moreover, the proof theories that have been or will be developed for the various argument-based semantics immediately apply to the systems that are an instance of these semantics.

On the other hand, the fully abstract nature of Dung's framework also leaves much to the developers of particular systems. In particular, they have to define the internal

structure of an argument, the ways in which arguments can conflict, and the origin of the defeat relation. In the next chapter a more concrete framework will be discussed in which these elements have been defined.

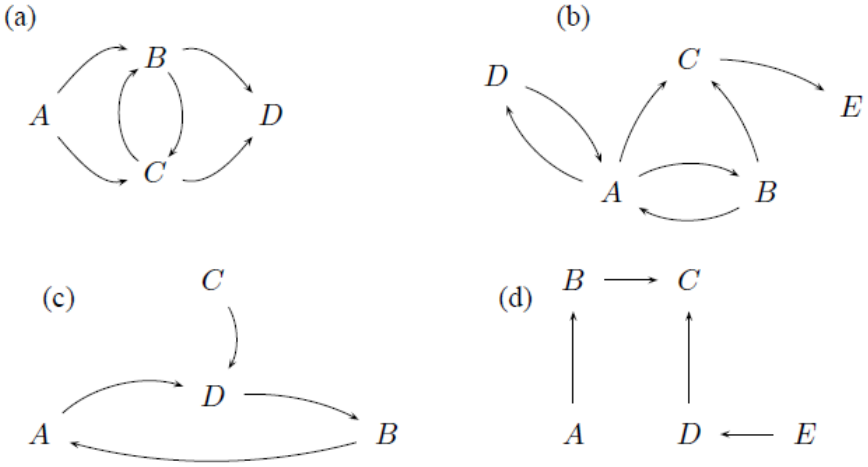
2.8 Exercises

EXERCISE 2.8.1 Determine, if possible, with Definition 2.1.2 which arguments are justified in the following two examples.



EXERCISE 2.8.2 Prove that if no argument of AF is undefeated, then $F^{AF}(\varnothing) = \varnothing$.

EXERCISE 2.8.3 Determine the grounded extension of the following defeat graphs. Show in each case its construction as in Proposition 2.2.4.



EXERCISE 2.8.4 Let

- $G(S) = \{A \in \mathcal{A} \mid A \text{ is not defeated by a member of } S\}$

1. Show that, for every set of arguments X, $F(X) = G^2(X)$ $[= G(G(X))]$.
2. Show that G is anti-monotonic. G is anti-monotonic if $A \subseteq B$ implies $G(B) \subseteq G(A)$.
3. Show on the basis of (2) that F is monotonic.
4. Let $\{G_i\}_{i \geq 0}$ be sets of arguments, such that

$$\begin{aligned} G_0 &=_{Def} \varnothing, \\ G_i &=_{Def} G(G_{i-1}). \end{aligned}$$

Show that $G_0 \subseteq G_2 \subseteq G_4 \subseteq \ldots \subseteq G_5 \subseteq G_3 \subseteq G_1$.

EXERCISE 2.8.5 Determine for each of the defeat graphs in Exercise 2.8.3 which arguments are justified, which are defensible and which are overruled, all according to grounded semantics.

EXERCISE 2.8.6 Prove that S is a stable extension iff $S = \{A \mid A \text{ is not defeated by } S\}$.

EXERCISE 2.8.7 Determine all status assignments in Examples 2.1.3, 2.1.4 and 2.3.8. Which of these assignments are maximal?

EXERCISE 2.8.8 Consider two status assignments $S = (In, Out)$ and $S' = (In', Out')$ to the same argumentation framework such that $In \subset In'$.

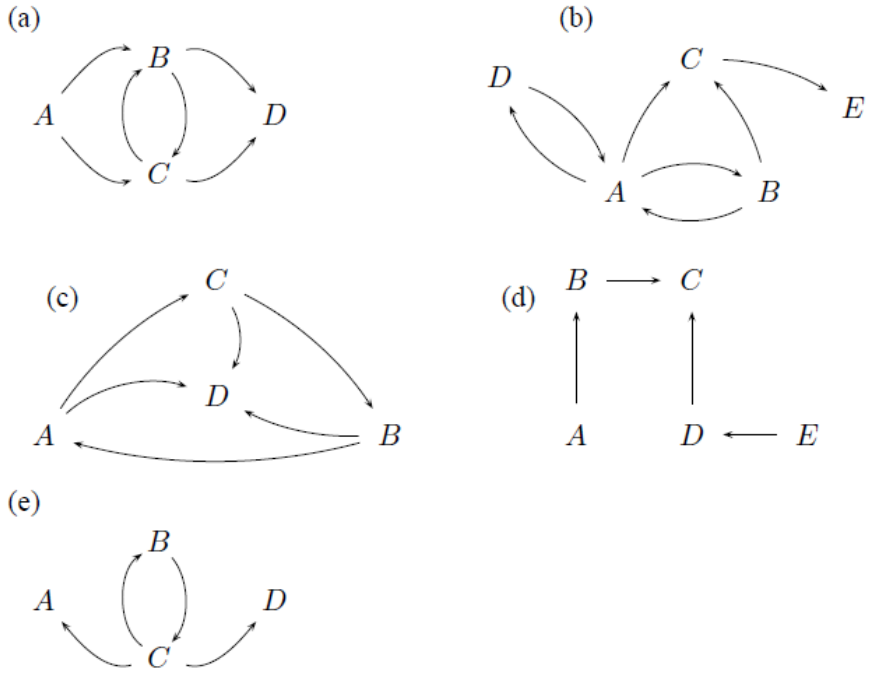
1. Does it hold that $Out \subseteq Out'$? If so, give the proof; if not, give a counterexample.
2. Does it hold that $Out \subset Out'$? Again, if so, give the proof; if not, give a counterexample.

EXERCISE 2.8.9 Give one or more alternative definitions of the notions of defensible and overruled arguments in preferred semantics. Verify for each definition whether it implies that each argument is either justified, or defensible, or overruled. If not, do you regard this as a flaw of your definition?

EXERCISE 2.8.10 Determine the admissible sets in Example 2.3.8. Which of these is or are maximally admissible?

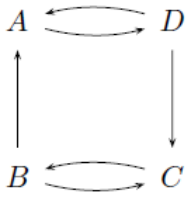
EXERCISE 2.8.11

1. Determine the preferred and stable extension(s) of the following defeat graphs.



2. Determine for each of the above defeat graphs, and with respect to each semantics, which arguments are justified, which are defensible and which are overruled.

EXERCISE 2.8.12 Consider four arguments A, B, C and D such that B strictly defeats A, D strictly defeats C, A and D defeat each other and B and C defeat each other.



Here is a natural-language version, in which the defeat relations are based on which argument uses the more specific of two conflicting defaults.

$A =$ Larry is rich because he is a public defender, public defenders are lawyers, and lawyers are rich;
$B =$ Larry is not rich because he is a public defender, and public defenders are not rich;
$C =$ Larry is rich because he lives in Hollywood, and people who live in Hollywood are rich;
$D =$ Larry is not rich because he rents in Hollywood, and people who rent in Hollywood are not rich.

1. Determine the grounded extension and the preferred extension(s) of this argumentation framework.

2. Determine in both cases which conclusions about Larry's richness are justified. Does the result agree with your intuitions?

EXERCISE 2.8.13 This exercise builds on Example 2.6.5. To see why preferred semantics can improve default logic, consider the default theory Δ_3 which is Δ_2 plus an extra default $\frac{:q}{q}$.

1. Determine the stable and preferred extensions of $AF(\Delta_3)$.

2. Explain why preferred semantics gives the better outcome.

EXERCISE 2.8.14

1. Consider a default theory $\Delta = (W, D)$ with

$$W = \varnothing$$

$$D = \left\{ \frac{:b}{a}, \frac{:e}{e}, \frac{a : c \wedge d}{c}, \frac{c : b}{b}, \frac{e : \neg a}{\neg d}, \frac{: \neg a}{\neg a} \right\}$$

and answer the following questions on the basis of the argumentation framework $AF(\Delta)$.

   (a) Construct a minimal argument for the conclusion b.

   (b) Construct all minimal arguments that defeat the argument found under (a).

(c) Is the argument found under (a) element of an admissible set?

(d) Is it in a preferred extension?

(e) Is it in the grounded extension?

EXERCISE 2.8.15 Verify that any failed finite process is a selfdefeating argument.

Chapter 3

Games for abstract argumentation semantics

So far mainly semantical aspects have been discussed, where the main focus was on characterising properties of *sets* of arguments, without specifying procedures for determining whether a given argument is a member of the set. In this chapter we shall go deeper into proof-theoretical, or procedural aspects of argumentation, where the chief concern is to investigate the status of *individual* arguments. This aspect of argumentation logics is less well-developed than its semantics; much research is ongoing or still to be carried out.

3.1 General ideas

The main question of this chapter is: given an argument from an abstract argumentation framework, how can its status be investigated? Several argumentation systems have tackled this problem in dialectical style. The common idea can be explained in terms of an argument game between two players, a proponent and an opponent of an argument. A dispute is an alternating series of moves by the two players. The proponent starts with an argument to be tested, and each following move consists of an argument that defeats (or in some cases strictly defeats) a move of the other party. The initial argument provably has a certain dialectical status if the proponent has a winning strategy, i.e., if he can win whatever moves the opponent makes.

The precise rules of the game depend on the semantics the game is meant to capture. A common winning criterion is that a player has won if s/he has made the other player run out of moves. However, other criteria are also possible. Other aspects on which choices have to be made are:

- Must moves strictly defeat their target or can they be weakly defeating?
- May moves be repeated?
- May players backtrack?
- May players defeat or be defeated by their own earlier moves?

These choices have to be made independently for both sides.

A natural idea in dialectical proof theories is that of dialectical asymmetry. The players of an argument game have different objectives: proponent wants to build a (dialectical) proof, while opponent wants to prevent proponent from doing so. In other

words, while proponent is constructive, opponent is destructive, and this leads to different rules for the two players. Moreover, the burden induced by these rules will be heavier for one player than for the other. Which player has the heavier burden depends on whether the reasoning is credulous or skeptical: in skeptical reasoning the heavier burden is on proponent, while in credulous reasoning it is on opponent.

Let us now make these informal observations more precise. A dialectical proof theory takes the form of an argument game regulating a *dispute* between two *players*, the proponent P and opponent O of an argument. If p is a player, then $\overline{p}$ denotes the other player. The players *move* alternatingly, moving one argument at each turn. The game has a *protocol* function for determining *legality* of moves, by defining at each point in a dispute which arguments can be moved. Finally, a *winning criterion* is a partial function that determines the winner of a dispute, if any. If one player wins, the other player loses, so the argument game is a so-called zero-sum game.

These notions are formally defined as follows (recall that, unless stated otherwise, we implicitly assume an arbitrary but fixed argumentation framework).

Definition 3.1.1 [Moves, disputes and protocols.] Given an argumentation framework $AF = (\mathcal{A}, \mathcal{D})$ we define the following notions.

- The set M of *moves* consists of all pairs (p, A) such that $p \in \{P, O\}$ and $A \in \mathcal{A}$; for any move (p, A) in M we denote p by $pl(m)$ and A by $s(m)$.
- The set of $M^{\leq\infty}$ of *disputes* is the set of all sequences from M and the set $M^{<\infty}$of *finite disputes* is the set of all finite sequences from M.
- A *protocol* is a function that specifies the *legal moves* at each stage of a dispute. Formally, a protocol is a function Pr with domain a nonempty subset D of $M^{<\infty}$ taking subsets of M as values. That is:
  - $Pr : D \longrightarrow Pow(M)$

  such that $D \subseteq M^{<\infty}$. The elements of D are called the *legal finite disputes*. The elements of $Pr(d)$ are called the moves allowed after d. If d is a legal dispute and $Pr(d) = \varnothing$, then d is said to be a *terminated* dispute. Pr must satisfy the following conditions for all finite disputes d and moves m:
  1. $d \in D$ and $m \in Pr(d)$ iff $d, m \in D$;
  2. if $m \in Pr(d)$ then $pl(m) = P$ if d is of even length, otherwise $pl(m) = O$.
- A *winning function* is a partial function of type $W : D \longrightarrow \{P, O\}$.

The crucial elements of this definition are the protocol and the winning criterion. Dialectical proof theories differ only on these two elements.

We now define an abstract game-theoretic notion of defeasible provability, which is the same for all dialectical proof theories. It is defined in terms of the notion of a strategy. A strategy for a player in a dispute game has the form of a tree of disputes that for each possible move of the other player specifies a unique reply.

Definition 3.1.2 [Strategies.]

1. A *strategy* for player p is a tree of disputes only branching after p's moves, and containing all legal replies of $\overline{p}$.

2. A strategy for p is *winning* iff p wins all disputes in the strategy.

If the winning criterion is that the other player has no legal moves, then it is easy to see that a winning strategy for a player is a strategy in which all branches end with a move by that player.

Defeasible provability is now defined as follows, parametrised by a protocol X.

Definition 3.1.3 [Provability.] An argument A is *defeasibly provable in the X-game* iff the proponent has a winning strategy in a dispute with as root the argument A that satisfies protocol X.

3.2 Dialectics for grounded semantics

In this section we discuss a proof theory for determining whether an argument is in the grounded extension of a given argumentation framework. Since a grounded extension only contains justified arguments, the dialectical asymmetry favours the opponent: her moves are allowed to be simply defeating[1], while proponent's moves must be strictly defeating. Moreover, the proponent is not allowed to repeat his arguments. Finally, backtracking is not allowed for both players.

Definition 3.2.1 [Proof theory for grounded semantics.] A dispute satisfies the *G-game* protocol iff it satisfies the following conditions.

1. Moves are legal iff in addition to Definition 3.1.1 they satisfy the following conditions.
   (a) Proponent does not repeat his moves; and
   (b) Proponent's moves (except the first) strictly defeat opponent's last move; and
   (c) Opponent's moves defeat proponent's last move.
2. A player wins a dispute iff the other player has no legal moves.

A dispute satisfying the protocol of the G-game is called a G-dispute.

Example 3.2.2 Let A, B, C and D be arguments such that B and D defeat A, and C defeats B. Then a G-dispute on A may run as follows:

P: A, O: B, P: C

In this dispute P attempts to show A justified. Both B and D defeat A, which means that O has two choices in response to A. O chooses to respond with B in the second move. Then C is the only argument defeating B, so that P has no choice than to respond with C in the third move. There are no arguments against C, so that O cannot move and loses the dispute.

However, this outcome is not inevitable for O; her loss was merely caused by her weak play. A dispute in which O follows an optimal strategy is

P: A, O: D

[1] When below we say that move m defeats move m' we mean that $s(m)$ defeats $s(m')$.

And P has no reply, so O wins. Concluding, in this example P has no winning strategy. The only reason why P wins the first dispute is that O chooses the wrong argument, viz. B, in response to A. In fact, O is in the position to win every game, provided it chooses the right moves. In other words, O possesses a winning strategy.

Example 3.2.3 To give an another example, consider two strategies for P as depicted in Figure 3.1. The tree on the left is based on an argumentation framework AF_1 with $\mathcal{A} = \{A, B, C, D, E, F, G\}$ and $\mathcal{D}$ as shown by the arrows. Here P has a winning strategy, since in all disputes O eventually runs out of moves; so argument A is provable on the basis of AF_1. The tree on the right is based on an extension of AF_1 into AF_2 by adding H, I and J to $\mathcal{A}$ and adding new defeat relations corresponding to the new arrows (the extension is shown inside the dotted box). This is not a winning strategy for P, since one dispute ends with a move by O; so (assuming P has no better strategy) A is not provable on the basis of AF_2.

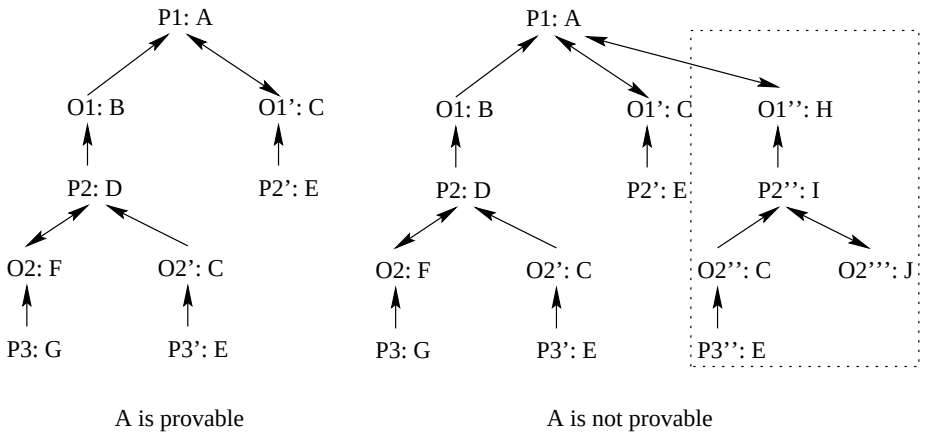


Figure 3.1: Two trees of proof-theoretical disputes.

Some words are in order on the non-repetition requirement of Definition 3.2.1 (condition 1a). This requirement does not change provability of any argument, since O will have a reply the second time iff she had a reply the first time. However, it avoids infinite disputes if $\mathcal{A}$ is finite, which is especially convenient for computational purposes. The same holds for the condition that P's arguments are strictly defeating; allowing them to be simply defeating does not change provability, but it avoids certain infinite disputes.

As for the relation between grounded semantics and its proof theory, the following proposition holds.

Proposition 3.2.4 [Soundness and completeness of the G-game.] An argument is in the grounded extension of an AF iff it is defeasibly provable on the basis of AF in the G-game.

Proof: (Sketch). We give a sketch of the proof for finitary AF's. Without this restriction the proof is more complicated. The restriction makes sense for computational purposes, since saying that an AF is finitary is equivalent to saying that each strategy based on AF has at most a finite number of branches.

$\Leftarrow$ (soundness):
Assume that P has a winning strategy W for A. Clearly, all of W's leaves A_n are in F^1, since they have no defeaters. But then in every branch of W, A_{n-2} is acceptable with respect to F^1 and so is in F^2. This can be repeated until the root of W is reached. □

$\Rightarrow$ (completeness):
Suppose A is in the grounded extension of AF. Then, since AF is finitary, there is a least number i such that $A \in F^i$. Then P has the following winning strategy if he begins a dispute with A. For each argument B defeating A moved by O, P can choose one argument C from F^{i-1} that strictly defeats B. This can be repeated for each argument defeating C, and so on, until P can choose an argument from F^1, which has no defeaters, so O has no legal reply. □

Note that completeness here does not imply semi-decidability (a logic is semi-decidable iff there exists an algorithm that can produce any provable formula): if the logic for constructing individual arguments is not decidable, then the search for counterarguments is in general not even semi-decidable, since this search is essentially a consistency check.

This completes the discussion of the dialectical proof theory for grounded semantics. We now turn to a dialectical proof theory for credulous reasoning, in particular for preferred semantics.

3.3 Dialectics for preferred semantics

In this section we present the so-called P- game[2], which serves as a credulous proof theory for preferred semantics, and was developed by Vreeswijk and Prakken (2000). For notational convenience we now denote defeat relations with $\leftarrow$. Throughout this section we will use the following example.

Example 3.3.1 The pair $\mathcal{A} = \langle X, \leftarrow \rangle$ with arguments

$$X = \{a, b, c, d, e, f, g, h, i, j, k, l, m, n, p, q\}$$

and $\leftarrow$ as indicated in Figure 3.2 is an example of an abstract argumentation framework. It accommodates a number of interesting cases, and will therefore be used as a running example throughout this chapter.

3.3.1 The basic ideas illustrated

Example 3.3.1 gives us some useful clues as to which features the argument game for preferred semantics should have. We are interested in credulous reasoning, so in testing membership of *some* extension. The argument game is based on the following idea. By definition, a preferred extension is a $\subseteq$-maximal admissible set. It is known that each admissible set is contained in a maximal admissible set (see the proof of Proposition 2.3.13), so the procedure comes down to trying to construct an admissible set 'around' the argument in question. If this succeeds, we know that the admissible set and hence the argument in question is contained in a preferred extension.

[2]The P in 'P-game should not be confused with the P denoting proponent.

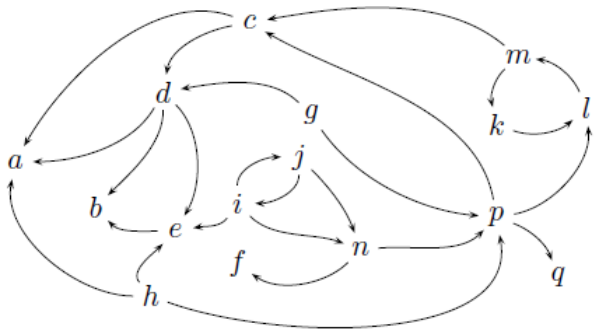
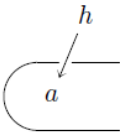


Figure 3.2: Defeat relations in the running example.

Suppose now we wish to investigate whether a is preferred, i.e., belongs to a preferred extension. We know that it suffices to show that the argument in question is admissible. The idea is to start with $S = \{a\}$ and, if a has defeaters, to find other arguments in order to complete S into an admissible set.

Example 3.3.2 (Straight failure). Consider the argument system of Figure 3.2, and suppose that P's task is to show that a is preferred. The first action of P is simply putting forward a:

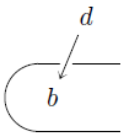
If a cannot be defeated, then $S = \{a\}$ is admissible, and P succeeds. However, since $a \leftarrow h$, O forwards h:



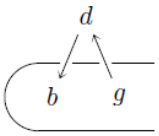
Now it is up to P to defend a by finding arguments against h. There are no such arguments, so that P fails to construct an admissible set 'around' a. So a is not admissible, hence not preferred.

Example 3.3.3 (Straight success). Suppose that P wants to show that b is admissible. The first action of P is putting forward b:

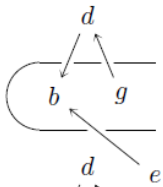
O defeats b with d:



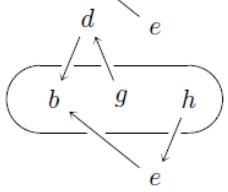
P defends this attack with g:



Since O 's attack on b with d has failed, O returns to b and defeats it again, this time with e:



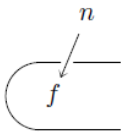
P defends b again, this time with h. Since O is unable to find other arguments against b, g or h, P may now close S:



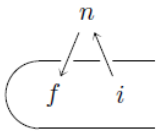
Example 3.3.4 (Even loop success). Suppose that P wants to show that f is admissible.

The first action of P is putting forward f:

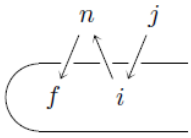
O defeats f with n:



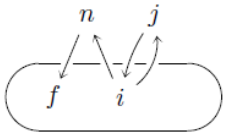
P defends this attack with i:



O defeats i with j:



P defends i with i itself (so that i is self-defending). O is unable to put forward other arguments that defeat f or i so that P closes S:

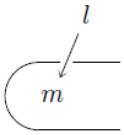


This example shows that P must be allowed to repeat his arguments, while O must be forbidden to repeat O 's arguments (at least in the same 'line of dispute'; see further below).

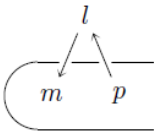
Example 3.3.5 (Odd loop failure). Suppose that P wants to show that m is admissible.

The first action of P is putting forward m:

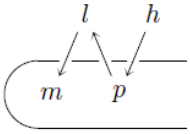
O defeats m with l:



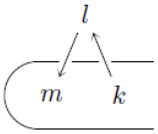
P defends this attack with p:



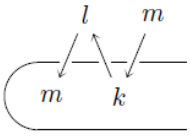
O defeats p with h:



P backtracks and removes p from S. He then tries to defend l with k instead:

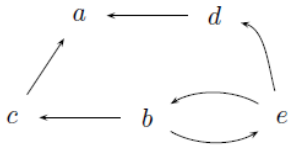


O defeats k with m (and, as a bonus, introduces an inconsistency in S):



P has no other arguments in response to l and m, so that he is unable to close S into an admissible set. So m is not contained in an admissible set. Note that we cannot allow P to reply to m with l, since otherwise the set that P is constructing 'around' m is not conflict-free, hence not admissible. So we must forbid P to repeat O 's moves. On the other hand, this example also shows that O should be allowed to repeat P 's moves, since such a repetition reveals a conflict in P 's position.

Example 3.3.6 (The need for backtracking). The next feature of our argument game is not illustrated by Figure 3.2 so we need a new example. Consider an argument system with five arguments a, b, c, d and e and defeat relations as shown in the graph.



This example shows that we must allow O to backtrack. Suppose P starts with a, O defeats a with d, and P defends a with e. If O now defeats e with b, P can defend e by repeating e itself. However, O can backtrack to a, this time defeating it with c, after which P can only defend a with b which repeats O, and in Example 3.3.5 we concluded that P must be forbidden to do so. So by backtracking O can reveal that P's position is not conflict-free.

Repetition

Let us summarise our observations about repetition of moves.

i. It makes sense for P to repeat himself (if possible), because O might fail to find or produce a new defeater of P's repeated argument. If so, then P's repetition closes a cycle of even length, of which P's arguments are admissible.

ii. It makes sense for O to repeat P (if possible), because thus she shows that P's collection of arguments is not conflict-free.

iii. P must not repeat O, because doing so would introduce a conflict into P's own collection of arguments.

iv. O must not repeat herself, because P has already shown to have adequate defense for O's previous arguments.

3.3.2 The P-game defined

We now turn to the formal definition of the argument game for preferred semantics. Let us fix some terminology.

- A *dispute line* is a dispute without backtracking moves.
- An *eo ipso* (meaning: "you said it yourself") is a move that uses a previous argument of the other player.

Definition 3.3.7 [A proof theory for preferred semantics.] A dispute satisfies the *P-game* protocol iff satisfies the following conditions.

1. Moves are legal iff in addition to Definition 3.1.1 they satisfy the following conditions.
   (a) A move by P responds to the previous move by O.
   (b) A move by O responds to some earlier move by P.
   (c) A move defeats the argument to which it responds.
   (d) P does not repeat O's moves.
   (e) O does not repeat O's moves in the same dispute line.
   (f) No two responses to the same move have the same content.
2. O wins a dispute iff she does an *eo ipso* or makes P run out of legal moves; otherwise P wins.

A dispute satisfying the rules of the P-game is called a P-dispute.

Note that an infinite dispute is won by P.

Since the P-game allows O to backtrack, during a P-dispute a tree of dispute lines is constructed. (By contrast, a G-dispute consists of only one dispute line, since in a G-dispute each argument replies to the immediately preceding move in the dispute.) Accordingly, there are two ways to display a P-dispute: as a *linear* structure, in the order in which the arguments are moved, and as a *tree* structure, where the edges indicate to which argument an argument replies. The reader should not confuse the tree form of

a single dispute with the tree form of a strategy: in the latter tree (cf. Definition 3.1.2) an edge between two arguments indicates that the child argument is moved immediately after the parent argument; in other words, each branch of a strategy tree is a complete dispute, possibly with backtracking moves, but displayed in linear form.

Proposition 3.3.8 [Soundness and completeness of the P-game.] An argument is in some preferred extension of an AF iff it is defeasibly provable on the basis of AF in the P-game

Proof: (Below we say that an argument a is *defended* in a dispute iff the dispute begins with a and is won by P.) By definition of preferred extensions it suffices to show that an argument is admissible iff it can be defended in every dispute.

First suppose that a can be defended in every dispute. This includes disputes in which O has opposed optimally. Let us consider such a dispute. Let A be the arguments that P used to defend a. (in particular $a \in A$.) If A is not conflict-free then $a_i \leftarrow a_j$ for some $a_i, a_j \in A$, and O would have done an *eo ipso*, which is not the case. If A is not admissible, then $a_i \leftarrow b$ for some $a_i \in A$ while $b \not\leftarrow A$. In that case, O would have used b as a winning argument, which is also not the case. Hence A is admissible.

Conversely, suppose that $a \in A$ with A admissible. Now P can win every dispute by starting with a, and replying with arguments from A only. (P can do this, because all arguments in A are acceptable wrt A.) As long as P picks his arguments from A, O cannot win by *eo ipso*, because A is conflict-free. So a can be defended in dispute.□

Finally, a drawback of the P-game is that in some cases proofs have to be infinite. This is obvious when an argument has an infinite number of defeaters, but even otherwise some proofs are infinite, as in the case of Example 2.2.5. Nevertheless, it is easy to verify that with a finite set of arguments all proofs are finite.

3.4 A simplification of the P-game

Applying the P-game as defined above can be quite complex, since it combines two kinds of trees: the tree of reply relations within a single P- game and the game tree in the game-theoretical sense, that is, the tree of all possible ways in which a game about a given argument can be played. Fortunately, a simplification is possible, since Wu (2012) has proved that the proponent has a winning strategy in the P-game just in case there exists a terminated game won by the proponent. Here 'terminated' means that the player to move cannot move any further legal move. Note that infinite games can also be terminated in this sense. The intuition behind this result is that since the opponent can freely backtrack in a single game, a single terminated game will already contain all possible ways the opponent can attack the proponent's arguments.

3.5 Exercises

EXERCISE 3.5.1 Consider an argumentation framework with the arguments $\{A-G\}$ and the following defeat relations: A and B defeat each other, E and G defeat each other, C defeats B, D defeats A, E defeats D, and F defeats D.

1. Draw the defeat graph.

2. Determine all strategies for P and O in a game for A according to grounded semantics. Indicate which of these strategies are winning.

EXERCISE 3.5.2

1. Change Definition 3.2.1 to the effect that the non-repetition rule is dropped, and P's arguments are allowed to be simply defeating. Give a dispute that is finite under the original definition but infinite under the new definition.

2. Answer the same question for the case that only the non-repetition rule is dropped.

3. Give a dispute that is infinite under the original definition.

EXERCISE 3.5.3

1. Investigate for the following arguments in Exercise 2.8.3 whether they can be proven justified with respect to grounded semantics. For each provable argument, give a winning strategy for P. For each argument that is not provable, show why P's strategies fail.

   (a) In (a): investigate A, B and D.

   (b) In (b): investigate C and E.

   (c) In (c): investigate A, B and C.

   (d) In (d): investigate C.

2. Answer the same question about defeat graph (e) of Exercise 2.8.11, for the arguments C and D.

3. For each argument under 1 that is provable, compare the structure of P's winning strategy with the construction of the grounded extension that you found in Exercise 2.8.3. How are they related?

EXERCISE 3.5.4 This exercise is a continuation of Exercise 2.8.14. Investigate whether the argument for b that you constructed in that exercise, is defeasibly provable in the G-game. If so, give a winning strategy for P.

EXERCISE 3.5.5 Verify that a proof in the P-game of A_1 in Example 2.2.5 has to be infinite.

EXERCISE 3.5.6 Show with an example that the P-game is incorrect as a proof theory for stable semantics.

EXERCISE 3.5.7 Investigate for all arguments in Exercise 2.8.11(b) whether they can be proven to be in some preferred extension. For each provable argument, give a terminated game won by P. For each argument that is not provable, explain why such a game does not exist.

Chapter 4

A framework for argumentation with structured arguments

4.1 Introduction

As explained above, Dung's (1995) abstract framework was an important advance in the formal study of argumentation. However, its fully abstract nature makes it less suitable for directly representing specific argumentation problems. It is best used as a tool for analysing particular argumentation formalisms and for developing a metatheory of such systems. When actual applications of argumentation-based inference have to be modelled, Dung's framework should be refined with accounts of the structure of arguments and the nature of the defeat relation. However, here too abstraction is still possible and worthwhile. This chapter instantiates Dung's abstract approach by assuming an unspecified logical language and by defining arguments as (directed acyclic) inference graphs formed by applying two kinds of inference rules, deductive (or 'strict') and defeasible rules'. As explained in Section 1.2, the notion of an argument as an inference graph naturally leads to three ways of attacking an argument: attacking a premise, attacking a conclusion and attacking an inference. To resolve such conflicts, preferences may be used, which leads to three corresponding kinds of defeat: undermining, rebutting and undercutting defeat. To characterise them, some minimal assumptions on the logical object language must be made, namely that certain well-formed formulas are a contrary or contradictory of certain other well-formed formulas. Apart from this the framework is still abstract: it applies to any set of inference rules, as long as it is divided into strict and defeasible ones, and to any logical language with a (possibly non-symmetric) negation connective.

The resulting framework unifies two ways to capture the fallibility of reasoning. Some, e.g. Bondarenko et al. (1997), locate the fallibility of arguments in the uncertainty of their premises, so that arguments can only be attacked on their premises. Others, e.g. Pollock (1994); Vreeswijk (1997), instead locate the fallibility of arguments in the riskiness of their inference rules: in these logics inference rules are of two kinds, being either deductive or defeasible, and arguments can only be attacked on their applications of defeasible inference rules. Vreeswijk (1993b, Ch. 8) called these two approaches *plausible* and *defeasible* reasoning: he described plausible reasoning as sound (i.e, deductive) reasoning on an uncertain basis, and defeasible reasoning as unsound (but still rational) reasoning on a solid basis. In his chapter 8, Vreeswijk attempted to combine both forms of reasoning in a single formalism, but since then most

formal accounts of argumentation have modelled either only plausible or only defeasible reasoning. The present framework again combines the two forms of reasoning but this time within the abstract setting of Dung (1995).

The account offered in this chapter further develops work undertaken in the European ASPIC project (Amgoud et al.; 2006; Caminada and Amgoud; 2007) and is more fully reported in (Prakken; 2010; Modgil and Prakken; 2013). It is based on work of John Pollock (1987; 1994) and Gerard Vreeswijk (1993b; 1997) on the structure of arguments, work of Pollock (1974; 1987) on notions of defeat and work of Prakken and Sartor (1997) and others on argumentation with prioritised rules. The proofs of the formal results stated in this chapter can be found in (Prakken; 2010; Modgil and Prakken; 2013). The text of this chapter is largely based on Modgil and Prakken (2014), which gives a tutorial introduction to the *ASPIC*$^+$ framework. In addition, some fragments are taken from the recent handbook chapter of Modgil and Prakken (2018).

4.2 Design choices and overview

People argue to remove doubt about a claim (Walton; 2006, p. 1), by giving reasons why one should accept the claim and by defending these reasons against criticism. The strongest way to remove doubt is to show that the claim deductively follows from indisputable grounds. A mathematical proof from the axioms of arithmetic is like this: its grounds are mathematical axioms, while its inferences are deductively sound. So such a proof cannot be attacked in any way: not on its grounds and not on its inferences. However, such perfection is not attainable in real life: our grounds may not be indisputable or they may provide less than conclusive support for their claim.

Suppose we believe that John was in Holland Park some morning and that Holland Park is in London. Then we can deductively reason from these beliefs, to conclude that John was in London that morning. So the reasoning cannot be attacked. However, perfection remains unattainable since the argument is still fallible: its grounds may turn out to be wrong. For instance, Jan may tell us that he met John in Amsterdam that morning around the same time. We now have a reason against our belief that John was in Holland Park that morning, since witnesses usually speak the truth. Can we retain our belief or must we give it up? The answer to this question determines whether we can accept that John was in London that morning.

Maybe we originally believed that John was in Holland Park for a reason. Maybe we went jogging in Holland Park and we saw John. We then have a reason supporting our belief that John was in Holland Park that morning, since we know that our senses are usually accurate. But we cannot be sure, since Jan told us that he met John in Amsterdam that morning around the same time. Perhaps our senses betrayed us this morning? But then we hear that Jan has a reason to lie, since John is a suspect in a robbery in Holland Park that morning and Jan and John are friends. We then conclude that the basis for questioning our belief that John was in Holland Park that morning (namely, that witnesses usually speak the truth and Jan witnesses John in Amsterdam) does not apply to witnesses who have a reason to lie. So our reason in support of our belief is undefeated and we accept it.

If we want to formalise a logic for argumentation, then this simple example (displayed in Figure 4.1) already suggests a number of issues we have to deal with. At least two further important design decisions have to be made: how can arguments be built, i.e., how can claims be supported with grounds, and how can arguments be attacked?

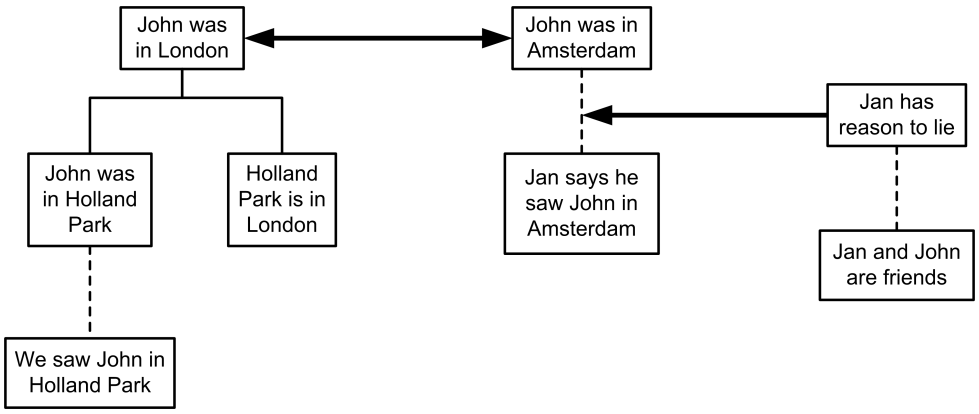


Figure 4.1: An informal example

We shall see that the answers to these two questions are related.

First, the claims and beliefs in our example were supported in various ways: in the first case we appealed to the principles of deductive inference when concluding that John was in London (visualised in Figure 4.1 with solid links). *ASPIC*$^+$ is therefore designed so that arguments can be constructed using deductive or *strict* inference rules that license deductive inferences from premises to conclusions. However, in the other two cases the reasoning from grounds to claim appealed to the reliability of, respectively, our senses and witnesses as sources of information. Should these kinds of support (inferences) from grounds to claims be modelled as deductive?

To help answer this question, consider that our informal example contains three ways of attacking an argument: 1) Our initial argument that John was in London was attacked by the witness argument on its ground, or *premise*, that John was in Holland Park that morning; 2) We then modified our initial argument by extending it with an additional argument for the attacked premise, but the extended argument was still attacked (by the witness argument) on the (now) intermediate conclusion that John was in Holland Park that morning; 3) Finally, we counterattacked the witness argument not on a premise or conclusion but on the reasoning from the grounds to the claim: namely, the inference step from the premise that Jan said he met John in Amsterdam that morning to the claim that John was in Amsterdam that morning (note that here we regard the principle that witnesses usually speak the truth as an inference rule).

Now, returning to the question whether all kinds of inference should be deductive, the second type of attack would not be possible on the deductively inferred intermediate conclusion since the nature of deductive support is that it is absolutely watertight: if one accepts all antecedents of a deductively valid inference rule, then one must also accept its consequent no matter what, on the penalty of being irrational. If the antecedents of a deductively valid inference rule are true, then its consequent must also be true. So if we have reason to believe that the conclusion of a deductive inference is not true, then there must be something wrong with its premises (which may in turn be the conclusions of subarguments). It is for this very same reason that the third type of attack, on the deductive inferential step itself, is also not possible.

ASPIC$^+$ is therefore designed to comply with the common-sense and philosophically argued position (Pollock (1995, p.41); Pollock (2009, p. 173)) advocating the rationality of supporting claims with grounds that do not deductively entail them. In other words, the fallibility of an argument need not only be located in its premises, but

can also be located in the inference steps from premises to conclusion (visualised in Figure 4.1 with dashed links). Thus, arguments in *ASPIC*$^+$ can be constructed using *defeasible* inference rules, and arguments can be attacked on the application of such defeasible inference rules, in keeping with the interpretation that the premises of such a rule presumptively, rather than deductively, support their conclusions,

However, some would argue that the second and third type of attacks can be simulated using only deductive rules (specifically the deductive rules of classical logic) by augmenting the antecedents of these rules with normality premises. For example, with regard to the second type of attack, could we not say that our argument claiming that John was in Holland Park that morning since we saw him there has an implicit premise *our senses functioned normally*, and that the argument that John was in Amsterdam that morning in fact attacks this implicit premise, rather than its claim, thus reducing attacks on conclusions to attacks on premises? With regard to the third type of attack, could we not say that instead of attacking the defeasible inference step from Jan's testimony to the claim that John was in Amsterdam, we could model this step as deductive, and then add the premise that normally witnesses speak the truth, and then direct the attack at this premise? In other words, can we reduce attacks on inferences to attacks on premises?

In answer to these questions, we first note that some have argued that such deductive simulations are prone to yielding counterintuitive results. This is a topic that we will return to and examine in more detail in Section 4.4.7. Second, we claim that there is some merit in modelling the everyday practice of 'jumping to defeasible conclusions' and of considering arguments for contradictary conclusions. This is especially important given that one of the argumentation paradigm's key strengths is its characterisation of formal logical modes of reasoning in a way that corresponds with human modes of reasoning and debate.

The above discussion introduced the notion of *fallible* premises that can be attacked. However *ASPIC*$^+$ also wants to allow you to distinguish premises that are axiomatic and so cannot be attacked. We discuss the uses of such premises in Section 4.4, but for the moment we can summarise by saying that *ASPIC*$^+$ arguments can be constructed from fallible and infallible premises (respectively called *ordinary* and *axiom* premises in Section 4.3), and strict and defeasible inference rules, and that arguments can be attacked on their ordinary premises, the conclusions of defeasible inference rules, and the defeasible inference steps themselves. Finally, a key feature of the *ASPIC*$^+$ framework is that it accommodates the use of preferences over arguments, so that an attack from one argument to another only succeeds (as a defeat) if the attacked argument is not stronger than (strictly preferred to) the attacking argument, according to some given preference relation. The justified *ASPIC*$^+$ arguments are then evaluated with respect to the Dung framework relating *ASPIC*$^+$ arguments by the defeat relation.

4.3 The framework defined: Special case with 'ordinary' negation

In this section we present the basis definitions of the *ASPIC*$^+$ framework. Note that in this section we present a special case of *ASPIC*$^+$, in which conflict is based on the standard classical notion of negation, and then in Section 4.5 we replace negation by a more general notion of conflict between formulae.

4.3.1 Argumentation systems, knowledge bases, and arguments

To use *ASPIC*$^+$, you need to provide the following information. You must choose a *logical language* $\mathcal{L}$ closed under negation $\neg$ (which we later replace with a more general notion of conflict). You must then provide two (possibly empty) sets of *strict* ($\mathcal{R}_s$) and *defeasible* ($\mathcal{R}_d$) inference rules. If you provide a non-empty set of defeasible rules, you then need to also specify which well-formed formulas in $\mathcal{L}$ correspond to (i.e., name) which defeasible rule in $\mathcal{R}_d$. To do the latter requires specifying a partial function n from $\mathcal{R}_d$ to $\mathcal{L}$. These names can then by used when attacking arguments on defeasible inference steps. Informally, $n(r)$ is a wff in $\mathcal{L}$ which says that the defeasible rule $r \in \mathcal{R}$ is applicable, so that an argument claiming $\neg n(r)$ attacks the inference step in the corresponding rule[1].

The above is summarised in the following formal definition:

Definition 4.3.1 [**Argumentation systems**] An *argumentation system* is a triple $AS = (\mathcal{L}, \mathcal{R}, n)$ where:

- $\mathcal{L}$ is a nonempty logical language with a unary negation symbol $\neg$.
- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules of the form $\{\varphi_1, \ldots, \varphi_n\} \rightarrow \varphi$ and $\{\varphi_1, \ldots, \varphi_n\} \Rightarrow \varphi$ respectively (where φ_i, φ are meta-variables ranging over wff in $\mathcal{L}$), and $\mathcal{R}_s \cap \mathcal{R}_d = \varnothing$. $\varphi_1, \ldots, \varphi_n$ are called the *antecedents* and φ the *consequent* of the rule.[2]
- n is a partial function such that $n : \mathcal{R}_d \longrightarrow \mathcal{L}$.

If there is no danger for confusion, we will sometimes write the sequence of antecedents of a strict or defeasible rule as a set. Furthermore, we write $\psi = -\varphi$ just in case $\psi = \neg\varphi$ or $\varphi = \neg\psi$ (we will sometimes informally say that formulas φ and $-\varphi$ are each other's negation). Note that $-$ is not part of the logical language $\mathcal{L}$ but a metalinguistic function symbol to obtain more concise definitions.

It is important to stress here that *ASPIC*$^+$'s strict and defeasible inference rules are *not* object-level formulae in the language $\mathcal{L}$, but are meta to the language, allowing one to deductively, respectively defeasibly, infer the rule's consequent from the rule's antecedents. Such inference rules may range over arbitrary formulae in the language, in which case they will, as usual in logic, be specified as *schemes*. For example, a scheme for strict inference rules capturing modus ponens for the material implication of classical logic can be written as $\alpha, \alpha \supset \beta \rightarrow \beta$[3], where α and β are metavariables for wff in $\mathcal{L}$. Alternatively, strict or defeasible inference rules may be domain-specific in that they reference specific formulae, as in the defeasible inference rule concluding that an individual flies if that individual is a bird: $Bird \Rightarrow Flies$. We will further discuss these distinct uses of inference rules in Section 4.4.

If you want to use *ASPIC*$^+$, then an argumentation system is not all you have to specify: you must also specify from which body of information the premises of an argument can be taken. We call this a knowledge base, and as discussed in Section 4.2, distinguish ordinary premises, which are uncertain and so can be attacked, and premises that are axioms, hence certain, and so cannot be attacked.

[1] n is a partial function since you may want to enforce that some defeasible inference steps cannot be attacked.

[2] Below the brackets around the antecedents will be omitted.

[3] In this chapter we use $\supset$ to denote the material implication connective of classical logic.

Definition 4.3.2 [**Knowledge bases**] A *knowledge base* in an $AS = (\mathcal{L}, \mathcal{R}, n)$ is a set $\mathcal{K} \subseteq \mathcal{L}$ consisting of two disjoint subsets $\mathcal{K}_n$ (the *axioms*) and $\mathcal{K}_p$ (the *ordinary premises*).

ASPIC$^+$ leaves you fully free to choose your language, what is an axiom and what is an ordinary premise and how you specify your strict and defeasible rules. However some care needs to be taken in making these choices, to ensure that the result of argumentation is guaranteed to be well-behaved. By 'well-behaved' we mean that the desirable properties proposed by Caminada and Amgoud (2007) are satisfied; for example, that the conclusions of arguments in the same extension are mutually consistent (we will define below what this means) and are closed under application of strict inference rules (whatever you can derive from your conclusions of arguments in an extension, with strict rules alone, is already a conclusion of an argument in that extension). In Section 4.4 we present some theorems which tell you how you can make your choices in such a way that the result is guaranteed to be well-behaved. These theorems will talk about two notions of consistency, namely, direct and indirect consistency. Indirect consistency is defined in terms of the closure of a set of well-formed formulas under application of strict inference rules. Informally, the strict closure of a set of wff is the set itself plus everything that can be derived from it when only applying strict rules.

Definition 4.3.3 [**Consistency and strict closure**] For any $S \subseteq \mathcal{L}$, let the closure of S under strict rules, denoted $Cl(S)$, be the smallest set containing S and the consequent of any strict rule in $\mathcal{R}_s$ whose antecedents are in $Cl(S)$. Then a set $S \subseteq \mathcal{L}$ is

- *directly consistent* iff $\nexists\ \psi, \varphi \in S$ such that $\psi = -\varphi$
- *indirectly consistent* iff $Cl(S)$ is directly consistent.

We call the combination of an argumentation system and a knowledge base an argumentation theory:

Definition 4.3.4 [**Argumentation theory**] An *argumentation theory* is a tuple AT = $(AS, \mathcal{K})$ where AS is an argumentation system and $\mathcal{K}$ is a knowledge base in AS.

ASPIC$^+$ arguments are now defined relative to an argumentation theory $AT = (AS, \mathcal{K})$, and chain applications of the inference rules from AS into directed acyclic inference graphs, starting with elements from the knowledge base $\mathcal{K}$ (if no premise is used more than once, then the graph will be a tree). In what follows, for a given argument, the function `Prem` returns all the formulas of $\mathcal{K}$ (called *premises*) used to build the argument, `Conc` returns its conclusion, `Sub` returns all its sub-arguments, `DefRules` returns all the defeasible rules of the argument and `TopRule` returns the last inference rule used in the argument.

Definition 4.3.5 [Argument] An *argument* A on the basis of an argumentation theory with a knowledge base $\mathcal{K}$ and an argumentation system $(\mathcal{L}, \mathcal{R}, n)$ is any structure obtainable by applying one or more of the following steps finitely many times:

1. φ if $\varphi \in \mathcal{K}$ with: $\texttt{Prem}(A) = \{\varphi\}$, $\texttt{Conc}(A) = \varphi$, $\texttt{Sub}(A) = \{\varphi\}$, $\texttt{DefRules}(A)$ = $\varnothing$, $\texttt{TopRule}(A)$ = undefined.

2. $A_1, \ldots A_n \rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments such that there exists a strict rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$ in $\mathcal{R}_s$.
   $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,
   $\texttt{Conc}(A) = \psi$,
   $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$.
   $\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n)$,
   $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots \texttt{Conc}(A_n) \rightarrow \psi$

3. $A_1, \ldots A_n \Rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments such that there exists a defeasible rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$.
   $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,
   $\texttt{Conc}(A) = \psi$,
   $\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$,
   $\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n) \cup \{\texttt{Conc}(A_1), \ldots \texttt{Conc}(A_n) \Rightarrow \psi\}$,
   $\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots \texttt{Conc}(A_n) \Rightarrow \psi$.

For any argument A we define $\texttt{Prem}_n(A) = \texttt{Prem}(A) \cap \mathcal{K}_n$ and $\texttt{Prem}_p(A) = \texttt{Prem}(A) \cap \mathcal{K}_p$. Moreover, each of the functions Func in this definition is also defined on sets of arguments $S = \{A_1, \ldots, A_n\}$ as follows: $\texttt{Func}(S) = \texttt{Func}(A_1) \cup \ldots \cup \texttt{Func}(A_n)$. Note, finally, that the $\rightarrow$ and $\Rightarrow$ symbols are overloaded to denote both inference rules and arguments.

Example 4.3.6 Consider a knowledge base in an argumentation system with $\mathcal{L}$ consisting of $p, q, r, s, t, u, v, w, x, d_1, d_2, d_3, d_4, d_5, d_6$ and their negations, with $\mathcal{R}_s = \{s_1, s_2\}$ and $\mathcal{R}_d = \{d_1, d_2, d_3, d_4, d_5, d_6\}$, where

d_1:	$p \Rightarrow q$	d_4:	$u \Rightarrow v$	s_1:	$p, q \rightarrow r$
d_2:	$s \Rightarrow t$	d_5:	$v, x \Rightarrow \neg t$	s_2:	$v \rightarrow \neg s$
d_3:	$t \Rightarrow \neg d_1$	d_6:	$s \Rightarrow \neg p$		

Moreover, $\mathcal{K}_n = \{p\}$ and $\mathcal{K}_p = \{s, u, x\}$. Note that in presenting the example, we have informally used names d_i to refer to defeasible inference rules. We now define the n function that formally assigns wff d_i to such rules, i.e., for any rule informally referred to as d_i, we have that $n(d_i) = d_i$, so that '$n(d_1) = d_1$' is a shorthand for $n(p \Rightarrow q) = d_1$. In further examples we will often specify the n function in the same way.[4]

An argument for r (i.e., with conclusion r) is displayed in Figure 4.2, with the premises at the bottom and the conclusion at the top of the argument graph (which in this case is a tree). In this and the next figure, the type of a premise is indicated with a superscript and defeasible inferences, underminable premises and rebuttable conclusions are displayed with dotted lines. The figure also displays the formal structure of the argument. We have that

$\texttt{Prem}(A_3) =$	$\{p\}$	$\texttt{DefRules}(A_3) =$	$\{d_1\}$
$\texttt{Conc}(A_3) =$	r	$\texttt{TopRule}(A_3) =$	s_1
$\texttt{Sub}(A_3) =$	$\{A_1, A_2, A_3\}$		

[4] In our further examples we will often leave the logical language $\mathcal{L}$ and the n function implicit, trusting that they will be obvious.

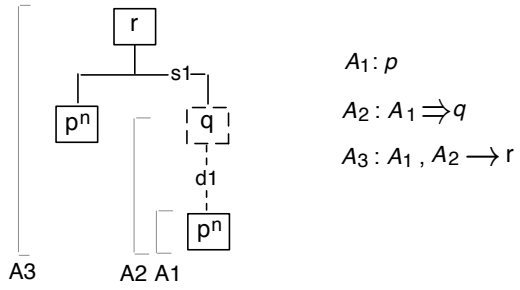


Figure 4.2: An argument

The distinction between two kinds of inference rules and two kinds of premises motivates a distinction into four kinds of arguments.

Definition 4.3.7 [Argument properties] An argument A is *strict* if $\texttt{DefRules}(A) = \varnothing$; *defeasible* if $\texttt{DefRules}(A) \neq \varnothing$; *firm* if $\texttt{Prem}(A) \subseteq \mathcal{K}_n$; *plausible* if $\texttt{Prem}(A) \cap \mathcal{K}_p \neq \varnothing$. We write $S \vdash \varphi$ if there exists a strict argument for φ with all premises taken from S, and $S \mid\!\sim \varphi$ if there exists a defeasible argument for φ with all premises taken from S.

Example 4.3.8 In Example 4.3.6 the argument A_1 is both strict and firm, while A_2 and A_3 are defeasible and firm. Furthermore, we have that $\mathcal{K} \vdash p$, $\mathcal{K} \mid\!\sim q$ and $\mathcal{K} \mid\!\sim r$.

In logic-based approaches to argumentation (see Section 4.4.6 below) arguments are often required to be minimal in that no proper subset of their premises should logically (according to the adopted base logic) imply the conclusion. In the *ASPIC*$^+$ context such a constraint would be fine for applications of strict rules. However, minimality cannot be required for application of defeasible inference rules, since defeasible rules that are based on more information may well make an argument stronger. For example, *Observations done in ideal circumstances are usually correct* is stronger than *Observations are usually correct*.

Another requirement of logic-based approaches, namely, that an argument's premises have to be consistent, can optionally be imposed in basic *ASPIC*$^+$, leading to two variants of the basic framework. We define a special class of arguments whose premises are indirectly consistent. In this way *ASPIC*$^+$ can be used as a framework for reconstructing logic-based argumentation formalisms, as we will further discuss in Section 4.4.6.

Definition 4.3.9 [consistent arguments] An argument A is *consistent* iff $\texttt{Prem}(A)$ is indirectly consistent.

4.3.2 Attack and defeat

Recall that *ASPIC*$^+$ is meant to generate Dung-style abstract argumentation frameworks, that is, a set of arguments with a binary relation of defeat. Having defined arguments above, we now define the attack relation and then, as discussed in Section 4.2, we apply preferences to determine the defeat relation (in fact Dung called his relation "attack" but we reserve this term for the basic notion of conflict, to which we then apply preferences).

Attack

We now first present the three ways in which arguments in *ASPIC*$^+$ can be in conflict, that is, three kinds of attack. In short, arguments can be attacked on a conclusion of a defeasible inference (rebutting attack), on a defeasible inference step itself (undercutting attack), or on an ordinary premise (undermining attack). As discussed in Section 4.2, that arguments cannot be attacked on their strict inferences goes without saying. We also discussed why arguments cannot be attacked on the conclusions of strict inferences: if the antecedents of a deductively valid inference rule are true, then its consequent must also be true no matter what. So if we have reason to believe that the conclusion of a deductive inference is not true, then there must be something wrong with the claims from which it is drawn. In Section 4.4.4 we will give a second reason why arguments cannot be attacked on conclusions of strict inferences. In short, this is because if we allow such attacks, then consistency and strict closure of conclusions cannot be guaranteed.

To define undercutting attack, the function n of an AS is used, which assigns to elements of $\mathcal{R}_d$ a well-formed formula in $\mathcal{L}$. Recall that informally, $n(r)$ (where $r \in R_d$) means that r is applicable. Then an argument using r is undercut by any argument with conclusion $-n(r)$.

Definition 4.3.10 [attacks] A *attacks* B iff A *undercuts*, *rebuts* or *undermines* B, where:

- A *undercuts* argument B (on B') iff $\texttt{Conc}(A) = -n(r)$ for some $B' \in \texttt{Sub}(B)$ such that B''s top rule r is defeasible.
- A *rebuts* argument B (on B') iff $\texttt{Conc}(A) = -\varphi$ for some $B' \in \texttt{Sub}(B)$ of the form $B''_1, \ldots, B''_n \Rightarrow \varphi$.
- Argument A *undermines* B (on φ) iff $\texttt{Conc}(A) = -\varphi$ for an ordinary premise φ of B.

This definition allows for a distinction between direct and indirect attack: an argument can be indirectly attacked by directly attacking one of its proper subarguments. This distinction will turn out to be crucial for a proper application of preferences to resolve attacks.

Example 4.3.11 In our running example argument A_3 cannot be undermined, since all its premises are axioms. A_3 can potentially be rebutted on A_2, with an argument for $\neg q$. However, the argumentaton theory of our example does not allow the construction of such a rebuttal. Likewise, A_3 can potentially be undercut on A_2, with an argument for $\neg d_1$. Our example does allow the construction of such an undercutter, namely:

B_1: s
B_2: $B_1 \Rightarrow t$
B_3: $B_2 \Rightarrow \neg d_1$

Argument B_3 has an ordinary premise s, so it can be undermined on B_1 with an argument for $\neg s$:

C_1: u
C_2: $C_1 \Rightarrow v$
C_3: $C_2 \rightarrow \neg s$

Note that since C_3 has a strict top rule, argument B_1 *does not* in turn rebut C_3.

Argument B_3 can potentially be rebut or undercut on either B_2 or B_3, since both of these subarguments of B_3 have a defeasible top rule. Our argumentation theory only allows for a rebutting attack on B_2:

C_1: u
C_2: $C_1 \Rightarrow v$
D_3: x
D_4: $C_2, D_3 \rightarrow \neg t$

All relevant arguments and attacks are displayed in Figure 4.3.

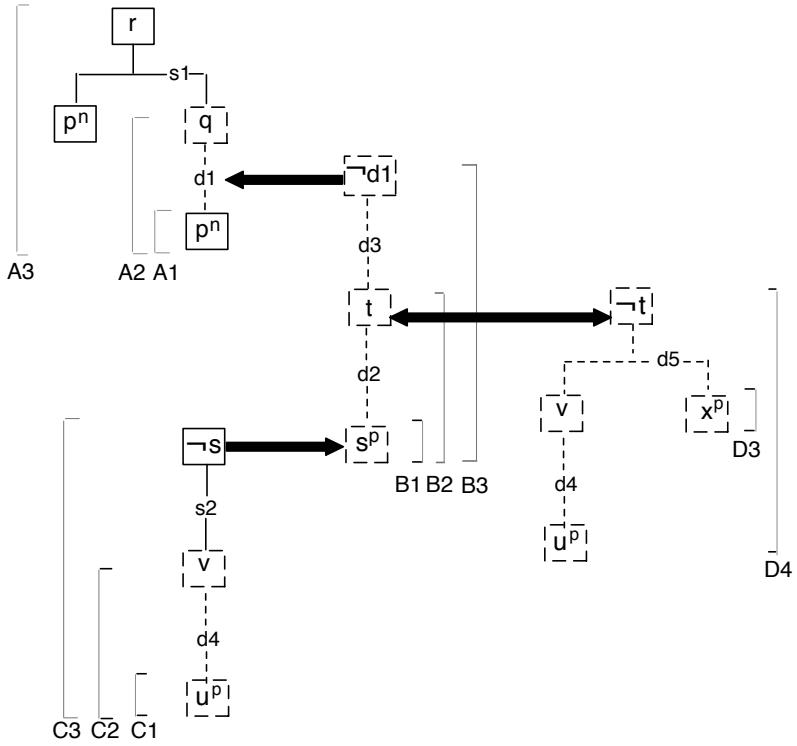


Figure 4.3: Attacks

Defeat

The attack relation tells us which arguments are in conflict with each other: if two arguments are in conflict then they cannot both be justified. However, Definition 2.2.1's notion of the acceptability of arguments is based on the notion that one argument can be used as a counter-argument to another. In general, an argument A can be used as a counter-argument to B, if A *successfully attacks*, i.e., defeats, B. Whether an attack from A to B (on its sub-argument B') succeeds as a defeat, may depend on the relative strength of A and B', i.e., whether B' is *strictly stronger than, or strictly preferred* to A. Note that only the success of undermining and rebutting attacks is contingent on preferences; undercutting attacks succeed as defeats independently of any preferences (see Modgil and Prakken (2013) for a discussion as to why this is the case).

Where do these preferences come from? Again, *ASPIC*$^+$ allows you to make any choice you like. All that *ASPIC*$^+$ as a framework wants is that you as a user give a binary ordering $\preceq$ on the set of all arguments that can be constructed on the basis of an argumentation theory. Then, as usual, if $A \preceq B$ and $B \not\preceq A$ then B is strictly preferred to A (denoted $A \prec B$). Also, if $A \preceq B$ and $B \preceq A$ then $A \approx B$. We will later identify some conditions under which argument orderings are well-behaved in that they promote consistency and strict closure of conclusions. We will also define two example argument orderings that satisfy these conditions. However, for now all we need for defining *ASPIC*$^+$'s defeat relation is the attack relation and a preference ordering over arguments.

How should the preference ordering be applied to resolve attacks? At first sight, it would seem that *ASPIC*$^+$ can be taken to generate a so-called *preference-based argumentation framework* (PAF) in the sense of Amgoud and Cayrol (2002), that is, a triple consisting of the set of arguments, the attack relation and the argument ordering. That A defeats B could then be defined as A attacks B and $A \not\prec B$. However, this does not work, for two reasons. First, PAFs do not recognise that undercutting attacks succeed irrespective of preferences. More seriously, PAFs cannot express how and at which points arguments attack each other, and yet this is crucial for a proper application of preferences to attack relations. Prakken (2012); Modgil and Prakken (2013) have shown that the use of PAFs leads to violation of the rationality postulates of subargument closure and consistency (see further Section 4.4.4 below) in cases where *ASPIC*$^+$ with the following definition satisfies these postulates.

Definition 4.3.12 [Successful rebuttal, undermining and defeat]

- A *successfully rebuts* B if A rebuts B on B' and $A \not\prec B'$.
- A *successfully undermines* B if A undermines B on φ and $A \not\prec \varphi$.
- A *defeats* B iff A undercuts or successfully rebuts or successfully undermines B.

The success of rebutting and undermining attacks thus involves comparing the conflicting arguments at the points where they conflict; that is, by comparing those arguments that are in a *direct* rebutting or undermining relation with each other. The definition of successful undermining exploits the fact that an argument premise is also a subargument.

Example 4.3.13 In our running example two argument orderings are relevant for whether attacks are successful: between B_1 and C_3 and and between B_2 and D_4. Note that the undercutting attack of B_3 on A_2 (and thereby on A_3) succeeds as a defeat irrespective of the argument ordering between B_3 and A_2. The undermining attack of C_3 on B_1 succeeds if $C_3 \not\prec B_1$. If $B_2 \approx D_4$ or their relation is undefined then these two arguments defeat each other, while D_4 strictly defeats B_3. If $D_4 \prec B_2$ then B_2 strictly defeats D_4 while if $B_2 \prec D_4$ then D_4 strictly defeats both B_2 and B_3.

Let us now put all these elements together; that is the arguments and attacks defined on the basis of an argumentation theory, and a preference ordering over the arguments:

Definition 4.3.14 Let AT be an *argumentation theory* (AS, KB). A *(c-)structured argumentation framework (*(c-)SAF*)* defined by AT, is a triple $(\mathcal{A}, \mathcal{C}, \preceq)$ where

- In a SAF, $\mathcal{A}$ is the set of all arguments on the basis of KB in AS;

- In a c-SAF, $\mathcal{A}$ is the set of all consistent arguments on the basis of KB in AS;
- $\preceq$ is a preference ordering on $\mathcal{A}$;
- $(X, Y) \in \mathcal{C}$ iff X attacks Y.

Example 4.3.15 In our running example $\mathcal{A} = \{A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, C_3, D_3, D_4\}$, while $\mathcal{C}$ is such that B_3 attacks both A_2 and A_3, argument C_3 attacks all of B_1, B_2, B_3, argument D_4 attacks both B_2 and B_3 and, finally, B_2 attacks D_4.

4.3.3 Generating Dung-style abstract argumentation frameworks

We are now ready to instantiate a Dung framework with *ASPIC*$^+$ arguments and the *ASPIC*$^+$ defeat relation.

Definition 4.3.16 [**Argumentation frameworks**] An *abstract argumentation framework (AF) corresponding to a (c-)SAF* $= (\mathcal{A}, \mathcal{C}, \preceq)$ is a pair $(\mathcal{A}, \mathcal{D})$ such that $\mathcal{D}$ is the defeat relation on $\mathcal{A}$ determined by $(\mathcal{A}, \mathcal{C}, \preceq)$.

The justified arguments of the above defined AF are then defined under the various semantics of Chapter 2.

It is now also possible to define a consequence notion for well-formed formulas. Several definitions are possible. The following definition directly uses the notions of justified, defensible and overruled arguments from Chapter 2: (here an S-justified (S-defensible, S-overruled) argument is an argument that is justified (defensible, overruled) according to semantics S):

Definition 4.3.17 [The status of conclusions] For every semantics S and for every (c-)structured argumentation framework *(c-)SAF* with corresponding abstract argumentation framework AF, and every formula $\varphi \in \mathcal{L}_{AT}$:

1. φ is *S-justified* in *(c-)SAF* if and only if there exists an S-justified argument on the basis of AF with conclusion φ;
2. φ is *S-defensible* in *(c-)SAF* if and only if φ is not S-justified in SAF and there exists an S-defensible argument on the basis of AF with conclusion φ;
3. φ is *S-overruled* in *(c-)SAF* if and only if it is not S-justified or S-defensible in SAF and there exists an S-overruled argument on the basis of AF with conclusion φ.

Example 4.3.18 In our running example, if D_4 strictly defeats B_2, then we have a unique extension in all semantics which at least contains the set $S = \{A_1, A_2, A_3, C_1, C_2, C_3, D_3, D_4\}$. If in addition C_3 does not defeat B_1, then the extension also contains B_1. In both cases this yields that wff r is sceptically justified.

Alternatively, if B_2 strictly defeats D_4, then the status of r depends on whether C_3 defeats B_1. If it does, then we again have a unique extension in all semantics consisting of the set S, so r is sceptically justified. By contrast, if C_3 does not defeat B_1, we obtain a unique extension with A_1, B_1, B_2, B_3, C_1, C_2, C_3 and D_3, so r is neither sceptically nor credulously justified.

Finally, if B_2 and D_4 defeat each other, then the outcome again depends on whether C_3 defeats B_1. If it does, then the situation is as in the previous case – a unique extension S – but if C_3 does not defeat B_1, then the grounded extension consists of A_1, B_1, C_1-C_3, D_3. So in the latter case, in grounded semantics r is neither sceptically nor credulously justified. However, in preferred and stable semantics we then obtain two alternative extensions: the first contains D_4 while the second instead contains B_2 and B_3 and so excludes A_2 and A_3. So in the latter case r is credulously, but not sceptically justified under stable and preferred semantics.

Note that the first condition of Definition 4.3.17 is equivalent to

1. φ is *S-justified* in *(c-)SAF* if and only if there exists an argument with conclusion φ that is contained in all S-extensions of AF.

Thus this definition does not allow that different extensions contain different arguments for a skeptical conclusion and therefore does not capture floating conclusions (see Section 2.2). The following alternative definition does capture floating conclusions.

Definition 4.3.19 [Justified conclusions (possibly floating)]

1. φ is *S-f-justified* in *(c-)SAF* if and only if all S-extensions of AF contain an argument with conclusion φ.

4.3.4 More on argument orderings

A well studied use of preferences in the non-monotonic logic literature is based on the use of priority orderings over formulae in the language or defeasible inference rules. If *ASPIC*$^+$ is to be used as a framework for giving argumentation-based characterizations of non-monotonic formalisms augmented with priorities, then it needs to provide an account of how these priority orderings can be 'lifted' to preferences over arguments. Now the first thing to note is that if your use of *ASPIC*$^+$ involves using defeasible inference rules and ordinary premises, then both may come equipped with priority orderings $\leq$ on $\mathcal{R}_d$ and $\leq'$ on $\mathcal{K}_p$. We assume that these priority orderings are distinct to allow for the ontological nature of the rules and premises to be distinct. For example, the ordinary premises may represent the content of percepts from sensors or of witness testimonies, whose priority ordering reflects the relative reliability of the sensors, respectively witnesses. The defeasible rules may, for example, be prioritized based on probabilistic strength, on temporal precedence (defeasible rules acquired later are preferred to those acquired earlier), on the basis of principles of legal precedence, and so on. The challenge is to then define a preference over arguments A and B based on the priorities over their constituent ordinary premises *and* defeasible rules.

We now define two argument preference orderings, called the weakest-link and last-link orderings. These orderings are in turn based on priority orderings $\leq$ on $\mathcal{R}_d$ and $\leq'$ on $\mathcal{K}_p$, where as usual, $X <^{(\prime)} Y$ iff $X \leq^{(\prime)} Y$ and $Y \not\leq^{(\prime)} X$ (note that we may represent orderings in terms of the strict counterpart they define). However, these priorities relate individual defeasible rules, respectively ordinary premises, whereas when comparing two arguments, we want to compare them on the (possibly non-singleton) *sets of* rules/premises that these arguments are constructed from. So, to define these argument preferences, we need to first define an ordering over *sets* of rules/premises. We will denote this ordering with $\lhd_s$. For technical reasons we interpret it as strict preference; that is, $\Gamma \lhd_s \Gamma'$ means that Γ' is strictly preferred over Γ.

Note that for any sets of defeasible rules/ordinary premises Γ and Γ', we intuitively want that:

1) if Γ is the empty set, it cannot be that $\Gamma \lhd_{\mathtt{s}} \Gamma'$;

2) if Γ' is the empty set, it must be for any non-empty Γ that $\Gamma \lhd_{\mathtt{s}} \Gamma'$.

In other words, arguments that have no defeasible rules (ordinary premises) are, modulo the premises (rules), strictly stronger than (preferred to) arguments that have defeasible rules (ordinary premises). Hence the following definition explicitly imposes these constraints, and then gives two alternative ways of defining $\lhd_{\mathtt{s}}$; the so called `Elitist` and `Democratic` ways (i.e., $\mathtt{s} = \mathtt{Eli}$ and `Dem` respectively). `Eli` compares sets on their minimal and `Dem` on their maximal elements.

Definition 4.3.20 [Orderings $\lhd_{\mathtt{s}}$] Let Γ and Γ' be finite sets[5]. Then $\lhd_{\mathtt{s}}$ is defined as follows:

1. If $\Gamma = \varnothing$ then it cannot be that $\Gamma \lhd_{\mathtt{s}} \Gamma'$;
2. If $\Gamma' = \varnothing$ and $\Gamma \neq \varnothing$ then $\Gamma \lhd_{\mathtt{s}} \Gamma'$;
   else, assuming a preordering $\leq$ over the elements in $\Gamma \cup \Gamma'$, then if :
3. $\mathtt{s} = \mathtt{Eli}$:
   $\Gamma \lhd_{\mathtt{Eli}} \Gamma'$ if $\exists X \in \Gamma$ s.t. $\forall Y \in \Gamma', X < Y$.
   else, if:
4. $\mathtt{s} = \mathtt{Dem}$:
   $\Gamma \lhd_{\mathtt{Dem}} \Gamma'$ if $\forall X \in \Gamma, \exists Y \in \Gamma', X < Y$.

Henceforth, we will assume that $\lhd_{\mathtt{Eli}}$ is used to compare sets of rules/premises.

Now the **last-link principle** strictly prefers an argument A over another argument B if the last defeasible rules used in B are strictly less preferred ($\lhd_{\mathtt{s}}$) than the last defeasible rules in A or, in case both arguments are strict, if the premises of B are strictly less preferred than the premises of A. The concept of 'last defeasible rules' is defined as follows.

Definition 4.3.21 [Last defeasible rules] Let A be an argument.

- $\mathtt{LastDefRules}(A) = \varnothing$ iff $\mathtt{DefRules}(A) = \varnothing$.
- If $A = A_1, \ldots, A_n \Rightarrow \phi$, then $\mathtt{LastDefRules}(A) = \{\mathtt{Conc}(A_1), \ldots, \mathtt{Conc}(A_n) \Rightarrow \phi\}$, else $\mathtt{LastDefRules}(A) = \mathtt{LastDefRules}(A_1) \cup \ldots \cup \mathtt{LastDefRules}(A_n)$.

A simple example with more than one last defeasible rule is with $\mathcal{K} = \{p; q\}$, $\mathcal{R}_s = \{r, s \rightarrow t\}$ and $\mathcal{R}_d = \{p \Rightarrow r;\ q \Rightarrow s\}$. Then for the argument A for t we have that $\mathtt{LastDefRules}(A) = \{p \Rightarrow r;\ q \Rightarrow s\}$.

The above definition is now used to compare pairs of arguments as follows:

Definition 4.3.22 [Last link principle] Let A and B be two arguments. Then $A \prec B$ iff:

1. $\mathtt{LastDefRules}(A) \lhd_{\mathtt{s}} \mathtt{LastDefRules}(B)$; or

[5]Notice that it suffices to restrict $\lhd$ to finite sets since *ASPIC*$^+$ arguments are assumed to be finite (in Definition 4.3.14) and so their sets of ordinary premises/defeasible rules must be finite.

2. $\texttt{LastDefRules}(A)$ and $\texttt{LastDefRules}(B)$ are empty and $\texttt{Prem}_p(A) \lhd_{\texttt{s}} \texttt{Prem}(_pB)$.

Moreover, $A \preceq B$ iff $A \prec B$ or $A = B$.

Because of this definition, the last-link ordering $\preceq$ is in fact a *strict partial ordering*, i.e., it is *transitive* (If $A \preceq B$ and $B \preceq C$ then $A \preceq C$) and *antisymmetric* (if $A \preceq B$ and $B \preceq A$ then $A = B$).

Example 4.3.23 Suppose in our running example that $u <' s$, $x <' s$, $d_2 < d_5$ and $d_4 < d_2$. Applying the last-link ordering, we must, to check whether C_3 defeats B_1, compare $\texttt{LastDefRules}(C_3) = \{d_4\}$ with $\texttt{LastDefRules}(B_1) = \varnothing$. Clearly, $\{d_4\} \lhd_{\texttt{Eli}} \varnothing$, so $C_3 \prec B_1$, so C_3 does not defeat B_1. Next, to check the conflict between B_2 and D_4 we compare $\texttt{LastDefRules}(B_2) = \{d_2\}$ with $\texttt{LastDefRules}(D_4) = \{d_5\}$. Since $d_2 < d_5$ we have that $\texttt{LastDefRules}(B_2) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(D_4)$, so D_4 strictly defeats B_2.

The **weakest-link principle** considers not the last but all uncertain elements in an argument. Recall that in the following definition, $\texttt{Prem}_p(A) = \texttt{Prem}(A) \cap \mathcal{K}_p$.

Definition 4.3.24 [Weakest link principle] Let A and B be two arguments. Then $A \prec B$ iff

1. If both B and A are strict, then $\texttt{Prem}_{\texttt{p}}(A) \lhd_{\texttt{s}} \texttt{Prem}_{\texttt{p}}(B)$, else;
2. If both B and A are firm, then $\texttt{DefRules}(A) \lhd_{\texttt{s}} \texttt{DefRules}(B)$, else;
3. $\texttt{Prem}_{\texttt{p}}(A) \lhd_{\texttt{s}} \texttt{Prem}_{\texttt{p}}(B)$ and $\texttt{DefRules}(A) \lhd_{\texttt{s}} \texttt{DefRules}(B)$

Moreover, $A \preceq B$ iff $A \prec B$ or $A = B$.

Like the last-link ordering, the weakest-link ordering is also a strict partial ordering.

Example 4.3.25 If in our running example we apply the weakest-link ordering, then we must, to check whether C_3 defeats B_1, first compare $\texttt{Prem}_p(C_3) = \{u\}$ with $\texttt{Prem}_p(B_1) = \{s\}$. Since $u <' s$ we have that $\texttt{Prem}_p(C_3) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_1)$. Then we must compare $\texttt{DefRules}(C_3) = \{d_4\}$ with $\texttt{DefRules}(B_1) = \varnothing$. We have as above that $\{d_4\} \lhd_{\texttt{Eli}} \varnothing$. So $C_3 \prec B_1$ and so C_3 does not defeat B_1. Next, to check the conflict between B_2 and D_4 we must first compare $\texttt{Prem}_p(B_2) = \{s\}$ with $\texttt{Prem}_p(D_4) = \{u, x\}$. Since both $u <' s$ and $x <' s$ we have that $\texttt{Prem}_p(D_4) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. We must then compare $\texttt{DefRules}(B_2) = \{d_2\}$ with $\texttt{DefRules}(D_4) = \{d_4, d_5\}$. Since $d_4 < d_2$ we now have that $\texttt{DefRules}(D_4) \lhd_{\texttt{Eli}} \texttt{DefRules}(B_2)$. So $D_4 \prec B_2$ and B_2 strictly defeats D_4.

We next discuss with two examples when the last-, respectively, weakest-link ordering may be more suitable. Consider first the following example on whether people misbehaving in a university library may be denied access to the library.[6]

Example 4.3.26 Let $\mathcal{K}_p = \{Snores;\ Professor\}$, $\mathcal{R}_d =$

$\{Snores \Rightarrow_{d_1} Misbehaves;$
$Misbehaves \Rightarrow_{d_2} AccessDenied;$
$Professor \Rightarrow_{d_3} \neg AccessDenied\}$.

[6] In all examples below, sets that are not specified are assumed to be empty.

Assume that $Snores <' Professor$ and $d_1 < d_2,\ d_1 < d_3,\ d_3 < d_2$, and consider the following arguments.

A_1:	$Snores$	B_1:	$Professor$
A_2:	$A_1 \Rightarrow Misbehaves$	B_2:	$B_1 \Rightarrow \neg AccessDenied$
A_3:	$A_2 \Rightarrow AccessDenied$		

Let us apply the ordering to the arguments A_3 and B_2. The rule sets to be compared are $\texttt{LastDefRules}(A_3) = \{d_2\}$ and $\texttt{LastDefRules}(B_2) = \{d_3\}$. Since $d_3 < d_2$ we have that $\texttt{LastDefRules}(B_2) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(A_3)$, hence $B_2 \prec A_3$. So A_3 *strictly* defeats B_2 (i.e., A_3 defeats B_2 but B_2 does not defeat A_3). We therefore have that A_3 is justified in any semantics, so we conclude $AccessDenied$.

With the weakest-link principle the ordering between A_3 and B_2 is different. Both A and B are plausible and defeasible so we are in case (3) of Definition 4.3.24. Since $Snores <' Professor$, we have that $\texttt{Prem}_p(A_3) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. Furthermore, the rule sets to be compared are now $\texttt{DefRules}(A_3) = \{d_1, d_2\}$ and $\texttt{DefRules}(B_2) = \{d_3\}$. Since $d_1 < d_3$ we have that $\texttt{DefRules}(A_3) \lhd_{\texttt{Eli}} \texttt{DefRules}(B_2)$. So now we have that $A_3 \prec B_2$. Hence B_2 now strictly defeats A_3 and we conclude instead that $\neg AccessDenied$.

Which outcome in this example is better? Some have argued that the last-link ordering gives the better outcome since the conflict really is between the two legal rules about whether someone may be denied access to the library, while d_1 just provides a sufficient condition for when a person can be said to misbehave. The existence of a conflict on whether someone may be denied access to the library is in no way relevant for the issue of whether a person misbehaves when snoring. More generally, it has been argued that for reasoning with legal (and other normative) rules the last-link ordering is appropriate.

However, an example can be given of exactly the same form but with the legal rules replaced by empirical generalisations, and in that case intuitions seem to favour the weakest-link ordering:

Example 4.3.27 Let $\mathcal{K}_p = \{BornInScotland;\ FitnessLover\}$, $\mathcal{R}_d =$

$\{BornInScotland \Rightarrow_{d_1} Scottish;$
$Scottish \Rightarrow_{d_2} LikesWhisky;$
$FitnessLover \Rightarrow_{d_3} \neg LikesWhisky\}$.

Assume that $BornInScotland <' FitnessLover$ and $d_1 < d_2,\ d_1 < d_3,\ d_3 < d_2$, and consider the following arguments.

A_1:	$BornInScotland$	B_1:	$FitnessLover$
A_2:	$A_1 \Rightarrow Scottish$	B_2:	$B_1 \Rightarrow \neg LikesWhisky$
A_3:	$A_2 \Rightarrow LikesWhisky$		

This time it seems reasonable to conclude $\neg LikesWhisky$, since the epistemic uncertainty of the premise and d_1 of A_3 should propagate to weaken A_3. And this is the outcome given by the weakest-link ordering. So it could be argued that for epistemic reasoning the weakest-link ordering is appropriate.

4.4 Ways to use the framework

As should be clear by now, *ASPIC*$^+$ is not a system but a framework for specifying systems. *ASPIC*$^+$ leaves you fully free to make choices as to the logical language, the

strict and defeasible inference rules, the axioms and ordinary premises in your knowledge base, and the argument preference ordering. In this section we discuss various more or less principled ways to make your choices, and then show specific uses of *ASPIC*$^+$. We also present a modified version of *ASPIC*$^+$ motivated by some problems that arise if the language and strict rules are used to encode the full expressiveness of classical logic.

4.4.1 Choosing strict rules and axioms

Domain specific strict inference rules

When designing your *ASPIC*$^+$ system, you can specify domain specific strict inference rules, as illustrated by the following example (based on Example 4 of Caminada and Amgoud 2007) in which the strict inference rules capture definitional knowledge, namely, that bachelors are not married.[7]

Example 4.4.1 Let $\mathcal{R}_d = \{d_1, d_2\}$ and $\mathcal{R}_s = \{s_1, s_2\}$, where:

$$\begin{array}{llll} d_1 = & WearsRing \Rightarrow Married & s_1 = & Married \rightarrow \neg Bachelor \\ d_2 = & PartyAnimal \Rightarrow Bachelor & s_2 = & Bachelor \rightarrow \neg Married \end{array}$$

Finally, let $\mathcal{K}_p = \{ WearsRing, PartyAnimal\}$. Consider the following arguments.

$$\begin{array}{llll} A_1: & WearsRing & B_1: & PartyAnimal \\ A_2: & A_1 \Rightarrow Married & B_2: & B_1 \Rightarrow Bachelor \\ A_3: & A_2 \rightarrow \neg Bachelor & B_3: & B_2 \rightarrow \neg Married \end{array}$$

We have that A_3 rebuts B_3 on its subargument B_2 while B_3 rebuts A_3 on its subargument A_2. Note that A_2 does not rebut B_3, since B_3 applies a strict rule; likewise for B_2 and A_3.

Notice that in the above example, the rules s_1 and s_2 are 'transpositions' of each other, and $\mathcal{R}_s$ is 'closed under transposition', in the following sense:

Definition 4.4.2 [Closure under transposition] A strict rule s is a *transposition* of φ_1, …, $\varphi_n \rightarrow \psi$ iff s = φ_1, …, φ_{i-1}, $-\psi$, φ_{i+1}, …, $\varphi_n \rightarrow -\varphi_i$ for some $1 \leq i \leq n$.

The set $\mathcal{R}_s$ of strict rules is *closed under transposition* iff for all rules r in $\mathcal{R}_s$ the transposition of r is also in R_s. The closure under transposition of a set S of rules is denoted as $Cl_{tp}(S)$ or simply as $Cl(S)$ if there is no danger for confusion. An argumentation theory is said to be closed under transposition iff its set $\mathcal{R}_s$ is closed under transposition.

In general it is a good idea to ensure that your theory is closed under transposition. Proponents of this idea argue that this follows from the intuitive meaning of a strict rule as capturing deductive, that is, perfect inference: a strict rule $q \rightarrow \neg s$ expresses that if q is true, then this guarantees the truth of $\neg s$, no matter what. Hence, if we have s, then q cannot hold, otherwise we would have $\neg s$. In general, if the negation of the consequent of a strict rule holds, then we cannot have all its antecedents, since if we had

[7] In the examples that follow we may use terms of the form s_i, d_i or f_i, to identify strict or defeasible inference rules or items from the knowledge base. We will assume that the d_i names are those assigned by the n function of Definition 4.3.1; sometimes we will attach these names to the $\Rightarrow$ symbol. Note that the s_i and f_i names have no formal meaning and are for ease of reference only.

all of them, then its consequent would hold. This is the very meaning of a strict rule. So it is very reasonable to include in $\mathcal{R}_s$ the transposition of a strict rule that is in $\mathcal{R}_s$. A second reason for ensuring closure under transposition is that it ensures satisfaction of Caminada and Amgoud (2007)'s rationality postulates, as illustrated later in Section 4.4.4.

Strict inference rules and axioms based on deductive logics

Some find the use of domain-specific strict inference rules rather odd: why not instead express them as material implications in $\mathcal{L}$ and put them in the knowledge base as axiom premises? These people want to reserve the strict inference rules for general patterns of deductive inference, since they say that this is what inference rules are meant for in logic. (Below we will see that the same issue arises with regard to the choice of defeasible rules, but we ignore that issue for the moment). *ASPIC*$^+$ allows you to do this by basing your strict inference rules (and axioms) on a deductive logic of your choice. You can do so by choosing a semantics for your choice of $\mathcal{L}$ with an associated monotonic notion of semantic consequence, and then filling $\mathcal{R}_s$ with rules that are sound with respect to that semantics. For example, suppose you want it to conform to classical logic: you want to choose a standard propositional (or first-order) language, and you want that arguments can contain any classically valid inference step over this language. In *ASPIC*$^+$ you can achieve this in two ways, a crude way and a sophisticated way.

A crude way is to simply put all valid propositional (or first-order) inferences over your language of choice in $\mathcal{R}_s$. So if you have chosen a propositional language, then you define the content of $\mathcal{R}_s$ as follows. (where $\vdash_{PL}$ denotes standard propositional-logic consequence). For any finite $S \subseteq \mathcal{L}$ and any $\varphi \in \mathcal{L}$:[8]

$$S \rightarrow \varphi \in \mathcal{R}_s \text{ if and only if } S \vdash_{PL} \varphi$$

In fact, with this choice of $\mathcal{R}_s$, strict parts of an argument don't need to be more than one step long. For example, if rules $S \rightarrow \varphi$ and $\varphi \rightarrow \psi$ are in $\mathcal{R}_s$, then $S \cup \{\varphi\} \rightarrow \psi$ will also be in $\mathcal{R}_s$. Note also that using this method your strict rules will be closed under transposition, because of the properties of classical logic. The proof is easy: suppose $p \rightarrow q$ is in $\mathcal{R}_s$ for some p and q. Then we know that $p \vdash_{PL} q$, so (by the deduction theorem for classical logic) $\vdash_{PL} p \supset q$ so (by the properties of $\vdash_{PL}$) we have $\vdash_{PL} -q \supset -p$ so (by the other half of the deduction theorem) we have $-q \vdash_{PL} -p$, so (by choice of $\mathcal{R}_s$) $-q \rightarrow -p \in \mathcal{R}_s$.

Let us illustrate the crude approach with a variation on Example 4.4.1. We retain the defeasible rules d_1 and d_2 but we replace the domain-specific strict rules s_1 and s_2 with a single material implication $Married \supset \neg Bachelor$ in $\mathcal{K}_n$. Moreover, we put all propositionally valid inferences over our language in $\mathcal{R}_s$. Then the arguments change as follows:

A_1:	$WearsRing$	B_1:	$PartyAnimal$
A_2:	$A_1 \Rightarrow Married$	B_2:	$B_1 \Rightarrow Bachelor$
A_3:	$Married \supset \neg Bachelor$	B_3:	$Married \supset \neg Bachelor$
A_4:	$A_2, A_3 \rightarrow \neg Bachelor$	B_4:	$B_2, B_3 \rightarrow \neg Married$

Now A_4 rebuts B_4 on B_2 while B_4 rebuts A_4 on A_2.

[8] Although antecedents of rules formally are sequences of formulas, we will sometimes abuse notation and write them as sets.

A sophisticated way to base the strict part of *ASPIC*$^+$ on a deductive logic of your choice is to build an existing axiomatic system for your logic into *ASPIC*$^+$. You can include its axiom(s) (typically a handfull) in $\mathcal{K}_n$ and its inference rule(s) (typically just one or a few) in $\mathcal{R}_s$. For example, there are axiomatic systems for classical logic with just four axioms and just one inference rule, namely, modus ponens (i.e, $\varphi \supset \psi, \varphi \rightarrow \psi$)[9]. With this choice of $\mathcal{R}_s$ strict parts of an argument could be very long, since in logical axiomatic systems proofs of even trivial validities might be long. However, this difference with the crude way is not very big, since if we want to be crude, we must, to know whether $S \rightarrow \varphi$ is in $\mathcal{R}_s$, first construct a propositional proof of φ from S.

With the sophisticated way of building classical logic into our argumentation system, argument A_4 in our example stays the same, since modus ponens is in $\mathcal{R}_s$. However, argument B_4 will change, since modus tollens is not in $\mathcal{R}_s$. In fact, B_4 will be replaced by a sequence of strict rule applications, together being an axiomatic proof of $\neg Married$ from $Married \supset \neg Bachelor$ and $Bachelor$.

Which approach is more natural? We think that the crude way is more like how people reason: people often summarise chunks of deductive reasoning in one step. But if you want to implement such reasoning on a computer, then the crude and sophisticated way do not differ much.

However, note that in the sophisticated method, closure under transposition may not hold; our example above does not contain modus tollens (that is, $\varphi \supset \psi, -\psi \rightarrow -\varphi$). But we have already argued that the contrapositive reasoning yielded by the inclusion of transpositions is a desirable feature. Is this a problem for this method? No, since this reasoning can also be enforced without explicitly requiring transpositions of rules. Recall that $S \vdash \varphi$ was defined as 'there exists a strict argument for φ with all premises taken from S'. Now it turns out that if $\vdash$ contraposes, then this is just as good as closure of the strict rules under transposition. Contraposition of $\vdash$ means that if $S \vdash \varphi$, then if we replace one element s of S with $-\varphi$, then $-s$ is strictly implied:

Definition 4.4.3 [Closure under contraposition] An argumentation theory is *closed under contraposition* iff for all $S \subseteq \mathcal{L}$, $s \in S$ and ϕ, if $S \vdash \phi$, then $S \backslash \{s\} \cup \{-\phi\} \vdash -s$.

Now the point is that if $\vdash$ corresponds to classical provability (as we have made it by our choice of axioms and inference rules), then $\vdash$ does indeed contrapose. Again, as will be discussed in Section 4.4.4, closure under contraposition also ensures satisfaction of rationality postulates.

We end this section by stating a quite general result on a class of logics that, if embedded in *ASPIC*$^+$, ensures closure of the strict rules under contraposition. In Amgoud and Besnard (2009) the idea was introduced to base argumentation logics on so-called Tarskian abstract logics. Very briefly, abstract logics assume just some unspecified logical language $\mathcal{L}$ and a consequence operator over this language, which to each subset of $\mathcal{L}$ assigns a subset of $\mathcal{L}$ (its logical consequences). Tarski then assumed a number of constraints on Cn, which we need not repeat here. Finally, Tarski defined a set $S \subseteq \mathcal{L}$ as *consistent* iff $Cn(S) \neq \mathcal{L}$.

Now Amgoud and Besnard (2009)'s idea was to define an argument as a pair (S, p) where $S \subseteq \mathcal{L}$ and $p \in \mathcal{L}$, where S is consistent, $p \in Cn(S)$ and S is minimal in satisfying all these conditions. In *ASPIC*$^+$ Tarski's notion of an abstract logic can be used to generate the strict rules, via the following constraint (for any finite S):

[9] As explained above, this strictly speaking is not a rule but a scheme or rules, with meta variables ranging over $\mathcal{L}$.

$S \rightarrow p \in \mathcal{R}_s$ iff $p \in Cn(S)$

It turns out that any AT with this choice of strict rules satisfies closure under contraposition. Strictly speaking, this only holds under some assumptions on the relation between the Cn function and *ASPIC*$^+$'s negation (note that Tarski did not make any assumption on the syntax of $\mathcal{L}$), but these assumptions are quite natural. For the details we refer the reader to Section 5.2 of Modgil and Prakken (2013).

4.4.2 Self-defeat, contamination and a variant of *ASPIC*$^+$

In this subsection we discuss two ways in which arguments can be self-defeating. The second way will motivate a variant of *ASPIC*$^+$ that avoids the problems created by it.

Self-defeat

In Chapter 2, Section 2.2 we said that a proper analysis of self-defeating arguments must make the structure of arguments explicit. Now that we have done so, we can explain why this is needed. In the present framework two types of self-defeating arguments are possible: *serial self-defeat* occurs when an argument defeats one of its earlier steps, while *parallel self-defeat* occurs when the contradictory conclusions of two or more arguments are taken as the premises for $\bot$. It turns out that parallel self-defeating can cause problems if argumentation systems are not carefully defined, particularly if they include standard propositional logic.

The following example explains why serial self-defeat does not cause problems.

Example 4.4.4 Consider the following version of the argument scheme from witness testimony plus an undercutter in case the witness is incredible:

$d_w(x, \varphi)$: $\texttt{Says}(x, \varphi) \Rightarrow \varphi$
$u_w(x, \varphi)$: $\texttt{Incredible}(x) \rightarrow \neg d_w(x, \varphi)$

Now suppose that $\mathcal{K}_p$ contains $\texttt{Says}(John, \text{“}\texttt{Incredible}(John)\text{”})$. Then we have

A_1: $\texttt{Says}(John, \text{“}\texttt{Incredible}(John)\text{”})$
A_2: $A_1 \Rightarrow \texttt{Incredible}(John)$
A_3: $A_2 \rightarrow \neg d_w(John, \text{“}\texttt{Incredible}(John)\text{”})$

Argument A_3 is self-defeating since it undercuts itself on A_2. In both preferred and grounded semantics there is a unique extension $E = \{A_1\}$. Arguably this is the desired outcome, since suppose witness John also says something completely unrelated, say, 'the suspect stabbed the victim with a knife' if the self-defeating argument A_3 were overruled, the argument that can be constructed for 'the suspect stabbed the victim with a knife' would be justified since all its defeaters are overruled, while yet it is based on a statement of a witness who says of himself that he is incredible.

The following abstract example illustrates the problems that can be caused by parallel self-defeat.

Example 4.4.5 Let $\mathcal{R}_d = \{p \Rightarrow q;\ r \Rightarrow \neg q;\ t \Rightarrow s\}$ and $\mathcal{K} = \{p, r, t\}$ while $\mathcal{R}_s$ consists of all propositionally valid inferences. Then:

A_1: p $\quad A_2$: $A_1 \Rightarrow q$
B_1: r $\quad B_2$: $B_1 \Rightarrow \neg q$ $\quad C$: $A_2, B_2 \rightarrow \neg s$
D_1: t $\quad D_2$: $D_1 \Rightarrow s$

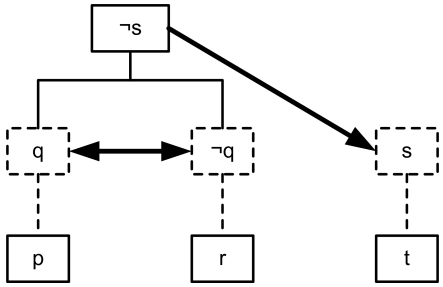


Figure 4.4: Illustrating trivialisation

Figure 4.4 displays these arguments and their attack relations. Argument C attacks D_2. Whether C defeats D_2 depends on the argument ordering but plausible argument orderings are possible in which $C \not\prec D_2$ and so C defeats D_2. This is problematic, since s can be any formula, so any defeasible argument unrelated to A_2 or B_2, such as D_2, can, depending on the argument ordering, be defeated by C. Clearly, this is extremely harmful, since the existence of just a single case of mutual rebutting defeat, which is very common, could trivialise the system. For instance, in this example neither of A_2 nor B_2 are in the grounded extension, since they defeat each other. But then the grounded extension does not defend D_2 against C and therefore does not contain D_2. This problem is sometimes called the *contamination* or *trivialisation problem*.

(Actually, if examples of parallel 'self-defeat' are translated into a Dung-style abstract argumentation framework, there are no abstract self-defeating arguments. Nevertheless, intuitively, this is a case of self-defeat, which is why it is discussed in this section.)

Grooters and Prakken (2016) propose the following formalisation of the property of trivialisation.

Definition 4.4.6 [Trivialising argumentation systems] An argumentation system AS is *trivialising* iff for all $\varphi, \psi \in \mathcal{L}$ and all knowledge bases $\mathcal{K}$ such that $\{\varphi, \neg\varphi\} \subseteq \mathcal{K}$ a strict argument on the basis of $\mathcal{K}$ can be constructed in AS with conclusion ψ.

The argumentation system in our example is clearly trivialising since $\mathcal{R}_s$ contains strict rules $\varphi, \neg\varphi \rightarrow \psi$ for all $\varphi, \psi \in \mathcal{L}$.

Concluding this subsection, there are good reasons to believe that the two types of self-defeating arguments should be treated differently: while arguments based on parallel self-defeat should always be overruled or should not be constructible, arguments with serial self-defeat should retain their force to prevent other arguments from being justified or defensible.

A variant of *ASPIC*$^+$ that avoids trivialisation

Grooters and Prakken (2016)[10] propose a modification of *ASPIC*$^+$ called *ASPIC** that avoids trivialisation of argumentation systems if the language and strict rules encode full classical logic. *ASPIC** imposes two additional constraints on the construction of arguments: (1) strict rules can only be applied to classically consistent sets of formulas, and (2) strict rules cannot be chained. This rules out C_1 in Example 4.4.5 and, moreover, rules out other problematic examples.

[10]This section reuses and adapts some fragments from their paper.

ASPIC* in fact adopts an inconsistency-tolerant variant of classical logic as the source of strict rules, namely, the so-called weak consequence relation originally proposed by Rescher and Manor (1970). Its basic idea is that a sentence weakly follows from a set S of sentences if it classically follows from at least one consistent subset of S. Weak consequence over a standard propositional language is formally defined as follows.

Definition 4.4.7 [**Weak consequence relation,** $\vdash_W$] $\Gamma \vdash_W \alpha$ if and only if there is a maximal consistent subset Δ of Γ such that $\Delta \vdash \alpha$ in classical logic.

Note that the word 'maximal' is in fact not required, since according to Lindenbaum's Lemma every consistent set of formulas can be extended into a maximally consistent one.

It is easy to see that $\{a, \neg a\} \vdash_W b$ does not hold, because $\{a, \neg a\}$ is not a maximal consistent subset of $\{a, \neg a\}$. Excluding the chaining of strict rules is motivated by the fact that weak consequence does not satisfy the Cut rule:

[Cut] $\Gamma, \alpha \vdash_W \beta$ and $\Gamma \vdash_W \alpha$, then $\Gamma \vdash_W \beta$.

For a counterexample, consider the set $\Gamma = \{a, \neg a \wedge b\}$. Then $\Gamma \vdash_W b$ and $\Gamma, b \vdash_W a \wedge b$, while it is not the case that $\Gamma \vdash_W a \wedge b$.

Since the **Cut** rule does not hold, a naive instantiation of *ASPIC*$^+$'s strict rules with weak consequence would not avoid explosion, as shown in the following example:

Example 4.4.8 Consider the following knowledge base $\mathcal{K}_p = \{p, \neg p, r\}$, $\mathcal{K}_n = \varnothing$, instantiate the strict rules with all valid inferences from finite sets in the logic W and let $\mathcal{R}_d = \varnothing$. Then the following arguments can be constructed:

$A_1 : p$ $\quad A_2 : A_1 \rightarrow p \vee \neg r$
$B : \neg p$ $\quad C : A_2, B \rightarrow \neg r$
$D : r$

Argument C concludes with $\neg r$.

The underlying reason for this problem is that the **Cut** rule does not hold for $\vdash_W$, so that in our example $\mathcal{K}_p \nvdash_W \neg r$. So if we want *ASPIC*$^+$'s strict part to behave according to $\vdash_W$, chaining of strict rules should be excluded.

Accordingly, the definition of an argument is in *ASPIC** as follows.

Definition 4.4.9 [**Argument*** **in** ***ASPIC****] An *argument** A on the basis of a knowledge base $\mathcal{K} = (\mathcal{K}, \preceq)$ in an argumentation system $(\mathcal{L}, \mathcal{R}, n, \preceq')$ is any structure obtainable by applying one or more of the following steps finitely many times:

1. φ if $\varphi \in \mathcal{K}$ with: $\texttt{Prem}(A) = \{\varphi\}$, $\texttt{Conc}(A) = \varphi$, $\texttt{Sub}(A) = \{\varphi\}$, $\texttt{DefRules}(A)$ = $\varnothing$, $\texttt{TopRule}(A)$ = undefined.

2. $A_1, \ldots, A_n \rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments* with a defeasible top rule or are from $\mathcal{K}$ and such that there exists a strict rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$ in $\mathcal{R}_s$.
   $\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,

$\texttt{Conc}(A) = \psi$,
$\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$,
$\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n)$,
$\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \rightarrow \psi$.

3. $A_1, \ldots, A_n \Rightarrow \psi$ if $A_1, \ldots, A_n$ are arguments* such that there exists a defeasible rule $\texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$ in $\mathcal{R}_d$.
$\texttt{Prem}(A) = \texttt{Prem}(A_1) \cup \ldots \cup \texttt{Prem}(A_n)$,
$\texttt{Conc}(A) = \psi$,
$\texttt{Sub}(A) = \texttt{Sub}(A_1) \cup \ldots \cup \texttt{Sub}(A_n) \cup \{A\}$,
$\texttt{DefRules}(A) = \texttt{DefRules}(A_1) \cup \ldots \cup \texttt{DefRules}(A_n) \cup \{\texttt{Conc}(A_1), \ldots \texttt{Conc}(A_n) \Rightarrow \psi\}$,
$\texttt{TopRule}(A) = \texttt{Conc}(A_1), \ldots, \texttt{Conc}(A_n) \Rightarrow \psi$.

Arguments* are just a special case of 'normal' arguments. Therefore, all definitions for (sets of) arguments are the same in case the term argument can be replaced by argument* without problems. *Attack* and *defeat* are just the same for arguments*. *Structured* and *abstract argumentation frameworks* for the *ASPIC** framework are just the same except that they only contain arguments*. Accordingly, the notions of justified and defensible arguments* and conclusions are still defined as in Section 4.3.3.

Grooters and Prakken (2016) prove for the case with any reasonable argument ordering that trivialisation is avoided by *ASPIC**. They also modify the rationality postulates of strict closure and indirect consistency (see Section 4.4.4) to exclude chaining of strict rules and they then prove that the modified postulates are satisfied under the same conditions as the original ones for *ASPIC*$^+$.

4.4.3 Choosing defeasible inference rules

Let us return to the question of how to choose the defeasible rules. Can we derive them from a logic of our choice just as we can derive the strict rules from a logic of our choice if we want to? This is controversial. Some philosophers argue that all rule-like structures that we use in daily life are "inference licences" and so cannot be expressed in the logical object language. In this view, all that can be done is apply them to formulas from $\mathcal{L}$ to support new formulas from $\mathcal{L}$. That is, these philosophers see all defeasible generalisations as inference rules, whether they are domain-specific or not.

Others (usually logicians) take a more standard-logic approach (e.g. Kraus et al. (1990); Pearl (1992)). They say that all contingent knowledge should be expressed in the object language, so they reject the idea of domain-specific defeasible inference rules (for the same reason they don't like domain-specific strict rules). They would introduce a new connective into $\mathcal{L}$, let us write it as $\rightsquigarrow$, where they informally read $p \rightsquigarrow q$ as something like "If p then normally/typically/usually q". They then want to give a model-theoretic semantics for this connective just as logicians give a model-theoretic semantics for all connectives. The main difference is that such semantics for defeasible conditionals do not look at *all* models of a theory to check whether it entails a formula (as semantics for deductive logics do) but only to a *preferred class* of models of the theory (for example, all models where things are as normal as possible). They would then add a strict inference rule $S \rightarrow \varphi$ to $\mathcal{R}_s$ just in case φ is true in *all* models of S, while they would add a defeasible inference rule $S \Rightarrow \varphi$ to $\mathcal{R}_d$ just in case φ is true in all *preferred* models of S but *not* in all models of S.

Now what inference rules for $\rightsquigarrow$ could result from such an approach? On two things there is consensus between logicans: modus ponens for $\rightsquigarrow$ is defeasibly but not deductively valid, so the rule $\varphi \rightsquigarrow \psi, \varphi \Rightarrow \psi$ should go into $\mathcal{R}_d$. There is also consensus that contraposition for $\rightsquigarrow$ is deductively invalid, so the rule $\varphi \rightsquigarrow \psi \rightarrow -\psi \rightsquigarrow -\varphi$ should *not* go into $\mathcal{R}_s$. However, here the consensus ends. Should the defeasible analogue of this rule go into $\mathcal{R}_d$ or not? Opinions differ at this point[11].

Let us illustrate the difference between the two approaches with a further variation on Example 4.4.1. Above we used the approach where all defeasible generalisations are inference rules. We now replace the two domain-specific defeasible inference rules d_1 and d_2 with two object-level conditionals expressed in $\mathcal{L}$ and now add them to $\mathcal{K}_p$:

$$WearsRing \rightsquigarrow Married$$
$$PartyAnimal \rightsquigarrow Bachelor$$

Moreover, we add defeasible modus ponens for $\rightsquigarrow$ to $\mathcal{R}_d$:

$$\mathcal{R}_d = \{\varphi \rightsquigarrow \psi, \varphi \Rightarrow \psi\}$$

The arguments then change as follows (assuming the crude way of incorporating classical logic):

A_1:	$WearsRing$	B_1:	$PartyAnimal$
A_2:	$WearsRing \rightsquigarrow Married$	B_2:	$PartyAnimal \rightsquigarrow Bachelor$
A_3:	$A_1, A_2 \Rightarrow Married$	B_3:	$B_1, B_2 \Rightarrow Bachelor$
A_4:	$Married \supset \neg Bachelor$	B_4:	$Married \supset \neg Bachelor$
A_5:	$A_3, A_4 \rightarrow \neg Bachelor$	B_5:	$B_3, B_4 \rightarrow \neg Married$

Now A_5 rebuts B_5 on B_3 while B_5 rebuts A_5 on A_3.

Concluding, if you want, you can base at least some of your choices concerning defeasible inference rules on model-theoretic semantics for nonmonotonic logics. However, it is an open question whether a model-theoretic semantics is the *only* criterion by which we can choose our defeasible rules. Some have based their choice on other criteria, since they do not primarily see defeasible rules as logical inference rules but as principles of human cognition or rational action, so that they should be based on foundations other than semantics. For example, John Pollock based his defeasible reasons on his account of epistemology (the part of philosophy that studies how we can obtain knowledge). Others have based their choice of defeasible reasons on the study of argument schemes in informal argumentation theory. We give examples of both these approaches in Section 4.4.5.

Naming defaults in first-order languages

We finally illustrate some subtleties of the naming convention for defeasible rules. If domain-specific defeasible rules are defined over a first-order language, then the same notational naming convention is often used as for defaults in default logic. A rule with free variables is used as a scheme for all its ground instances, that is, for all its instances in which the variable x is replaced by a ground term from $\mathcal{L}$. Moreover, the scheme is often given a name $d(x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are all free variables that occur in the scheme. Such a name allows the formulation of undercutters to a rule. Consider, for example:

[11] See Chapter 4 of Caminada (2004) for a very readable overview of the discussion.

$d(x)$: $\texttt{Bird}(x) \Rightarrow \texttt{Flies}(x)$

Then schemes for undercutters can be written as follows:

$u(x)$: $\texttt{Penguin}(x) \Rightarrow \neg d(x)$

To see how this naming convention can be used, consider the following knowledge base:

$K_n = \{\forall x(\texttt{Penguin}(x) \supset \texttt{Bird}(x))\}$
$K_p = \{\texttt{Penguin}(Tweety), \texttt{Bird}(Polly)\}$

Then two arguments can be constructed for the conclusions that Tweety and Polly can fly (the strict rules are assumed to be all valid first-order inferences):

A_1: $\texttt{Penguin}(Tweety)$
A_2: $\forall x(\texttt{Penguin}(x) \supset \texttt{Bird}(x))$
A_3: $A_1, A_2 \rightarrow \texttt{Bird}(Tweety)$
A_4: $A_3 \Rightarrow \texttt{Flies}(Tweety)$

B_1: $\texttt{Bird}(Polly)$
B_2: $B_1 \Rightarrow \texttt{Flies}(Polly)$

However, only for Tweety can an undercutter be constructed:

C_1: $\texttt{Penguin}(Tweety)$
C_2: $C_1 \Rightarrow \neg d(Tweety)$

The point is that $d(x)$ is not a rule name but a rule name scheme, and only for its instance $d_1(Tweety)$ can an undercutter be constructed. If, by contrast, the birds-fly rule had been named with d, then applying the undercutter for Tweety would also block the default for Polly, which is clearly undesirable.

4.4.4 Satisfying rationality postulates

We are now in a position to state under what conditions *ASPIC*$^+$ satisfies Caminada and Amgoud (2007)'s four rationality postulates. These are listed below (it is helpful to refer to concepts defined in Definition 4.3.3 when reading these postulates), adapted to the *ASPIC*$^+$ framework.[12]

Definition 4.4.10 [Rationality postulates for *ASPIC*$^+$] Let *(c-)SAF* $= (\mathcal{A}, \mathcal{C}, \preceq)$ be an *ASPIC*$^+$ (c-)structured argumentation framework defined by an *ASPIC*$^+$ AT with $AS = (\mathcal{L}, \mathcal{R}, n)$ and $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p$. Let AF be the abstract argumentation framework corresponding to *(c-)SAF* and let $T \in \{\text{complete, preferred, grounded, stable}\}$. Then:

- *(c-)SAF* satisfies the *closure under subarguments postulate* iff for all T-extensions E of AF it holds that if an argument A is in E then all subarguments of A are in E;
- *(c-)SAF* satisfies the *direct consistency postulate* iff for all T-extensions E of AF it holds that $\texttt{Conc}(E)$ is directly consistent;
- *(c-)SAF* satisfies the *indirect consistency postulate* iff for all T-extensions E of AF it holds that $\texttt{Conc}(E)$ is ndirectly consistent;

[12] Caminada and Amgoud (2007) also propose postulates for the intersection of extensions and their conclusion sets, but since their satisfaction directly follows from satisfaction of the postulates for individual extensions, these postulates will below be ignored.

- *(c-)SAF* satisfies the *strict closure postulate* iff for all T-extensions E of AF it holds that $\texttt{Conc}(E) = Cl_{\mathcal{R}_s}(\texttt{Conc}(E))$.

The first postulate, closure under subarguments, holds unconditionally for the present framework.

Proposition 4.4.11 Let $(\mathcal{A}, \mathcal{D})$ be an abstract argumentation framework as defined in Definition 4.3.16 and E any of its grounded, preferred or stable extensions. Then

- for all $A \in E$: if $A' \in \texttt{Sub}(A)$ then $A' \in E$;
- $\texttt{Conc}(E) = Cl_{\mathcal{R}s}(\texttt{Conc}(E))$.

The two consistency postulates do not hold in general.

Example 4.4.12 A simple counterexample to consistency is with two defeasible rules d_1: $\Rightarrow p$ and d_2: $\Rightarrow q$ and a strict rule $p \rightarrow \neg q$, where $d_1 < d_2$. Then with the weakest- or last-link ordering the argument for $\neg q$ does not defeat the argument for q so in all semantics we have a single extension with both arguments.

We now discuss under which conditions the consistency postulates are satisfied.

Depending on the choices outlined in Section 4.4.1, the first requirement for satisfying the consistency postulates is that your argumentation theory is closed under transposition or contraposition. This is because if neither property is satisfied, then since strict rule applications cannot be attacked, direct consistency may then be violated. This can be illustrated with the first version of Example 4.4.1. Suppose we only have one strict rule, namely, s_1. we cannot construct B_3, since B_3 applies the now missing rule s_2. We still have that A_3 rebuts B_2. Suppose now that $d_1 < d_2$ and we apply the last-link argument ordering. Then A_3 does not defeat B_2. In fact, no argument in the example is defeated, so we end up with a single extension in all semantics, which contains arguments for both $Bachelor$ and $\neg Bachelor$ and so violates direct and indirect consistency.

However, with transposition this bad outcome is avoided: if we also have s_2, then argument B_3 can be constructed, which rebuts A_3 on A_2. Again applying the preference $d_1 < d_2$ with the last-link ordering, we have that B_3 strictly defeats A_2. Again we have a unique extension in all semantics, containing all arguments except A_2 and A_3. This extension does not violate consistency.

Example 4.4.13 Consider Example 4.3.6. As discussed in Example 4.3.18, if the argument ordering is such that C_3 does not defeat B_1, then both arguments will be in the same extension, which thus violates consistency since the conclusions of these arguments contradict each other. However, if the transposition $s \rightarrow \neg v$ of $v \rightarrow \neg s$ is added to $\mathcal{R}_s$, then B_1 can be continued to an argument for $\neg v$, which successfully rebuts C_3 on C_2, excluding the consistency-violating extensions.

Some say that the above violation of consistency, before inclusion of the transposed rule, arises because *ASPIC*$^+$ forbids attacks on strictly derived conclusions. Consistency would not be violated if B_2 was allowed to attack $A3$ in the first version of Example 4.4.1. However, apart from the reasons discussed in Section 4.2, there is another reason for prohibiting attacks on strictly derived conclusions: if they are allowed, then extensions may not be strictly closed or indirectly consistent, even if the strict rules are

closed under transposition. To see why, suppose we changed *ASPIC*$^+$'s definitions to allow attacks on strict conclusions, so that B_2 attacks A_3, A_2 attacks B_3, and A_3 and B_3 attack each other in Example 4.4.1. Suppose also that all knowledge-base items and all defeasible rules in the example are of equal preference, and suppose we apply the weakest- or last-link argument ordering. Then all rebutting attacks in the example succeed. But then the set $\{A_1, A_2, B_1, B_2\}$ is admissible and is in fact both a stable and preferred extension. But this violates the rationality postulates of strict closure and indirect consistency. The extension contains an argument for $Bachelor$ but not for $\neg Married$, which strictly follows from it by rule s_2. Likewise, the extension contains an argument for $Married$ but not for $\neg Bachelor$, which strictly follows from it by rule s_1. So the extension is not closed under strict rule application. Moreover, the extension is indirectly inconsistent, since its strict closure contains both $Married$ and $\neg Married$, and both $Bachelor$ and $\neg Bachelor$.

Other requirements for satisfying the consistency postulates are that the axioms $\mathcal{K}_n$ are indirectly consistent (axiom consistency) and the preference ordering is *reasonable*. The rationale for requiring the former is self-evident. A reasonable argument ordering essentially amounts to requiring that: 1) arguments that are both strict and firm are strictly preferred over all other arguments; 2) the strength (and implied relative preference) of an argument is determined exclusively by the defeasible rules and/or ordinary premises; 3) the preference ordering is acyclic, and if $B \prec A$ then it must be that $B' \prec A$ where B' is some maximal fallible (i.e., defeasible or plausible) sub-argument of B (for example in our running example C_2 but not C_1 is a maximal fallible argument of C_3). We refer the reader to Modgil and Prakken (2013) for the technical definition of a reasonable ordering; suffice to say that it has been shown that the weakest- and last-link argument orderings of Section 4.3.4 are reasonable.

We are now in a position to state an important result proved in Modgil and Prakken (2013) that if your *(c-)SAF* is *well-defined*, in that its argumentation theory satisfies axiom consistency, and transposition or contraposition, and your argument preference ordering is reasonable, then the consistency postulates are satisfied by the *ASPIC*$^+$ framework as defined in Section 4.3.

Theorem 4.4.14 Let $(\mathcal{A}, \mathcal{D})$ be an abstract argumentation framework corresponding to a well-defined *(c-)SAF* and let E be any of its grounded, preferred or stable extensions. Then

- $\texttt{Conc}(E)$ is consistent;
- $Cl_{\mathcal{R}_s}(\texttt{Conc}(E))$ is consistent.

Finally, note that if you do not include any strict rules or axiom premises in your argumentation theory, then the requirement that your *(c-)SAF* be well defined obviously does not apply, but it is also worth noting that the preference ordering need *not* be reasonable in order that all four rationality postulates be satisfied (indeed no assumptions as to the properties of the preference ordering are required in this case).

4.4.5 Using *ASPIC*$^+$ to model argument schemes

We concluded Section 4.4.3 by remarking on the use of defeasible inference rules as principles of cognition in John Pollock's work and as argument schemes in informal

argumentation theory. We now illustrate how both approaches can be formalised in *ASPIC*$^+$ and how strict inference rules can also be accommodated when doing so.

Let us first look in more detail at John Pollock's work. He formalised defeasible rules for reasoning patterns involving perception, memory, induction, temporal persistence and the statistical syllogism, as well as undercutters for these reasons.

In *ASPIC*$^+$ his principles of perception and memory can be written as follows:

$d_p(x, \varphi)$: $\quad \texttt{Sees}(x, \varphi) \Rightarrow \varphi$
$d_m(x, \varphi)$: $\quad \texttt{Recalls}(x, \varphi) \Rightarrow \varphi$

In fact, these defeasible inference rules are schemes for all their ground instances (that is, for any instance where x and φ are replaced by ground terms denoting a specific perceiving agent and a specific perceived state of affairs). Therefore, their names $d_p(x, \varphi)$ and $d_m(x, \varphi)$ as assigned by the n function are in fact also schemes for names. A proper name is obtained by instantiating these variables by the same ground terms as used to instantiate these variables in the scheme. Thus it becomes possible to formulate undercutters for one instance of the scheme (say for Jan who saw John in Amsterdam) while leaving another instance unattacked (say for Bob who saw John in Holland Park). Note, finally, that these schemes assume a naming convention for formulas in a first-order language, since φ is a term in the antecedent while it is a well-formed formula in the consequent. In the remainder we will leave this naming convention implicit.

Now undercutters for d_p state circumstances in which perceptions are unreliable, while undercutters of d_m state conditions under which memories may be flawed. For example, a well-known cause of false memories of events is that the memory is distorted by, for instance, seeing pictures in the newspaper or watching a TV programme about the remembered event. A general undercutter for distorted memories could be

$u_m(x, \varphi)$: $\quad \texttt{DistortedMemory}(x, \varphi) \Rightarrow \neg d_m(x, \varphi)$

combined with information such as

$\forall x, \varphi(\texttt{SeesPicturesAbout}(x, \varphi) \supset \texttt{DistortedMemory}(x, \varphi))$

Pollock's epistemic inference schemes are in fact a subspecies of argument schemes. The notion of an argument scheme was developed in philosophy and is currently an important topic in the computational study of argumentation. Argument schemes are stereotypical non-deductive patterns of reasoning, consisting of a set of premises and a conclusion that is presumed to follow from them. Uses of argument schemes are evaluated in terms of critical questions specific to the scheme. An example of an epistemic argument scheme is the scheme from the position to know (Walton; 1996, pp. 61–63):

A is in the position to know whether P is true
A asserts that P is true

P is true

Walton gives this scheme three critical questions:

1. Is A in the position to know whether P is true?
2. Did A assert that P is true?
3. Is A an honest (trustworty, reliable) source?

A natural way to formalise reasoning with argument schemes is to regard them as defeasible inference rules and to regard critical questions as pointers to counterarguments.

For example, in the scheme from the position to know questions (1) and (2) point to underminers (of, respectively, the first and second premise) while questions (3) points to undercutters (the exception that the person is for some reason not credible).

Accordingly, we formalise the position to know scheme and its undercutter as follows:

$d_w(x,\varphi)$: $\texttt{PositionToKnow}(x,\varphi), \texttt{Says}(x,\varphi) \Rightarrow \varphi$
$u_w(x,\varphi)$: $\neg\texttt{Credible}(x) \Rightarrow \neg d_w(x,\varphi)$

We will now illustrate the modelling of both Pollock's defeasible reasons and Walton's argument schemes with our example from Section 4.2, focusing on a specific class of persons who are in the position to know, namely, witnesses. In fact, witnesses always report about what they observed in the past, so they will say something like "I remember that I saw that John was in Holland Park". Thus an appeal to a witness testimony involves the use of three schemes: first the position to know scheme is used to infer that the witness indeed remembers that he saw that John was in Holland Park, then the memory scheme is used to infer that he indeed saw that John was in Holland Park, and finally, the perception scheme is used to infer that John was indeed in Holland Park. Now recall that John was a suspect in a robbery in Holland Park and that Jan testifed that he saw John in Amsterdam on the same morning, while Jan is a friend of John. Suppose now we also receive information that Bob read newspaper reports about the robbery in which a picture of John was shown. One way to model this in *ASPIC*$^+$ is as follows.

The knowledge base consists of the following facts (since we don't want to dispute them, we put them in $\mathcal{K}_n$):

f_1: $\texttt{PositionToKnow}(Bob, \texttt{Recalls}(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John))))$
f_2: $\texttt{Says}(Bob, \texttt{Recalls}(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John))))$
f_3: $\texttt{SeesPicturesAbout}(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John)))$
f_4: $\forall x,\varphi.(\texttt{SeesPicturesAbout}(x,\varphi) \supset \texttt{DistortedMemory}(x,\varphi))$
f_5: $\forall x.\texttt{InHollandPark}(x) \supset \texttt{InLondon}(x)$
f_6: $\texttt{PositionToKnow}(Jan, \texttt{Recalls}(Jan, \texttt{Sees}(Jan, \texttt{InAmsterdam}(John))))$
f_7: $\texttt{Says}(Jan, \texttt{Recalls}(Jan, \texttt{Sees}(Jan, \texttt{InAmsterdam}(John))))$
f_8: $\texttt{Friends}(Jan, John)$
f_9: $\texttt{SuspectedRobber}(John)$
f_{10}: $\forall x,y,\varphi.\texttt{Friends}(x,y) \wedge \texttt{SuspectedRobber}(y) \wedge \texttt{InvolvedIn}(y,\varphi) \supset \neg\texttt{Credible}(x)$
f_{11}: $\texttt{InvolvedIn}(John, \texttt{Recalls}(Jan, \texttt{Sees}(Jan, \texttt{InAmsterdam}(John))))$
f_{12}: $\forall x\neg(\texttt{InAmsterdam}(x) \wedge \texttt{InLondon}(x))$

Combining this with the schemes from perception, memory and position to know, we obtain the following arguments (for reasons of space we don't list separate lines for arguments that just take an item from $\mathcal{K}$).

A_3: $f_1, f_2 \Rightarrow_{dw} \texttt{Recalls}(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John)))$
A_4: $A_3 \Rightarrow_{dm} \texttt{Sees}(Bob, \texttt{InHollandPark}(John))$
A_5: $A_4 \Rightarrow_{dp} \texttt{InHollandPark}(John)$
A_7: $A_5, f_5 \rightarrow \texttt{InLondon}(John)$

This argument is undercut (on A_4) by the following argument applying the undercutter for the memory scheme:

B_3: $f_3, f_4 \rightarrow \texttt{DistortedMemory}(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John)))$
B_4: $B_3 \Rightarrow_{um} \neg d_m(Bob, \texttt{Sees}(Bob, \texttt{InHollandPark}(John)))$

Moreover, A_7 is rebutted (on A_5) by the following argument:

C_3: $f_6, f_7 \Rightarrow_{dw} \texttt{Recalls}(Jan, \texttt{Sees}(Jan, \texttt{InAmsterdam}(John)))$
C_4: $C_3 \Rightarrow_{dm} \texttt{Sees}(Jan, \texttt{InAmsterdam}(John))$
C_5: $C_4 \Rightarrow_{dp} \texttt{InAmsterdam}(John)$
C_8: $C_5, f_5, f_{12} \rightarrow \neg\texttt{InHollandPark}(John)$

This argument is also undercut, namely, on C_3 based on the undercutter of the position to know scheme:

D_5: $f_8, f_9, f_{10}, f_{11} \rightarrow \neg\texttt{Credible}(Jan)$
D_6: $D_5 \Rightarrow_{uw} \neg d_w(Jan, \texttt{Recalls}(Jan, \texttt{Sees}(Jan, \texttt{InAmsterdam}(John))))$

Finally, C_8 is rebutted on C_5 by the following continuation of argument A_7:

A_8: $A_5, f_5, f_{12} \rightarrow \neg\texttt{InAmsterdam}(John)$

A_8 is in turn undercut by B_4 (on A_4) and rebutted by C_8 (on A_5).

The example is displayed in Figure 4.5.

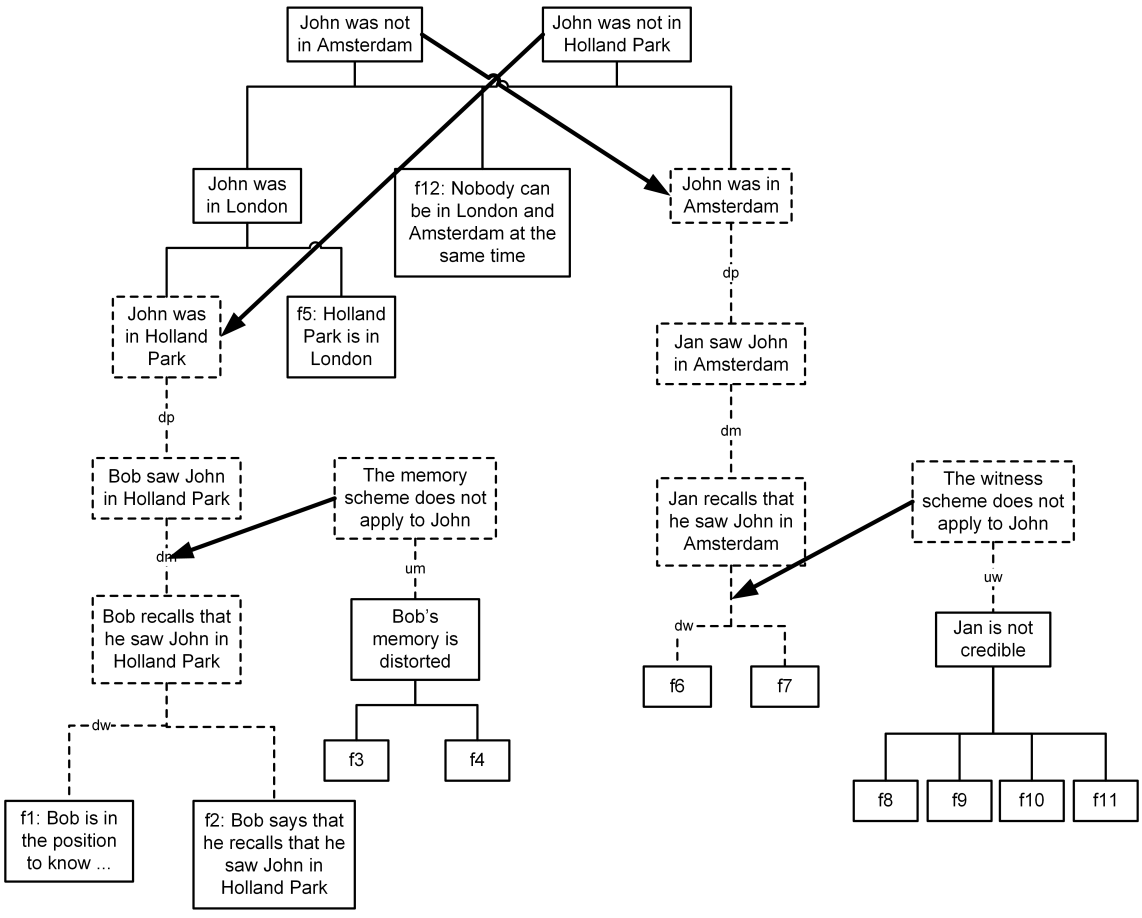


Figure 4.5: A formalised example

Because of the two undercutting arguments, neither of the testimony arguments are credulously or sceptically justified in any semantics. Let us see what happens if we do

not have the two undercutters. Then we must apply preferences to the rebutting attack of C_8 on A_5 and to the rebutting attack of A_8 on C_5. As it turns out, the same preferences have to be applied in both cases, namely, those between the three defeasible-rule applications in the respective arguments. And this is what we intuitively want.

Finally, we note that counterarguments based on critical questions of argument schemes may themselves apply argument schemes. For example, we may believe that Jan and John are friends because another witness told our so. Or we may believe that Holland Park is in London because a London taxi driver told us so (an application of the so-called expert testimony scheme).

4.4.6 Instantiations with no defeasible rules

All that has been said so far about ways to choose the strict rules applies irrespective of whether you also want to include defeasible rules in your argumentation system. In fact, *ASPIC*$^+$ allows you to only use strict inference rules. Principled ways to do so are to base the strict rules on classical logic or indeed on any Tarskian consequence relation. In this way, *ASPIC*$^+$ extends the classical-logic approach of Besnard and Hunter (2009) and the abstract-logic approach of Amgoud and Besnard (2009), by providing guidelines for using preferences to resolve inconsistencies in classical logic or any other underlying Tarksian logic. The use of preferences is of particular importance in such contexts, since in these contexts the stable and preferred extensions of Dung frameworks simply correspond to the maximal consistent subsets of the instantiating theories (Amgoud and Besnard; 2013). One thus needs some 'extra-logical' means, such as preferences, to resolve inconsistencies.

The idea is as follows. Given a set S of wff in some language $\mathcal{L}$ and a Tarksian consequence relation Cn over $\mathcal{L}$ (note that classical consequence is such a Tarskian consequence relation), we let the axioms and defeasible inference rules be empty, and the strict rules defined as indicated in Section 4.4.1, namely, as $S \rightarrow p \in \mathcal{R}_s$ iff $p \in Cn(S)$, for any finite $S \subseteq \mathcal{L}$. Furthermore, in keeping with the above mentioned classical, and more general Tarskian Logic approaches, we assume all arguments to be consistent and, moreover, their premise sets subset-minimal in applying their conclusion. So argumentation theories defined this way define *(c-)SAFs* instead of *(c-)SAFs*.

For this special case all *ASPIC*$^+$ arguments are strict, so all attacks are undermining attacks. In Modgil and Prakken (2013) it was shown that these *ASPIC*$^+$ reconstructions of Tarskian and classical approaches are equivalent to the originals if these originals use a form of undermining attack. Moreover, the result stated in Section 4.4.1 – that any *ASPIC*$^+$ AT with the strict rules derived from a Tarskian logic satisfies closure under contraposition — then implies that without preferences these reconstructions are well-defined and thus satisfy the rationality postulates. Moreover, if these reconstructions are extended with a reasonable argument ordering, then this result also holds for the case with preferences. Thus the *ASPIC*$^+$ framework has in fact been used to extend both the classical-logical approach of Besnard and Hunter (2009) and the more general Tarskian approach of Amgoud and Besnard (2009) with preferences in a way that satisfies all rationality postulates of Caminada and Amgoud (2007).

Note, finally, that for these instantiations there is no contamination problem, since if arguments have contradictory conclusions, then they also have mutually inconsistent premises, so they cannot be in *(c-)SAFs*.

4.4.7 Illustrating uses of *ASPIC*$^+$ with and without defeasible rules

In this section we compare respective uses of *ASPIC*$^+$ with and without defeasible rules in more detail. We first say more about the arguments of some that classical-logic simulations of defeasible rules may yield counterintuitive results. Let us assume a classical-logic instantiation of *ASPIC*$^+$ as defined in Section 4.4.6 and formalise natural-language generalisations 'If P then normally Q' as material implications $P \supset Q$ put in $\mathcal{K}_p$. The idea is that since $P \supset Q$ is an ordinary premise, its use as a premise can be undermined in exceptional cases. Observe that by classical reasoning we then have a strict argument for $\neg Q \supset \neg P$. Some say that this is problematic. Consider the following example: 'Anyone who is a man usually has no beard', so (strictly) 'Anyone who has a beard usually is not a man'. This strikes some as counterintuitive, since we know that virtually everyone who has a beard is a man, so the contraposition of 'If P then normally Q' cannot be deductively valid[13].

A more refined classical approach is to give the material implication an extra normality condition N, which informally reads as 'everything is normal as regards P implying Q', and which is also put in $\mathcal{K}_p$. The idea then is that exceptional cases give rise to underminers of N. However, $(P \wedge N) \supset Q$ also deductively contraposes, namely, as $(\neg Q \wedge N) \supset \neg P$, so it seems that we still have the controversial deductive validity of contraposition for generalisations (in the beard and men example the contraposition of the rule with the added normality condition would read: 'Anyone who has a beard and all is normal regarding men and having beards, usually is not a man' !).

So far we only discussed reasons for belief but argumentation is often about what to do, prefer or value (what philosophers often call *practical reasoning*). Here too it has been argued on philosophical grounds that reasons for doing, preferring or valuing cannot be expressed in classical logic since they do not contrapose. This view can, of course, not be based on a statistical semantics for such reasons, since statistics only applies to reasoning about what is the case (what philosophers often call *epistemic reasoning*). Space limitations prevent us from giving more details about these philosophical arguments.

We next illustrate two different ways to use *ASPIC*$^+$ with a detailed example. Both ways use classical logic in their strict part and use explicit preferences, but only the second way uses defeasible inference rules. The first way instead expresses defeasible generalisations as material implications with normality assumptions. The example will shed further light on the issue whether empirical generalisations can be represented in classical logic, and it will also motivate the use of axiom premises. Our example is a well-known one from the literature on nonmonotonic logic. Suppose a defeasible reasoner accepts all following natural-language statements are true. For the generalisations (1) and (2) this means that the reasoner accepts that they hold in general but that they may have exceptions.

[13] One way to argue why classical simulations may give counter-intuitive results is to recall that a number of researchers provide statistical semantics for defeasible inference rules. These semantics regard a defeasible rule of the form $P \Rightarrow Q$ as a qualitative approximation of the statement that the conditional probability of Q, given P, is high. The laws of probability theory then tell us that this does not entail that the conditional probability of $\neg P$, given $\neg Q$, is high. The problem with the classical-logic approach is then that it conflates this distinction by turning the conditional probability of Q given P into the unconditional probability of $P \supset Q$, which then has to be equal to the unconditional probability of $\neg Q \supset \neg P$.

(1) Birds normally fly
(2) Penguins normally don't fly
(3) All penguins are birds
(4) Penguins are abnormal birds with respect to flying
(5) Tweety is a penguin

A defeasible reasoner then wants to know what can be concluded from this information about whether Tweety can fly. It seems uncontroversial to say that any defeasible reasoner will conclude that Tweety can fly.

We now formalise these statements with the just-explained method to represent empirical generalisations as material implications with explicit normality assumptions. We use a classical-logic instantiation of *ASPIC*$^+$ with preferences as defined above in Section 4.4.6.

(1) $bird \wedge \neg ab_1 \supset canfly$
(2) $penguin \wedge \neg ab_2 \supset \neg canfly$
(3) $penguin \supset bird$
(4) $penguin \supset ab_1$
(5) $penguin$

Let us first add these formulas to $\mathcal{K}_p$. The idea now is that the normality assumptions of a defeasible reasoner are expressed as additional statements $\neg ab_1$ and $\neg ab_2$, also added to $\mathcal{K}_p$. We then define the preference ordering on $\mathcal{K}_p$ such that all of (1-5) are strictly preferred over any of these two assumptions and that $\neg ab_1 <' \neg ab_2$.

We can then construct many arguments on the issue whether Tweety can fly. Note that $\{1, 2, 3, 4, 5\} \cup \{\neg ab_1, \neg ab_2\}$ is minimally inconsistent, so if we take any single element out, the rest can be used to build an argument against it. This means that we can formally build arguments not just against the two normality assumptions but also against any of (1-5) (note the similarity with the fact that, as noted above, in classical-logic argumentation without preferences the stable and preferred extensions corespond to maximal consistent subsets of the knowledge base). With the weakest- or last-link ordering we do obtain the intuitive conclusion $\neg canfly$, but the fact that arguments against any of (1-5) can be built may be regarded as somewhat odd, since we just noted that a defeasible reasoner accepts (1-5) as given and is only interested in what follows from them.

Let us therefore move (1-5) to the axioms $\mathcal{K}_n$, so that they cannot be attacked. Then we have just a few arguments on the issue whether Tweety can fly: we have an argument $\{1, 2, 3, 4, 5\} \cup \{\neg ab_2\} \rightarrow \neg canfly$, which has one attacker, namely, $\{1, 2, 3, 5\} \cup \{\neg ab_1\} \rightarrow ab_2$. However, with the weakest- or last link principle this attacker does not defeat it target, since we have $\neg ab_1 <' \neg ab_2$. Hence $\neg canfly$ is justified in any semantics. So at first sight it would seem that the classical-logic approach enriched with axiom premises adequately models reasoning with empirical generalisations.

However, this approach still has some things to explain, as can be illustrated by changing our example a little: above it was given as a matter of fact that Tweety is a penguin but in reality the particular 'facts' of a problem are often not simply given but derived from information sources (sensors, testimonies, databases, the internet, and so on). Now in reality none of these sources is fully reliable, so inferring facts from them can only be done under the assumption that things are normal. So let us change the example by saying that Tweety was observed to be a penguin and that animals that are observed to be penguins *normally* are penguins. We change 5 to $5'$ and we add 6 to $\mathcal{K}_n$:

(5') $observed\_as\_penguin$
(6) $observed\_as\_penguin \wedge \neg ab_3 \supset penguin$

Moreover, we add $\neg ab_3$ to $\mathcal{K}_p$. We can still build an argument that Tweety cannot fly, namely, $\{1, 2, 3, 4, 5'\} \cup \{\neg ab_2, \neg ab_3\} \rightarrow \neg canfly$. However, we can also build an attacker of this argument, namely $\{1, 2, 3, 4, 5', 6\} \cup \{\neg ab_1, \neg ab_2\} \rightarrow ab_3$. We can still obtain the intuitive outcome by preferring the assumption $\neg ab_3$ over the assumption $\neg ab_1$. However, some have argued that this is an ad-hoc solution, since there would be no general principle on which such a preference can be based. The heart of the problem, they say, is the fact that the material implication satisfies contraposition, a property which, as we just mentioned, can be argued to be too strong for defeasible generalisations. In reality a defeasible reasoner would not even construct an argument against $penguin$. As can be easily checked, the same issues arise if we put (1-4,5',6) in $\mathcal{K}_p$ while we then have our old issue back that arguments can be constructed against any element of $\mathcal{K}_p$.

Concluding so far, those who want to model 'default reasoning' in classical argumentation have to explain why arguments as the one for ab_3 can be constructed and why it does not defeat the argument for $\neg canfly$ (or alternatively, why the latter conclusion is not justified). Moreover, if they apply the first version of this approach, by putting all of $\{1, 2, 3, 4, 5', 6\}$ in $\mathcal{K}_p$, then they also have to explain why arguments against any of these premises can be constructed and whether these arguments succeed as defeats.

Let us next formalise the example with domain-specific defeasible rules and with the strict rules still corresponding to classical logic.

d_1: $bird \Rightarrow canfly$
d_2: $penguin \Rightarrow \neg canfly$
d_3: $observed\_as\_penguin \Rightarrow \neg penguin$
f_1: $penguin \supset bird$
f_2: $penguin \supset \neg r_1$
f_3: $observed\_as\_penguin$

It now does not matter whether we put the facts in $\mathcal{K}_n$ or $\mathcal{K}_p$, nor does it matter which priorities we define on $\mathcal{K}_p$ or $\mathcal{R}_d$. We have the following arguments:

A_1: $observed\_as\_penguin$ $\quad$ B_1: $A_2 \Rightarrow \neg canfly$
A_2: $A_1 \Rightarrow penguin$
A_3: $penguin \supset bird$
A_4: $A_2, A_3 \Rightarrow canfly$ $\quad$ C_1: $A_2 \Rightarrow \neg r_1$

Note also that no argument can be built against the conclusion $penguin$. We have that A_4 and B_1 rebut each other while C_1 undercuts A_4. Whatever the argument ordering between A_4 and B_1, we thus obtain that the conclusion $\neg canfly$ is justified in any semantics.

Concluding, the classical modelling of this example is simpler in that it only uses classical inference and does not have to rely on the notion of a defeasible inference rule. On the other hand, to obtain the intuitive outcome it needs more preferences than the modelling with defeasible rules, while the issue arises on which grounds these preferences can be stated. Moreover, if the classical approach regards all knowledge as fallible in principle, then it generates many more arguments than perhaps intuitively expected, at least many more than in the modelling with defeasible rules.

4.4.8 Representing facts

ASPIC$^+$ allows you to represent facts in various ways, each with their pros and cons. *Disputable facts* φ can either be put as such in $\mathcal{K}_p$ or as defeasible rules $\Rightarrow \varphi$ with empty antecedents. An advantage of including disputable facts in $\mathcal{K}_p$ is that thus *ASPIC*$^+$ captures classical and abstract-logic argumentation with preferences as special cases. On the other hand, if disputable facts φ are represented as defeasible rules $\Rightarrow \varphi$, then the definition of the weakest- and last-link argument orderings becomes simpler, since then only sets of defeasible rules need to be compared. In addition, this choice removes the need for undermining attack, which simplifies the definitions of attack and defeat.

Undisputable facts φ can either be put as such in $\mathcal{K}_n$ or as strict rules $\rightarrow \varphi$ with empty antecedents. This choice does not make a difference for the weakest- or last-link argument ordering, since these orderings disregard axiom premises and strict rules. However, a disadvantage of representing undisputable fact φ as strict rules $\rightarrow \varphi$ is that then the strict rules do not express a logic any more, so the above-mentioned theorems on definitions of $\mathcal{R}_s$ in terms of Tarskian abstract logics do not apply any more.

4.4.9 Summary

We have seen that *ASPIC*$^+$ allows you to make any choice of axioms, strict and defeasible rules you like. You can choose domain-specific strict and/or defeasible inference rules, and you can choose logical strict and/or defeasible inference rules, for any deductive and/or nonmonotonic logic of your choice, good or bad. You can add logical axioms to $\mathcal{K}_n$ but you can also add any other information to $\mathcal{K}_n$ that you don't want to put up for discussion. You can also base your defeasible rules on informal accounts of argument schemes. All that *ASPIC*$^+$ tells you is how arguments can be built with your rules of choice, how they can be attacked, and how these attacks can be resolved, given an argument ordering of your choice. Moreover, we have some theorems about *ASPIC*$^+$ that inform you about some properties of your choices.

4.5 Generalising negation in *ASPIC*$^+$

The notion of an argumentation system in Section 4.3.1, assumed a language $\mathcal{L}$ closed under negation ($\neg$), where the standard classical interpretation of $\neg$ licenses a symmetric notion of conflict based attack, so that an argument consisting of an ordinary premise ϕ or with a defeasible top rule concluding ϕ, *symmetrically* attacks an argument consisting of an ordinary premise $\neg\phi$ or with a defeasible top rule concluding $\neg\phi$. However, the *ASPIC*$^+$framework as presented in Prakken (2010); Modgil and Prakken (2013), accommodates a more general notion of conflict, by defining an argumentation system to additionally include a function $^-$ that, for any wff $\psi \in \mathcal{L}$, specifies the set of wff's that are in conflict with ψ. With this idea, which is taken from assumption-based argumentation (Bondarenko et al.; 1997; Dung et al.; 2009), one can define both an asymmetric and symmetric notion of conflict-based attack. More formally:

Definition 4.5.1 $^-$ is a function from $\mathcal{L}$ to $2^{\mathcal{L}}$, such that:

- φ is a *contrary* of ψ if $\varphi \in \overline{\psi}$, $\psi \notin \overline{\varphi}$;
- φ is a *contradictory* of ψ (denoted by '$\varphi = -\psi$'), if $\varphi \in \overline{\psi}$, $\psi \in \overline{\varphi}$;
- each $\varphi \in \mathcal{L}$ has at least one contradictory.

Note that classical negation is now a special case of the symmetric contradictory relation: $\alpha \in \overline{\beta}$ iff α is of the form $\neg\beta$ or β is of the form $\neg\alpha$ (i.e., for any wff α, α and $\neg\alpha$ are contradictories). Modgil and Prakken (2013) then redefine Definition 4.3.3's notion of direct consistency so that a set S is *directly consistent* iff $\nexists\ \psi, \varphi \in S$ such that $\psi \in \overline{\varphi}$. Also, $\texttt{Conc}(A) \in \overline{\varphi}$ ($\texttt{Conc}(A) \in \overline{n(r)}$) replaces $\texttt{Conc}(A) = -\varphi$ ($\texttt{Conc}(A) = -n(r)$) in Definition 4.3.10's definition of attacks.

With this, one can reconstruct assumption-based argumentation (ABA) in *ASPIC*$^+$ as shown by Prakken (2010), since as just noted, ABA also generalises the notion of conflict through the use of a $^-$ function. To summarise, an *ASPIC*$^+$reconstruction of ABA will have empty sets of defeasible rules and axiom premises, and consist of ordinary premises and strict rules (respectively corresponding to the assumptions and rules in an ABA theory). Then, for every ordinary premise α, one specifies that:

1. there is a β in $\mathcal{L}$ such that β is a contrary or contradictory of α

2. α is not the conclusion of a strict inference rule (corresponding to so called 'flat' ABA theories)

Then, without the use of preference relation, a correspondence can be shown between ABA and *ASPIC*$^+$. One benefit of this is that one can then identify conditions under which ABA satisfies rationality postulates (by requiring, for instance, that the strict rules are closed under transposition).

The rationale for these more general notions of conflict and attack is two-fold. Firstly, one can for pragmatic reasons state that two formulae are in conflict, rather than requiring that one implies the negation of another; for example, assuming a predicate language with the binary '$<$' relation, one can state that any two formulae of the form $\alpha < \beta$ and $\beta < \alpha$ are contradictories. Secondly, the $^-$ function allows for an asymmetric notion of negation. This in turn is required for modelling negation as failure (as in logic programming). Using the negation as failure symbol $\sim$ (also called 'weak' negation, in contrast to the 'strong' negation symbol $\neg$), then $\sim\alpha$ denotes the negation of α under the assumption that α is not provable (i.e., the negation of α is assumed in the absence of evidence to the contrary). It is not then meaningful to assert that such an assumption brings into question (and so initiates an attack on) the evidence whose very absence is required to make the assumption in the first place. In other words, if A is an argument consisting of the premise $\sim\alpha$, and B concludes α (the contrary of $\sim\alpha$), then B attacks A, but not vice versa. Furthermore, since the very construction of A is invalidated by evidence to the contrary, i.e., B, then such attacks succeed as defeats *independently* of preferences.

To accommodate the notion of contrary, and attacks on contraries succeeding as defeats independently of preferences, we further modify Definition 4.3.10 to distinguish the special cases where $\texttt{Conc}(A)$ is a contrary of φ, in which case we say that A *contrary rebuts* B and A *contrary undermines* B, and then modify Definition 4.3.12 so that:

- A successfully rebuts B if A contrary rebuts B, or A rebuts B on B' and $A \not\prec B'$.

- A successfully undermines B if A contrary undermines B, or A undermines B on ϕ and $A \not\prec \phi$.

Following on from the discussion in Section 4.4.4, one can then show (Modgil and Prakken; 2013) that with the additional notion of contrary, satisfaction of the four rationality postulates not only requires that the argument theory satisfy axiom consistency, and transposition or contraposition, but also that it is *well formed* in the following sense:

Definition 4.5.2 An argumentation theory is *well-formed* if the following holds: if ϕ is a contrary of ψ then $\psi \notin \mathcal{K}_n$ and ψ is not the consequent of a strict rule.

To illustrate the use of negation as failure, suppose you want your arguments to be built from a propositional language that includes both $\neg$ and $\sim$. One could then define $\mathcal{L}$ as a language of propositional literals, composed from a set of propositional atoms $\{a, b, c, \ldots\}$ and the symbols $\neg$ and $\sim$. Then:

- α is a *strong literal* if α is a propositional atom or of the form $\neg\beta$ where β is a propositional atom (strong negation cannot be nested).
- α is a wff of $\mathcal{L}$, if α is a strong literal or of the form $\sim \beta$ where β is a strong literal (weak negation cannot be nested).

Then $\alpha \in \overline{\beta}$ iff (1) α is of the form $\neg\beta$ or β is of the form $\neg\alpha$; or (2) β is of the form $\sim \alpha$ (i.e., for any wff α, α and $\neg\alpha$ are contradictories and α is a contrary of $\sim \alpha$). Finally, for any $\sim \alpha$ that is in the antecedent of a strict or defeasible inference rule, one is required to include $\sim \alpha$ in the ordinary premises.

Consider now Example 4.3.6, where we now have that $u \in \overline{\sim u}$, and we replace the rule $d_4 : u \Rightarrow v$ with d_4': $\sim u \Rightarrow v$, and add $\sim u$ to the ordinary premises: $\mathcal{K}_p = \{\sim u, s, u, x\}$. Then, the arguments C_3 and D_4 are now replaced by arguments C_3' and D_4' each of which contain the sub-argument $E : \sim u$ (instead of $C_1 : u$). Then $C_1 : u$ contrary undermines, and so defeats, C_3' and D_4' on $\sim u$.

We finally note that according to Toni (2014) the philosophy behind ABA is to translate preferences and defeasible rules into ABA rules plus ABA assumptions, so that rebutting and undercutting attack and the application of preferences all reduce to premise attack. The idea of this is to keep the formal theory simpler and to make the technical machinery of ABA available for other approaches. In line with this philosophy, Dung and Thang (2014) have shown that their rule-based systems, which are a special case of *ASPIC*$^+$ with no knowledge base and no preferences, can be translated into ABA instantiations. They do this by translating every defeasible rule $p_1, \ldots, p_n \Rightarrow q$ as a strict rule $d_i, p_1, \ldots, p_n, not\neg q \rightarrow q$, where

- $d_i = n(p_1, \ldots, p_n \Rightarrow q)$ in *ASPIC*$^+$;
- $d_i, not\neg q \in \mathcal{A}$ (i.e., they are ABA assumptions);
- $q = \overline{not\neg q}$ and for all φ: $\varphi = \overline{\neg\varphi}$ and $\neg\varphi = \overline{\varphi}$

Dung and Thang (2014) then show (on the assumption that *ASPIC*$^+$ rule names do not occur as antecedents or consequents in *ASPIC*$^+$ rules), that for grounded, preferred and stable semantics the resulting ABA framework validates the same conclusions as the original *ASPIC*$^+$ SAF. We agree that this approach has its merits but note that it is an open question whether *ASPIC*$^+$ can in its full generality be translated into ABA. Also, as we noted above, we claim that there is also some merit in having a theory with explicit notions of rebutting and undercutting attack and preference application, namely, if the aim is to formalise modes of reasoning in a way that corresponds with human modes of reasoning and debate.

4.6 Variants of rebutting attack

Several papers have considered alternative definitions of rebutting attack in which an argument can under specific conditions also be rebutted on the conclusions of strict inferences.

4.6.1 Unrestricted rebuts

In *ASPIC*$^+$ as presented so far, arguments can only be rebutted on conclusions of defeasible-rule applications. Caminada and Amgoud (2007) call this *restricted rebut*. They also study *unrestricted rebut*, which allows rebuttals on the conclusion of a strict inference provided that at least one of the argument's subarguments is defeasible. Their replacement of restricted with unrestricted rebut leads to a variant of their simplified version of *ASPIC*$^+$ (which is in fact equivalent to Dung and Thang (2014)'s rule-based systems). They prove that for grounded semantics the rationality postulates are (under the usual conditions) satisfied but they provide a counterexample for stable and preferred semantics, presented above in Section 4.4.1 with a modification of Example 4.4.1.

Caminada et al. (2014) argue in favour of unrestricted rebut on the grounds that this would lead to more natural presentations of dialogues. They argue that when applying argumentation in dialogical settings, the notion of restricted rebuts sometimes forces agents to commit to statements they have insufficient reasons to believe. In abstract terms, suppose an agent Ag_1 submitting an argument A whose top rule is a strict rule $s_1 = \alpha_1, \ldots, \alpha_n \rightarrow \alpha$, where for $i = 1 \ldots n$, α_i is an ordinary premise in A or the head of a defeasible rule in A. Now suppose Ag_2 has an argument B that defeasibly concludes $\neg\alpha$. Since B does not rebut A on α, then to attack A requires that Ag_2 construct, for some $i = 1 \ldots n$, an argument B' that extends B *and the arguments concluding* α_j, $j \neq i$, with the transposition $s_1^i = \alpha_1, \ldots, \alpha_{i-1}, \neg\alpha, \alpha_{i+1}, \alpha_n \rightarrow \neg\alpha_i$. But then Ag_2 is forced to commit to her interlocutors' arguments concluding α_j, $j \neq i$, for which she has no reasons to believe.

Caminada et al. (2014) give the following concrete example.

> John: *"Bob will attend conferences AAMAS and IJCAI this year, as he has papers accepted at both conferences."*
> Mary: *"That won't be possible, as his budget of £1000 only allows for one foreign trip."*

Formally, this discussion could be modelled using an argumentation theory with $\mathcal{R}_d \supseteq \{\text{accA} \Rightarrow \text{attA};\ \text{accI} \Rightarrow \text{attI};\ \text{budget} \Rightarrow \neg(\text{attA} \wedge \text{attI})\}$ and $\mathcal{R}_s \supseteq \{\rightarrow \text{accA};\ \rightarrow \text{accI};\ \rightarrow \text{budget};\ \text{attA}, \text{attI} \rightarrow \text{attA} \wedge \text{attI}\}$.

A direct formalisation of the above arguments is then:

J_1: $\rightarrow$ accA
J_2: $J_1 \Rightarrow$ attA
J_3: $\rightarrow$ accI
J_4: $J_3 \Rightarrow$ attI
J_5: $J_3, J_4 \rightarrow$ attA $\wedge$ attI

M_1: $\rightarrow$ budget
M_2: $M_1 \Rightarrow \neg$(attA $\wedge$ attI)

In *ASPIC*$^+$, Mary's argument does *not* attack John's argument, since the conclusion Mary wants to attack (attA $\wedge$ attI) is the consequent of a strict rule. Mary can only

attack John's argument by attacking the consequent of one of the defeasible rules, that is, by uttering one of the following two statements.

> Mary$'$: *"Bob can't attend AAMAS because he will attend IJCAI, and his budget does not allow him to attend both."*
> Mary$''$: *"Bob can't attend IJCAI because he will attend AAMAS, and his budget does not allow him to attend both."*

The associated formal counterarguments are as follows.[14]

M_1:	$\rightarrow$ budget		
M_2:	$M_1 \Rightarrow \neg(\text{attA} \wedge \text{attI})$		
J_3:	$\rightarrow$ accI	J_1:	$\rightarrow$ accA
J_4:	$J_3 \Rightarrow$ attI	J_2:	$J_1 \Rightarrow$ attA
M_5':	$M_2, J_4 \rightarrow \neg$attA	M_5'':	$M_2, J_2 \rightarrow \neg$attI

According to Caminada et al. (2014) the problem with this is that Mary does not know which of the two conferences Bob will attend, but *ASPIC*$^+$ with restricted rebut forces her to assert that Bob will attend one or the other. They argue that from the perspective of commitment in dialogue (Walton and Krabbe; 1995), this is unnatural.

Caminada et al. (2014) then define a restricted version of basic *ASPIC*$^+$ as presented above in Section 4.3 – which they call *ASPIC*$^-$ – that substitutes strict rules with empty antecedents for axiom premises, and defeasible rules with empty antecedents for ordinary premises. Moreover, *ASPIC*$^-$ allows unrestricted rebuts on the conclusions of strict rules. They then show that under the assumption of a *total* ordering on the defeasible rules, and assuming either the `Elitist` or `Democratic` set comparisons used in defining weakest- or last-link preferences, all of Caminada and Amgoud (2007)'s rationality postulates are satisfied for well-defined $SAFs$, but only for the grounded semantics. They have thus generalised Caminada and Amgoud (2007)'s results for some specific cases with preferences.

4.6.2 Weak rebuts and an alternative view on the rationality postulates

Prakken (2016) studies a weaker version of unrestricted rebut, motivated by the general observation that deductive inferences may weaken an argument. His argument is that when a deductive inference is made from the conclusions of at least two 'fallible' (defeasible or plausible) subarguments, the deductive inference can be said to aggregate the degrees of fallibility of the individual arguments to which it is applied. This in turn means that the deductive inference may be less preferred than either of these subarguments, so that a successful attack on the deductive inference does not necessarily imply a successful attack on one of its fallible subarguments. And this in turn means that there can be cases where it is rational to accept a set of arguments that is not strictly closed and that violate indirect consistency. Note that this line of reasoning does not apply to cases where a deductive inference is applied to at most one fallible subargument: then the amount of fallibility of the new argument is exactly the same as the amount of fallibility of the single fallible argument to which the deductive inference is applied. Accordingly, Prakken (2016) defines *weak rebut* as allowing rebuttals on the conclusion of a strict inference, provided that the strict inference is applied to at least two fallible

[14] Assuming $\mathcal{R}_s$ ito be closed under transposition, the fact that $\mathcal{R}_s$ contains $\text{attA}, \text{attI} \rightarrow \text{attA} \wedge \text{attI}$ implies that $\mathcal{R}_s$ also contains $\neg(\text{attA} \wedge \text{attI}), \text{attI} \rightarrow \neg\text{attA}$ and $\text{attA}, \neg(\text{attA} \wedge \text{attI}) \rightarrow \neg\text{attI}$.

subarguments. Moreover, he argues that there are cases where argument orderings cannot be required to satisfy all properties of a so-called reasonable argument ordering as defined by Modgil and Prakken (2013).

Prakken (2016) illustrates this with the lottery paradox, a well-known paradox from epistemology, first discussed by Kyburg (1961). Imagine a fair lottery with one million tickets and just one prize. If the principle is accepted that it is rational to accept a proposition if its truth is highly probable, then for each ticket T_i it is rational to accept that T_i will not win while at the same time it is rational to accept that exactly one ticket will win. If we also accept that everything that deductively follows from a set of rationally acceptable propositions is rationally acceptable, then we have two rationally acceptable propositions that contradict each other: we can join all individual propositions $\neg T_i$ into a big conjunction $\neg T_1 \wedge \ldots \wedge \neg T_{1,000,000}$ with one million conjuncts, which contradicts the certain fact that exactly one ticket will win.

Many views on this paradox exist. Prakken (2016) wants to formalise the view that for each individual ticket it is rational to accept that it will not win while at the same time it is not rational to accept the conjunction of these acceptable beliefs. He considers the following modelling of the lottery paradox in *ASPIC*$^+$. Let $\mathcal{L}$ be a propositional language built from the set of atoms $\{T_i \mid 1 \leq i \leq 1,000,000\}$. Then let X denote a well-formed formula $X_1 \veebar \ldots \veebar X_{1,000,000}$ where $\veebar$ is exclusive or and where each X_i is of one of the following forms:

- If $i = 1$ then $X_i = T_1 \wedge \neg T_2 \wedge \ldots \wedge \neg T_n$
- If $i = n$ then $X_i = \neg T_1 \wedge \neg T_2 \wedge \ldots \wedge \neg T_{n-1} \wedge T_n$
- Otherwise $X_i = \neg T_1 \wedge \ldots \wedge \neg T_{i-1} \wedge T_i \wedge \neg T_{i+1} \wedge \ldots \wedge \neg T_n$

Next we choose $\mathcal{K}_p = \{\neg T_i \mid 1 \leq i \leq 1,000,000\}$, $\mathcal{K}_n = \{X\}$, $\mathcal{R}_s$ as consisting of all propositionally valid inferences from finite sets and $\mathcal{R}_d = \varnothing$.

The following arguments are relevant for any i such that $1 \leq i \leq 1,000,000$.

$$\neg T_i \qquad \text{and} \quad \neg T_1, \ldots, \neg T_{i-1}, \neg T_{i+1}, \ldots, \neg T_{1,000,000}, X \rightarrow T_i \text{ (call it } A_i)$$

Making $\neg T_i$ justified for all i requires for all i that $A_i \prec \neg T_i$, to prevent A_i from defeating $\neg T_i$. Then we have a single extension in all semantics containing arguments for all conclusions $\neg T_i$ but not for their conjunction.

Prakken then proposes a definition of a *weakly reasonable* argument ordering according to which applying a strict rule to the conclusion of a single argument A to obtain an argument A' does not change the 'preferedness' of A' compared to A. This is reasonable in general, since A and A' have exactly the same set of fallible elements (ordinary premises and/or defeasible inferences). He then proposes weakened versions of the postulates of strict closure and indirect consistency, according to which these properties are only required to hold for subsets of extensions with at most one fallible argument. He also proposes a notion of *weak rebut*, according to which an argument can be rebutted on a strict top rule provided it has at least two fallible subarguments. He then proves that if weak rebut is allowed in addition to restricted rebut and argument orderings are required to be weakly reasonable, then the original postulate of direct consistency plus the weakened postulates of strict closure and indirect consistency are satisfied if AT is closed under contraposition or transposition and $\texttt{Prem}(A) \cup \mathcal{K}_n$ is indirectly consistent.

Prakken (2016) concludes with some general observations on the relation between deduction and justification. He argues to have shown that preservation of truth (the definition of deductively valid arguments) does not imply preservation of rational acceptance, since truth and rational acceptance are different things. However, he also argues that deduction still plays an important role in argumentation. Deductive inference rules are still available as argument construction rules and if an argument with a strict top rule has no attackers or all its attackers are less preferred, then the argument may still be sceptically justified. The specifics of the adopted argument ordering are essential here. For instance, in the lottery paradox the argument ordering might allow that application of the conjunction rule to a small number of conclusions $\neg T_i$ is still sceptically justified.

4.7 Conclusion

In this chapter we presented *ASPIC*$^+$, a framework for structured argumentation based on two ideas: that conflicts between arguments are sometimes resolved with explicit preferences, and that arguments are built with two kinds of inference rules: strict, or deductive rules, which logically entail their conclusion, and defeasible rules, which only create a presumption in favour of their conclusion. The second idea implies that *ASPIC*$^+$ does not primarily see argumentation as inconsistency handling in a given 'base' logic: conflicts between arguments may not only arise from the inconsistency of a knowledge base but also from the defeasibility of the reasoning steps in an argument.

ASPIC$^+$ is not a system but a framework for specifying systems. A main objective is to identify conditions under which instantiations of ASPIC+ satisfy logical consistency and closure properties. We first discussed *ASPIC*$^+$'s philosophical underpinnings. We then illustrated the main definitions with examples and we presented some more and less principled ways to instantiate the framework. We also briefly discussed how *ASPIC*$^+$ captures several other approaches as special cases. As we saw above, the *ASPIC*$^+$ framework can be instantiated in many different ways. We have already discussed some of these ways and their properties. We hope that in due course more 'best practices' in using *ASPIC*$^+$ will emerge.

Finally, two implementations are available online of instantiations of *ASPIC*$^+$ with domain-specific inference rules and with rule priorities:

- TOAST (http://toast.arg-tech.org);
- PyArg (https://pyarg.npai.science.uu.nl/).

4.8 Exercises

In the following exercises an argument ordering is called *simple* if it holds that $A \prec B$ iff A is plausible or defeasible while B is strict and firm, and $A \approx B$ otherwise.

EXERCISE 4.8.1 Consider an argumentation system in which $\mathcal{R}_s$ consists of all valid propositional and first-order inferences from finite sets, and with as knowledge base

$\mathcal{K}_n = \{\forall x(Px \supset Qx)\}$
$\mathcal{K}_p = \{Pa, \forall x(Qx \supset Rx)\}$

1. Construct a consistent argument A for Ra.
2. Identify $\texttt{Prem}(A)$, $\texttt{Conc}(A)$, $\texttt{Sub}(A)$, $\texttt{DefRules}(A)$ and $\texttt{TopRule}(A)$.
3. What is in terms of Definition 4.3.7 the type of this argument?

EXERCISE 4.8.2 Consider the following example of a civil legal case. Assume that in a medical malpractice case, a doctor is liable for compensation if the patient was injured because of the doctor's negligence, and that if a patient is injured in a non-risky operation, this is negligence. We also have that an appendicitis operation generally is a non-risky operation but that operations on patients with bad blood circulation are generally risky. Assume finally, that a given patient was injured in an appendicitis operation and that two medical tests gave contradicting results on whether the patient had bad blood circulation. One way to represent this is with the following facts and domain-specific defeasible rules: $\mathcal{R}_s = \mathcal{K}_p = \varnothing$, $\mathcal{R}_d = \{r_1\text{-}r_6\}$ while $\mathcal{K}_n = \{f_1\text{-}f_4\}$.

r_1:	*injury*, *negligence* $\Rightarrow$ *compensation*	f_1:	*injury*
r_2:	*injury*, $\neg$ *risky operation* $\Rightarrow$ *negligence*	f_2:	*appendicitis*
r_3:	*appendicitis* $\Rightarrow \neg$ *riskyOperation*	f_3:	*medicalTest1*
r_4:	*badCirculation* $\Rightarrow$ *riskyOperation*	f_4:	*medicalTest2*
r_5:	*medicalTest1* $\Rightarrow$ *badCirculation*		
r_6:	*medicalTest2* $\Rightarrow \neg$ *badCirculation*		

1. Construct all arguments on the basis of this argumentation theory and their attack relations.
2. Specify the following for all arguments X: $\texttt{Prem}(X)$, $\texttt{Conc}(X)$, $\texttt{Sub}(X)$, $\texttt{DefRules}(X)$ and $\texttt{TopRule}(X)$.
3. Suppose that $r_3 < r_4$ and $r_5 < r_6$. Determine the defeat relations with the elitist last-link ordering.
4. Determine the grounded extension of the SAF defined by the above argumentation theory and the argument ordering induced by the preference relation of (b).
5. Determine the preferred extension(s).
6. Move f_3 and f_4 from $\mathcal{K}_n$ to $\mathcal{K}_p$ and assume also that $f_4 <' f_3$. Answer again questions (b-d) but now for the elitist weakest-link ordering.

EXERCISE 4.8.3 Consider the following argumentation theory with:

$\mathcal{R}_s = \{r \rightarrow s; v, x \rightarrow \neg q\}$;
$R_d = \{$
$\quad p, q \Rightarrow r,$
$\quad s \Rightarrow t,$
$\quad u \Rightarrow v,$
$\quad w \Rightarrow \neg u\}$
$\mathcal{K}_n = \{x\}$
$\mathcal{K}_p = \{p, q, u, w\}$ Evaluate the following questions relative to the *c-SAF* induced by this example.

1. Construct an argument for t, construct all its attackers and all attackers of its attackers.
2. Verify the status of t according to grounded semantics, assuming the empty argument ordering.
3. Assume now the following preference orderings $\leq$ on $\mathcal{R}_d$ and $\leq'$ on $\mathcal{K}_p$:

   $w \Rightarrow \neg u < u \Rightarrow v$

   $q <' u$

   $w <' u$

   Verify how the answer to question (2) changes for the elitist last-link ordering.
4. Answer the same question for the elitist weakest-link ordering.
5. Replace $\mathcal{R}_s$ by its closure under transposition and answer questions (1-4) again.

EXERCISE 4.8.4 Give the abstract argumentation framework corresponding to Figure 4.5.

EXERCISE 4.8.5 Consider the following argumentation theory with:

$\mathcal{R}_s = \{p, q \rightarrow r,\ t \rightarrow \neg d_1\}$,
$R_d = \{$
$\quad d_1\colon p \Rightarrow q$,
$\quad d_2\colon s \Rightarrow t$,
$\quad d_3\colon u \Rightarrow v$,
$\quad d_4\colon v \Rightarrow \neg t\}$
$\mathcal{K}_p = \{p, s, u\}$

With orderings $\leq$ on R_d and $\leq'$ on $\mathcal{K}_p$ such that $d_2 < d_4$, $d_3 < d_2$ and $u <' s$.

1. Verify the status of r according to preferred semantics, assuming the weakest-link ordering on arguments.
2. Answer the same question assuming the last-link ordering on arguments.

EXERCISE 4.8.6 Consider the following, equally strong defaults

1. Persons born in The Netherlands are typically Dutch.
2. Persons with a Norwegian name are typically Norwegian.
3. Persons who are Dutch or Norwegian typically like ice skating.

and the following facts:

4. Brigt Rykkje was born in the Netherlands
5. Brigt Rykkje has a Norwegian name.
6. Nobody is both Dutch and Norwegian.

Evaluate the following questions relative to the *c-SAF* induced by this example.

1. Translate this information into an argumentation theory of which $\mathcal{R}_s$ consists of all valid propositional and first-order inferences from finite sets and $\mathcal{R}_d$ consists of the defeasible inference scheme for $\rightsquigarrow$ from Section 4.4.3.

2. Assume that the argument ordering is determined by the last-link principle. We want to know whether Brigt Rykkje likes ice skating. Construct all arguments that are relevant for this proposition and determine whether the conclusion that Brigt Rykkje likes ice skating is justified in grounded semantics.

3. Answer the same question for preferred semantics.

4. Answer the same question for f-justification in preferred semantics.

EXERCISE 4.8.7 Formalise the example of Exercise 2.8.12 as an argumentation theory with domain-specific defeasible rules in a way that satisfies your intuitions about this example.

EXERCISE 4.8.8 Consider the argumentation theory of Example 4.4.5.

1. Verify the status of argument D_2 for s in grounded semantics.

2. Verify the status of argument D_2 for s in preferred semantics.

EXERCISE 4.8.9 Let $\mathcal{R}_s = \{p \rightarrow q; p \rightarrow r; p, r \rightarrow s\}$.

1. Determine $Cl_{tp}(R_s)$.

2. Determine whether with $Cl_{tp}(R_s)$ it holds that $\{p\} \vdash s$.

3. Determine whether with $Cl_{tp}(R_s)$ it holds that $\{-s\} \vdash -p$.

EXERCISE 4.8.10 Let $\mathcal{R}_s = \{p \rightarrow q;\ \neg q \rightarrow r;\ r \rightarrow \neg p;\ \neg r \rightarrow q;\ p \rightarrow \neg r\}$ and let $^-$ correspond to classical negation.

1. Is an argumentation theory with $\mathcal{R}_s$ closed under transposition?

2. Is an argumentation theory with $\mathcal{R}_s$ closed under contraposition?

EXERCISE 4.8.11 Consider the following argumentation theory with:

$\mathcal{R}_s$ = $Cl_{tp}(S)$, where $S = \{q, r \rightarrow \neg t\}$ (NB: Cl_{tp} stands for 'closure under transposition'). $R_d = \{$
d_1: $p \Rightarrow q$,
d_2: $p, q \Rightarrow r$,
d_3: $s \Rightarrow t\}$
$\mathcal{K}_p = \{p, s\}$

With an ordering $\leq$ on R_d such that $d_3 < d_1$ and $d_2 < d_3$. Evaluate the following questions relative to the *c-SAF* induced by this example.

1. Verify the status of t and $\neg t$ according to preferred semantics, assuming the last-link ordering on arguments.

2. Specify the following for all arguments X that you constructed in your answer: $\texttt{Prem}(X)$, $\texttt{Conc}(X)$, $\texttt{Sub}(X)$, $\texttt{DefRules}(X)$, $\texttt{LastDefRules}(X)$ and $\texttt{TopRule}(X)$.

EXERCISE 4.8.12 Consider an argumentation theory AT in *ASPIC** with $\mathcal{L}$ being a propositional language and with:

$\mathcal{R}_s$ consisting of all valid inferences of propositional logic from finite sets according to $\vdash_W$;
$R_d = \{$
$\quad d_1\colon p \Rightarrow r,$
$\quad d_2\colon q \Rightarrow s,$
$\quad d_3\colon s \Rightarrow t\}$
$\mathcal{K}_n = \{p, q, \neg r \vee \neg t\}$
$\mathcal{K}_p = \varnothing$
and with a strict partial ordering on R_d such that $d_2 < d_1 < d_3$.

1. Construct an argument* for $\neg r$.
2. Verify the status of $\neg r$ according to grounded semantics, assuming the weakest-link argument ordering.
3. Verify the status of $\neg r$ according to grounded semantics, assuming the last-link argument ordering.

You may display the arguments* by drawing pictures, but make sure that the pictures are unambiguous.

EXERCISE 4.8.13 [15] Let $(\mathcal{L}, {}^-, \mathcal{R}, n)$ be an argumentation system where:

- $\mathcal{L}$ is a language of propositional literals, composed from a set of propositional atoms $\{a, b, c, \ldots\}$ and the symbols $\neg$ and $\sim$ respectively denoting strong and weak negation (i.e., negation as failure). α is a strong literal if α is a propositional atom or of the form $\neg\beta$ where β is a propositional atom (strong negation cannot be nested). α is a wff of $\mathcal{L}$, if α is a strong literal or of the form $\sim \beta$ where β is a strong literal (weak negation cannot be nested).
- $\alpha \in \overline{\beta}$ iff (1) α is of the form $\neg\beta$ or β is of the form $\neg\alpha$; or (2) β is of the form $\sim \alpha$ (i.e., for any wff α, α and $\neg\alpha$ are contradictories and α is a contrary of $\sim \alpha$).
- $\mathcal{R}_s = \{t, q \rightarrow \neg p\}$, $\mathcal{R}_d = \{\sim s \Rightarrow t; r \Rightarrow q; a \Rightarrow p\}$
- $n(\sim s \Rightarrow t) = d_1$, $n(r \Rightarrow q) = d_2$, $n(a \Rightarrow p) = d_3$

Furthermore, $\mathcal{K}$ is the knowledge base such that $\mathcal{K}_n = \varnothing$ and $\mathcal{K}_p = \{a, r, \neg r, \sim s\}$.

1. Construct all arguments on the basis of this argumentation theory.
2. Determine the attack relations.
3. Assume that the argument ordering $\preceq$ is defined in terms of preorderings $\leq$ on defeasible rules and $\leq'$ on ordinary premises. Assume that $r \Rightarrow q < a \Rightarrow p$ (i.e., $d_2 < d_3$) and $\neg r <' r; \neg a \approx' r; \sim s <' \neg r$. Determine the defeat relations with the elitist last link ordering.

[15] Adapted from S. Modgil & H. Prakken, A general account of argumentation with preferences. *Artificial Intelligence* 195 (2013): 361–397.

4. Add the transpositions of $t, q \rightarrow \neg p$ to $\mathcal{R}_s$. Which new arguments, attacks and defeats are now generated?

EXERCISE 4.8.14 Consider the same language $\mathcal{L}$ as in Exercise 4.8.13 but let now $\mathcal{R}_s = \{\sim a \rightarrow b\}$, $\mathcal{R}_d = \{b \Rightarrow_{d_1} \neg c;\ \Rightarrow_{d_2} c;\ c \Rightarrow_{d_3} a\}$ (here the names of the defaults are attached to $\Rightarrow$), $\mathcal{K}_n = \varnothing$ and $\mathcal{K}_p = \{\sim a\}$. Finally, assume a partial preorder $<$ on $\mathcal{R}_d$ such that that $d_2 < d_1$ and $d_1 < d_3$.

1. Determine the arguments and their attack relations.
2. Determine which attacks succeed as defeats with the elitist last-link ordering.
3. Determine the grounded extension of the resulting abstract argumentation theory.
4. Determine the preferred extension(s) of this abstract argumentation theory.

EXERCISE 4.8.15 Consider an argumentation theory in *ASPIC** in which $\mathcal{R}_s$ consists of all valid propositional inferences from finite sets according to $\vdash_W$, $\mathcal{R}_d = \mathcal{K}_n = \varnothing$ and $\mathcal{K}_p =$

$\{\neg ab \supset \neg guilty,$
$murder \supset guilty,$
$murder,$
$\neg ab\}.$

1. Verify whether $guilty$ is justified according to grounded semantics, assuming a simple argument ordering.
2. Then specify a partial preorder on $\mathcal{K}_p$ such that with the elitist weakest-link argument ordering $guilty$ is justified according to grounded semantics.
3. Alternatively to (b), move one or more formulas from $\mathcal{K}_p$ to $\mathcal{K}_n$ such that $guilty$ becomes justified as a result of the change.

Chapter 5

Preferences, support, graduality: abstract versus structured approaches

In Chapter 4 we discussed how Dung's (1995) abstract approach to argumentation is instantiated by the *ASPIC*$^+$ framework by specifying the structure of arguments and the nature of the defeat relation. In this chapter we review work employing an alternative approach, consisting in not *instantiating* Dung's notions but *extending* them with new notions. We will in particular review work that extends abstract argumentation frameworks with preference and support relations between arguments. At the end of this chapter we will also briefly comment on another recent development, developing gradual notions of argument evaluation as alternatives to Dung's (1995) semantics. An important theme in our discussion will be that it is dangerous to extend or modify Dung's (1995) frameworks or semantics in the abstract, without considering the structure of arguments and the nature of their relations, since this creates the danger that implicit assumptions are made at the abstract level that do not hold in general for instantiations.

5.1 Preference-based argumentation frameworks

The approach to extend argumentation frameworks at the abstract level was first applied for preferences. Amgoud and Cayrol (1998) added to AFs a preference relation on $\mathcal{A}$, resulting in *preference-based argumentation frameworks* ($PAFs$), which are a triple $\mathcal{A}, \mathcal{C}, \preceq)$, where $\mathcal{C}$ is an attack relation on $\mathcal{A}$. An argument A then *defeats* an argument B if A attacks B and $A \not\prec B$. Thus each PAF generates an AF of the form $(\mathcal{A}, \mathcal{D})$, to which Dung's theory of AFs can be applied. At first sight, this looks very similar to the treatment of preferences in *ASPIC*$^+$, but there is a crucial difference, since in *ASPIC*$^+$ the structure of arguments is crucial in determining how preferences must be applied to attacks. Since $PAFs$ do not specify the structure of arguments, they cannot model various subtle differences at this point.

To start with, there are reasonable notions of attack that result in defeat irrespective of preferences, such as *ASPIC*$^+$'s undercutting attack. A framework that does not make the structure of arguments explicit cannot distinguish between preference-dependent and preference-independent attacks. At first sight it might seem that this problem can

be solved by allowing two abstract kinds of attack, called preference-dependent and preference-independent attack, and to apply the argument ordering only to the first type of attack. However, this solution still faces problems, since it cannot recognise that in general the question which preference must be used to resolve an attack depends on the structure of arguments.

Consider the following example in *ASPIC*$^+$, with $\mathcal{K}_n = \mathcal{K}_a = \varnothing; \mathcal{K}_p = \{p, q\}$, $\mathcal{R}_s = \varnothing$, $\mathcal{R}_d = \{p \Rightarrow r; q \Rightarrow \neg r; \neg r \Rightarrow s\}$, where the contrariness relation over $\mathcal{L}$ corresponds to classical negation in the obvious way. We then have the following arguments:

A_1 = p $\quad\quad$ B_1 = q
A_2 = $A_1 \Rightarrow r$ $\quad\quad$ B_2 = $B_1 \Rightarrow \neg r$
$\quad\quad\quad\quad\quad\quad$ B_3 = $B_2 \Rightarrow s$

We have that A_2 and B_2 attack each other and A_2 attacks B_3, since it directly rebuts its subargument B_2 (see Figure 5.1).

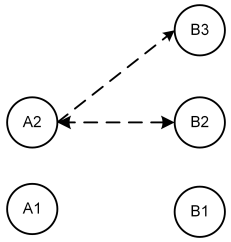


Figure 5.1: The attack graph

Assume that the defeasible rules are ordered as follows: $q \Rightarrow \neg r < p \Rightarrow r$, $p \Rightarrow r < \neg r \Rightarrow s$ and let us apply the last-link argument ordering, which orders arguments according to the preferences of their last-applied defeasible rules (this ordering is, for instance, suitable for reasoning with legal rules). Then the following argument ordering is generated: $B_2 \prec A_2$ since $q \Rightarrow \neg r < p \Rightarrow r$, and $A_2 \prec B_3$ since $p \Rightarrow r < \neg r \Rightarrow s$. A PAF modelling then generates the following single defeat relation: A_2 defeats B_2 (see Figure 5.2). Then we have a single extension (in whatever semantics), namely, $\{A_1, B_1, A_2, B_3\}$. So not only A_2 but also B_3 is justified. However, this

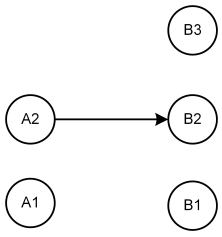


Figure 5.2: The PAF defeat graph

violates Caminada and Amgoud (2007)'s rationality postulate of subargument closure of extensions, since B_3 is in the extension while its subargument B_2 is not. The cause of the problem is that the PAF modelling of this example cannot recognise that the reason why A_2 attacks B_3 is that A_2 directly attacks B_2, which is a subargument of B_3.

So the PAF modelling fails to capture that in order to check whether A_2's attack on B_3 succeeds, we should compare A_2 not with B_3 but with B_2, as happens in *ASPIC*$^+$. Now since $B_2 \prec A_2$ we also have that A_2 defeats B_3 (see Figure 5.3), so in *ASPIC*$^+$ the single extension (in whatever semantics) is $\{A_1, B_1, A_2, B_3\}$ and we have that A_2 is justified and both B_2 and B_3 are overruled, so closure under subarguments is respected. Moreover, recall that *ASPIC*$^+$ always satisfies this postulate.

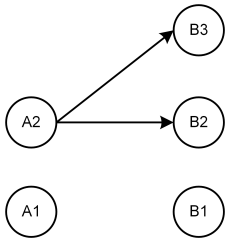


Figure 5.3: The *ASPIC*$^+$ defeat graph

The lesson that can be learned from this example is that in general the choice of preference to resolve an attack depends on the structural nature of the attack, and the problem with $PAFs$ is that they cannot model the structural nature of attacks.

5.2 Bipolar argumentation frameworks

A second way in which argumentation frameworks have been extended at the abstract level is by adding a support relation between arguments. This results in so-called **Bipolar argumentation frameworks** ($BAFs$), which formally are a triple $(\mathcal{A}, \mathcal{D}, \mathcal{S})$ where $\mathcal{A}$ and $\mathcal{D}$[1] are defined as above and where $\mathcal{S}$ is a support relation on $\mathcal{A}$. Cayrol and Lagasquie-Schiex (2009) define a *sequence of supports* for argument B by argument A as a sequence $A\mathcal{S}B_1, \ldots \mathcal{S}B_n\mathcal{S}B$. Often the semantics of $BAFs$ is defined in terms of constraints on the defeat relation given sets of defeat and support relations between arguments, specifying which defeat relations should also hold given the initial defeat and support relations. Arguments are then evaluated by applying a given Dung-style semantics to AFs that contain all additional defeats induced by these constraints. The following constraints are the most important ones that have been considered in the literature. Accordingly, we will call them the 'standard semantics' for $BAFs$. A semantics of $BAFs$ can use any subset of these constraints. Given a $BAF = (\mathcal{A}, \mathcal{D}, \mathcal{S})$:

- there is a *supported defeat* from A to B iff there exists an argument C such that there is a sequence of supports from A to C and $(C, B) \in \mathcal{D}$;
- there is a *secondary defeat* from A to B iff there exists an argument C such that there is a sequence of supports from C to B and $(A, C) \in \mathcal{D}$;
- there is an *extended defeat* from A to B iff there exists an argument C such that there is a sequence of supports from C to A and $(C, B) \in \mathcal{D}$;

[1] Like in much other work on abstract approaches, the literature on $BAFs$ usually speaks of an attack relation, but for reasons explained above we will speak of defeat.

- there is a *mediated defeat* from A to B iff there exists an argument C such that there is a sequence of supports from B to C and $(A, C) \in \mathcal{D}$.

The question arises which semantics is suitable or 'good'. It turns out that for support this issue is far more subtle than for defeat. For defeat the main intuitive constraints are that if A defeats B than A and B cannot be accepted together, while, moreover, if the choice is between A and B, then A must be accepted. From these basic intuitions the notions of defence, conflict-freeness and admissibility naturally follow and these notions are the essence of all semantics for AFs; although variations between semantics are still possible, there differences do not depend on the nature of the defeat relation. With support this is different, as we will see now.

Cohen et al. (2018) have carried out a systematic study of semantics for different support relations in the context of *ASPIC*$^+$, which we now briefly summarise. Cohen et al. first define four ways in which *ASPIC*$^+$ arguments can support each other. They are illustrated below in Figure 5.4. The first is the *ASPIC*$^+$ **proper subargument relation** between arguments. The second notion of support is a notion of argument accrual: two different arguments A and B **conclusion-support** each other if they have the same final conclusion. A third notion is premise support. Argument A **premise-supports** another argument B if A's final conclusion is a premise of B. Fourth, a variant of conclusion support is intermediate support: if A conclusion-supports a proper subargument of another argument B that does not equal a premise of B, then A **intermediate-supports** B.

Cohen et al. then consider three semantics for $BAFs$ in terms of the AFs generated by three alternative sets of defeat constraints. **General support** adds all supported and secondary defeats to the original defeat relation, **deductive support** adds all supported and mediated defeats and **necessary support** adds all extended and secondary defeats where, moreover, the underlying support relation is irreflexive and transitive. These semantics are formally defined as follows.

Definition 5.2.1 [*AFs* **associated with** *BAFs*] For any $BAF = (\mathcal{A}, \mathcal{D}, \mathcal{S})$, the AF associated with BAF under semantics S is $(\mathcal{A}, \mathcal{D} \cup \mathcal{D}^+)$ where:

1. $\mathcal{D}^+$ is the set of all supported and secondary defeats given BAF if $S =$ general support;

2. $\mathcal{D}^+$ is the set of all supported and mediated defeats given BAF if $S =$ deductive support;

3. $\mathcal{D}^+$ is the set of all extended and secondary defeats given BAF if $S =$ necessary support, where $\mathcal{S}$ is irreflexive and transitive.

Cohen et al. then investigate whether their four notions of support in *ASPIC*$^+$ can be related to these three BAF semantics. They do this for each of the three *ASPIC*$^+$ defeat relations separately. For simplicity they assume no preferences, so that all *ASPIC*$^+$ attacks succeed as defeats. For each $AF = (\mathcal{A}, \mathcal{D})$ induced by an *ASPIC*$^+$ instantiation and a particular *ASPIC*$^+$ attack relation (undermining, rebutting or undercutting attack), they first consider the $BAF = (\mathcal{A}, \mathcal{D}, \mathcal{S}_s)$ where $\mathcal{S}_s$ is the support relation on $\mathcal{A}$ according to *ASPIC*$^+$ support type s (proper-subargument, conclusion, premise or intermediate support). Then for each of the three BAF semantics x (general, deductive

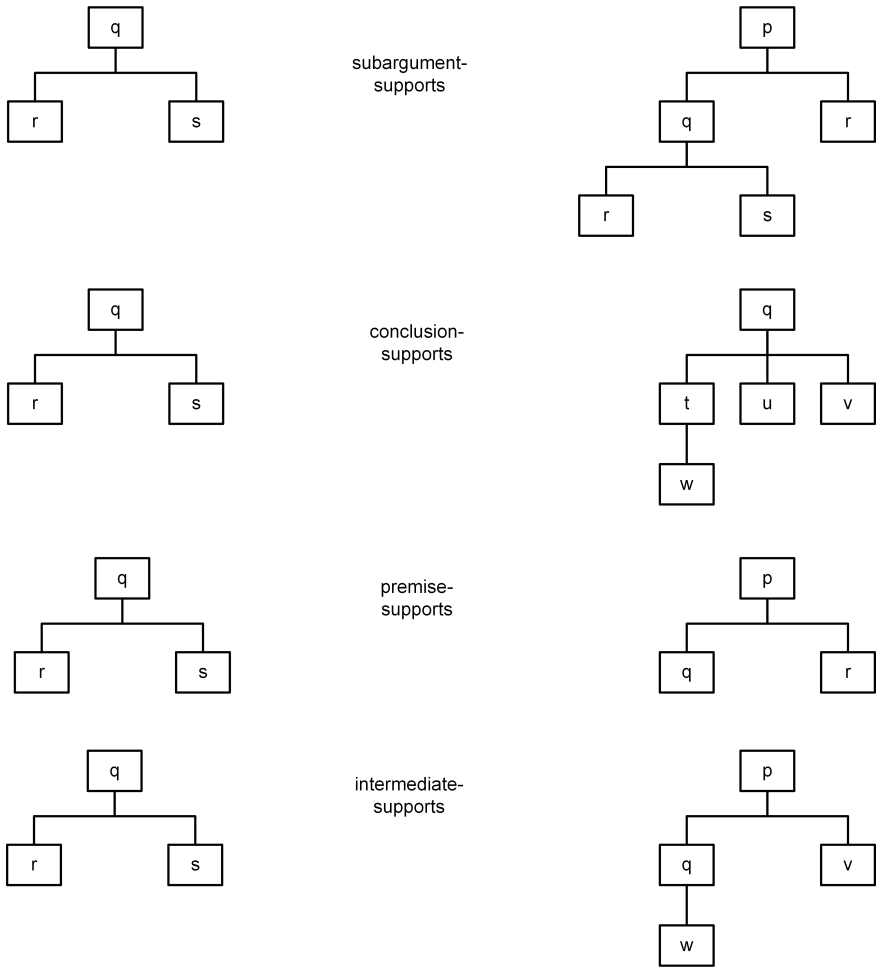


Figure 5.4: Cohen et al.'s (2018) four kinds of support in *ASPIC*$^+$.

or necessary) they consider $BAF_x = (\mathcal{A}, \mathcal{D}_x, \mathcal{S}_s)$, where $\mathcal{D}_x$ adds to $\mathcal{D}$ all defeats induced by the constraints of semantics x. They then compare for each *ASPIC*$^+$-induced $BAF = (\mathcal{A}, \mathcal{D}, \mathcal{S}_s)$ and each corresponding $BAF_x = (\mathcal{A}, \mathcal{D}_x, \mathcal{S}_s)$ the sets $\mathcal{D}$ and $\mathcal{D}_x$. The semantics x is an abstraction of support type s just in case $(\mathcal{A}, \mathcal{D})$ and $(\mathcal{A}, \mathcal{D}_x)$ have the same extensions.

Cohen et al.'s findings on this question are largely negative. They identify only one full correspondence, between *ASPIC*$^+$ proper-subargument support and $BAFs$ for necessary support, regardless of the type of *ASPIC*$^+$ defeat. The conclusion we can draw from their findings is that the choice of semantics for $BAFs$ to a large extent depends on the nature of the support relation, so that whether a particular semantics is appropriate in a particular context cannot be determined if that nature is not specified.

5.3 Gradual notions of argument acceptability

Another recent trend in the formal study of argumentation is the development of gradual notions of argument acceptability. These notions are proposed as alternatives to extension-based notions that are defined on top of the theory of abstract or bipolar ar-

gumentation frameworks. The gradual notions are often motivated by a discontent with the fact that extension-based notions of acceptability only allow for rather coarse distinctions between degrees of acceptability. Pollock (2002) was, to our knowledge, the first who addressed this issue and proposed a formalisation of gradual "justification". The current developments arguably go back to Cayrol and Lagasquie-Schiex (2005) and gained momentum with publications like Amgoud and Ben-Naim (2013).

Although the new developments are very interesting and the formal achievements have been impressive, there are also reasons to take a step back. To start with, there is a need to reflect on which notions or aspects of argument acceptability, or argument strength, are modelled, and why proposed semantics or proposed sets of principles for those semantics are good. What is needed is a conceptual or philosophical underpinning of the formal ideas and constructs. Furthermore, almost all work builds on abstract or bipolar argumentation frameworks and thus does not give explicit formal accounts of the nature of arguments and their relations, while yet this may be relevant when evaluating the formal proposals.

Below a classification of three aspects of argument strength is proposed based on philosophical insights, in particular Aristotle's distinction between logic, dialectic and rhetoric. It is then argued that when developing or evaluating gradual accounts of argument strength it is essential to be explicit about which aspect of argument strength is modelled and about the adopted interpretation of the arguments and their relations.

5.3.1 Kinds of argument strength

In classifying aspects of argument strength it is natural to take Aristotle's famous distinction between logic, dialectic and rhetoric as starting point. Very briefly, *logic* concerns the validity of arguments given their form, *dialectic* is the art of testing ideas through critical discussion and *rhetoric* deals with the principles of effective persuasion (van Eemeren et al.; 2014, Section 1.4). Accordingly, we distinguish between logical, dialectical and rhetorical argument strength, where logical argument strength in turn divides into two aspects: inferential and contextual argument strength.

Inferential argument strength is about how well an argument's premises support its conclusion considering only the argument itself. Example criteria for argument strength are that arguments with only deductive inferences are stronger than arguments with defeasible inferences, or that arguments with only non-attackable premises are stronger than arguments with attackable premises.

Contextual argument strength is about how well the conclusion of an argument is supported in the context of a given set of arguments. Formal frameworks like Dung's theory of abstract argumentation frameworks and *ASPIC*$^+$ formalise this kind of argument strength. The reader might wonder why contextual strength is not called dialectical strength, since after all, determining an argument's contextual strength as defined here involves the comparison of argument and counterargument. Yet this is not truly dialectical, since the just-mentioned formalisms do not model principles of critical discussion but define structural relations between (sets of) arguments on the basis of a given body of information.

Rhetorical argument strength looks at how capable an argument is to persuade other participants in a discussion or an audience. Persuasiveness essentially is a psychological notion; although principles of persuasion may be formalised, their validation as principles of successful persuasion is ultimately psychological.

Dialectical argument strength looks at how challengeable an argument is in the context of a critical discussion. In (Zenker et al.; 2020, pp. 657) this is formulated as

> (...) the (un)availability of participant moves that constrain further interlocutor moves. Minimally, argument strength thus is a function of the (un)availability of non-losing future participant moves. In this sense, the strongest proponent-argument leaves no further opponent-move except concession (i.e., retraction of either a standpoint or of critical doubt), and the weakest proponent argument constrains no opponent-move, given the "move-space".

Thus conceived, an important aspect of dialectical strength is the degree of vulnerability of an argument in that how many attacks are allowed in a given state that decrease the argument's contextual status. This reflects an intuition that many decision makers are aware of, namely, to justify one's decisions as sparsely as possible, in order to minimise the chance of successful appeal.

5.3.2 Be explicit about which aspects of argument strength are modelled

In developing a gradual argumentation semantics, it is important to be explicit about which aspects of argument strength are modelled. The aspects serve different purposes, so principles or definitions that are good for one aspect may not be good for another aspect. Consider, for example, the two arguments A and B in Figure 5.5, where A

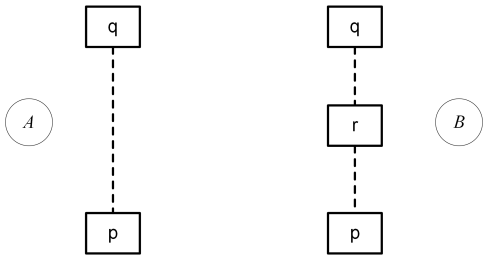


Figure 5.5: The *ASPIC*$^+$ defeat graph

defeasibly infers q from p while B first defeasibly infers r from p and then defeasibly infers q from r. Consider a definition of dialectical strength capturing that having fewer attackable elements is dialectically better and a definition of rhetorical strength that captures that a larger overlap of an argument's elements with the audience's beliefs is rhetorically better. Even without formalising these notions it is obvious that argument A is dialectically stronger than argument B, since A has one attackable element less than B. However, if the audience accepts that p defeasibly implies r and that r defeasibly implies q but not that p defeasibly implies q, then B is rhetorically stronger than A since it shares some elements with the background theory while A does not. This illustrates

that while sparsely justifying one's claims or decisions may be dialectically good, it may at the same time make an argument less persuasive.

5.3.3 Be explicit about the nature of arguments and their relations

We next illustrate how the nature of arguments and their relations can be relevant for the strength of arguments. We first consider support relations. Recall the four kinds of support in *ASPIC*$^+$ discussed in Section 5.2. Clearly, when support corresponds to (proper) subargument support, it makes no sense to regard supporters as strengthening the supported argument. Logically, the number of supports of an argument is then just a measure of its inferential complexity while dialectically, having more supporters may make an argument more vulnerable to attack and thus weaker. Similar observations hold for premise support. This notion may be useful in debate contexts, where there usually is no global knowledge base from which the debaters construct their arguments. In

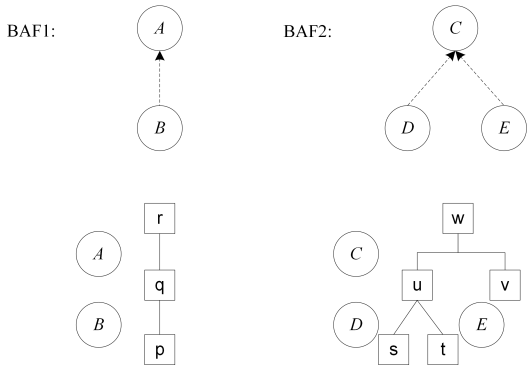


Figure 5.6: Is having more supporters better?

such a context it might be argued that, everything else being equal, a premise-supporter strengthens the supported argument. However, even then the nature of the arguments and their relations matters. Consider Figure 5.6, with two bipolar frameworks in the top row and two instantiations of these frameworks in the bottom row (in BAF1 and BAF2 the dashed arguments depict support relations between arguments). According to the idea that, everything else being equal, having more supporters is better, argument C on the top right is better supported than argument A on the top left since C has two premise-supporters while A has just one. However, as shown in the bottom row, all of A's premises (namely, q) are supported while only one of C's two premises is supported, so dialectically and perhaps also rhetorically A might just as well be regarded as better supported than C. Or imagine that D does not premise-support C on u but on v: then both A and C have all their premises supported, so there seems no reason to prefer C over A. Concluding, even in applications in which it makes sense to regard premise-supporters as, everything else being equal, strengthening the supported argument in some sense, it is important to take the structure of arguments and the nature of their relations into account.

We next illustrate the importance of being explicit about whether an argument is attackable or not. Consider the Cardinality Precedence principle that having fewer defeaters makes an argument stronger (Amgoud and Besnard; 2013) and consider AF1

with A being defeated by B and AF2 with C being defeated by D and E (Figure 5.7; in the AFs in Figures 5.7 and 5.8 the solid arrows depict defeat relations). According

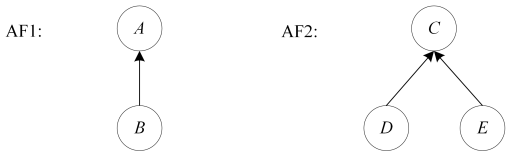


Figure 5.7: Is having fewer defeaters better?

to Cardinality Precedence argument A is stronger than argument C. However, if B is not attackable while D and E are attackable, then it is not obvious why this should be the case. For example, from the point of view of dialectical strength C is arguably dialectically stronger than A since C can still be made in by adding new arguments and defeats while for A this cannot happen.

Another example of why the distinction between attackable and non-attackable arguments matters concerns the principle that having more defenders makes an argument stronger. (An argument A defends an argument B if A defeats a defeater of B.) Consider the AFs displayed in Figure 5.8. According to the gradual semantics of Grossi

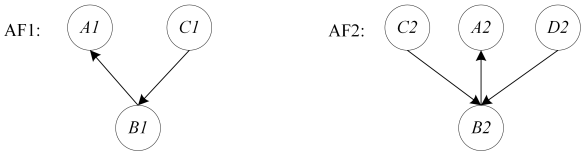


Figure 5.8: Is having more defenders better?

and Modgil (2015, 2019), A_2 is justified to a higher degree than A_1, since A_2 has two defenders (C_2 and D_2) while A_1 has only one defender (C_1). However, if C_1 is unattackable while C_2 and D_2 are attackable then it is not obvious why this has to be so, whatever aspect of argument strength is modelled.

5.4 Exercises

EXERCISE 5.4.1 In Exercise 4.8.5 assume that $d_2 < d_1$ and $d_4 < d_2$ and apply the last-link ordering.

1. Specify the resulting preference-based argumentation framework.
2. Verify the status of r according to preferred semantics.
3. What is the answer to (2) in *ASPIC*$^+$?

EXERCISE 5.4.2 In Example 4.3.6 delete s_2, assume that $d_2 < d_5$ and $d_5 < d_3$ and apply the last-link ordering.

1. Specify the resulting preference-based argumentation framework.
2. Verify the status of r according to preferred semantics.
3. What is the answer to (2) in *ASPIC*$^+$?

EXERCISE 5.4.3 Consider the *ASPIC*$^+$-style SAF in Figure 5.9, where all premises are ordinary and there are no preferences, and in which only direct defeat relations are considered (namely, the symmetric one between arguments B and H). Consider then

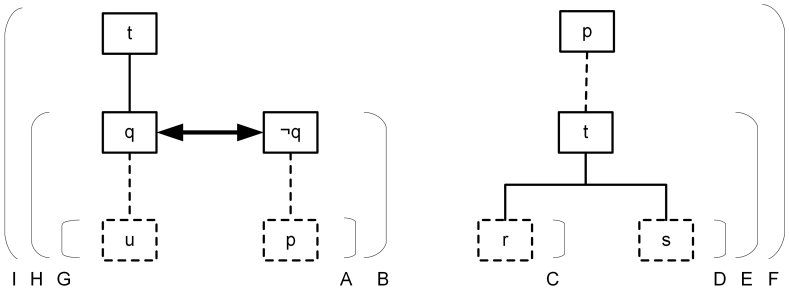


Figure 5.9: An *ASPIC*$^+$-style *SAF*

the AF corresponding to the SAF, again with only direct defeat relations, where the arguments are named with the capitals in the figure.

1. Construct for each of the four kinds of support relations defined in Section 5.2 for *ASPIC*$^+$ the BAF by adding all support relations of that kind to the AF.

2. Specify for each of the four $BAFs$ constructed in your answer to (1) and for each of the four constraints on $\mathcal{D}$ defined in Section 5.2 for $BAFs$ which additional defeat relations are induced by these constraints.

3. Construct for each of the four $BAFs$ from your answer to (1) combined with each of the three semantics for $BAFs$ of Definition 5.2.1 the AF associated with the *BAF*. So you have to construct $4 \times 3 = 12$ AFs.

4. Determine the grounded labelling for each of the 12 AFs constructed in your answer to (3).

Chapter 6

Dynamics of argumentation

6.1 Introduction

In this chapter aspects of the dynamics of argumentation are discussed while abstracting from the procedural context in which argumentation takes place. For example, when discussing methods for extending or revising argumentation frameworks, we disregard the question whether such a change is allowed according to the rules of debate (for example, whether certain types of evidence are admissible or whether claims made earlier can be retracted). The procedural aspects of argumentation are discussed in Chapter 7.

The study of information dynamics in argumentation concerns the nature and effects of change operations on a given argumentation state. This work is motivated by several application scenario's, such as:

- Adjudication dialogues like in legal procedure, where two adversaries aim to persuade an adjudicator of the dispute (judge or jury).
- Debates in parliament or similar bodies that have to vote on proposals, where members try to persuade each other to vote for or against the various proposals.
- Any individual or group of individuals interested in a debate and wanting to evaluate it from his/her/their point of view.

In dynamic contexts, adding new arguments clearly makes sense but adding attacks or defeats only seems to make sense when these attacks involve at least one new argument. Deleting attacks or defeats makes sense when interpreted as applying preferences to decide that a given attack relation does not succeed as defeat. Finally, deleting arguments makes sense in contexts where elements of arguments can be retracted by a participant or can be rejected by an adjudicator without stating a counterargument. An example of rejection by an adjudicator is in legal dialogues, where a judge can, for example, reject a factual premise since it has not been sufficiently backed by evidence and must therefore be ignored by the rules of legal procedure.

Most current work on argumentation dynamics concerns abstract approaches to argumentation. In particular the following operations on abstract argumentation frameworks have been studied:[1] addition or deletion of (sets of) arguments (e.g. Baumann

[1]Note that most of the literature on argumentation dynamics uses the term 'attack' for what in this reader (also in the present chapter) is called 'defeat'.

(2012); Baumann and Brewka (2010); Cayrol et al. (2010)) and addition or deletion of (sets of) defeat relations (e.g. Modgil (2006); Baroni and Giacomin (2007); Bisquert et al. (2013)). This work then studies preservation and enforcement properties. Preservation is about the extent to which the current status of arguments is preserved under change, while enforcement concerns the extent to which desirable outcomes can or will be obtained by changing a framework.

The work on abstract argumentation dynamics disregards the structure of arguments and the nature of their conflicts, which in dynamic settings is a serious limitation. For example, abstract models of argumentation dynamics do not recognise that some arguments are not attackable (such as strict-and-firm arguments in *ASPIC*$^+$) or that some defeats cannot be deleted (for example between arguments that were determined to be equally strong), or that the deletion of one argument implies the deletion of other arguments (when the deleted argument is a subargument of another), or that the deletion or addition of one defeat implies the deletion or addition of other defeats (for example, defeating an argument implies that all arguments of which it is a subargument are also defeated). All this means that formal results on preservation and enforceability are only relevant for very specific cases and do not cover many realistic situations in argumentation.

Accordingly, the purpose of this chapter is twofold:

1. to introduce the current research on the dynamics of argumentation;
2. to warn against naive work at the abstract level.

6.2 Work on preservation properties: resolution semantics

The first work on preservation properties concerned so-called resolution semantics. Here the focus is on deleting defeat relations as a way to express a preference of one argument over another: that a defeat from A on B is deleted means that A is regarded as inferior to B so that A's attack on B does not succeed as defeat. This idea was introduced for abstract argumentation by Modgil (2006) and further developed by Baroni and Giacomin (2007).

6.2.1 Abstract resolution semantics

Given an abstract argumentation framework $AF = (\mathcal{A}, \mathcal{D})$ (where $\mathcal{A}$ is a set of *arguments* and $\mathcal{D}$ a binary *defeat relation* on $\mathcal{A}$), a resolution $AF' = (\mathcal{A}, \mathcal{D}')$ is such that $\mathcal{D}'$ replaces one or more symmetric defeats in $\mathcal{D}$ by an asymmetric relation in $\mathcal{D}'$. More precisely:

Definition 6.2.1 [**Resolutions**] An argumentation framework $AF' = (\mathcal{A}, \mathcal{D}')$ is a *resolution* of an argumentation framework $AF = (\mathcal{A}, \mathcal{D})$ iff for all arguments A and B:

1. If $(A, B) \in \mathcal{D}$ and $(B, A) \notin \mathcal{D}$, then $(A, B) \in \mathcal{D}'$;
2. If $(A, B) \in \mathcal{D}$ and $(B, A) \in \mathcal{D}$, then $(A, B) \in \mathcal{D}'$ or $(B, A) \in \mathcal{D}'$;
3. If $(A, B) \in \mathcal{D}'$, then $(A, B) \in \mathcal{D}$.

A resolution $AF' = (\mathcal{A}, \mathcal{D}')$ is *partial* if there exist $A, B \in \mathcal{A}$ such that $A \neq B$ and $(A, B) \in \mathcal{D}'$ and $(B, A) \in \mathcal{D}'$; otherwise a resolution is *full*.

Then properties can be studied concerning the relations between the original status of an argument and its status in some or all resolutions. We will discuss some of these properties for grounded and preferred semantics.

Property 6.2.2 [**Left to Right Sceptical**] If X is a justified argument of $AF = (\mathcal{A}, \mathcal{D})$, then X is a justified argument of every full resolution $AF' = (\mathcal{A}, \mathcal{D}')$ of AF.

This property holds for grounded semantics but not for preferred semantics. For grounded semantics the intuition why the property holds is that arguments can only be in the grounded extension if all conflicts on which they depend are fully resolved. For a counterexample for preferred semantics let $\mathcal{A} = \{A, B, C, D\}$ such that A defeats B, B defeats C, C defeats A and A and D defeat each other. Then the unique preferred extension is $\{D\}$ but there exists a resolution with an empty preferred extension, namely, when the defeats of D on A is deleted.

Property 6.2.3 [**Right to Left Sceptical**] If X is a justified argument of every full resolution $AF' = (\mathcal{A}, \mathcal{D}')$ of $AF = (\mathcal{A}, \mathcal{D})$, then X is a justified argument of AF.

This property holds for preferred semantics but not for grounded semantics. For preferred semantics the intuition why the property holds is that each preferred extension is already implicitly a full resolution. For a counterexample for grounded semantics let $\mathcal{A} = \{A, B, C, D\}$ such that A and B defeat each other, both A and B defeat C and C defeats D. Then there are two full resolutions: one in which the defeat of A on B is deleted and one in which the defeat of B on A is deleted. The first resolution yields the grounded extension $\{B, D\}$ while the second resolution yields the grounded extension $\{A, D\}$. So D is justified in all full resolutions. However, the initial grounded extension is empty.

Other preservation properties can be formulated by replacing one or both occurrences of 'justified' with 'defensible' and/or replacing occurrences of 'all' with 'some'. For example:

Property 6.2.4 [**Left to Right Credulous to Justified**] If X is a defensible argument of $AF = (\mathcal{A}, \mathcal{D})$, then X is a justified argument of some full resolution $AF' = (\mathcal{A}, \mathcal{D}')$ of AF.

This property does not hold for grounded semantics. The counterexample to *Right to Left Sceptical* also holds here.

6.2.2 Structured resolution semantics

When resolutions are intended to model the outcome of preference arguments, then the above-defined abstract study of resolutions has limited applicability (cf. Modgil and Prakken (2012)). Firstly, one must also account for the resolution of *asymmetric* attacks, since many argumentation formalisms, including *ASPIC*$^+$, apply preferences to deny the success of asymmetric attacks as defeats. Furthermore, some formalisms apply preferences so that *both* attacks in a symmetric attack fail to succeed as defeats. Third, sometimes resolutions of symmetric attacks are impossible; for example when two symmetrically attacking arguments are assigned equal strength.

Resolutions can also be impossible for another reason: preference relations have properties, so the addition of preferences to resolve one attack may imply further preferences and thereby make resolutions based on conflicting preferences impossible. Finally, resolutions are impossible if some attacks succeed irrespective of preferences (e.g., undercutters or contrary-underminers in *ASPIC*$^+$).

Such subtleties can only be fully appreciated in a setting where the structure of arguments and the nature of attack and the use of preference to define defeats is made explicit. To this end Modgil and Prakken (2012) study resolutions in the *ASPIC*$^+$ framework. They are interested in the case where given a (c-)SAF $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ and its defined defeat relation, what is the relationship, under different semantics, between the justified arguments of Δ and the justified arguments of $\Delta' = (\mathcal{A}, \mathcal{C}, \preceq')$, where Δ' is a resolution of Δ obtained by extending' $\preceq$ to the preference relation $\preceq'$. They assume that the preference relation on arguments is a partial preorder, that is, transitive and reflexive. Note that the above-defined last- and weakest-link argument orderings are special cases of a partial preorder with no symmetric relations between different arguments.

Definition 6.2.5 Let $\preceq$ be a partial preorder over a set Γ. Then $\preceq'$ *extends* $\preceq$ iff $\preceq \subseteq \preceq'$ and $\forall X, Y \in \Gamma$, $X \prec Y$ implies $X \prec' Y$.
Let $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$ be a *SAF*. Then $\Delta' = (\mathcal{A}, \mathcal{C}, \preceq')$ *preference-extends* Δ iff $\preceq'$ *extends* $\preceq$.

To motivate the definition of *extends*, recall that $\preceq$ is a partial preorder. Thus it does not in general suffice to define *extends* in terms of the condition $X \prec Y$ implies $X \prec' Y$ alone (although it does suffice for the weakest- and last-link ordering). To see why, suppose $X \preceq Y$ and $Y \preceq X$, which implies $X \approx Y$; that is they are effectively assigned the same strength. Hence, it might be that $\preceq'$ preserves the strict preferences in $\preceq$, but $X \not\preceq Y$ and $Y \not\preceq X$. But we certainly want to preserve the assignment of equal strength to X and Y. On the other hand, it does not suffice to define *extends* in terms of the condition $\preceq \subseteq \preceq'$ alone. This is because given only $X \preceq Y$ and so $X \prec Y$, we want that this strict preference be preserved in the extended argument ordering. However, if $X \preceq' Y$ and $Y \preceq' X$, then this strict preference would not be preserved.

It is straightforward to then show that if $(\mathcal{A}, \mathcal{C}, \preceq')$ *preference-extends* $(\mathcal{A}, \mathcal{C}, \preceq)$, and $\mathcal{D}'$ and $\mathcal{D}$ are the defeat relations respectively defined by $\preceq'$ and $\preceq$, then $\mathcal{D}' \subseteq \mathcal{D}$.

Now the notion of a preference-based resolution can be defined:

Definition 6.2.6 Let $\Delta' = (\mathcal{A}, \mathcal{C}, \preceq')$ be a SAF that *preference-extends* $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$, and let $\mathcal{D}'$ and $\mathcal{D}$ be defeat relations respectively defined by $\preceq'$ and $\preceq$. Then

- Δ' is a *preference-based resolution* of Δ iff $\mathcal{D}' \subset \mathcal{D}$.
- Δ' is a *full preference-based resolution* of Δ iff Δ' is a preference-based resolution of Δ and there exists no preference-based resolution $\Delta'' = (\mathcal{A}, \mathcal{C}, \preceq'')$ with induced defeat relations $\mathcal{D}''$ such that $\mathcal{D}'' \subset \mathcal{D}'$.

Below we will assume that the argument ordering $\preceq$ is an elitist weakest- or last-link ordering induced by partial preorders $\leq$ on $\mathcal{R}_d$ and $\leq'$ on $\mathcal{K}_p$. Moreover, we will only consider preference-based resolutions that extend $\leq$ and $\leq'$ in the sense of Definition 6.2.5.

Next the preservation properties for preference-based resolutions are restated as follows:

Property 6.2.7 [**Left to Right Sceptical**] If X is a justified argument of $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$, then X is a justified argument of every full preference-based resolution $\Delta' = (\mathcal{A}, \mathcal{C}, \preceq')$ of Δ.

Property 6.2.8 [**Right to Left Sceptical**] If X is a justified argument of every full preference-based resolution $\Delta' = (\mathcal{A}, \mathcal{C}, \preceq')$ of $\Delta = (\mathcal{A}, \mathcal{C}, \preceq)$, then X is a justified argument of Δ.

Now it is crucial to note that the results on the abstract version of resolution semantics are not inherited by these definitions in the context of *ASPIC*$^+$, so they have to be verified again. Modgil and Prakken (2012) prove that grounded semantics still fails *Right to Left Sceptical* and that it still satisfies *Left to Right Sceptical* but the latter only for or finitary frameworks. Moreover, they prove that preferred semantics still fails *Left to Right Sceptical* but, remarkably, it now also fails *Right to Left Sceptical*, while this property holds for abstract resolution semantics. We will not give their counterexample here but only remark that it is due to the fact that some preferences entail other preferences by the properties of partial preorders, so that not all resolutions that are possible in abstract resolution semantics as defined in Modgil (2006); Baroni and Giacomin (2007) are possible in preference-based resolution semantics as defined by Modgil and Prakken (2012). This shows that the nature of attack and defeat is relevant when defining resolution semantics.

6.3 Work on enforcement properties: expansions of argumentation frameworks

We next discuss work on enforcement properties, and we do so in the context of the theory of expansions of argumentation framework. This concerns contexts where new arguments and possibly new defeats involving new arguments can be added. Almost all current work on expansions and enforcement is in abstract argumentation. We will now first review the abstract notion of expansions as introduced by Baumann and Brewka (2010).

6.3.1 Abstract theory of expansions

Baumann and Brewka (2010) define expansions of AFs as follows.

Definition 6.3.1 [**Expansions**] An abstract argumentation framework AF' is an *expansion* of an abstract argumentation framework $AF = (A, \mathcal{D})$ iff $AF' = (\mathcal{A} \cup \mathcal{A}', \mathcal{D} \cup \mathcal{D}')$ for some nonempty $\mathcal{A}'$ disjoint from $\mathcal{A}$ such that for all $A, B \in \mathcal{A}$: if $(A, B) \in \mathcal{D}'$ then $A \in \mathcal{A}'$ or $B \in \mathcal{A}'$.

Thus expansions add new arguments and possibly new defeat relations[2].

A central enforcement result of Baumann and Brewka (2010) is the following one (restricted to the types of semantics considered in this reader).

Theorem 6.3.2 For $T \in \{\text{complete, preferred, grounded, stable}\}$, for any $AF = (\mathcal{A}, \mathcal{D})$ and for any conflict-free $C \subset \mathcal{A}$ unequal to a T-extension of AF, there exists an expansion AF' of AF such that $C \subset E$ for some T-extension E of AF', where the expansion can be chosen such that E is the unique T-extension of AF'.

[2] Baumann and Brewka (2010) call such expansions *normal* and also define other types of expansions.

Their proof of this result shows how a single argument can be added that defeats all arguments in AF outside C.

Now it is crucial to note that this proof depends on the implicit assumption that such an expansion can always be constructed, but this may not be the case. For example, as remarked above, in an *ASPIC*$^+$ instantiation not all arguments may be attackable, and in Section 6.3.2 below we will see that the result depends on more implicit assumptions. For now we give a simple abstract example (also given by Baumann and Brewka), with an AF consisting of two arguments A_2 and A_1 where A_2 defeats A_1. According to Baumann and Brewka the AF can be expanded by adding some A_3 defeating A_2 but if A_2 is unattackable since it is a strict-and-firm *ASPIC*$^+$ argument, then there will be no such expansion.

For these reasons Prakken (2023) refines the above notion of an expansion, by making expansions relative to a given background *universal argumentation framework* and by distinguishing between expansions that are allowed and those that are not allowed. Both refinements are useful for avoiding implicit assumptions at the abstract level that are not always satisfied by instantiations.

Definition 6.3.3 [**Argumentation frameworks in a universal *AF***] Given a universal argumentation framework $UAF = (\mathcal{A}^u, \mathcal{D}^u)$, an *argumentation framework in* UAF is any $AF = (\mathcal{A}, \mathcal{D})$ such that $\mathcal{A} \subseteq \mathcal{A}^u$ and $\mathcal{D} \subseteq \mathcal{D}^u_{|\mathcal{A}\times\mathcal{A}}$.

The fact that $\mathcal{D}$ is not required to equal $\mathcal{D}^u_{|\mathcal{A}\times\mathcal{A}}$ is to allow for instantiations with systems like *ASPIC*$^+$ that use preferences to resolve attacks, similar to in resolution semantics (see Section 6.2).

The notion of a universal argumentation framework can be used for expressing, for instance, whether an argument can be defeated. However, it cannot be used for, for instance, ensuring that implied defeats are added. For such constraints we must also distinguish between allowed and not allowed expansions. While expansions can be disallowed for dialogical or procedural reasons, for present purposes it is especially relevant that underlying structured accounts of argumentation may disallow expansions, as we will see in Section 6.3.2 for *ASPIC*$^+$.

Definition 6.3.4 [**Expansions given a universal argumentation framework**] Let AF = $(\mathcal{A}, \mathcal{D})$ and AF' be two abstract argumentation frameworks in UAF. Then AF' is an *expansion* of AF given UAF if AF' = $(\mathcal{A} \cup \mathcal{A}', \mathcal{D} \cup \mathcal{D}')$ for some nonempty $\mathcal{A}'$ disjoint from $\mathcal{A}$ such that for all $A, B \in \mathcal{A}$: if $(A, B) \in \mathcal{D}'$ then $A \in \mathcal{A}'$ or $B \in \mathcal{A}'$.

Let $X_{UAF}(AF)$ be the set of all expansions of AF given UAF. Then the set of *allowed expansions* of AF given UAF is some designated subset of $X_{UAF}(AF)$.

6.3.2 Expansions in *ASPIC*$^+$

In this section we instantiate Definitions 6.3.3 and 6.3.4 for *ASPIC*$^+$. This requires a specification of how the UAF can be generated by a universal *structured* argumentation framework to which it corresponds. Since a *SAF* is in *ASPIC*$^+$ determined by an argumentation theory, we must also specify the notion of a universal argumentation theory.

Universal Structured Argumentation Frameworks

A UAF is now defined as corresponding to a universal structured argumentation framework, which is in turn defined by a universal argumentation theory. Together, they define the space of possible knowledge bases, possible sets of inference rules and possible argument orderings and thus define the space of possible argumentation frameworks.

Definition 6.3.5 [Universal Argumentation theories and universal structured *AFs*] A *universal argumentation theory* is a tuple $UAT = ((\mathcal{L}^u, \mathcal{R}_s^u \cup \mathcal{R}_d^u, n^u), \mathcal{K}_n^u \cup \mathcal{K}_p^u)$ where all elements are defined as for *ASPIC*$^+$ argumentation theories except that $\mathcal{K}_n^u$ and $\mathcal{K}_p^u$ do not have to be disjoint. Then a *universal structured argumentation framework* defined by UAT is a tuple $USAF = (\mathcal{A}^u, \mathcal{C}^u, \preceq^u)$ defined according to Definition 4.3.14, where $\preceq^u$ is an empty preference ordering on $\mathcal{A}^u$. A $UAF = (\mathcal{A}^u, \mathcal{D}^u)$ that is the abstract argumentation framework corresponding to some given $USAF$ is denoted by $sUAF$.

Example 6.3.6 Consider a *UAT* with

$\mathcal{L}^u = \{p, \neg p, q, \neg q, r, \neg r, d, \neg d, d', \neg d'\}$,
$\mathcal{R}_s^u = \varnothing$,
$\mathcal{R}_d^u = \{p \Rightarrow q; \neg r \Rightarrow \neg q\}$,
$n^u = \{(p \Rightarrow q, d), (\neg r \Rightarrow \neg q, d')\}$,
$\mathcal{K}_n^u = \{p\}$,
$\mathcal{K}_p^u = \mathcal{L}^u$

Note that the sets $\mathcal{R}_s^u$ and $\mathcal{R}_d^u$ of a UAT are not required to contain all well-formed strict, respectively, defeasible rules over $\mathcal{L}^u$. This is to allow for instantiations where the strict rules are defined by a logical interpretation of $\mathcal{L}^u$ and/or the defeasible rules correspond to some recognised set of argument schemes. The limiting case where $\mathcal{R}_s^u$ and $\mathcal{R}_d^u$ do contain all well-formed rules over $\mathcal{L}$ is suitable for applications where the choice of strict and/or defeasible rules is fully free, as, for instance, in online debate settings. For similar reasons $\mathcal{K}_p^u$ and K_p^u are not required but are allowed to equal $\mathcal{L}^u$. The reason why $\mathcal{K}_n^u$ and K_p^u can overlap is to allow that the type of a premise is unspecified until determined when constructing an AT in UAT. Accordingly, to keep the notion of an argument on the basis of a UAT well-defined, we now assume that in Definition 4.3.5(1) it is explicitly indicated whether a premise is taken from $\mathcal{K}_n^u$ or from $\mathcal{K}_p^u$. Finally, the idea behind the choice of $\preceq^u$ as the empty ordering is that a universal SAF does not commit to any way to resolve preference-dependent conflicts. Note that the empty ordering induces the greatest set of defeat relations in that every attack succeeds as defeat. Commitments on how conflicts should be resolved can be expressed in the specification of a SAF in a $USAF$, by adopting any nonempty argument ordering. At the abstract level this was captured in Definition 6.3.3 in the use of $\subseteq$ instead of $=$ in the requirement that $\mathcal{D} \subseteq \mathcal{D}^u_{|\mathcal{A}\times\mathcal{A}}$. The structural counterpart of this definition looks as follows.

Definition 6.3.7 [Argumentation theories and structured *AFs* in a universal *AT*] An *argumentation theory in* a given UAT is an *ASPIC*$^+$ argumentation theory $AT = ((\mathcal{L}, \mathcal{R}_s \cup \mathcal{R}_d, n), \mathcal{K}_n \cup \mathcal{K}_p)$ where

- $\mathcal{L} \subseteq \mathcal{L}^u$;

- $\mathcal{R} \subseteq \{S \rightsquigarrow \varphi \in \mathcal{R}^u \mid S \subseteq \mathcal{L} \text{ and } \varphi \in \mathcal{L}\}$;
- $\mathcal{K}_n \subseteq \mathcal{K}_n^u$;
- $\mathcal{K}_p \subseteq \mathcal{K}_p^u$;
- $n = n^u \cap \{(r, \varphi) \mid r \in \mathcal{R}_d\}$.

An argumentation theory in UAT is is *logic-based* iff $\mathcal{R}_s = \{S \rightarrow \varphi \in \mathcal{R}_s^u \mid S \subseteq \mathcal{L}$ and $\varphi \in \mathcal{L}\}$.

A *structured argumentation framework in* UAT is a structured argumentation framework $SAF = (\mathcal{A}, \mathcal{C}, \preceq)$ defined by an AT in UAT for some ordering $\preceq$ on $\mathcal{A}$.

Logic-based ATs are called thus since they accept all strict inference rules from UAT that can be expressed over their language. Consider, for example, a UAT with $\mathcal{L}$ a propositional language and $\mathcal{R}_s = \{S \rightarrow \varphi \mid S \subseteq \mathcal{L}$ and S is finite and $\varphi \in \mathcal{L}$ and $S \vdash \varphi\}$ where $\vdash$ denotes propositional-logical consequence. Then all logic-based AT's in UAT allow for deductive reasoning with the full power of propositional logic over their language.

Example 6.3.6 (Cont) Consider the following AT in the above UAT:

$\mathcal{L} = \{p, \neg p, q, \neg q, r, \neg r, d\}$,
$\mathcal{R}_s = \varnothing$,
$\mathcal{R}_d = \mathcal{R}_d^u$,
$n = n^u$,
$\mathcal{K}_n = \varnothing$,
$\mathcal{K}_p = \{p, r\}$

Combined with $\preceq = \varnothing$ (or any argument ordering) the *SAF* defined by this *AT* contains three arguments:

A_1: p $\quad$ A_2: $A_1 \Rightarrow q$ $\quad$ B: r

and no attack relations; see also Figure 6.1 below on the left. The corresponding *AF* equals $(\{A_1, A_2, B\}, \varnothing)$.

Allowed expansions

So far all we have done is instantiating the notion of an AF in a UAF for *ASPIC*$^+$. The next step is to define the *allowed expansions* of an AF that corresponds to a SAF in a universal argumentation theory. The main task is to ensure that the result of such an expansion still corresponds to a structured AF in the universal argumentation theory, in order to respect the structural constraints imposed by *ASPIC*$^+$. Since the idea of expansions as originally proposed by Baumann and Brewka (2010) is that information is only added and not deleted, a natural way to achieve this is to require that expansions correspond to a SAF that expand (in a sense to be defined) the SAF to which the expanded AF corresponds. This is directly stated by the following definition.

Definition 6.3.8 [**Allowed expansions**] Consider any AF in a given $sUAF$ that corresponds to a $SAF = (\mathcal{A}, \mathcal{C}, \preceq)$ in UAT defined by $AT = ((\mathcal{L}, \mathcal{R}, n), \mathcal{K})$, and consider any AF' in $sUAF$ that expands AF. Then AF' is an *allowed expansion* of AF given UAF iff AF' corresponds to a $SAF' = (\mathcal{A}', \mathcal{C}', \preceq')$ in UAT defined by $AT' = ((\mathcal{L}', \mathcal{R}', n'), \mathcal{K}')$ such that:

1. $\mathcal{L} \subseteq \mathcal{L}'$;
2. $\mathcal{R} \subseteq \mathcal{R}'$;
3. $\mathcal{K}_n \subseteq \mathcal{K}'_n$ and $\mathcal{K}_p \subseteq \mathcal{K}'_p$;
4. for $\preceq'$ it holds that
   (a) $\preceq'$ is of the same type as $\preceq$;
   (b) $\preceq \subseteq \preceq'$;
   (c) $A \prec' B$ if $A \prec B$;
5. if AT is logic-based then AT' is logic-based.

The first condition on $\preceq'$ assumes that the argument ordering $\preceq$ of a SAF comes with a definition of its type, such as, for example, the definitions of a weakest- or last link ordering. The second and third condition together say that for arguments in $\mathcal{A}_{AF}$ it can only differ from $\preceq$ in that it turns undefined into defined relations; so it has to respect the strict and equality relations in $\preceq$. In this case we say that $\preceq'$ *extends* $\preceq$.

For specific application scenario's further constraints on expansions can be defined. For example, in knowledge-based applications (such as systems for medical diagnosis or crime investigation) the general knowledge can be required to be fixed across expansions, which can only add specific observations (such as the results of medical tests on a person who is ill, or of searches for evidence predicted by a crime scenario). Moreover, in dialogical settings, the dialoge protocol may impose constraints, such as admissibility of evidence or of types of arguments in legal procedures (for example, in some systems of criminal law analogical applications of criminal provisions are not allowed).

Example 6.3.6 (Cont) Suppose that someone wants to extend AT in a way that makes B overruled. Then given *UAT* this can only be done by adding r to $\mathcal{K}_p$ and letting $\preceq'=\{(r, \neg r)\}$. This results in

$\mathcal{L}' = \mathcal{L}, \mathcal{R}'_s = \mathcal{R}_s, \mathcal{R}'_d = \mathcal{R}_d, n' = n, \mathcal{K}'_n = \mathcal{K}_n,$
$\mathcal{K}'_p = \{p, q, \neg r\}$
$\preceq'=\preceq \cup\{(r, \neg r)\}$

The arguments and direct attacks in the *SAF'* defined by AT' are visualised in Figure 6.1 on the right. Combined with $\preceq'$ the corresponding AF' is as visualised on the right of

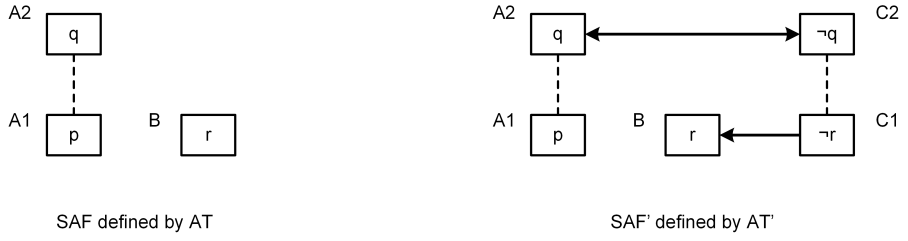


Figure 6.1: *SAF* defined by *AT* and *SAF*' defined by *AT*'

the figure. As regards the corresponding A's, the AF corresponding to AT is visualised on the left of the Figure 6.2. At first sight, it would seem that the addition of $\neg r$ to AT

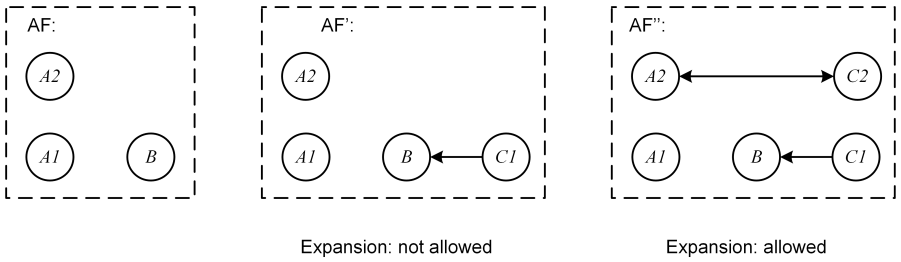


Figure 6.2: Allowed and not allowed abstract expansions

at the abstract level results in expanding with only argument C_1 from Figure 6.1 and a single defeat relation from C_1 to B. This would yield AF' in Figure 6.2. However, this expansion is not allowed: since the rule $\neg r \Rightarrow \neg q$ is in AT', argument C_2 must also be added. Moreover, on the basis of $\preceq'$, which contains no preference between C_2 and A_2, a mutual defeat relation between A_2 and C_2 has to be added. The result is AF'' in Figure 6.2. Assuming $\preceq$ and $\preceq'$ are of the same type, it is easy to see that AF'' is an allowed expansion of AF. Note that if no preference was added, then the defeat relation between B and C_1 would also be mutual while, moreover, a defeat relation from B to C_2 would have to be added.

Example 6.3.6 further illustrates that purely abstract accounts of expansions like Baumann and Brewka (2010) implicitly make assumptions that are not in general satisfied by instantiations.

6.4 Properties

We next investigate some properties of the above definitions. To start with, it holds that each allowed expansion adds at least one rule or one premise, otherwise it contains no new arguments. Next, since an expansion that is allowed according to Definition 6.3.8 corresponds to a SAF, it by definition satisfies closure under argument construction, under the subargument relation and under the constraints that *ASPIC*$^+$ imposes on the defeat relation. For example, it satisfies the constraint that if A defeats B and B is a subargument of C, then A defeats C (in the literature on bipolar argumentation frameworks called *closure under secondary attacks*; see Section 5.2 above).

Prakken (2023) proves some weaker counterparts of Theorem 6.3.2. However, for present purposes it is more interesting to show why Theorem 6.3.2 does not hold in general for the *ASPIC*$^+$ instantiation. This is for various reasons.

Not all arguments are attackable In Section 6.2.2 we already observed that Theorem 6.3.2 depends on the assumption that all arguments are attackable.

No conflict-free set of defenders Consider the AF inside the dotted box in Figure 6.3 based on an AT with $\mathcal{K}_n = \varnothing$, $\mathcal{K}_p = \{p, \neg p\}$, $\mathcal{R}_s = \{p \rightarrow \neg q; \neg p \rightarrow \neg q\}$, $\mathcal{R}_d = \{\Rightarrow q\}$. Assume, furthermore, that on the basis of UAT only one defeater D: $r \rightarrow \neg p$ of B and one defeater E: $\neg r \rightarrow p$ of C can be constructed, where $r, \neg r \in \mathcal{K}_p^u, \notin K_n^u$. Then there is no expansion with a conflict-free set of defenders of A, since for any $\preceq$ it holds that D defeats E or E defeats D. So A cannot be made justified or defensible in any expansion.

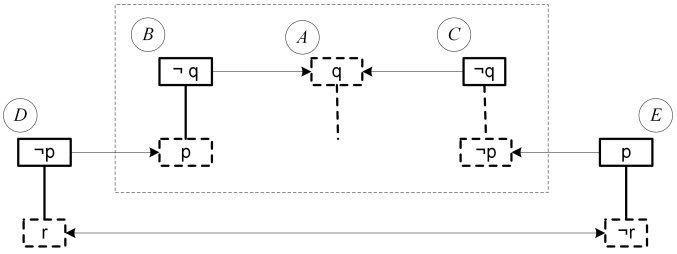


Figure 6.3: No conflict-free set of defenders.

All defenders imply a defeater Consider the example in Figure 6.4 based on an AT with $\mathcal{K}_n = \{q, s\}$, $\mathcal{K}_p = \{p\}$, $\mathcal{R}_s = \{s \rightarrow \neg p; r \rightarrow \neg p; r \rightarrow \neg d_2\}$, $\mathcal{R}_d = \{q \Rightarrow_{d1} r; s \Rightarrow_{d2} \neg p\}$ where the subscripts of $\Rightarrow$ denote the rule names. The AF contains A, B, C, D and all their subarguments, where both B and C defeat A and D defeats C. Assume that $A \prec B$ and $A \prec C$. Note that A is not in any T-extension for $T =$ grounded or complete or preferred, since it is defeated by B which is undefeated. The question is whether A can be made part of all $T-$extensions of some allowed expansion of AF. All such expansions must add a defeater E of B's subargument for r. Assume that on the basis of UAT a single undefeated argument E exists that defeats B but no defeater of C other than D exists. For instance, UAT could differ from AT only in that it also contains a strict rule $s \rightarrow \neg q$. Then any expansion defeating B contains E so also D is strictly defeated (on its subargument for r). But then C is defended and prevents A from being in any T-extension of the expansion. Hence no expansion exists in which A is in any T-extension. Dung (1995) calls arguments like E, which

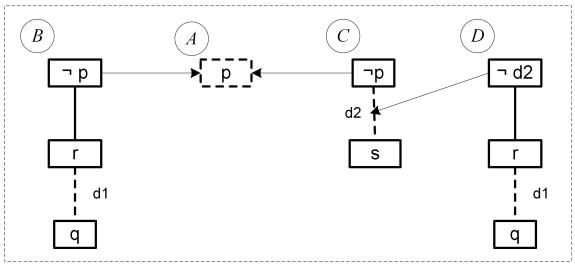


Figure 6.4: All defenders imply a defeater.

both defend and indirectly defeat an argument, *controversial arguments*. This example illustrates another assumption underlying Theorem 6.3.2, namely, that a defeat from a new to an old argument has no side effects in that the new argument also defeats other old arguments that are relevant to the status of an argument in the set that should be in an extension of the expansion. In other words, it is not the case in general that a set S' can be found such that $S \cup S'$ is admissible.

Further implicit assumptions underlying Theorem 6.3.2 are visualised in Figure 6.5, where the dotted boxes contain AFs while the entire graphs are $UAFs$. For the three abstract examples we leave it to the reader to verify that instantiations for *ASPIC*$^+$ exist.

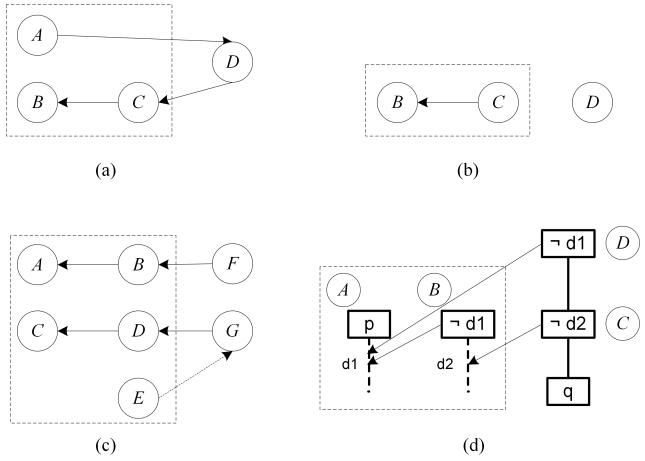


Figure 6.5: Further assumptions underlying Theorem 6.3.2.

No undefeated defenders Figure 6.5(a) refutes the assumption that always an undefeated expansion can be found with an AF with $\mathcal{A} = \{A, B, C\}$, where C defeats B and $USAF$ contains just one argument that defeats C, namely, D but which is defeated by A. Then there is no expansion that makes $\{B\}$ included in any extension.

No defenders Figure 6.5(b) refutes the assumption that always a defender of any argument in S exists in UAF.

No allowed way to extend the argument ordering Another assumption underlying Theorem 6.3.2 is that $\preceq$ can always be extended in any way. Counterexamples to this assumption can be constructed for the same reason as in Modgil and Prakken (2012) for resolution semantics in *ASPIC*$^+$ (see Section 6.2.2 above): properties of the argument ordering, such as transitivity, may make that adding explicit preferences to resolve a conflict in a desired way implies the addition of implicit preferences that prevent resolving another conflict in the desired way.

Effects of implied arguments Finally, Figure 6.5(d) illustrates the possible effects of implied arguments. Consider a logic-based AT with $\mathcal{K}_n = \mathcal{K}_p = \varnothing$, $\mathcal{R}_s = \{\neg d_2 \rightarrow \neg d_1; q \rightarrow \neg d_2\}$, $\mathcal{R}_d = \{\Rightarrow_{d1} p; \Rightarrow_{d2} \neg d_1\}$ and where UAT has $q \in \mathcal{K}_p^u$ and $q \rightarrow \neg d_2 \in \mathcal{R}_s^u$. Consider then the AF in Figure 6.5(d) and assume that $sUAF$ further only contains q, C and D. No expansion can make $\{A\}$ included in a T-extension for any T, since adding C (the only defender of A against B) also adds D to the expansion, which defeats A.

6.5 Conclusion

In this chapter we discussed two formal accounts of argumentation dynamics. Resolution semantics models and studies the addition of preferences to resolve conflicts between arguments, while the theory of expansions models and studies the addition of arguments and defeat relations to argumentation frameworks. We saw that both kinds

of operations on argumentation frameworks model realistic argumentation phenomena, but that their proper modelling must take the structure of arguments and the nature of their relations into account, on the penalty of implicitly making assumptions that are not always satisfied by instantiations of abstract approaches. In other words, abstraction does not imply generality. As we saw in Chapter 5, this observation is not confined to resolution semantics or expansion theory but holds more generally for any way of extending or modifying Dung's theory of abstract argumentation frameworks in the abstract.

6.6 Exercises

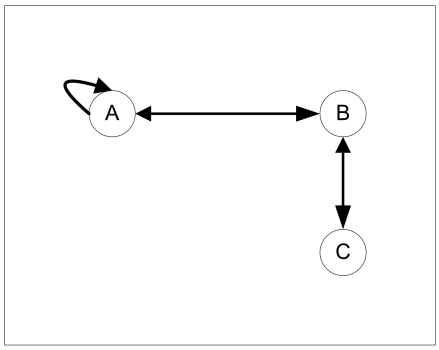
EXERCISE 6.6.1 Consider an AF such that A and B defeat each other, B defeats C and C defeats A.

1. Is B justified in preferred semantics?
2. Is B justified in all full resolutions in preferred semantics?

EXERCISE 6.6.2 Consider again Exercise 4.8.11(a,b,e).

1. Is D is justified in some and/or in all full resolutions in grounded semantics?
2. Is D is justified in some and/or in all full resolutions in preferred semantics?

EXERCISE 6.6.3 Consider the following abstract argumentation framework AF:



1. Specify all full resolutions of AF, drawing them as graphs.
2. Give two full resolutions of AF in which B is a member of all stable extensions.

EXERCISE 6.6.4 Consider again Example 4.3.6 and its development until 4.3.25. Does the status of r change in some resolutions? Answer this question both for the elitist weakest- and for the elitist last-link ordering.

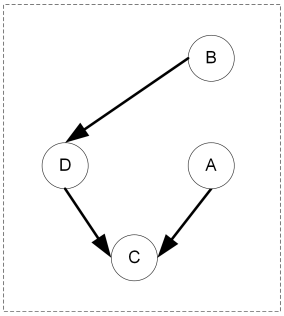
EXERCISE 6.6.5 let $\mathcal{R}_d = \mathcal{K}_n = \varnothing$, let $\mathcal{R}_s$ consist of all valid propositional inferences from finite sets and let $\mathcal{K}_p = \{p, q, \neg(p \wedge q\}$. Assume that $p <' \neg(p \wedge q)$ and $p <' q$ and apply the elitist weakest link ordering.

1. Is the argument $A = \neg(p \wedge q)$ justified in grounded semantics?

2. Is the argument $A = \neg(p \wedge q)$ justified in all full preference-based resolutions in grounded semantics?

EXERCISE 6.6.6 Create *ASPIC*$^+$ instantiations of the examples in Figure 6.5(a,b,c).

EXERCISE 6.6.7 Consider the following abstract argumentation framework AF:



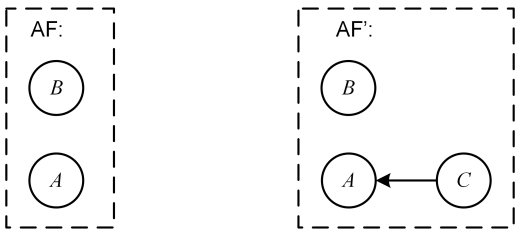
1. For $T \in \{$complete, preferred, grounded, stable$\}$, is there an expansion of AF according to Definition 6.3.1 for which $\{B, C\}$ is included in the unique T-extension?

2. Consider the following *ASPIC*$^+$ instantiation of this AF:

   A: $\Rightarrow_{d2} \neg d_1$
   B: $A \Rightarrow_{d4} \neg d_3$
   C: $\Rightarrow_{d1} p$
   D: $\Rightarrow_{d3} \neg p$

   where $d_1 < d_3$ and the last-link argument ordering is applied. Is there an allowed expansion of AF in the sense of Definition 6.3.8 that only adds a defeater of A and for which $\{B, C\}$ is included in the unique T-extension?

EXERCISE 6.6.8 Consider the following abstract argumentation frameworks AF and AF':



1. Is AF' an expansion of AF according to Definition 6.3.1 in the reader?

2. Suppose AF is an AF in a $sUAF$ that corresponds to a $SAF = (\mathcal{A}, \mathcal{C}, \preceq)$ in UAT defined by $AT = ((\mathcal{L}, \mathcal{R}, n), \mathcal{K})$, where:

   $\mathcal{L}^u = \{p, \neg p, q, \neg q, r, \neg r\}$,
   $\mathcal{R}^u_s = \{p \rightarrow q; r \rightarrow \neg p\}$,
   $\mathcal{K}^u_p = \mathcal{L}^u$,
   $\mathcal{R}^u_d = \mathcal{K}^u_n = \varnothing$

where

$\mathcal{L} = \mathcal{L}^u$,
$\mathcal{R}_s = \mathcal{R}_s^u$,
$\mathcal{K}_p = \{p\}$,
$\mathcal{R}_d = \mathcal{K}_n = \varnothing$

and where $\preceq^u = \preceq = \varnothing$.

Let $A = p$, $B = p \rightarrow q$, $C = r \rightarrow \neg p$. Is AF' an allowed expansion of AF according to to Definition 6.3.8 in the reader?

EXERCISE 6.6.9 Consider an AT with $\mathcal{K}_n = \{p, s\}$, $\mathcal{K}_p = \{u\}$, $\mathcal{R}_s = \varnothing$, $\mathcal{R}_d = \{p \Rightarrow_{r1} q; q \Rightarrow_{r2} r; s \Rightarrow_{r3} \neg r; q \Rightarrow_{r4} t; u \Rightarrow_{r5} \neg t; \neg u \Rightarrow_{r6} v\}$ (the subscripts of $\Rightarrow$ denote the rule names in $\mathcal{L}$), assume that $r_3 < r_2$ and $r_3 < r_5$ and let $\preceq$ be the elitist weakest-link argument ordering.

1. Construct the AF corresponding to the *SAF* defined by AT and $\preceq$.
2. Assume, furthermore, that $\mathcal{K}_n^u = \mathcal{K}_n$, $K_p^u = \{u, \neg u\}$, $\mathcal{R}_s^u = \mathcal{R}_s$ and $\mathcal{R}_d^u = \mathcal{R}_d$.
   (a) Is an AF' which adds $E : \neg u$ to $\mathcal{A}_{AF}$ and $(\{D_1, E), (E, D_1)\}$ to $\mathcal{D}_{AF}$ an allowed expansion of AF?
   (b) Is there an allowed expansion of AF in which arguments for $\neg r$ and t are both justified in grounded semantics?

Chapter 7

Dialogue systems for agent interaction with argumentation

This chapter is about formal dialogue systems for agent interaction with argumentation. The main focus is on so-called persuasion dialogues, in which two or more participants try to resolve a difference of opinion by arguing about the tenability of one or more claims or arguments, each trying to persuade the other participants to adopt their point of view. Dialogue systems for persuasion regulate what utterances the participants can make and under which conditions they can make them, what the effects of their utterances are on their propositional commitments, when a dialogue terminates and what the outcome of a dialogue is. Good dialogue systems regulate all this in such a way that conflicts of view can be resolved in a way that is both fair and effective.

The term 'persuasion dialogue' was introduced into argumentation theory by Douglas Walton (Walton; 1984) as part of his influential classification of dialogues into six types according to their goal (see also e.g. Walton and Krabbe (1995)). While *persuasion* aims to resolve a difference of opinion, *negotiation* tries to resolve a conflict of interest by reaching a deal, *information seeking* aims at transferring information, *deliberation* wants to reach a decision on a course of action, *inquiry* is aimed at "growth of knowledge and agreement" and *quarrel* is the verbal substitute of a fight. This classification is not meant to be exhaustive and leaves room for dialogues of mixed type, such as a negotiation that can shift to an embedded persuasion if the negotiating agents disagree about a relevant matter of fact.

The modern study of formal dialogue systems for persuasion probably started with two publications by Charles Hamblin (Hamblin; 1970, 1971). Initially, the topic was studied only within philosophical logic and argumentation theory. From the early nineteen nineties the study of persuasion dialogues was taken up in several fields of computer science. In Artificial Intelligence logical models of commonsense reasoning have been extended with formal models of persuasion dialogue as a way to deal with resource-bounded reasoning. In artificial intelligence & law interest in dialogue systems arose when researchers realised that legal reasoning is bound not only by the rules of logic and rational inference but also by those of fair and effective procedure. Persuasion was here seen as an appropriate model of legal procedures. Finally, in the field of multi-agent systems dialogue systems have been incorporated into models of rational agent interaction. To fulfill their own or joint goals, intelligent agents often need to interact with other agents. When they pursue joint goals, the typical modes of interaction are information seeking and deliberation and when they self-interestedly

pursue their own goals, they often interact by way of negotiation. In all these cases the dialogue can shift to persuasion. For example, in information-seeking a conflict of opinion could arise on the credibility of a source of information, in deliberation the participants may disagree about likely effects of plans or actions and in negotiation they may disagree about the reasons why a proposal is in one's interest; also, in all three cases the participants may disagree about relevant factual matters.

To delineate the precise scope of this chapter, it is useful to discuss what is the subject matter of dialogue systems. According to Carlson (1983) dialogue systems define the principles of coherent dialogue. In his words, whereas logic defines the conditions under which a proposition is true, dialogue systems define the conditions under which an utterance is appropriate. And the leading principle here is that an utterance is appropriate if it furthers the goal of the dialogue in which it is made. So, for instance, an utterance in a persuasion should contribute to the resolution of the conflict of opinion that triggered the persuasion, and an utterance in a negotiation should contribute to reaching agreement on a reallocation of scarce resources. Thus according to Carlson the principles governing the meaning and use of utterances should not be defined at the level of individual speech acts but at the level of the dialogue in which the utterance is made. Carlson therefore proposes a game-theoretic approach to dialogues, in which speech acts are viewed as moves in a game and rules for their appropriateness are formulated as rules of the game. Virtually all work on formal dialogue systems for persuasion follows this approach and therefore the discussion in this chapter will assume a game format of dialogue systems. It should be noted that the term *dialogue system* as used in this chapter only covers the rules of the game, i.e., which moves are allowed; it does not cover principles for playing the game well, i.e., strategies and heuristics for the individual players. Of course, the latter are also important in the study of dialogue, but they will be treated as being external to dialogue systems and instead of aspects of models of dialogue participants.

This chapter is organised as follows. First in Section 7.1 an example persuasion dialogue will be presented, to give a feel for what persuasion dialogues are and to provide material for illustration and comparison in the subsequent discussions. Then in Section 7.2 the general layout of dialogue systems is described and in Section 7.3 some common elements of dialogue systems for persuasion are discussed. Finally, in Section 7.4 two particular dialogue systems for persuasion are discussed. Exercises can be found at the end of the chapter.

7.1 An example persuasion dialogue

The following example persuasion dialogue exhibits some typical features of persuasion and will be used in this chapter to illustrate different degrees of expressiveness and strictness of the various persuasion systems.

`Paul:` My car is safe. (*making a claim*)
`Olga:` Why is your car safe? (*asking grounds for a claim*)
`Paul:` Since it has an airbag, (*offering grounds for a claim*)
`Olga:` That is true, (*conceding a claim*) but this does not make your car safe. (*stating a counterclaim*)
`Paul:` Why does that not make my care safe? (*asking grounds for a claim*)
`Olga:` Since the newspapers recently reported on airbags expanding without cause.

(*stating a counterargument by providing grounds for the counterclaim*)
`Paul:` Yes, that is what the newspapers say (*conceding a claim*) but that does not prove anything, since newspaper reports are very unreliable sources of technological information. (*undercutting a counterargument*)
`Olga:` Still your car is still not safe, since its maximum speed is very high. (*alternative counterargument*)
`Paul:` OK, I was wrong that my car is safe.

This dialogue illustrates several features of persuasion dialogues.

- Participants in a persuasion dialogue not only exchange arguments and counterarguments but also express various propositional attitudes, such as claiming, challenging, conceding or retracting a proposition.
- As for arguments and counterarguments it illustrates the following features.
  - An argument is sometimes attacked by constructing an argument for the opposite conclusion (as in Olga's two counterarguments) but sometimes by saying that in the given circumstances the premises of the argument do not support its conclusion (as in Paul's counterargument). This is the distinction between rebutting and undercutting counterarguments.
  - Counterarguments are sometimes stated at once (as in Paul's undercutter and Olga's last move) and are sometimes introduced by making a counterclaim (as in Olga's second and third move).
  - Natural-language arguments sometimes leave elements implicit. For example, Paul's second move arguably leaves a commonsense generalisation 'Cars with airbags usually are safe' implicit.
- As for the structure of dialogues, the example illustrates the following features.
  - The participants may return to earlier choices and move alternative replies: in her last move Olga states an alternative counterargument after she sees that Paul had a strong counterattack on her first counterargument. Note that she could also have moved the alternative counterargument immediately after her first, to leave Paul with two attacks to counter.
  - The participants may postpone their replies, sometimes even indefinitely: by providing her second argument why Paul's car is not safe, Olga postpones her reply to Paul's counterattack on her first argument for this claim; if Paul fails to successfully attack her second argument, such a reply might become superfluous.

7.2 General layout of dialogue systems

In this section the general layout of dialogue systems is described. Dialogue systems have a *dialogue goal* and at least two *participants*, who can have various *roles*. Dialogue systems have two languages, a *topic language* $\mathcal{L}_t$ governed by a *logic*, and a *communication language* $\mathcal{L}_c$ wrapped around the topic language. The communication language defines which utterances, or *speech acts* can be made about elements of subsets of the topic language. The heart of a dialogue system is formed by a *protocol*,

specifying the allowed moves at each point in a dialogue, the *effect rules*, specifying the effects of utterances on the participants' commitments, and the *outcome rules*, defining the outcome of a dialogue. Two kinds of protocol rules are sometimes separately defined, viz. *turntaking* and *termination* rules.

Some of these elements will now be defined formally in a generalisation of Definition 3.1.1. In the rest of this chapter this specification will be used when describing systems from the literature; in consequence, their appearance in this text may differ from their original presentation.

Definition 7.2.1 (Some elements of dialogue systems) Let $\mathcal{L}_c$ be a communication language with $\mathcal{L}_t$ its topic language.

- The set of *dialogues* defined by $\mathcal{L}_c$, denoted by $M^{\leq\infty}$, is the set of all sequences from $\mathcal{L}_c$, and the set of *finite dialogues*, denoted by $M^{<\infty}$, is the set of all finite sequences from $\mathcal{L}_c$. For any dialogue $d = m_1, \ldots, m_n, \ldots$, the subsequence $m_1, \ldots, m_i$ is denoted with d_i.

- A set of *effect rules* C for $\mathcal{L}_c$ specifies for each utterance $\varphi \in \mathcal{L}_c$ its effects on the commitments of the participants. These rules are specified as functions

  - $C_a : M^{<\infty} \longrightarrow Pow(\mathcal{L}_t)$

- A *protocol* Pr for $\mathcal{L}_c$ specifies the allowed (or 'legal') moves at each stage of a dialogue. Formally, A *protocol* on $\mathcal{L}_c$ is a function Pr with as domain a nonempty subset D of $M^{<\infty}$ taking subsets of $\mathcal{L}_c$ as values. That is:

  - $Pr : D \longrightarrow Pow(\mathcal{L}_c)$

  such that $D \subseteq M^{<\infty}$. The elements of D are called the *legal finite dialogues*. The elements of $Pr(K, d)$ are called the moves allowed after d given K. If d is a legal dialogue and $Pr(K, d) = \varnothing$, then d is said to be a *terminated* dialogue. Pr must satisfy the following condition: for all finite dialogues d and moves m, $d \in D$ and $m \in Pr(K, d)$ iff $d, m \in D$.

  It is useful (although not strictly necessary) to explicitly distinguish elements of a protocol that regulate turntaking and termination:

  - A *turntaking* function is a function $T : D \longrightarrow Pow(\mathcal{A})$. A *turn* of a dialogue is defined as a maximal sequence of moves in the dialogue in which the same player is to move. Note that T can designate more than one player as to-move next.
  - *Termination* is above defined as the case where no move is legal. Accordingly, an explicit definition of termination should specify the conditions under which Pr returns the empty set.

Note that no relations are assumed between a participant's commitments and beliefs. Commitments are an agent's publicly declared standpoints, which may or may not coincide with the agent's internal beliefs. For instance, an accused in a criminal trial may publicly defend his innocence while he knows he is guilty.

Definition 7.2.2 (Some protocol types)

- A protocol has a *public semantics* iff the set of legal moves is always independent from the agents' belief bases.
- A protocol Pr is *fully deterministic* if Pr always returns a singleton or the empty set. It is *deterministic in* $\mathcal{L}_c$ if the set of moves returned by Pr at most differ in their content but not in their speech act type.
- A protocol is *unique-move* if the turn shifts after each move; it is *multiple-move* otherwise.

Paul and Olga (ct'd): The protocol in our running example clearly is multiple-move.

7.3 Persuasion

We now discuss some common elements of systems for persuasion dialogue.

7.3.1 Communication languages and commitment rules for persuasion

As for the communication language and commitment rules of systems for persuasion dialogue, some common elements can be found throughout the literature. We list the most common speech acts, with their informal meaning and the various ways they are named in the literature.[1]

- *claim* φ (assert, statement, ...). The speaker asserts that φ is the case.
- *why* φ (challenge, deny, question, ...) The speaker challenges that φ is the case and asks for reasons why it would be the case.
- *concede* φ (accept, admit, ...). The speaker admits that φ is the case.
- *retract* φ (withdraw, no commitment, ..) The speaker declares that he is not committed (any more) to φ. Retractions are 'really' retractions if the speaker is committed to the retracted proposition, otherwise it is a mere declaration of non-commitment (for example, in reply to a question).
- φ *since* S (argue, argument, ...) The speaker provides reasons why φ is the case. Some protocols do not have this move but require instead that reasons be provided by a *claim* φ or *claim* S move in reply to a *why* ψ move (where S is a set of propositions). Also, in some systems the reasons provided for φ can have structure, for example, of a proof three or a deduction.
- *question* φ (...) The speaker asks another participant's opinion on whether φ is the case.

Paul and Olga (ct'd): In this communication language our example from Section 7.1 can be more formally displayed as follows:

[1] To make this chapter more uniform, the present terminology will be used even if the original publication of a system uses different terms.

Table 7.1: Locutions and typical replies

Locutions	Replies
claim φ	*why* φ, *claim* $\overline{\varphi}$, *concede* φ
why φ	φ *since* S, *retract* φ
concede φ	
retract φ	
φ *since* S	*why* ψ ($\psi \in S$), *concede* ψ ($\psi \in S$), φ' *since* S'

P_1: *claim* safe
O_2: *why* safe
P_3: safe *since* airbag
O_4: *concede* airbag
O_5: *claim* $\neg$ safe
P_6: *why* $\neg$ safe
O_7: $\neg$ safe *since* newspaper: "explode"
P_8: *concede* newspaper: "explode"
P_9: so what *since* $\neg$ newspapers reliable
O_{10}: $\neg$ safe *since* high max. speed
P_{11}: *retract* safe

As for the commitment rules, the following ones seem to be uncontroversial and can be found throughout the literature. (Below pl denotes the speaker of the move and s denotes the speech act performed in the move; effects on the other parties' commitments are only specified when a change is effected.)

- If $s(m) = \mathit{claim}(\varphi)$ then $C_{pl}(d, m) = C_{pl}(d) \cup \{\varphi\}$
- If $s(m) = \mathit{why}(\varphi)$ then $C_{pl}(d, m) = C_{pl}(d)$
- If $s(m) = \mathit{concede}(\varphi)$ then $C_{pl}(d, m) = C_{pl}(d) \cup \{\varphi\}$
- If $s(m) = \mathit{retract}(\varphi)$ then $C_{pl}(d, m) = C_{pl}(d) \ - \{\varphi\}$
- If $s(m) = \varphi$ *since* S then $C_{pl}(d, m) \supseteq C_{pl}(d) \cup S$

The rule for *since* uses $\supseteq$ since such a move may commit to more than just the premises of the moved argument. For instance, in Prakken (2005) the move also commits to φ, since arguments can also be moved as counterarguments instead of as replies to challenges of a claim. And in some systems that allow incomplete arguments, such as Walton and Krabbe (1995), the move also commits the speaker to the material implication $S \rightarrow \varphi$.

Paul and Olga (ct'd): According to these rules, the commitment sets of Paul and Olga at the end of the example dialogue are

- $C_P(d_{11}) \supseteq$ {airbag, newspaper: "explode", $\neg$ newspapers reliable}
- $C_O(d_{11}) \supseteq$ {$\neg$ safe, airbag, newspaper: "explode", high max. speed}

Speech act types often come with typical replies. Table 7.1 lists the typical replies of the common speech acts listed above.

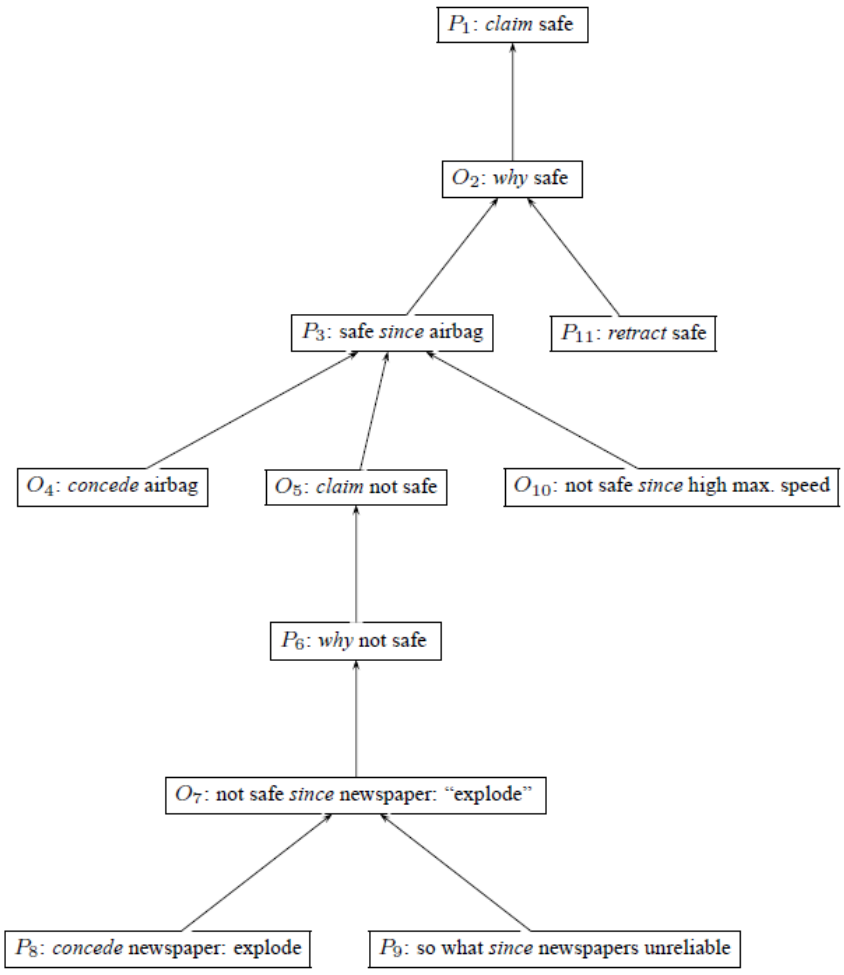


Figure 7.1: Reply structure of the example dialogue.

Paul and Olga (ct'd): In terms of this table our running example can now be displayed as in Figure 7.1, where the boxes stand for moves and the links for reply relations.

A table like the above one induces another distinction between dialogue protocols.

Definition 7.3.1 A dialogue protocol is *unique-reply* if at most one reply to a move is allowed throughout a dialogue; otherwise it is *multiple-reply*.

Of course, this distinction can be made fully precise only for systems that formally incorporate the notion of replies.

Paul and Olga (ct'd): The protocol governing our running example is multiple-reply, as illustrated by the various branches in Figure 7.1.

7.3.2 Types of protocol rules

According to their subject matter, several types of protocol rules can be distinguished. Some rules regulate a participant's *consistency*. This can be about *dialogical* consistency, such as a rule that each move must leave the speaker's commitments consistent or a rule that upon demand a speaker must resolve such an inconsistency. Or it can

be about a participant's *internal* consistency, such as the use of so-called assertion and acceptance attitudes (see Sections 7.3.3 and 7.4.1 below). For instance, a protocol rule could say that a participant may claim or accept a proposition only if his belief base contains a justified argument for the claim.

Other rules are about *dialogical coherence*, such as the rules that require a non-initial move to be an appropriate reply to some earlier move (see e.g. the table above). Yet other rules are about the *dialogical structure*, such as the termination rules and the rules that make the protocol a unique- or multiple move protocol, a unique- or -multiple reply protocol, or an immediate- or non-immediate-response protocol.

7.3.3 Assertion and acceptance attitudes

Sometimes so-called 'assertion and acceptance attitudes' are incorporated into persuasion protocols, which specify how an agent must choose between various otherwise legal moves given the information that the agent has available. We discuss the attitudes defined in Parsons et al. (2003), generalising them to any argument-based logic. In particular, we define them relative to an implicitly assumed argumentation theory AT as defined in Chapter 4, assuming that each argument has a conclusion, and also assuming a preference ordering on arguments. The idea is that AT contains all arguments that can be constructed on the basis of the information with which an agent reasons internally.

Definition 7.3.2 (Assertion and acceptance attitudes) An agent can have one of the following three *assertion attitudes*.

- A *confident* agent can assert any proposition for which he can construct an argument.
- A *careful* agent can assert any proposition p for which he can construct an argument and cannot construct a stronger argument for $-p$[2].
- A *thoughtful* agent can assert any proposition for which he can construct a justified argument.

(Here it is assumed that moving an argument implies that both its premises and conclusion are asserted.) An agent can have one of the following three *acceptance attitudes*.

- A *credulous* agent can accept any proposition for which he can construct an argument.
- A *cautious* agent can accept any proposition p for which he can construct an argument and cannot construct a stronger argument for $-p$.
- A *skeptical* agent can accept any proposition for which he can construct a justified argument.

It can be debated whether such attitudes must be part of a protocol or of a participant's heuristics. According to one approach, a dialogue protocol should only enforce coherence of dialogues; according to another approach, it should also enforce rationality of the agents engaged in a dialogue. The second approach allows protocol rules to refer to an agent's internal belief base and therefore such protocols do not have a public

[2]Here $-p$ is a contradictory of p in the sense of Definition 4.3.1

semantics (in the sense defined above in Section 7.2). The first approach does not allow such protocol rules and instead studies assertion and acceptance attitudes as an aspect of dialogical behaviour of agents.

7.3.4 Roles of commitments

Commitments can serve several purposes in dialogue systems (though particular systems may not use all of them). One role is in enforcing a participant's dialogical consistency, for instance, by requiring him to keep his commitments consistent at all times or to make them consistent upon demand. Another role is to enlarge the hearer's means to construct arguments. For instance, in Parsons et al. (2003)'s use of assertion and acceptance attitudes, they are applied relative to the agents' internal beliefs plus the other participant's commitments (see further Section 7.4.1 below). A third role of commitments is to determine termination and outcome of a dialogue, such as in the above definition of pure persuasion. For example, in two-party pure persuasion the proponent wins as soon as the opponent concedes his main claim while the opponent wins as soon as the proponent has retracted his main claim. Finally, commitments can determine certain 'dialectical obligations', as in a protocol rule that a participant's commitments must be consistent, or in a protocol rule that a commitment must be supported with an argument when it is challenged.

7.3.5 The role of the logic

The logic of most philosophical persuasion-dialogue systems is monotonic (usually standard propositional logic), while of most AI & Law and MAS-systems it is nonmonotonic. The logic of a persuasion-dialogue system can serve several purposes (though again particular systems may not use all of them). Firstly, it can be used in determining consistency of a participant's commitments. For this purpose a monotonic logic must be used. Secondly, it can be used to determine whether the reasons given by a participant for a challenged proposition indeed imply the proposition. When the logic is monotonic, the sense of 'imply' is obvious; when the logic is nonmonotonic, 'imply' means 'being an argument' in argument-based logics and (roughly) 'being a nonmonotonic consequence from the premises alone' in other nonmonotonic logics. Not all protocols require the reasons to be 'valid' in these senses. For instance, Walton and Krabbe (1995) allow the moving of incomplete arguments (but this still commits the speaker to the material implication *premises* $\rightarrow$ *conclusion*).

Note that this second use of a nonmonotonic logic does not yet exploit the nonmonotonic aspects of the logic. In argument-based terms, it only focuses on how arguments can be constructed, not on how they can be attacked by counterarguments. This is different in a third use of the logic, viz. to determine whether a participant respects his assertion or acceptance attitude: as we have just seen, most of these attitudes are defined in terms of counterarguments and/or defeasible consequence.

However, even if the full power of a nonmonotonic logic is used, it is still possible to distinguish between *internal* and *external* use of the logic. In Parsons et al. (2003) the nonmonotonic aspects of their (argument-based) logic are only used in verifying compliance with the assertion and acceptance attitudes; as we will see in Section 7.4.1, no other protocol rule refers to the notion of a counterargument. In particular, there is no rule allowing the attack of a moved argument by a counterargument. Also, the logic is not used in defining the outcome of a dialogue. Consequently, (if the attitudes

are regarded as heuristics and therefore external to a dialogue system), in these systems defeasible argumentation takes place only *within* an agent and not *between* agents. By contrast, in the system of Prakken (2005) (See Section 7.4.2) the moving of counterarguments in dialogues is allowed.

One external use of argumentation logics is to formulate dialogical notions of soundness and completeness. For example:

- A protocol is *sound* if whenever at termination p is accepted, p is justified by the participants' joint knowledge bases.
- A protocol is *weakly* complete if whenever p is justified by the participants' joint knowledge bases, there is a legal dialogue at which at termination p is accepted.
- A protocol is *strongly* complete if whenever p is justified by the participants' joint knowledge bases, all legal dialogues terminate with acceptance of p.

Similar notions can be defined relative to the joint theory constructed during a dialogue, while the notions can also be made conditional on particular agent strategies and heuristics.

7.4 Two systems

To illustrate the general discussion and some of the main design options, now two persuasion protocols will be discussed and applied to our running example.

7.4.1 Parsons, Wooldridge & Amgoud (2003)

In a series of papers Parsons, Wooldridge & Amgoud have developed an approach to specify dialogue systems for various types of dialogues. We base our discussion on Parsons et al. (2003), focusing on their system for persuasion dialogue.

The system is for dialogues between two players called White (W) and Black (B) on a single topic. The player who starts a dialogue is its proponent and the other player must, depending on her acceptance attitude, declare at her first move whether she is negative or doubtful towards the topic or wants to concede it. The participants have their own, possibly inconsistent belief base Σ. The players are assumed to adopt an assertion and an acceptance attitude, which they must respect throughout the dialogue. The attitudes are defined relative to their internal belief base (which remains constant throughout a dialogue) plus the commitment set of the other player (which may vary during a dialogue). The communication language $\mathcal{L}_c$ consists of claims, challenges, and concessions; it has no explicit reply structure but the protocol largely conforms to Table 7.1. Claims can concern both individual propositions and sets of propositions.

The logic of $\mathcal{L}_t$ is a special case of the ASPIC framework of Chapter 4[3] with $\mathcal{L}_t$ being a propositional language, with only strict rules, being the set of all classically valid inferences from finite sets, and with only ordinary premises. Moreover, arguments must have consistent premises. Defeat relations are defined according to the weakest-link principle in terms of a global total priority relation on $\mathcal{L}_t$, which they express

[3] In fact, there are some minor differences, but for ease of explanation and comparison we replace the logic adopted by Parsons et al. (2003) with the one described here.

in terms of levels, where level 1 contains the most preferred formulas. Defeasible inference is then defined with grounded semantics.

In dialogues, arguments cannot be moved as such but only implicitly as *claim* S replies to challenges of another claim φ, such that S is consistent and $S \vdash \varphi$. The logic is used to verify this condition and whether the players comply with their assertion and acceptance attitudes. The logic is not used externally. Finally, the commitment rules are standard and commitments are only used to enlarge the player's belief base with the other player's commitments; they are not used to constrain move legality or to define the dialogue's outcome.

The use of preferences involves some subtleties when applied to verify an assertion or acceptance attitude. As noted above, at any stage in a dialogue an agent a must reason with his own belief base Σ_a plus the commitments $C_{\overline{a}}(d)$ that the other party has in d. So W must define a total ordering on $\Sigma_W \cup C_B(d)$ while B must define a total ordering on $\Sigma_B \cup C_W(d)$. In practice these orderings may well be different but Parsons et al. (2003) still assume that the players agree on the ordering on $\mathcal{L}_t$. This may be justified by regarding the ordering on which the players agree as composed from their individual orderings. Several ways exist to define an overall preference ordering in terms of individual orderings (for example, p_i is overall preferred to p_j just in case both players prefer p_i to p_j, otherwise p_i and p_j are equal) but below we will abstract from such ways and simply assume that there is a unique ordering on $\mathcal{L}_t$ on which the agents agree.

We now present the formal definition of the persuasion protocol, which in fact defines a state transition diagram.

Definition 7.4.1 (PWA persuasion protocol) A move is legal iff it does not repeat a move of the same player, and satisfies the following procedure:

1. W claims φ (assuming W's assertion attitude allows it).
2. B concedes φ if its acceptance attitude allows, if not B claims $-\varphi$ if its assertion attitude allows it, or otherwise challenges φ.
3. If B claims $-\varphi$, then goto 2 with the roles of the players reversed and $-\varphi$ in place of φ.
4. If B has challenged, then:
   (a) W claims S, an argument for φ;
   (b) Goto 2 for each $s \in S$ in turn.
5. B concedes φ if its acceptance attitude allows, or the dialogue terminates.

Dialogues *terminate* as specified in condition 5, or when the move required by the procedure cannot be made, or when the player-to-move has conceded all claims made by the hearer.

No explicit win and loss functions are defined, but the possible outcomes are defined in terms of the propositions claimed by one player and conceded by the other.

To comment on this protocol, note first that in (4b) it is ambiguous in the case where S contains more than one premise, since it is unclear whether the turn shifts as soon as the first premise has been replied to or not. In the latter case, the protocol is multi-move, since a player may reply to each premise in turn. However, for simplicity we will below

assume that the turn shifts after the first reply to a *claim* S move; in this interpretation the protocol is unique move, except that after one premise is conceded, the next premise may immediately be replied to. Also, in both interpretations the protocol is unique-reply except that each element of a *claim* S move can be separately challenged or conceded. The protocol is deterministic in $\mathcal{L}_c$ but not fully deterministic, since if a player can construct more than one argument for a challenged claim, he has a choice which argument to play. Finally, the semantics of the protocol is not public, since agents have to comply with their assertion and acceptance attitudes, and these are partly defined in terms of their internal beliefs.

Let us first consider some simple dialogues that fit this protocol.

Example 7.4.2 First, let $\Sigma_W = \{p\}$ and $\Sigma_B = \varnothing$. Then the only legal dialogue is:

- W_1: *claim* p, B_1: *concede* p.

B_1 is B's only legal move, whatever its acceptance attitude, since after W_1, B must reason from $\Sigma_B \cup C_W(d_1) = \{p\}$ so that B can construct the trivial argument $(\{p\}, p)$. Here the dialogue terminates.

This example illustrates that the fact that the players must reason with the commitments of the other player makes that they can learn from each other. However, the following example illustrates that the same mechanism sometimes makes them learn too easily.

Example 7.4.3 Assume $\Sigma_W = \{q, q \supset p\}$ and $\Sigma_B = \{\neg q\}$, where all formulas are of the same preference level.

- W_1: *claim* p.

Now whatever her acceptance attitude, B has to concede p since she can construct the trivial argument $(\{p\}, p)$ for p while she can construct no argument for $\neg p$. Yet B has a defeater for W's only argument for p, namely, $(\{\neg q\}, \neg q)$, which defeats $(\{q, q \supset p\}, p)$. So even though p is not justified on the basis of the agents' joint knowledge, W_1 can win a dialogue about p.

This example thus illustrates that if the players have to reason with the other player's commitments, one player can sometimes 'force' an opinion onto the other player by simply making a claim. A possible solution to this problem is to restrict the information with which agent reason to their internal belief bases plus their own commitments. The following example illustrates another reason why this may be better.

Example 7.4.4 Consider next $\Sigma_W = \{q, q \supset p\}$ and $\Sigma_B = \{\neg p\}$, where q and $q \supset p$ are preferred over $\neg p$. Let W be thoughtful and skeptical and B careful. Then:

- W_1: *claim* p.

Since B must now reason with p, the continuation depends on the preference level of p. In fact, the protocol turns out to be problematic here. Since the players agree on the preference ordering, it seems reasonable to give p the same level as the level of the support of the strongest argument that can be constructed for p. However, the problem is that at this point in the dialogue B does not know which arguments W can construct for p. Let us sidestep this problem for the moment and let us first assume that p is preferred over $\neg p$. Then B must concede p whatever her acceptance attitude is. If, by contrast $\neg p$ is preferred over p, then a credulous agent must still concede p but a cautious and skeptical agent must instead proceed by claiming $\neg p$:

- B_1: *claim* $\neg p$.

Now W must apply clause (2) of the protocol, with $\varphi = \neg p$. Note that W must now reason with $\Sigma_W \cup \{\neg p\}$. He finds that he cannot accept $\neg p$ since his counterargument $(\{q, q \supset p\}, p)$ is acceptable since it is preferred over its only attacker $(\{\neg p, q \supset p\}, \neg q)$. Therefore, clause (2) requires him to assert p. However, the non-repetition rule makes this impossible, so that the dialogue terminates without agreement.

This example also illustrates that even if a proposition is defeasibly implied by $\Sigma_W \cup \Sigma_B$, it may not be agreed upon by the players (note that p is justified on the basis of this information). In fact, it also illustrates that sometimes there are no legal dialogues that agree upon such an implied proposition.

Paul and Olga (ct'd): Finally, our running example can be modelled in this approach as follows. Let us give Paul and Olga the following beliefs:

$\Sigma_W = \{\text{airbag, airbag} \supset \text{safe}, \neg(\text{newspaper} \supset \neg \text{ safe})\}$
$\Sigma_B = \{\text{newspaper, high-speed, newspaper} \supset \neg \text{ safe, high-speed} \supset \neg \text{ safe}\}$

(Note that Paul's undercutter must now be formalised as the negation of Olga's material implication.) Assume that all these propositions are equally preferred. We must also make some assumptions on the players' assertion and acceptance attitudes. Let us first assume that Paul is thoughtful and skeptical while Olga is careful and cautious, and that they only reason with their own beliefs and commitments.

P_1: *claim* safe $\qquad$ O_2: *claim* $\neg$ safe

Olga could not challenge Paul's main claim as in the example's orginal version, since she can construct an argument for the opposite claim '$\neg$ safe', while she cannot construct an argument for 'safe'. So she had to make a counterclaim. Now since players may not repeat moves, Paul cannot make the move required by the protocol and his assertion attitude, namely, claiming 'safe', so the dialogue terminates without agreement.

Let us now assume that the players must also reason with each others commitments. Then the dialogue evolves as follows:

P_1: *claim* safe $\qquad$ O_2: *concede* safe

Olga has to concede, since she can use Paul's commitment to construct the trivial argument $(\{\text{safe}\}, \text{safe})$, while her own argument for '$\neg$ safe' is not stronger. So here the dialogue terminates with agreement on 'safe', even though this proposition is not acceptable on the basis of the players' joint beliefs.

So far, neither of the players could develop their arguments. To change this, assume now that Olga is also thoughtful and skeptical, and that the players reason with each others commitments. Then:

P_1: *claim* safe $\qquad$ O_2: *why* safe

Olga could not concede, nor could she state her argument for $\neg$ safe since it is not preferred over its attacker $(\{\text{safe}\},\text{safe})$. So she had to challenge.

P_3: *claim* $\{\text{airbag, airbag} \supset \text{safe}\}$

Now Olga can create a (trivial) argument for 'airbag' by using Paul's commitments, but she can also create an argument for its negation by using her own beliefs. Neither of these arguments is acceptable, so she must challenge again. Likewise for the second premise, so:

O_4: *why* airbag
P_5: *claim* {airbag} O_6: *why* airbag $\supset$ safe
P_7: *claim* {airbag $\supset$ safe}

Here the nonrepetition rule makes the dialogue terminate without agreement. Note that only Paul could develop his arguments. To give Olga a chance to develop her arguments, let us make her careful and skeptical while the players still reason with each others commitments. Then:

P_1: *claim* safe O_2: *claim* $\neg$ safe

In the new dialogue state Paul's argument for 'safe' is not acceptable any more, since it is not preferred over its attacker ($\{\neg$ safe$\}$, $\neg$ safe). So he must challenge.

P_3: *why* $\neg$ safe O_4: *claim* {newspaper, newspaper $\supset \neg$ safe }

Although Paul can construct an argument for Olga's first premise, namely, ($\{\neg$(newspaper $\supset \neg$ safe'$\}$, safe), it is not acceptable since it is not preferred over its attacker based on Olga's second premise. So he must challenge.

P_5: *why* newspaper O_6: *claim* {newspaper}

Olga had to reply with a (trivial) argument for her first premise, after which Paul cannot repeat his challenge, so here the nonrepetition rule again makes the dialogue terminate without agreement. In this dialogue only Olga could develop her arguments (although she could not state her second counterargument).

In conclusion, the PWA persuasion protocol leaves little room for choice and exploring alternatives. Also, it induces one-sided dialogues in that at most one side can develop their arguments for a certain issue. The above examples also suggest that if a claim is accepted, it is accepted in the first 'round' of moves (but this should be formally verified). On the other hand, the strictness of the protocol induces short dialogues which are guaranteed to terminate, which is good for efficiency reasons. Also, without the requirement to respect the assertion and acceptance attitudes the protocol would be much more liberal while still enforcing some coherence.

7.4.2 Prakken (2005)

In Prakken (2005) a framework for specifying two-party persuasion dialogues about a single dialogue topic is presented, which is then instantiated with some example protocols. The participants have proponent and opponent role, and their beliefs are irrelevant to the protocols. The framework largely abstracts from the communication language, except for an explicit reply structure. It also largely abstracts from the logical language and the logic, except that the logic is assumed to conform to the format of the framework of this reader's Chapter 4 with Dung (1995)'s grounded semantics. The logic is used to verify whether a moved argument is logically constructible, to allow for explicit counterarguments, and to verify whether these arguments defeat their targets.

A main motivation of the framework is to ensure focus of dialogues while yet allowing for freedom to move alternative replies and to postpone replies. This is achieved with two main features of the framework. Firstly, an explicit reply structure on $\mathcal{L}_c$ is assumed, where each move either *attacks* or *surrenders to* its target. An example $\mathcal{L}_c$ of this format is displayed in Table 7.2. This enables the second feature of the framework, namely, an 'any-time' notion of winning that is defined in terms of a notion of *dialogical status* of moves.

Table 7.2: An example L_c in Prakken's framework

Acts	Attacks	Surrenders
claim φ	*why* φ	*concede* φ
argue A	*why* φ ($\varphi \in \texttt{Prem}(A)$) *argue* B (B defeats A)	*concede* φ ($\varphi \in \texttt{Prem}(A)$) *concede* φ ($\varphi = \texttt{Conc}(A)$)
why φ	*argue* A ($\varphi = \texttt{Conc}(A)$)	*retract* φ
concede φ		
retract φ		

Accordingly, particular communication languages must satisfy the following format.

Definition 7.4.5 (Dialogues) The set $\mathcal{L}_c$ of *moves* is defined as $\mathbb{N} \times \{P, O\} \times L_c \times \mathbb{N}$, where the four elements of a move m are denoted by, respectively:

- $id(m)$, the *identifier* of the move,
- $pl(m)$, the *player* of the move,
- $s(m)$, the *speech act* performed in the move,
- $t(m)$, the *target* of the move.

When $t(m) = id(m')$ we say that m replies to m' in d and that m' is the target of m in d. Abusing notation we sometimes let $t(m)$ denote a move instead of just its identifier. When $s(m)$ is an attacking (surrendering) reply to $s(m')$ we also say that m is an attacking (surrendering) reply to m'.

All protocols are further assumed to satisfy the following basic conditions for all moves m_i and all legal finite dialogues d. Note that these protocol rules only state necessary conditions for legality of moves; they can be completed in many ways with further conditions.

If $m \in Pr(d)$, then:

R_1: $t(m) = 0$ iff $m = m_1$.

R_2: If $t(m) \neq 0$ then $t(m) = i$ for some m_i preceding m in d.

R_3: $pl(m) \in T(d)$.[4]

R_4: If $t(m) \neq 0$ then $s(m)$ is a reply to $s(t(m))$ according to L_c.

R_5: If m replies to m', then $pl(m) \neq pl(m')$.

R_6: If there is an m' in d such that $t(m) = t(m')$ then $s(m) \neq s(m')$.

R_7: For any $m' \in d$ that surrenders to $t(m)$, m is not an attacking counterpart of m'.

R_8: If $d = d_0$ then $s(m)$ is of the form *claim* φ or *argue* A.

[4]Recall that $T(d)$ denotes the player(s) whose turn it is to move in d.

R_1 gives the first move a 'dummy' target; together with R_2 it says that all moves except the first reply to some earlier move in the dialogue. Rule R_3 says that the player of a move must be to move according to the turntaking function. R_4 says that a replying move must pick the reply to its target from Table 7.2. R_5 says that a player can only reply to the other player's moves. R_6 makes sure that a new reply to the same target has a different content. Rule R_7 says that once a move is surrendered, it may not be attacked any more (an attacking counterpart of a surrendering move is any attacking move that replies to the same target as the surrendering move). Finally, R_8 says that each dialogue begins with a claim or argument. The claim or conclusion of the argument is the dialogue's topic.

To define the dialogical status of a move first the notion of a surrendered move must be defined. A complication here is that surrendering to a premise of an argument does not yet mean that the argument is surrendered, since if the argument is defeasible; it can still be attacked with a counterargument even if all of its premises are conceded. Therefore, the notion of a surrendered move is defined as follows.

Definition 7.4.6 A move m in a dialogue d is *surrendered* in d iff

- if m is an *argue* A move then it has a *concede* φ reply in d, where $\varphi = \texttt{Conc}(A)$;
- else m has a surrendering reply in d.

The *dialogical status* of a move is now recursively defined as follows, exploiting the reply structure of dialogues.

Definition 7.4.7 [Dialogical status of moves] All attacking moves in a finite dialogue d are either *in* or *out* in d. Such a move m is *in* iff

1. m is surrendered in d; or else
2. all attacking replies to m are *out*

Otherwise m is *out*.

We can now define an 'anytime' outcome function for dialogues (whether or not they are terminated).

Definition 7.4.8 [The current winner of a dialogue]

- The status of the initial move m_1 of a dialogue d is *in favour of* $P(O)$ and *against* $O(P)$ iff m_1 is *in* (*out*) in d. We also say that m_1 favours, or is against p.
- $w_t(d) = p$ (i.e., player p *currently wins* dialogue d on topic t) iff m_1 of d favours p. Furthermore, $l_t(d) = p$ iff $w_t(d) = \overline{p}$.

The framework defined thus far allows for a structural notion of relevance that ensures focus while yet leaving the desired degree of freedom: a move is *relevant* just in case making its target *out* would make the speaker the current winner.

Definition 7.4.9 [Relevance] An attacking move in a dialogue d is *relevant* iff it changes the dialogical status of d's initial move. A surrendering move is relevant iff its attacking counterparts are relevant.

Note that, if not surrendered, an irrelevant target can become relevant again later in a dialogue, viz. if a player returns to a dialogue branch from which s/he has earlier retreated.

To illustrate these definitions, consider Figure 7.2 (where + means *in* and - means *out*). The dialogue tree on the left is the situation after P_7. The tree in the middle shows the dialogical status of the moves when O has continued after P_7 with O_8, replying to P_5: this move does not affect the status of P_1, so O_8 is irrelevant. Finally, the tree on the right shows the situation where O has instead continued after P_7 with O'_8, replying to P_7: then the status of P_1 has changed, so O'_8 is relevant.

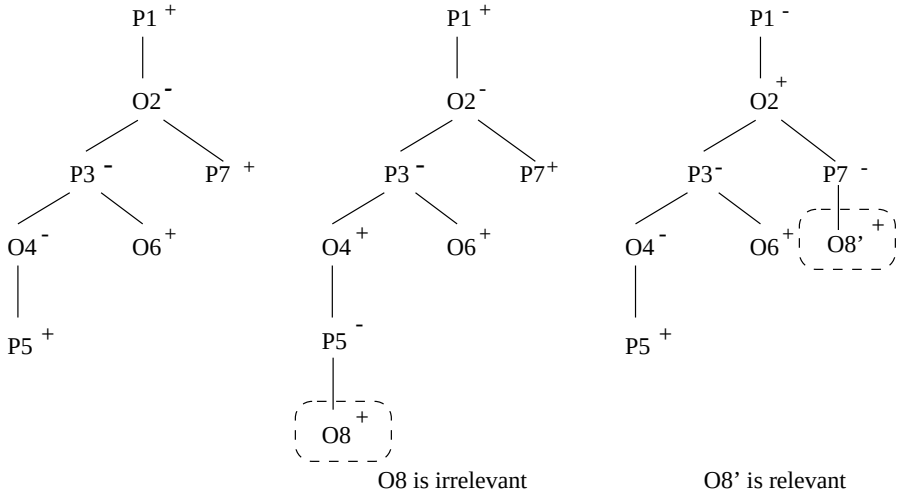


Figure 7.2: Dialogical status and relevance.

As for dialogue structure, the framework allows for all kinds of protocols. The instantiations presented in Prakken (2005) are all multi-move and multi-reply. One of them has the communication language of Table 7.2 and has one additional protocol rule, viz. that each move be relevant, while the turn shifts as soon as the player-to-move has succeeded in becoming the current winner. Protocols with this protocol and turntaking rule are called *protocols for relevant dialogue*. Together, these rules imply that each turn consists of zero or more surrenders followed by one attacker. Within these limits postponement of replies is allowed, sometimes even indefinitely.

We next discuss some examples in terms of a logic within the framework of Chapter 4 combined with grounded semantics. The connective $\rightsquigarrow$ is governed by defeasible modus ponens as in Section 4.4.3 above. We assume that the logic supports arguments about preferences, so that the definition of an overall preference ordering on the basis of the players' individual preferences is in fact the result of the dialogue. The example below should speak for itself so no formal definitions about the logic will be given. Consider two agents with the following belief bases (rule connectives are tagged with a rule name, which is needed to express rule priorities in the object language)

$\Sigma_P = \{q, q \rightsquigarrow_{r_1} p,\ q \wedge s \rightsquigarrow_{r_3} r_1 > r_2\}$
$\Sigma_O = \{r, r \rightsquigarrow_{r_2} \neg p\}$.

Then the following is a legal dialogue:[5]

[5]From now on we will, when the internal structure of the reasoning within an argument does not matter,

- P_1: *claim* p, O_1: *why* p, P_2: p *since* $q, q \rightsquigarrow p$, O_2: *concede* $q \rightsquigarrow p$, O_3: *why* q.

At this point P has four allowed moves, viz. retracting p, retracting q, giving an argument for q or giving a second argument for p. Note that the set of allowed moves is not constrained by P's belief base. If the dialogue terminates here since P withdraws from it then O has won since P_1 is *out*.

The dialogue may also evolve as follows. The first three moves are as above and then:

- O_2: $\neg p$ *since* $r, r \rightsquigarrow \neg p$
  P_3: $r_1 > r_2$ *since* $q, s, q \wedge s \rightsquigarrow r_1 > r_2$

P_3 is a priority argument which in the underlying logic makes P_2 strictly defeat O_2 (note that the fact that s is not in P's own knowledge base does not make the move illegal). At this point, P_1 is *in*; the opponent has various allowed moves, viz. challenging or conceding any premise of P_2 or P_3, moving a counterargument to P_3 or a second counterargument to P_2, conceding one of these two arguments, and conceding P's initial claim.

This example shows that the participants have much more freedom in this system than in the one of Parsons et al. (2003). The downside of this is that dialogues can be much longer, and that the participants can prevent losing by simply continuing to challenge premises of arguments of the other participant. One way to tackle such 'filibustering' is to introduce a third party who may reverse the burden of proof after a challenge: the challenger of φ then has to provide an argument for $\overline{\varphi}$.

Another drawback of Prakken's approach is that not all dialogues that can be found in natural language conform to an explicit reply structure. For instance, in legal cross-examination dialogues the purpose of the cross-examiner is to reveal an inconsistency in the testimony of a witness. Typically, questions by cross-examiners do not indicate from the start what they are aiming at, as in

> *Witness*: Suspect was at home with me that day.
> *Prosecutor*: Are you a student?
> *Witness*: Yes.
> *Prosecutor*: Was that day during summer holiday?
> *Witness*: Yes.
> *Prosecutor*: Aren't all students away during summer holiday?

Paul and Olga (ct'd): Let us finally model our running example in this protocol. Figure 7.3 displays the dialogue tree, where moves within solid boxes are *in* and moves within dotted boxes are *out*. As can be easily checked, this formalisation captures all aspects of our original version, except that arguments have to be complete and that counterarguments cannot be introduced by a counterclaim. (But other instantiations of the framework may be possible without these limitations.)

7.5 Conclusion

In this chapter we have discussed two systems for persuasion dialogue in terms of a formal specification of the main elements of such systems. In the literature a number of

write *argue* A moves as φ *since* S, where φ is A's conclusion and S are A's premises.

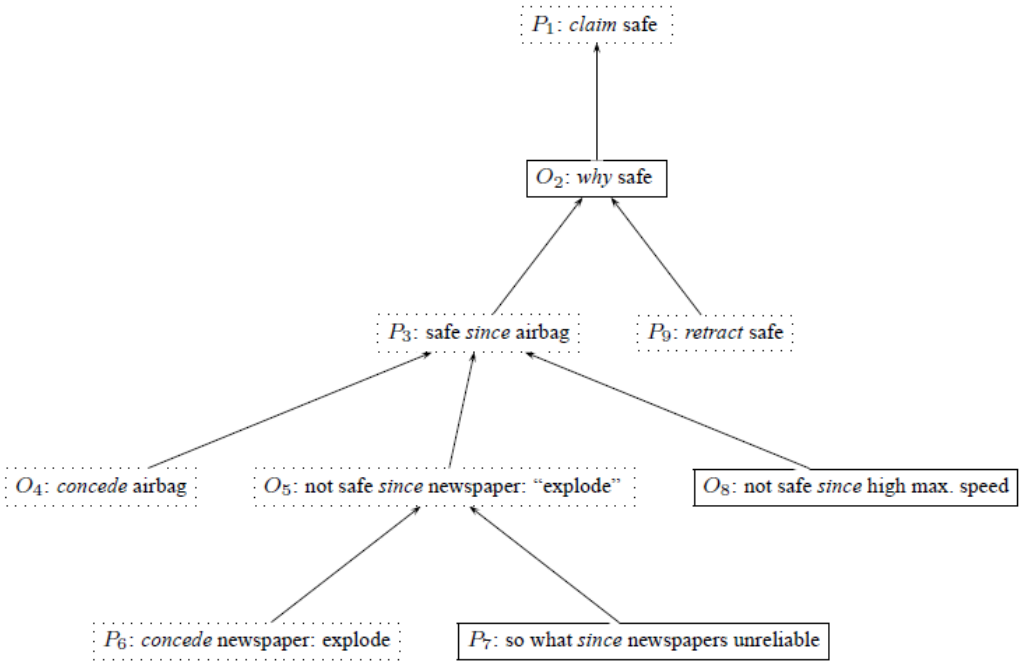


Figure 7.3: The example dialogue in Prakken's approach.

interesting dialogue-game protocols for persuasion have been proposed, some of which have been applied in insightful case studies or applications. However, a consensus on many issues is stil lacking. As a consequence, there is still little work on formally relating the various systems or on a general framework for designing persuasion protocols, and a formal metatheory of systems is still in its early stages. These are some of the main issues that should be tackled in future research. Some other issues are the study of strategies and heuristics for individual participants and how these interact with the protocols to yield certain properties of dialogues, a similar study of varying degrees of cooperativeness of participants, and the integration of persuasion systems with systems for other types of dialogues. Perhaps the main challenge in tackling all these issues is how to reconcile the need for flexibility and expressiveness with the aim to enforce coherent dialogues. The answer to this challenge may well vary with the nature of the context and application domain, and a precise description of the grounds for such variations would provide important insights in how dialogue systems for persuasion can be applied.

7.6 Exercises

7.6.1 On Sections 7.1 – 7.3

EXERCISE 7.6.1 Assume a dialogue protocol for two players P and O within the setup of Definition 7.2.1 with $\mathcal{L}_t$ the language of propositional logic, $\mathcal{L}_c$ the communication language from Table 7.1 and with the following protocol rules.

- P starts the dialogue with a move of the form *claim* φ.
- After the first move, the players take turns after each move.

- Each non-initial move replies to the previous move of the other player with a reply from Table 7.1.
- Each move of the form ϕ *since* S corresponds to a valid argument with premises S and conclusion φ in a classical-logic instantiation of *ASPIC*$^+$ with consistent and only ordinary premises.
- Each move φ' *since* S' that replies to a move φ *since* S is such that $\overline{\varphi} \in S$.

1. Give a terminated dialogue starting with *claim* q and won by P in which at least three different arguments constructible from the knowledge base $\mathcal{K}_p = \{p, p \supset q, r, r \supset \neg p\}$ are moved, where all formulas are of equal preference.
2. Answer the same question but now for a dialogue won by O.

EXERCISE 7.6.2 Assume the same protocol and commitment rules as in Exercise 7.6.1. Suppose both players have adopted an assertion and an acceptance attitude and that they both verify these attitudes relative to their own beliefs plus the other player's commitments. Assume, finally, that all arguments are equally strong. Then let the players have the following knowledge bases: $\mathcal{K}_P = \{p\}$ and $\mathcal{K}_O = \varnothing$.

1. If P starts with *claim* p, how will O reply?
2. Answer the same question if the players verify their attitudes with respect to their own knowledge base only.
3. Answer questions (1) and (2) again for $K_P = \{p \wedge q\}$ and $\mathcal{K}_O = \{\neg q\}$.

EXERCISE 7.6.3 Assume the same protocol and commitment rules as in Exercise 7.6.1. Suppose both players have adopted an assertion and an acceptance attitude and that they both verify these attitudes relative to their own beliefs only. Assume, finally, that all arguments are equally strong. Then let the players have the following knowledge bases: $\mathcal{K}_P = \{p \wedge q\}$ and $\mathcal{K}_O = \{\neg q \wedge r, \neg r\}$, and consider the following dialogue:

$P_1 =$ *claim* p
$O_1 =$ *why* p
$P_2 = p$ *since* $p \wedge q$
$O_2 = \neg(p \wedge q)$ *since* $\neg q \wedge r$
$P_3 =$ *concede* $\neg q \wedge r$

1. Which assertion and acceptance attitudes can P and O have?
2. Can we say anything about whether the protocol is sound?

7.6.2 On Parsons, Wooldridge & Amgoud (2003)

EXERCISE 7.6.4 Let $\Sigma_W = \{q, q \supset p\}$ and $\Sigma_B = \{\neg p, q \supset p\}$. Let the preference ordering $\preceq$ on formulas be:

$\Sigma_1 = \{q\}$
$\Sigma_2 = \{p, \neg p, \neg(q \supset p)\}$
$\Sigma_3 = \{q \supset p\}$

(Recall that this means that all formulas of level 1 are preferred over those over level 2 and 3 and that all formulas at level 2 are preferred over those at level 3.) Finally, assume that both players are thoughtful and skeptical and that these attitudes are verified relative to the speaker's beliefs and the hearer's commitments.

1. What is the dialectical status of p, $\neg p$ and $q \supset p$ on the basis of $\Sigma_W \cup \Sigma_B$ and $\preceq$?
2. Produce all legal dialogues on topic p. Determine the commitment sets of the players at termination. Are these sets consistent? And what is for each player the dialectical status of p and $\neg p$ on the basis of their internal beliefs plus their own commitments?
3. Assume now that the assertion and acceptance attitudes are verified relative to the speaker's beliefs and his own commitments, and answer again the previous question.

EXERCISE 7.6.5 Let $\Sigma_W = \{q, q \supset p\}$ and $\Sigma_B = \{q \supset p\}$ and let all formulas be of the same preference level. Assume that W is thoughtful and cautious while B is careful and skeptical and that both players reason with their own beliefs only.

1. Produce all legal dialogues on topic p.
2. Think of an acceptance attitude that allows a player to learn from the other agent but that avoids the problems as illustrated by Example 7.4.3.

EXERCISE 7.6.6 Let $\Sigma_W = \{p, p \supset q, q \supset r\}$ and $\Sigma_B = \{s, s \supset \neg q\}$. Let all formulas be of the same preference level. Assume that W is thoughtful and cautious while B is careful and skeptical and that both players reason with their own beliefs and with their commitments incurred by concessions. Assume finally that the players also apply the attitude that you defined in your answer to Exercise 7.6.5(2). Produce all legal dialogues on topic r if clause (4b) of the PWA protocol is applied in a depth-first fashion, i.e., if after each response to an element from S the other player may first respond to that response before the first player responds to the next element from S.

EXERCISE 7.6.7 Assume both players are thoughtful and skeptical.

1. Assume that these attitudes are verified relative to the speaker's beliefs and the hearer's commitments. Prove or refute:

   > If W and B agree on preference ordering $\preceq$ and at termination of dialogue d on topic t both $C_W(d) \vdash t$ and $C_B(d) \vdash t$, then t is justified on the basis of $\Sigma_W \cup \Sigma_B$ and $\preceq$.

2. Assume now that the assertion and acceptance attitudes are verified relative to the speaker's beliefs and commitments, and that the players also apply the attitude that you defined in your answer to Exercise 7.6.5(2). Answer the same question.

EXERCISE 7.6.8 Consider two agents W and B with knowledge bases $\Sigma_W = \{p, p \supset q\}$ and $\Sigma_B = \{\neg p\}$. Let all formulas be of equal preference, and assume that the agents are both thoughtful and skeptical and that these attitudes are verified relative to the speaker's knowledge base only. If W starts a dialogue with *claim* q, will W and B reach agreement on q?

7.6.3 On Prakken (2005)

EXERCISE 7.6.9 Prove that for each finite dialogue d there is a unique dialogical status assignment. Give a counterexample for infinite dialogues. (Hint: use results stated in Chapter 2.)

EXERCISE 7.6.10 Explain that a reply to a surrendered move is never relevant.

EXERCISE 7.6.11 Answer the following questions about Figure 7.3.

1. What are the relevant targets for O after P_7?
2. What are the relevant targets for P after O_8?
3. Assume at P_9 that P does not retract *safe* but instead moves another argument for *safe* in reply to O_2. What are then the relevant targets for O after P_9?

EXERCISE 7.6.12 Assume an instance of the dialogue framework of Prakken (2005) with the same argumentation logic as the dialogue system of Parsons, Wooldridge & Amgoud, with the communication language of Table 7.2, and with a protocol for relevant dialogue. Give a terminated dialogue starting with *claim* q and won by O in which at least three different arguments constructible from the knowledge base $\Sigma = \{p, p \supset q, r, r \supset \neg p\}$ are moved, where all formulas are of equal preference.

EXERCISE 7.6.13 Assume an instance of the dialogue framework of Prakken (2005) with the same argumentation logic as the dialogue system of Parsons, Wooldridge & Amgoud, with the communication language of Table 7.2, and with a protocol for relevant dialogue. Give a terminated dialogue starting with *claim* q and won by O in which at least three different arguments constructible from the knowledge base $\mathcal{K} = \mathcal{K}_n \cup \mathcal{K}_p$ are moved, where $\mathcal{K}_n = \varnothing$ and $\mathcal{K}_p = \{p \wedge q, \neg p\}$ and where $p \wedge q <' \neg p$.

Chapter 8

Legal argumentation with cases

8.1 Introduction

In this chapter several legal applications of argumentation formalisms will be discussed. These applications partly use formalisms introduced in earlier chapters, such as abstract argumentation frameworks and *ASPIC*$^+$, and partly use formalisms especially designed for legal argumentation with cases. The focus will especially be on formal models of legal case-based argumentation.

The law is both a rich test bed and and important application field for AI research. As a test bed, the law provides real, documented examples instead of artificial toy examples, and as an application field it may result in AI applications from which society as a whole, not just industry or consumers, can benefit. At first sight, one might think that such testing and application boils down to the use of techniques for knowledge representation and automated deduction. Once a legal text and a body of facts have been clearly represented in a formal language, the legal conclusions would follow from that representation as a matter of deduction. However, this view is too simplistic. For one thing it ignores that law is not just a conceptual or axiomatic system but has social objectives and social effects, which may require that a legal rule is overridden or changed. Moreover, legislators can never fully predict in which circumstances the law has to be applied, so legislation has to be formulated in general and abstract terms, such as 'duty of care', 'misuse of trade secrets' or 'intent', and qualified with general exception categories, such as 'selfdefence', 'force majeure' or 'unreasonable'. Such concepts and exceptions must be interpreted in concrete cases, which creates uncertainty and room for disagreement. This is reinforced by the fact that legal cases often involve conflicting interests of opposing parties. The prosecution in a criminal case wants the accused convicted while the accused wants to be acquitted. The plaintiff in a civil law suit wants to be awarded compensation for damages, while the defendant wants to avoid having to pay. The tax authority in a tax case wants to receive as much tax as possible, while the tax payer wants to pay as little as possible. All these aspects of the law, i.e., its orientation to future and not fully anticipated situations, the tension between the general terms of the law and the particulars of a case, and the adversarial nature of legal procedures, make that legal reasoning goes beyond the literal meaning of the legal rules and involves appeals to precedent, principle, policy and purpose, and involves the *attack* as well as the *construction* of arguments. A central notion then in the law is that of argumentation. Indeed, the formal and computational study of argumentation is an area of AI where AI-and-law researchers have not just *applied* AI techniques but where

they have also contributed significantly to their development.

A legal case has various aspects, each with their own modes of reasoning: determining the facts, classifying the facts under legal concepts or conditions, and deriving legal consequences from the thus classified facts. When determining the facts, the modes of reasoning are often probabilistic and may involve reasoning about causation and about mental attitudes such as intent. Classifying the facts under legal concepts involves interpretation of these legal concepts. Here the prevailing modes of reasoning are analogy, appeals to precedent or policy, and the balancing of interests. Finally, when deriving legal consequences from the classified facts, the main modes of reasoning are deductive but with room for nonmonotonic techniques to deal with exceptions to rules, either statutory or based on principle and purpose, and to choose between conflicting rules on the basis of the general hierarchy of legal systems, with rules from different sources.

8.2 Legal case-based argumentation for classification and interpretation

The concepts in the conditions of legal rules are often interpreted by referring to past cases in which these rules were applied. Case-based reasoning forms can be found in their most explicit form in common-law jurisdictions, in which judicial precedents can be legally binding beyond the decided case, so that court decisions legally constrain the decision in new cases. This leads to reasoning forms where interpretation rules are formulated by courts in the context of particular cases and are constantly refined and modified to fit new circumstances that were not taken into account in earlier decisions. These reasoning forms can to a lesser extent also be found in civil-law jurisdictions, since interpretations of the law by higher civil-law courts, even though strictly speaking not binding beyond the decided case, still tend to be followed by lower courts.

Much AI & law work on the interpretation of legal concepts centers around the notion of a factor, an idea going back to the HYPO system of Ashley (1990) and the CATO system of Aleven (2003). Factors are abstractions of fact patterns that favour (pro factors) or oppose (con factors) a conclusion. Factors are thus in an intermediate position between the specific facts of a case and the legal predicates to which such facts may be relevant. For example, in CATO, which like HYPO argues about misuse of trade secrets, some factors pro misuse are that a non-disclosure agreement was signed, that the plaintiff had made efforts to maintain secrecy and that the copied product was unique; and some factors con misuse are that disclosures were made by the plaintiff in negotiations and that the information was reverse-engineerable.

The HYPO and CATO systems are meant to model how lawyers make use of past decisions when arguing a case. They do not compute an 'outcome' or 'winner' of a dispute; instead they are meant to generate debates as they could take place between 'good' lawyers. HYPO generates disputes between a plaintiff and a defendant of a legal claim concerning misuse of a trade secret. Each move conforms to certain rules for analogizing and distinguishing precedents. These rules determine for each side which are the best cases to cite initially, or in response to the counterparty's move, and how the counterparty's cases can be distinguished. A case is represented as a set of factors for a decision and a set of factors against that decision, plus the decision that resolves the conflict between the competing factors. A case is citable for a side if it has the decision wished by that side and shares with the Current Fact Situation (CFS) at least

one factor which favors that decision. Thus citable cases do not have to exactly match the CFS, which is a way of coping with the case-specific nature of case law decisions. A citation can be countered by a counterexample, that is, by producing a citable case that has the opposite outcome. A citation may also be countered by distinguishing, that is, by indicating a factor in the CFS that is absent in the cited precedent and that supports the opposite outcome, or a factor in the precedent that is missing in the CFS and that supports the outcome of the cited case. HYPO also allows for multi-valued factors, called dimensions, to vary the degree to which a factor promotes a certain outcome. For example, a dimension is the number of people to which a trade secret has been disclosed, or the extent to which security measures were taken. A boolean factor is then a specific value of a dimension. Dimensions allow an additional way to distinguish a precedent, namely, on a shared pro-decision factor that more strongly favours the decision in the precedent than in the CFS.

CATO added to this a 'factor hierarchy', which expresses expert knowledge about the relations between the various factors: more concrete factors are classified according to whether they are a reason for or against the more abstract factors to which they are linked; links are given a strength (weak or strong), which can be used to solve certain conflicts. Thus the factor hierarchy can be used to explain why a certain decision was taken, which in turn facilitates debate on the relevance of differences between cases.

For instance, the hierarchy positively links the factor *Security measures taken* to the more abstract concept *Efforts to maintain secrecy*. Now if a precedent contains the first factor but the CFS lacks it, then not only can a citation of the precedent be distinguished on the absence of *Security measures taken*, but also this distinction can be emphasized by saying that thus no efforts were made to maintain secrecy. However, if the CFS also contains a factor *Agreed not to disclose information*, then the factor hierarchy enables downplaying this distinction, since it also positively links this factor to *Efforts to maintain secrecy*: the party that cited the precedent can say that in the current case, just as in the precedent, efforts were made to maintain secrecy. The factor hierarchy is not meant to be an independent source of information from which arguments can be constructed. Rather it serves as a means to *reinterpret* precedents: initially cases are in CATO, as in HYPO, still represented as one-step decisions; the factor hierarchy can only be used to argue that the decision was in fact reached by one or more intermediate steps.

While HYPO- and CATO-style work mainly focuses on rhetoric (modelling persuasive debates), other work addresses the logical question how precedents constrain decisions in new cases. An important idea here is that precedents are sources of preferences between factor sets and that these preferences are often justified by balancing underlying legal or societal values. This work has recently been extended to dimensions.

8.2.1 Factor-based models: basic notation and concepts

Consider a 'case base' (CB) with a large number of decided cases. A new case arises, which may be unlike any case in CB. How can the CB still be used in the new case?

Suppose that the new case and a given precedent share at least some similarities favouring the precedent's outcome. Then the following situations can arise.

- The new case and the precedent have exacty the same set of factors: $\Rightarrow$ unique answer (assuming 'consistency' of the CB).

- All differences between the new case and the precedent favour the precedent's outcome: $\Rightarrow$ unique answer (assuming 'consistency' of the CB) = a fortiori reasoning.
- Some differences between the new case and the precedent favour the opposite outcome than in the precedent. $\Rightarrow$ analogy, plus debate about whether the similarities or the differences should be decisive.

For modelling this kind of reasoning, the following is needed:

- knowledge about what are the relevant factors and whose side they favour;
- a language for representing cases;
- ways to cite precedents in support of a decision in a new case;
- ways to cite differences between a precedent and a new case.

The notation for the fist two of these elements is as follows. Let o and o' be two outcomes and Pro and Con be two disjoint sets of atomic propositions called, respectively, the *pro-* and *con* factors, i.e., the factors favouring, respectively, outcome o and o'. The variable s (for 'side') ranges over $\{o, o'\}$ and $\overline{s}$ denotes o' if $s = o$ while it denotes o if $s = o'$. We say that a set $F \subseteq Pro \cup Con$ favours side s (or F is pro s) if $s = o$ and $F \subseteq Pro$ or $s = o'$ and $F \subseteq Con$. For any set F of factors the set $F^s \subseteq F$ consists of all factors in F that favour side s. A *fact situation* is any subset of $Pro \cup Con$.

A case can then be represented as a triple $(pro(c), con(c), outcome(c))$ where $outcome(c') \in \{o, o'\}$. Moreover, if $outcome(c) = o$ then $pro(c) \subseteq Pro$ and $con(c) \subseteq Con$ and if $outcome(c) = o'$ then $pro(c) \subseteq Con$ and $con(c) \subseteq Pro$. Given all this, a case base CB is a set of cases. Elements of CB are called *precedents*.

8.2.2 Formalising persuasive debates with cases

HYPO generates dialogues between a plaintiff and a defendant, in which all moves cite or distinguish a case, assuming that o is that the plaintiff wins (π), so there was a misuse of a trade secret, and o' is that the defendant wins (δ), so there was no misuse of a trade secret. Below the variable p ranges over $\{\pi, \delta\}$.

Definition 8.2.1 [Citable cases, counterexamples, distinguishing] Given a case base CB and a fact situation F:

1. a case $c \in CB$ is *citable* by player p iff $outcome(c) = p$ and $pro(c) \cap F \neq \varnothing$;
2. a case c' is a *counterexample* to a case $c \in CB$ iff c and c' have opposite outcomes;
3. a case $c \in CB$ is *distinguishable* on factor f iff $f \in pro(c) \setminus pro(f)$ or $f \in con(f) \setminus con(c)$.

Then HYPO employs the following simple debate protocol.

Given a fact situation F:

1. The plaintiff starts with citing a citable precedent with outcome π;

2. The defendant cites all citable counterexamples to and distinguishes the precedent cited by plaintiff in all possible ways;

3. The plaintiff distinguishes each counterexample cited by defendant, after which the debate terminates.

HYPO has limited ways to compare precedents and define outcomes of a terminated-debate. First, given a fact situation, a precedent c_1 is more on point than a precedent c_2 if all factors that c_2 shares with the new case also occur in c_1 and c_1 shares additional factors with the new case.

Definition 8.2.2 [On-pointness] Let F be a fact situation and c_1 and c_2 two precedents. Then c_1 is *more on point* than c_2 given F ($c_1 >_F c_2$) iff $pro(c_2) \cup con(c_2) \cap F \subset pro(c_1) \cup con(c_1) \cap F$.

The debate protocol could be refined by requiring that the participants cite the most on point cases that are citable for them. Other ways to compare cases are possible and will be defined later on.

Given the definition of on-pointness, HYPO has two limited definitions of an outcome of a debate. A *relative outcome* is if the most on point precedent citable for side 1 is more on point than the most on point precedent citable for side 2: then side 1 has the 'better case' (but the court could still rule for the other side). An *absolute outcome* is that if one of the sides can make an 'a fortiori' argument, this side wins (assuming that the CB is 'consistent'. This is since in that case the other side can neither distinguish nor cite a counterexample. This notion is further explored in theories of precedential constraint, to be discussed below.

CATO includes HYPO but CATO adds background knowledge about why factors are pro or con a decision (the factor hierarchy). This yields two additional debate moves for emphasising and downplaying a distinction between cases.

The factor hierarchy is a tree of factors with

- the root is the ultimate decision *misuse of trade secrets*;
- a link between two nodes is in the direction of the root and is labelled either pro or con, where if the link is pro, the two factors favour the same side and if the link is con, the two factors favour different sides;
- the leaves are the original HYPO factors.

Figure 8.1, taken from publications on CATO, displays a snapshot of the CATO factor hierarchy plus a part of the CATO case base.

Definition 8.2.3 [Emphasising and downplaying a distinction] Consider a case base CB and a fact situation F.

1. Suppose precedent c_1 with decision p is distinguishable on pro p factor f, i.e., $f \in pro(c)$ but $f \notin F$.

   (a) Then the distinction can be *emphasised* by citing a more abstract factor g in the factor hierarchy with a path from f with only pro links (arguing that thus g is also lacking in F).

Table 1: Cases Used, Their Factors, Values and Outcomes

	Pro-P factors	Pro-D factors	Pro-P Values	Pro-D Values	Outcome
Group 1					
Arco		10 16 20		RE LM	D
Boeing	4 6 12 14 21	1 10	CA RE QM	RE	P
Bryce	4 6 18 21	1	CA RE MW	RE	P
CollegeWat	15 26	1	LM QM	RE	P
Den-Tal-Ez	4 6 21 26	1	CA RE QM	RE	P
Ecologix	21	1 19 23	CA	CA RE	D
Emery	18 21	10	CA MW	RE	P
Ferranti	2	17 19 20 27	QM	RE LM	D
Robinson	18 26	1 10 19	QM MW	RE	D
Sandlin		1 10 16 19 27		RE LM	D
Sheets	18	19 27	MW	RE	D
Space Aero	8 15 18	1 19	MW LM	RE	P
Televation	6 12 15 18 21	10 16	CA RE LM MW	RE LM	P
Yokana	7	10 16 27	QM	RE LM	D
Group 2					
FMC	4 6 7 12	10 11	CA RE QM	RE LM	P
KG	6 14 15 18 21	16 25	RE QM LM MW CA	LM	P
Mason	6 15 21	1 16	CA RE LM	RE LM	P
MineralDep	18	1 16 25	MW	RE LM	P
NationalRej	7 15 18	10 16 19 27	QM LM MW	RE LM	D
Reinforced	4 6 8 15 21	1	CA RE MW LM	RE	P
Scientology	4 6 12	10 11 20	CA RE	RE LM	D
Technicon	6 12 14 21	10 16 25	RE QM CA	RE LM	P
Trandes	4 6 12	1 10	CA RE	RE	P

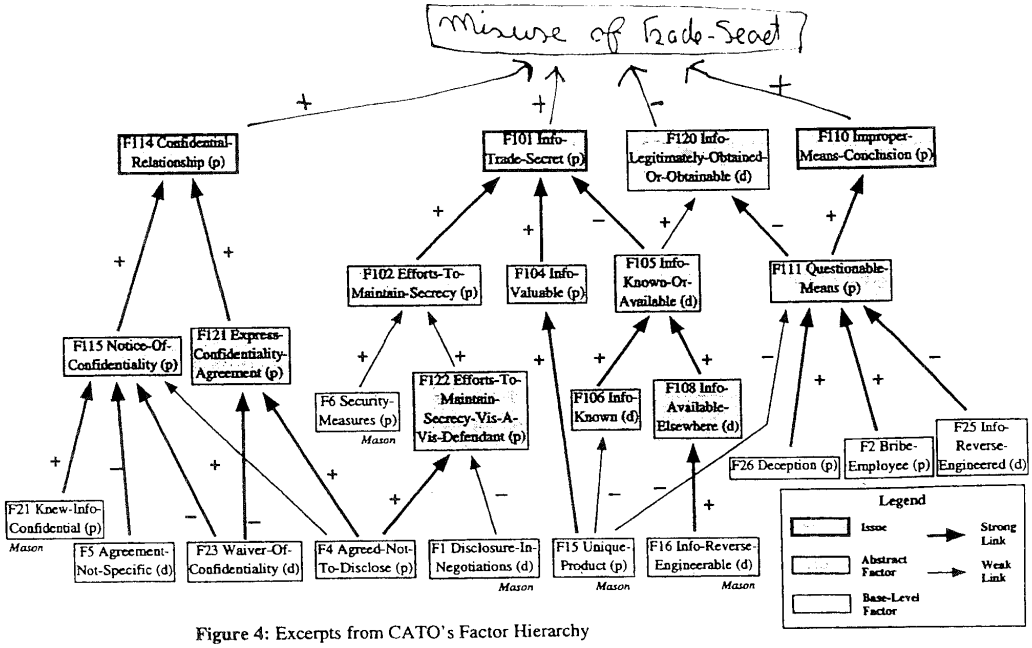


Figure 4: Excerpts from CATO's Factor Hierarchy

(b) And the distinction can be *downplayed* by pointing at a pro p factor f' in F with a path to pro p factor g with only pro links (thus arguing that g is still present in F). In this case we say that f' *substitutes* f.

2. Suppose c_1 with decision p is distinguishable on con p factor f, i.e., $f \in F$ but $f \notin con(c)$.

   (a) Then the distinction can be *emphasised* by citing a more abstract pro p factor g in the factor hierarchy with a con link from f to g (arguing that thus the more abstract pro p factor g is also lacking in F).

   (b) And the distinction can be *downplayed* by pointing at a pro p factor f' in F with a path to g with only pro links (thus arguing that g is still present in F). In this case we say that f' *cancels* f.

Consider the factor hierarchy in Figure 8.1. If the plaintiff cites a precedent containing F4 but the new case lacks F4, then the defendant can not only distinguish the citation of F4 but also emphasise the distinction by saying that thus the more abstract factors F122, F102 and F101 are also missing in the new case. If the new case contains the new factor F6, then the plaintiff can downplay this distinction by saying that F6 substitutes F4 so that the new case still contains F102 and F101.

Consider next a citation by the plaintiff where the new case contains F4 while the cited precedent lacks it but where the defendant distinguishes on the new pro-defendant factor F1. The defendant can emphasise the distinction by saying that thus F122, F102 and F101 are missing in the new case. Then the plaintiff can downplay by saying that F4 cancels the negative effect of F1 so that the new case still contains these more abstract factors.

8.2.3 Factor-based precedential constraint

We next summarise a factor-based model of precedential constraint originally introduced by Horty (2011).

Definition 8.2.4 [Preference relation on fact situations.] Let X and Y be two fact situations. Then $X \leq_s Y$ iff $X^s \subseteq Y^s$ and $Y^{\overline{s}} \subseteq X^{\overline{s}}$.

$X <_s Y$ is defined as usual as $X \leq Y$ and $Y \not\leq X$. This definition says that Y is at least as good for s as X iff Y contains at least all pro-s factors that X contains and Y contains no pro-$\overline{s}$ factors that are not in X.

Definition 8.2.5 [Precedential constraint with factors: result model.] Let CS be a case base and F a fact situation. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (X, Y, s)$ in CB such that $X \cup Y \leq_s F$. Moreover, deciding F for s is *allowed* iff deciding F for $\overline{s}$ is not forced.

This definition models *a fortiori reasoning* in that an outcome in a focus case is forced if a precedent with the same outcome exists such that all their differences make the focus case even stronger for their outcome than the precedent. Note that it can happen that both deciding F for s and for $\overline{s}$ is forced. This indicates that the case base is inconsistent. Formally, in the result model, a case-base is *inconsistent* if and only if for some fact situation F it holds that both deciding F for s and deciding F for $\overline{s}$ is forced.

As our running example we use a small part of the US trade secrets domain of the HYPO and CATO systems. We assume the following six factors along with whether they favour the outcome 'misuse of trade secrets' (π for 'plaintiff') or 'no misuse of trade secrets' (δ for 'defendant'): the defendant had obtained the secret by deceiving the plaintiff (π_1) or by bribing an employee of the plaintiff (π_2), the plaintiff had taken security measures to keep the secret (π_3, the product is not unique (δ_1), the product is reverse-engineerable (δ_2) and the plaintiff had voluntarily disclosed the secret to outsiders (δ_3). We assume the following precedents:

$c_1(\pi)$: *deceived*$_{\pi 1}$, *measures*$_{\pi 3}$, *not-unique*$_{\delta 1}$, *disclosed*$_{\delta 3}$
$c_2(\delta)$: *bribed*$_{\pi 2}$, *not-unique*$_{\delta 1}$, *disclosed*$_{\delta 3}$

Clearly, deciding a fact situation F for π is forced iff it has at least the π-factors $\{\pi_1, \pi_3\}$ and at most the δ-factors $\{\delta_1, \delta_3\}$ (by precedent c_1), since then we have $\{\pi_1, \pi_3\} \subseteq F^\pi$ and $F^\delta \subseteq \{\delta_1, \delta_3\}$. Likewise, deciding a fact situation for δ is forced iff it has at least the δ-factors $\{\delta_1, \delta_3\}$ and at most the π-factor $\{\pi_2\}$ (by precedent c_2).

Consider next the following fact situation:

F_1: *bribed*$_{\pi 2}$, *measures*$_{\pi 3}$, *reverse-eng*$_{\delta 2}$, *disclosed*$_{\delta 3}$

Comparing F_1 with c_1 we must check whether $\{\pi_1, \pi_3, \delta_1, \delta_3\} \leq_\pi \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is not the case, for two reasons. We have $\{\pi_1, \pi_3\} \not\subseteq F_1^\pi = \{\pi_2, \pi_3\}$ and we have $F_1^\delta = \{\delta_2, \delta_3\} \not\subseteq \{\delta_1, \delta_3\}$. Next, comparing with precedent c_2 we must check whether $\{\pi_2, \delta_1, \delta_3\} \leq_\delta \{\pi_2, \pi_3, \delta_2, \delta_3\}$. This is also not the case for two reasons. We have $\{\delta_1, \delta_3\} \not\subseteq F_1^\delta = \{\delta_2, \delta_3\}$ and we have $F_1^\pi = \{\pi_2, \pi_3\} \not\subseteq \{\pi_2\}$. So neither deciding F_1 for π nor deciding F_1 for δ is forced. Henceforth we will assume it was decided for π.

An alternative charactisation of precedential constraint is possible. The following definition says that a case decision expresses a preference for any pro-decision set containing at least the pro-decision factors of the case over any con-decision set containing at most the con-decision factors of the case. This allows *a fortiori* reasoning from a precedent adding pro-decision factors an/or deleting con-decision factors.

Definition 8.2.6 [Preferences from cases.] Let $(pro(c), con(c), s)$ be a case, CB a case base and X and Y sets favouring s and $\overline{s}$, respectively. Then

1. $Y <_c X$ iff $Y \subseteq con(c)$ and $X \supseteq pro(c)$;
2. $Y <_{CB} X$ iff $Y <_c X$ for some $c \in CB$.

In the result model, a case-base was said to be inconsistent if and only some fact situation both decisions are forced. In the reason model an alternative but equivalent definition is possible.

Definition 8.2.7 [(In)consistent case bases, reason model.] Let C be a case base with $<_{CB}$ the derived preference relation. Then CB is *inconsistent* if and only if there are factor sets X and Y such that $X <_{CB} Y$ and $Y <_{CB} X$. And CB is *consistent* if and only if it is not inconsistent.

The final definition says that deciding a case for a particular outcome is forced if that is the only way to keep the updated case base consistent.

Definition 8.2.8 [Precedential constraint with factors: reason model.] Let CB be a case base and $F = F^s \cup F^{\overline{s}}$ a fact situation. Then, given CB, deciding F for s is *allowed* iff $CB \cup \{(F^s, F^{\overline{s}}, s)\}$ is consistent. Moreover, deciding F for s is *forced* iff $CB \cup \{(F^s, F^{\overline{s}}, \overline{s})\}$ is inconsistent.

It can be shown that (for consistent case bases) Definitions 8.2.5 and 8.2.8 are (for forced decisions) equivalent.

In terms of these models of precedential constraint, formal definitions can be given of three concepts that are important in the common-law theory of precedent. Given a case base containing precedent c and a new case f such that $pro(c) \subseteq pro(f)$, a decision of f *follows* precedent c if it is the same decision as in c, it *distinguishes* precedent c if it is the opposite decision as in c and following the precedent c is allowed but not forced; and the decision in the new case *overrules* precedent c if if it is the opposite decision as in c and following the precedent c is forced. Note that following a precedent can change the law in that the next time the same fact situation arises, the decision is forced.

We next define a similarity relation on a case base given a focus case and prove a correspondence with the above factor-based model of precedential constraint. The similarity relation is defined in terms of the relevant differences between a precedent and the focus case. These differences are the situations in which a precedent can be distinguished in a HYPO/CATO-style approach with factors, namely, when the new case lacks some factors pro its outcome that are in the precedent or has new factors con its outcome that are not in the precedent.[1]

Definition 8.2.9 [Differences between cases with factors.] Let c be a case with fact situation C and outcome s and f a case with fact situation F. The set $D(c, f)$ of differences between c and f is defined as $C^s \setminus F^s \cup F^{\overline{s}} \setminus C^{\overline{s}}$.

Consider again our running example and consider first any focus case f with a fact situation that has at least the π-factors $\{\pi_1, \pi_3\}$ and at most the δ-factors $\{\delta_1, \delta_3\}$. Then $D(c, f) = \varnothing$. Likewise with any focus case f with a fact situation that has at least the δ-factors $\{\delta_1, \delta_3\}$ and at most the π-factor $\{\pi_2\}$. Next, let f be a focus case with fact situation F_1 and outcome π. We have

$D(c_1, f) = \{\textit{deceived}_{\pi 1}, \textit{reverse-eng}_{\delta 2}\}$
$D(c_2, f) = \{\textit{measures}_{\pi 3}, \textit{not-unique}_{\delta 1}\}$

The following result, which yields a simple syntactic criterion for determining whether a decision is forced, is proven by Prakken (2021a).

Proposition 8.2.10 Let CB be a case base CB and f a focus case with fact situation F. Then deciding F for s is forced given CB iff there exists a case c with outcome s in CB such that $D(c, f) = \varnothing$.

Clearly every case c such that $D(c, f) = \varnothing$ is citable for its outcome. Another result is that for any two cases with opposite outcomes that both have differences with the focus case, their sets of differences with the focus case are mutually incomparable (as with c_1 and c_2 in our running example).

Proposition 8.2.11 Let CB be a case base, f a focus case and c and c' two cases with opposite outcomes and with non-empty sets of differences with f. Then $D(c, f) \not\subseteq D(c', f)$ and $D(c', f) \not\subseteq D(c, f)$.

[1] The definition below is a simplification of but equivalent to a definition in Prakken (2021a) and is due to Wijnand van Woerkom (personal communication).

8.2.4 Dimension-based precedential constraint

In this section the above approach is adapted to dimensions, following Horty (2019). Formally, a *dimension* is a tuple $d = (V, \leq_o, \leq_{o'})$ where V is a set (of values) and $\leq_o$ and $\leq_{o'}$ two partial orders on V such that $v \leq_o v'$ iff $v' \leq_{o'} v$. Given a dimension d, a *value assignment* is a pair (d, v), where $v \in V$. The functional notation $v(d)$ denotes the value of dimension d. Then given a set D of dimensions, a *fact situation* is an assignment of values to all dimensions in D, and a *case* is a pair $c = (F, outcome(c))$ such that F is a fact situation and $outcome(c) \in \{o, o'\}$. Then a case base is as before a set of cases, but now explicitly assumed to be relative to a set D of dimensions in that all cases assign values to a dimension d iff $d \in D$. As for notation, $F(c)$ denotes the fact situation of case c and $v(d, c)$ denotes the value of dimension d in case c. Finally, $v \geq_s v'$ is the same as $v' \leq_s v$.

Note that the set of value assignments of a case is unlike the set of factors of a case not partitioned into two subsets pro and con the case's outcome. The reason is that unlike with factors, with value assignments it is often hard to say in advance whether they are pro or con the case's outcome. All that can often be said in advance is which side is favoured more and which side less if a value of a dimension changes, as captured by the two partial orders $\leq_s$ and $\leq_{s'}$ on a dimension's values.

In HYPO two of the factors from our running example are actually dimensions. *Security-Measures-Adopted* has a linearly ordered range, below listed in simplified form (where later items increasingly favour the plaintiff so decreasingly favour the defendant):

- *Minimal-Measures*, *Access-To-Premises-Controlled*, *Entry-By-Visitors-Restricted*, *Restrictions-On-Entry-By-Employees*

(For simplicity we will below assume that each case contains exactly one security measure; generalisation to multiple measures is straightforward by defining the orderings on sets of measures.). Moreover, *disclosed* has a range from 1 to some high number, where higher numbers increasingly favour the defendant so decreasingly favour the plaintiff. For the remaining four factors we assume that they have two values 0 and 1, where presence (absence) of a factor means that its value is 1 (0) and where for the pro-plaintiff factors we have $0 <_\pi 1$ (so $1 <_\delta 0$) and for the pro-defendant factors we have $0 <_\delta 1$ (so $1 <_\pi 0$).

Accordingly, we change our running example as follows, where if a factor that is now a two-valued dimension is not mentioned, its value equals 0.

$c_1(\pi)$: *deceived*$_{\pi 1}$, *measures* = *Entry-By-Visitors-Restricted*,
not-unique$_{\delta 1}$, *disclosed* = 10
$c_2(\delta)$: *bribed*$_{\pi 2}$, *measures* = *Minimal*,
not-unique$_{\delta 1}$, *disclosed* = 5
F_1: *bribed*$_{\pi 2}$, *measures* = *Access-To-Premises-Controlled*,
reverse-eng$_{\delta 2}$, *disclosed* = 20

In the dimension-based result model of precedential constraint a decision in a fact situation is forced iff there exists a precedent c for that decision such that on each dimension the fact situation is at least as favourable for that decision as the precedent. This idea is formalised with the help of the following preference relation between sets of value assignments.

Definition 8.2.12 [Preference relation on dimensional fact situations.] Let F and F' be two fact situations with the same set of dimensions. Then $F \leq_s F'$ iff for all $(d, v) \in F$ and all $(d, v') \in F'$ it holds that $v(d) \leq_s v'(d)$.

In our running example we have for any fact situation F' that $F(c_1) \leq_\pi F'$ iff F' has π_1 but not δ_3 and $v(F'$,*measures*$) \geq_\pi$ *Entry-By-Visitors-Restricted* and $v(F'$,*disclosed*$) \geq_\pi$ 20 (so ≤ 20). Likewise, $F(c_2) \leq_\delta F'$ iff F' has δ_1 but not π_1 and $v(F'$,*measures*$) =$ *Minimal* and $v(F'$,*disclosed*$) \geq_\delta 10$ (so ≥ 10).

Then adapting Definition 8.2.5 to dimensions is straightforward.

Definition 8.2.13 [Precedential constraint with dimensions.] Let CS be a case base and F a fact situation given a set D of dimensions. Then, given CB, deciding F for s is *forced* iff there exists a case $c = (F', s)$ in CB such that $F' \leq_s F$.

In our running example, deciding F_1 for π is not forced, for two reasons. First, $v(c_1, \mathit{deceived}) = 1$ while $v(F_1, \mathit{deceived}) = 0$ and for *deceived* we have that $0 <_\pi 1$. Second, $v(c_1$,*measures*$) =$ *Entry-By-Visitors-Restricted* while $v(F_1$,*measures*$) =$ *Access-To-Premises-Controlled* and *Access-To-Premises-Controlled* $<_\pi$ *Entry-By-Visitors-Restricted*. Deciding F_1 for δ is also not forced, since $v(c_2$,*measures*$) =$ *Minimal* while $v(F_1$,*measures*$) =$ *Access-To-Premises-Controlled* and *Minimal* $<_\delta$ *Access-To-Premises-Controlled*.

We next adapt Definition 8.2.9 to dimensions. Unlike with factors, there is no need to indicate whether a value assignment favours a particular side, since we have the $\leq_s$ orderings.

Definition 8.2.14 [Differences between cases with dimensions.] Let $c = (F(c), \mathit{outcome}(c))$ and $f = (F(f), \mathit{outcome}(f))$ be two cases. The set $D(c, f)$ of differences between c and f is defined as $D(c, f) = \{(d, v) \in F(c) \mid v(d, c) \not\leq_s v(d, f)$.

Let c be a precedent and f a focus case. Then this definition says that any value assignment in the precedent such that the value for the same dimension in the focus case is not at least as favourable for the outcome as in the precedent is a relevant difference. In our running example, we have:

$D(c_1, F_1) = \{(\mathit{deceived}, 1)$, (*reverse-eng*,0), $(\mathit{measures}$, *Entry-By-Visitors-Restricted*$)$, $(\mathit{disclosed}, 10)\}$
$D(c_2, F_1) = \{$(*not-unique*,1), $(\mathit{measures}, \mathit{Minimal})\}$

The following counterpart of Proposition 8.2.10 is proven by Prakken (2021a).

Proposition 8.2.15 Let, given a set D of dimensions, CB be a case base and f a focus case with fact situation F. Then deciding F for s is forced given CB iff there exists a case in CB with outcome s such that $D(c, f) = \varnothing$.

8.2.5 Reasoning about factors: purpose and value

The work on precedential constraint does not explain why factors are factors pro or con a decision. This question is addressed in a body of work initiated by Bench-Capon Bench-Capon (2002), who was inspired by Berman & Hafner Berman and Hafner (1993), who argued that often a factor can be said to favour a decision by virtue of the purposes served or values promoted by taking that decision because of the factor.

A choice in case of conflicting factors is then explained in terms of a preference ordering on the purposes, or values, promoted or demoted by the decisions suggested by the factors. Cases can then be compared in terms of the values at stake rather than on the factors they contain.

The role of purpose and value is often illustrated with some well-known cases from Anglo-American property law on ownership of wild animals that are being chased. In *Pierson* plaintiff was hunting foxes for sport on open land when defendant shot the chased fox and carried it away. The court held for defendant. In *Keeble* a pond owner placed a duck decoy in his pond with the intention to sell the caught ducks for a living. Defendant used a gun to scare away the ducks, for no other reason than to damage plaintiff's business. Here the court held for plaintiff. Finally, in *Young* both plaintiff and defendant were fishermen fishing in the open sea. Just before plaintiff closed his net, defendant came in and caught the fishes with his own net.

Let us assume that the task is to argue for a decision in *Young* on the basis of *Pierson* and *Keeble*. If cases are only compared on the factors they contain, then no ruling precedent can be found. *Pierson* shares with *Young* that plaintiff was on open land and that he had not yet caught the animal. Of these two factors, *Keeble* only shares the latter with *Young*, but in addition *Keeble* shares with *Young* that plaintiff was pursuing the animals for a living.

However, Berman & Hafner convincingly argue that skilled lawyers do not confine themselves to factor-based comparisons, but often frame their arguments in terms of the values that are at stake.[2] Let us apply this view to the above cases and assume that three values are at stake in these cases, viz. economic benefit for society ($Eval$), legal certainty ($Cval$), and the protection of property ($Pval$). Then a key idea is to specify how case decisions advance values.

- Deciding for a side because that side was hunting for a living advances $Eval$.
- Deciding for a side because that side was hunting on his own land advances $Pval$.
- Deciding for a side because that side had caught the animal advances $Pval$.
- Deciding for a side because the other side had not caught the animal advances $Cval$.

We can then say that *Pierson* was decided for defendant to promote legal certainty and since no values are served by deciding for plaintiff: he was not hunting for a living so economic benefit would not be advanced, and he had not yet caught the fox and was hunting on open land, so there are no property rights to be protected. Further, we can say that *Keeble* was decided for plaintiff since the value of economic benefit and the protection of property are together more important than the value of certainty. Thus *Keeble* also reveals part of an ordering of the values. Finally, in this interpretation of *Pierson* and *Keeble*, *Young* should be decided for defendant: the value of economic benefit does not support plaintiff since defendant was also fishing for his living, the value of protecting property does not apply since plaintiff had not yet caught the fish and was not on his own land, so the only value at stake is certainty, which is served by finding for the defendant. We now give a semiformal analysis of this analysis in the *ASPIC*$^+$ framework, leaving the logical language formally undefined and instead using streamlined natural language for expressing the premises and conclusions of the arguments. Recall that argument schemes are in *ASPIC*$^+$ modelled as defeasible inference rules. The first idea is that the specification of how case decisions advance values can

[2] Below we will use 'values' to cover also purposes, policies, interests etc.

be used in the following argument scheme.

Argument scheme from case decisions promoting values

Deciding *Current Pro* promotes set of values V_1
Deciding *Current Con* promotes set of values V_2
V_1 is preferred over V_2

Therefore (presumably), *Current* should be decided *Pro*.

Here *Pro* and *Con* are variables ranging over {*Plaintiff*,*Defendant*}. Another idea is that whether a set of values is preferred over another set of values, can be derived from a precedent (as in our example from *Keeble*).

Argument scheme from preference from precedent

Deciding *Precedent Pro* promotes set of values V_1
Deciding *Precedent Con* promotes set of values V_2
Precedent was decided *Pro*

Therefore (presumably), V_1^+ is preferred over V_2^-

Here the notation V_1^+ denotes any superset of V_1 of values while V_2^- denotes any subset of V_2. This notation captures *a fortiori* reasoning in that if in a new case deciding *Pro* promotes at least V_1 and possibly more values, while deciding *Con* promotes at most V_2, then the new case is even stronger for *Pro* than the precedent.

If it is also given that a proper superset of values is always preferred over a proper subset, then the first scheme directly applies to *Young*, since deciding *Young* for the defendant promotes $\{Pval, Eval\}$ while deciding *Young* for the plaintiff promotes $\{Eval\}$. However, imagine another new case in which deciding for the plaintiff promotes $\{Pval, Eval\}$ or a superset thereof, while deciding for the defendant promotes $\{Cval\}$: then the second scheme is needed to infer the preference of the first value set over the second (for instance, from *Keeble*), after which the first scheme can be applied to conclude that the plaintiff should win.

8.2.6 Arguing about rule change

What we have in fact done in the previous subsection is modelling legal case-based reasoning as what philosophers call practical reasoning, that is, reasoning about what to do. In particular, in both cases use is made of a variant of argument schemes of good and bad consequences of decisions for action. While in the previous section we applied this approach to factor-based reasoning, it can also be applied to other legal interpretation problems. In this subsection we illustrate this for the question whether a legal rule has to be modified by making an exception or not. In civil-law legal systems, in which rules are mainly created by legislation, this question can arise in parliamentary debate, where the debate will be about the good and bad consequences of enacting a rule, or about the legal, social or moral values promoted or demoted by enacting the rule. In common-law systems, where traditionally rules are gradually developed in a series of case-law decisions, the question arises in judicial decision-making in concrete cases, where the question is whether to follow or to distinguish a common-law rule. However, the types of arguments are often the same as in parliamentary settings.

We illustrate this with an American common law of contract case, the *Olga Monge v. Beebe Rubber Company* case, decided by the Supreme Court of New Hampshire (USA), February 28, 1974. In brief, the facts were that Olga Monge, according to the

court "a virtuous mother of three", was employed at will (that is, for an indefinite period of time) by Beebe Rubber Company. The relevant common law rule at that time said that every employment contract that specifies no duration is terminable at will by either party, which means that the employee can be fired for any reason or no reason at all. At some point, Olga Monge was fired for no reason by her foreman. Olga claimed that this was since she had refused to go out with him and she claimed breach of contract, arguing that the common law rule does not apply if the employee was fired in bad faith, malice, or retaliation. The court accepted that she was fired was that reason and was then faced with the problem whether to follow the old rule and decide that there was no breach of contract, or to distinguish the rule into a new rule by adding an exception in case the employee was fired in bad faith, malice, or retaliation, in order to decide that there was breach of contract. Here it is relevant that according to one common law theory of precedential constraint, courts can distinguish an old rule by adding an extra condition as long as the new rule still gives the same outcome in all precedent cases as the old rule.

The court decided to distinguish the old rule, on the following grounds:

> In all employment contracts, whether at will or for a definite term, the employer's interest in running his business as he sees fit must be balanced against the interest of the employee in maintaining his employment, and the public's interest in maintaining a proper balance between the two.
>
> (. . .)
>
> We hold that a termination by the employer of a contract of employment at will which is motivated by bad faith or malice or based on retaliation is not in the best interest of the economic system or the public good and constitutes a breach of the employment contract.

We now reconstruct this reasoning as practical reasoning with a variant of the above argument scheme from cases promoting values.

Argument scheme from decisions promoting values

Adopting *Proposal 1* promotes set of values V_1
Adopting *Proposal 2* promotes set of values V_2
V_1 is preferred over V_2

Therefore (presumably), *Proposal 1* should be adopted.

The two alternative decision options are to follow the old rule (proposal 2) or to distinguish it into the new rule by adding a condition 'unless the employee was fired in bad faith, malice, or retaliation' (proposal 1). In the present interpretation the court stated that following the old rule promotes the employer's interest in running his business as he sees fit (the single value in V_2) while it stated that distinguishing the old rule by adopting the new rule promotes the interest of the economic system and the public good (the two values comprising V_1). The court's decision to adopt the new rule then expresses a preference for V_1 over V_2. Note that the court does not give an explicit reason for this preference. The new rule is then applied with defeasible modus ponens to the facts of the case, leading to the final decision.

8.3 Exercises

Many exercises below are about Figure 8.1, in particular the case base Group 1 of Table 1 and the factor hierarchy.

8.3.1 Exercises on HYPO and CATO

EXERCISE 8.3.1 Consider Figure 8.1 and consider a new case with:

Pro-plaintiff factors:	F2, F4, F15, F21
Pro-defendant factors:	F16, F23

1. Which precedents are citable for the plaintiff?
2. Give all citable counterexamples for the defendant against Emery.
3. Give a counterexample from question 2 which can be distinguished by the plaintiff, and list the factors on which it can be distinguished.
4. Give a case that is citable for the plaintiff and then distinguishable for the defendant, after which the plaintiff can downplay the distinction. Show how the plaintiff can downplay it.

8.3.2 Exercises on precedential constraint

EXERCISE 8.3.2 For question 8.3.1, which of the citable precedents give rise to a preference with Definition 8.2.6 that can be used to argue that deciding for the plaintiff in the new case is forced?

EXERCISE 8.3.3 Consider the issue whether a child is allowed to watch TV after dinner (W). Assume the following factors:

Pro-W factors:	Done homework (D), at least 10 years old (T), Healthy (H)
Con-W factors:	Not-D, not-T, Not-H.

Assume the following precedent: Albert is 12 years old, he has done his homework but he has the flu. He was not allowed to watch TV after dinner.

1. Assume a new case with Betsy who is 9 years old, has done her homework and is healthy. Is the decision to let her watch TV after dinner allowed and/or forced by the precedent? And is the opposite decision allowed and/or forced by the precedent?
2. Assume a second new case with Carla who is 7 years old, has done her homework but is not healthy. Is the decision to let her watch TV after dinner allowed and/or forced by the precedent? And is the opposite decision allowed and/or forced by the precedent?
3. Imagine a third new case of Derek who is 15 years old, has done his homework but has a cold. Is the decision to let him watch TV after dinner allowed and/or forced by the precedent? And is the opposite decision allowed and/or forced by the precedent?

4. (Continuing 3) can you think of ways to distinguish Derek's case from the precedent?

EXERCISE 8.3.4 Express your answer to Question 8.3.3(1-3) in the reader with Definition 8.2.5. Identify for all new cases the relevant differences with the precedent, assuming they have the same outcome as the precedent.

EXERCISE 8.3.5 This question is about the table and figure on CATO in Section 8.2.2. Consider the following precedent and new case:

Case	Pro-plaintiff factors	Pro-defendant factors	Outcome
Prec	F2, F15, F21	F23	Plaintiff
New case	F2, F4, F21	F1, F25	?

1. Consider the factor hierarchy in Figure 4 in the reader. Find a way for the defendant to distinguish the precedent after which the plaintiff can downplay the distinction.

2. Verify whether deciding the new case Pro is forced, allowed but not forced, or not allowed. Express your answer with both Definition 8.2.5 and Definition 8.2.8.

3. Identify the relevant differences between the precedent and the new case, assuming they have the same outcome.

EXERCISE 8.3.6 Consider in the example on page 154 a new case

F_2: *deceived*$_{\pi 1}$, *measures* = *Access-To-Premises-Controlled*, *not-unique*$_{\delta 1}$,*reverse-eng*$_{\delta 2}$, *disclosed* = 15

1. Determine whether deciding F_2 for π or δ is forced on the basis of $CB = \{c_1, c_2\}$.

2. List the relevant differences between the precedents and the new case, assuming for each precedent that the new case has the same outcome.

EXERCISE 8.3.7 Model the problem domain of Question 8.3.3 of the reader with dimensions. Then answer questions (1-3) again, including the relevant differences between the precedent and the new cases.

EXERCISE 8.3.8 Consider the issue whether the fiscal domicile of a Dutch person who moved abroad for some time has changed, with the outcomes *changed* and *not changed*. Consider the following dimensions relevant to that issue, where *home* is whether the tax payer kept or gave up his Dutch home while being abroad, *employer* is whether during his stay abroad the tax payer had a Dutch or foreign employer or was self-employed, *duration* is the duration of the stay abroad in months, and *earnings* is the percentage of the tax-payer's income that was earned abroad during the stay:

home, with $V = \{$*given up*, *kept*$\}$ and *kept* $<_{changed}$ *given up*;
employer, with $V = \{$*Dutch*, *self-employed*, *foreign*$\}$ and *Dutch* $<_{changed}$ *self-employed*, *Dutch* $<_{changed}$ *foreign*;
duration, with V = the natural numbers and $x <_{changed} y$ iff $x < y$;
earnings, with $V = \{0, \ldots, 100\}$ and $x <_{changed} y$ iff $x < y$.

Consider furthermore the following case base:

c_1: *home* = *kept*; *employer* = *self-employed*; *duration* = 16; *earnings* = 60.
c_2: *home* = *gave up*; *employer* = *Dutch*; *duration* = 12; *earnings* = 80.

where *Outcome*(c_1) = *changed* and *Outcome*(c_2) = *not changed*. Consider finally the following fact situation

F: *home* = *kept*; *employer* = *foreign*; *duration* = 18; *earnings* = 60.

Is deciding F for *changed* forced, is deciding F for *not changed* forced, or are both decisions allowed? In your answer, specify the relevant differences between F and the two precedents.

Chapter 9

Answers to exercises from Chapters 1-8

9.1 Answer to exercise Chapter 1

EXERCISE 1.2.1

1. B and D are justified. B is reinstated by D.
2. A, C and E are justified. No argument is reinstated by D, since D is not justified. A and C are reinstated by E.

9.2 Answers to exercises Chapter 2

EXERCISE 2.8.1

(a): C is justified since it has no defeaters. B is not justified, since it is defeated by a justified argument, viz. by C. Therefefore, A is justified, since its only defeater, which is B, is not justified.
(b): The status of A and B cannot be determined: A is justified if and only if its only defeater, which is B, is not justifed. But B is not justified just in case A, which is its only defeater, is justified. Thus we enter a loop. And since the status of C depends on the status of its only defeater, which is B, the status of C cannot be determined either.

EXERCISE 2.8.2 Consider an arbitrary argument A. By assumption, there is an argument B such that B defeats A. So $A \in F(\varnothing)$ iff there is a $C \in \varnothing$ such that C defeats B. However, no such C exists, so $A \notin F(\varnothing)$. Since A was chosen arbitrarily, we can conclude that no argument is in $F(\varnothing)$. $\Box$.

EXERCISE 2.8.3

a:	b:	c:	d:
$F^0 = \varnothing$	$F^0 = \varnothing$	$F^0 = \varnothing$	$F^0 = \varnothing$
$F^1 = \{A\}$	$F^1 = F^0$	$F^1 = \{C\}$	$F^1 = \{A, E\}$
$F^2 = \{A, D\}$		$F^2 = \{C, B\}$	$F^2 = \{A, E, C\}$
$F^3 = F^2$		$F^3 = F^2$	$F^3 = F^2$

The grounded extensions are the fixed points of these sequences.

So the grounded extension is $\{A, D\}$.

EXERCISE 2.8.4

1. To show that $F(X) = G^2(X)$, for every set of arguments X, it turns out that it is easier to show that the complements of the two sets are equal. This has to do with quantifying over arguments. Thus, suppose $x \notin G^2(X)$. By definition of G this means that there exists a $y \in G(X)$ defeating x, i.e., $x \leftarrow y$. Since $y \in G(X)$, the argument y is not defeated by a member of X. Hence y shows that $x \notin F(X)$. Conversely, suppose that $x \notin F(X)$. Then x is defeated by a y that is not defeated by a $z \in X$. Thus x is defeated by a $y \in G(X)$, and hence $x \notin G^2(X)$.

2. The result that G is anti-monotonic follows from the fact that, if an argument is not defeated by a member of B, then it surely cannot be defeated by a member of any subset $A \subseteq B$.

3. Suppose $A \subseteq B$. Since G is anti-monotonic, it follows that $G(B) \subseteq G(A)$. Again by anti-monotonicity of G, we obtain $G^2(A) \subseteq G^2(B)$, which is equal to the expression $F(A) \subseteq F(B)$.

4. If $\{G_i\}_{i \geq 0}$ with $G_0 =_{Def} \varnothing$ and $G_i =_{Def} G(G_{i-1})$, then in particular

$$G_0 \subseteq G_1 \text{ and } G_0 \subseteq G_2. \tag{9.1}$$

   Now apply the anti-monotonicity of G to (9.1) repeatedly, to obtain the chain of inclusions desired.

EXERCISE 2.8.5
- (a): justified: A, D; overruled: B, C; defensible: none.
- (b): justified: none; overruled: none; defensible: all.
- (c): justified: B, C; overruled: A, D; defensible: none.
- (d): justified: A, C, E; overruled: B, D; defensible: none.

EXERCISE 2.8.6
$\Rightarrow$:
Consider any stable extension E, and consider first any argument A not defeated by E. Then $A \in E$. Consider next any argument B defeated by E. Then, since E is conflict-free, $B \notin E$. So $E = \{A \mid A \text{ is not defeated by } E\}$.□
$\Leftarrow$:
Let $E = \{A \mid A \text{ is not defeated by } E\}$. Clearly, E is conflict-free. Furthermore, for all A, if $A \notin E$, then E defeats A. So E is a stable extension.□

EXERCISE 2.8.7

- Example 2.1.3: There is just one status assignment, which is maximal:
  - $S_1 = (\{A, C\}, \{B\})$
- Example 2.1.4: There are three status assignments:

- $S_1 = (\varnothing, \varnothing)$
- $S_2 = (\{A\}, \{B\})$
- $S_3 = (\{B\}, \{A\})$

Only S_2 and S_3 are maximal.

- Example 2.3.8: There is just one status assignment, which is maximal:

  - $S_1 = (\varnothing, \varnothing)$

EXERCISE 2.8.8

1. Consider any $A \in Out$. Then there is a $B \in In$ defeating A. But also $B \in In'$, so that $A \in Out'$. So $Out \subseteq Out'$.

2. Consider any argument C such that $C \notin In$ but $C \in In'$.
   (i) Since $C \notin In$, there exists a $B \notin Out$ such that B defeats C.
   (ii) Consider next any such B that defeats C and is not in Out. Any such B must be in Out', otherwise C would not be in In'.
   Hence (from i and ii) there exists an argument that is in Out' but not in Out. Together with (1) this gives us that Out is a proper subset of Out'.

EXERCISE 2.8.9

- A is defensible iff is in *in* some but not all preferred status assignments, and A is overruled if A is *out* in all preferred status assignments. *This leaves open that there are arguments that neither justified, nor defensible, nor overruled. Cf. Example 2.3.8.*
- A is defensible iff is in *in* some but not all preferred status assignments, and A is overruled if there is no status assignment in which A is *in*. *With this definition all arguments are either justified, Xor defensible, Xor overruled.*

EXERCISE 2.8.10: The empty set, which is maximally admissible.

EXERCISE 2.8.11

1. (a) Preferred: $\{A, D\}$, also stable.

   (b) Preferred: $\{B, D, E\}$, also stable; $\{A, E\}$, also stable.

   (c) Preferred: $\varnothing$, no stable extensions.

   (d) Preferred: $\{A, C, E\}$, also stable.

   (e) (with slightly detailed explanation)
   (1) Preferred extensions:

   - $E_1 = \{A, B, D\}$
   - $E_2 = \{C\}$

   (2) Stable extensions. Both E_1 and E_2 are also stable extensions, since both sets defeat all arguments outside them. Furthermore, by Proposition 2.4.1 there are no other stable extensions.

2. (a) for preferred and stable semantics: A, D justified, B, C overruled.

(b) for preferred and stable semantics: E justified, C overruled, A, B, D defensible.

(c) for preferred semantics: neither is justified, defensible or overruled. For stable semantics: all are both justified and overruled.

(d) For preferred and stable semantics: A, C, E justified, B, D overruled.

(e) For preferred and stable semantics: all defensible

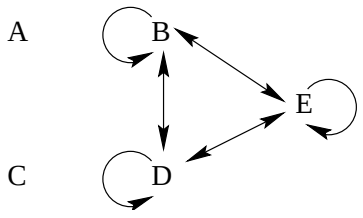
EXERCISE 2.8.12: The grounded extension is empty, while there are two preferred extensions, viz. $\{B, D\}$ and $\{A, C\}$. Note that one preferred extension concludes that Larry is rich, while the other concludes that Larry is not rich, so in bothsemantics no conclusion about Larry's richness is justified. Yet it may be argued that the conclusion that Larry is not rich is the intuitively justified conclusion, since all arguments for the opposite conclusion have a strict defeater. Anyone who adopts this analysis, will have to conclude that this example presents a problem for both grounded and preferred semantics. However, see Exercise 4.8.7 for a solution when the structure of arguments is made explicit.

EXERCISE 2.8.13

1. $AF(\Delta_3)$ contains five arguments:

   - $A = \varnothing$
   - $B = \frac{:p}{\neg p}$
   - $C = \frac{:q}{q}$
   - $D = \frac{:p}{\neg p}, \frac{:q}{q}$
   - $E = \frac{:q}{q}, \frac{:p}{\neg p}$

   The defeat graph is as follows:



   There is no stable extension, while there is one preferred extension, viz. $\{A, C\}$.

2. Since it recognizes that A and C should come out as justified, since they have no defeaters.

EXERCISE 2.8.14

1.

   (a) $A = \frac{:b}{a}, \frac{a:c \wedge d}{c}, \frac{c:b}{b}$

   (b)

   - $B = \frac{:e}{e}, \frac{e:\neg a}{\neg d}$
   - $C = \frac{:\neg a}{\neg a}, \frac{:b}{a}$

(c) Yes, for instance of $\{A\}$. Note that A defeats both B and C. Another admissible set is $\{A, A'\}$, where

- $A' = \frac{:b}{a}$

Note that A' also defeats both B and C.

(d) Yes, by (c) and the fact that every admissible set is contained in a preferred extension (see the proof of Proposition 2.3.13).

(e) No: the grounded extension is empty, since there is no undefeated argument. In particular, A' is defeated by C.

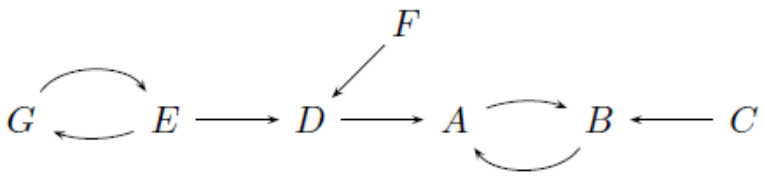
EXERCISE 2.8.15

Suppose A is finite and failed. Then $In(A) \cup Out(A) \neq \varnothing$, so $\varphi \in In(A)$ for some $\varphi \in Out(A)$. But then A defeats A.

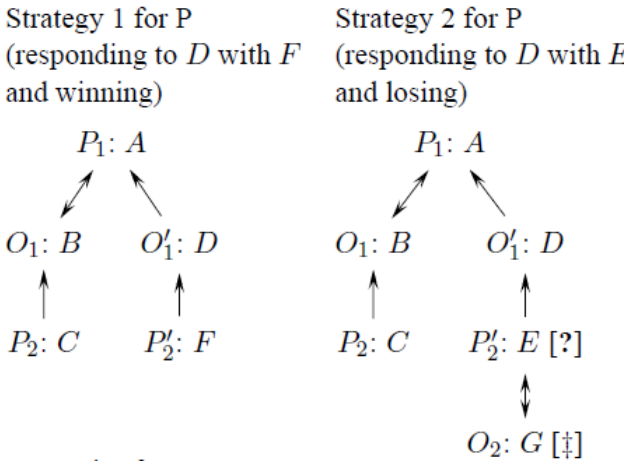
9.3 Exercises Chapter 3

EXERCISE 3.5.1

1. The defeat graph is:



2. We are asked to list all strategies of P an O. There are two strategies for P ("?" indicates an unfortunate move, "‡" indicates the move that leads to a loss for the other party):



There are two strategies for O:

Strategy 1 for O (responding to A with B and losing)

P_1: A ⟷ O_1: B ⟵ P_2: C [‡]

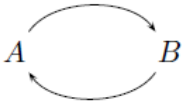
Strategy 2 for O (responding to A with D and losing)

P_1: A ⟵ O_1: D ⟵ P_2: E [?] ⟷ O_2: G

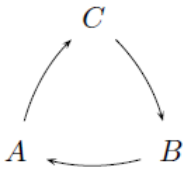
O_1: D ⟵ P_2: F [‡]

EXERCISE 3.5.2

1.



2.



3.

$A_1 \longleftarrow A_2 \longleftarrow A_3 \longleftarrow A_4 \longleftarrow A_5 \longleftarrow \cdots$

EXERCISE 3.5.3

(1a) P has winning strategies for A and D, but not for B:
- A winning strategy for A consists of putting forward A, after which O cannot respond because A has no defeaters.
- A winning strategy for B does not exist, because O can reply to B with A, after which P cannot move.
- A winning strategy for D is simple: put forward D; the only responses to D are B and C, which can both be countered with A, after which O cannot move.

(3) We make the comparison for the proof of A in graph (a):

$F^0 = \varnothing$
$F^1 = \{A\}$
$F^2 = \{A, D\}$

Compared to a won dialogue on D, the order of stating A and D is reversed. With F, we start with the undefeated arguments and at each iteration add the arguments reinstated by the arguments added at the previous iteration. In a dialectical proof, P starts

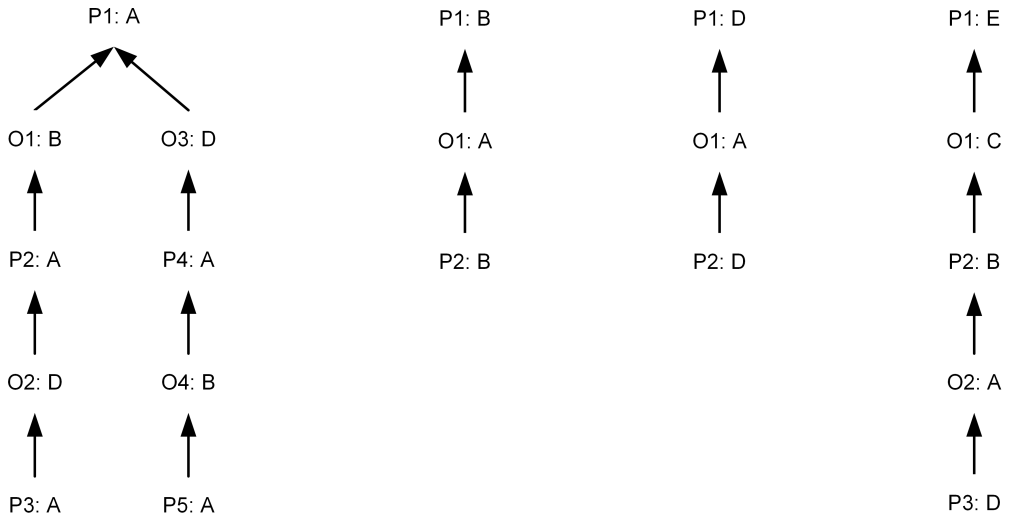


Figure 9.1: P's won games

with an argument from F^i where i may be greater than 1, and at each next turn P moves an argument from F^{i-1} that can reinstate the argument of the previous move.

EXERCISE 3.5.4 The argument $A == \frac{:b}{a}, \frac{a:c \wedge d}{c}, \frac{c:b}{b}$ is not provably justified, since O can reply with $C = \frac{:\neg a}{\neg a}, \frac{:b}{a}$, after wich P has no strictly defeating reply.

EXERCISE 3.5.5 P successively moves $A_1, A_3, \ldots, A_{2i-1}, A_{2i+1}, \ldots$ and O successively moves $A_2, A_4, \ldots, A_{2i}, A_{2i+2}, \ldots$ so they wil never repeat their own argument. And P always uses 'odd' arguments while O always uses 'even' arguments, so they will never repeat each other's argument. Finally, since the defeat chain is infinite, they will always have a new move.

EXERCISE 3.5.6 The simplest example is with two arguments A and B such that A defeats itself and there are no other defeat relations. B is provable since O has no reply if P starts with B, but this argumentation framework has no stable extensions.

EXERCISE 3.5.7 All arguments except argument C are provable. Figure 9.1 contains terminated games won by P for A, B, D and E. Argument C is not provable since O can always win by replying with B, and if P replies with A, then O can repeat A as an alternative reply to C, making an *eo ipso* move.

9.4 Exercises Chapter 4

EXERCISE 4.8.1

1. The following argument for Ra can be created.

A_1: $\forall x(Px \supset Qx)$
A_2: Pa
A_3: $A_1, A_2 \rightarrow Qa$
A_4: $\forall x(Qx \supset Rx)$
A_5: $A_3, A_4 \rightarrow Ra$

2. $\texttt{Prem}(A) = \{Pa, \forall x(Px \supset Qx), \forall x(Qx \supset Rx)\}$
   $\texttt{Conc}(A) = Ra$
   $\texttt{Sub}(A) = \{A_1, A_2, A_3, A_4, A_5\}$
   $\texttt{DefRules}(A) = \varnothing$
   $\texttt{TopRule}(A) = Qa, \forall x(Qx \supset Rx) \rightarrow Ra$

3. The argument is strict and plausible.

EXERCISE 4.8.2.

1. We have the following arguments:

   A_1: *injury*
   A_2: *appendicitis*
   A_3: $A_2 \Rightarrow \neg$ *riskyOperation*
   A_4: A_1, $A_3 \Rightarrow$ *negligence*
   A_5: $A_1, A_4 \Rightarrow$ *compensation*

   B_1: *medicalTests1*
   B_2: $B_1 \Rightarrow$ *badCirculation*
   B_3: $B_2 \Rightarrow$ *riskyOperation*

   C_1: *medicalTests2*
   C_2: $C_1 \Rightarrow \neg$ *badCirculation*

   Their attack relations are shown in Figure 9.2.

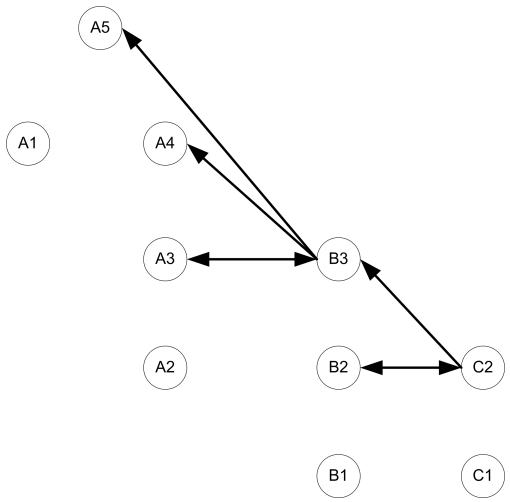


Figure 9.2: Abstract attack graph

2. - $\texttt{Prem}(A_1) = \{f_1\}$, $\texttt{Conc}(A_1) = injury$, $\texttt{Sub}(A_1) = \{A_1\}$, $\texttt{DefRules}(A_1) = \varnothing$ and $\texttt{TopRule}(A_1) =$ undefined.
   - $\texttt{Prem}(A_2) = \{f_2\}$, $\texttt{Conc}(A_2) = appendicitis$, $\texttt{Sub}(A_2) = \{A_2\}$, $\texttt{DefRules}(A_2) = \varnothing$ and $\texttt{TopRule}(A_2) =$ undefined.
   - $\texttt{Prem}(A_3) = \{f_2\}$, $\texttt{Conc}(A_3) = \neg riskyOperation$, $\texttt{Sub}(A_3) = \{A_2, A_3\}$, $\texttt{DefRules}(A_3) = \{r_3\}$ and $\texttt{TopRule}(A_3) = r_3$.

- $\texttt{Prem}(A_4) = \{f_1, f_2\}$, $\texttt{Conc}(A_4) = \mathit{negligence}$, $\texttt{Sub}(A_4) = \{A_1, A_2, A_3, A_4\}$, $\texttt{DefRules}(A_4) = \{r_2, r_3\}$ and $\texttt{TopRule}(A_4) = r_2$.
- $\texttt{Prem}(A_5) = \{f_1, f_2\}$, $\texttt{Conc}(A_5) = \mathit{compensation}$, $\texttt{Sub}(A_5) = \{A_1, A_2, A_3, A_4, A_5\}$, $\texttt{DefRules}(A_5) = \{r_1, r_2, r_3\}$ and $\texttt{TopRule}(A_5) = r_1$.
- $\texttt{Prem}(B_1) = \{f_3\}$, $\texttt{Conc}(B_1) = \mathit{medicalTests1}$, $\texttt{Sub}(B_1) = \{B_1\}$, $\texttt{DefRules}(B_1) = \varnothing$ and $\texttt{TopRule}(B_1) =$ undefined.
- $\texttt{Prem}(B_2) = \{f_3\}$, $\texttt{Conc}(B_2) = \mathit{badCirculation}$, $\texttt{Sub}(B_2) = \{B_1, B_2\}$, $\texttt{DefRules}(B_2) = \{r_5\}$ and $\texttt{TopRule}(B_2) = r_5$.
- $\texttt{Prem}(B_3) = \{f_3\}$, $\texttt{Conc}(B_3) = \mathit{riskyOperation}$, $\texttt{Sub}(B_3) = \{B_1, B_2, B_3\}$, $\texttt{DefRules}(B_3) = \{r_4, r_5\}$ and $\texttt{TopRule}(B_3) = r_4$.
- $\texttt{Prem}(C_1) = \{f_4\}$, $\texttt{Conc}(C_1) = \mathit{medicalTests2}$, $\texttt{Sub}(C_1) = \{C_1\}$, $\texttt{DefRules}(C_1) = \varnothing$ and $\texttt{TopRule}(C_1) =$ undefined.
- $\texttt{Prem}(C_2) = \{f_4\}$, $\texttt{Conc}(C_2) = \neg \mathit{badCirculation}$, $\texttt{Sub}(C_2) = \{C_1, C_2\}$, $\texttt{DefRules}(C_2) = \{r_6\}$ and $\texttt{TopRule}(C_2) = r_6$.

3. We have that $\texttt{LastDefRules}(A_3) = \{r_3\}$ while $\texttt{LastDefRules}(B_3) = \{r_4\}$ and since $r_3 < r_4$ we have that $\texttt{LastDefRules}(A_3) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(B_3)$, so $A_3 \prec B_3$, so B_3 strictly defeats A_3.

   Moreover, we have that $\texttt{LastDefRules}(B_2) = \{r_5\}$ while $\texttt{LastDefRules}(C_2) = \{r_6\}$ and since $r_5 < r_6$ we have that $\texttt{LastDefRules}(B_2) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(C_2)$, so $B_2 \prec C_2$, so C_2 strictly defeats B_2.

   The other attack relations succeed as defeats. The resulting defeat relations are shown in Figure 9.3.

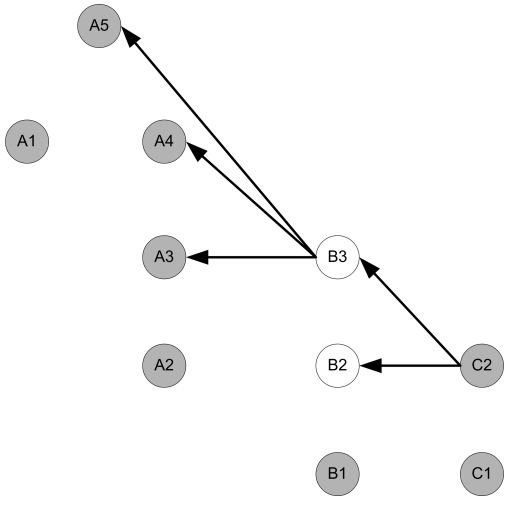


Figure 9.3: Abstract argumentation framework with grounded and unique preferred labelling

4. Figure 9.3 shows the grounded labelling: the arguments that are *in* are coloured gray, the arguments that are *out* are coloured white. The grounded extension consists of all arguments that are labelled *in*.

5. Since the abstract argumentation theory depicted in Figure 9.3 is finite and has no cycles, all semantics give the same result. But we also have $\texttt{DefRules}(A_3) = \{r_3\}$ while $\texttt{DefRules}(B_3) = \{r_4, r_5\}$. Since the priority relation between r_3 and r_5 is undefined, we have that the preference relation between $\texttt{DefRules}(A_3)$ and $\texttt{DefRules}(B_3)$ is also undefined. So the grounded extension is also the unique preferred (and stable) extension.

6. We now also have that $\texttt{Prem}_p(A_3) = \varnothing$ while $\texttt{Prem}_p(B_3) = \{f_3\}$ so we have that $\texttt{Prem}_p(B_3) \lhd_{\texttt{Eli}} \texttt{Prem}_p(A_3)$. But we also have $\texttt{DefRules}(A_3) = \{r_3\}$ while $\texttt{DefRules}(B_3) = \{r_4, r_5\}$. Since the priority relation between r_3 and r_5 is undefined, we have that the preference relation between $\texttt{DefRules}(A_3)$ and $\texttt{DefRules}(B_3)$ is also undefined. Then we have that $A_3 \not\preceq B_3$ and $B_3 \not\preceq A_3$ so A_3 and B_3 now defeat each other.

   Moreover, we now also have that $\texttt{Prem}_p(B_2) = \{f_3\}$ while $\texttt{Prem}_p(C_2) = \{f_4\}$, so since $f_4 <' f_3$, we have that $\texttt{Prem}_p(C_2) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. Then we have that $B_2 \not\preceq C_2$ and $C_2 \not\preceq B_2$ so B_2 and C_2 now also defeat each other. So the defeat relations now equal the attack relations as displayed in Figure 9.2. Then there are three preferred labellings: the original one displayed in Figure 9.3 and two new ones displayed in, respectively Figure 9.4 and Figure 9.5. The two new preferred extensions consist, respectively, of the sets of argument labelled *in* in these two preferred labellings.

   The grounded labelling now makes A_1, A_2, B_1 and C_1 *in* and the remaining arguments undecided. So the grounded extension is $\{A_1, A_2, B_1, C_1\}$.

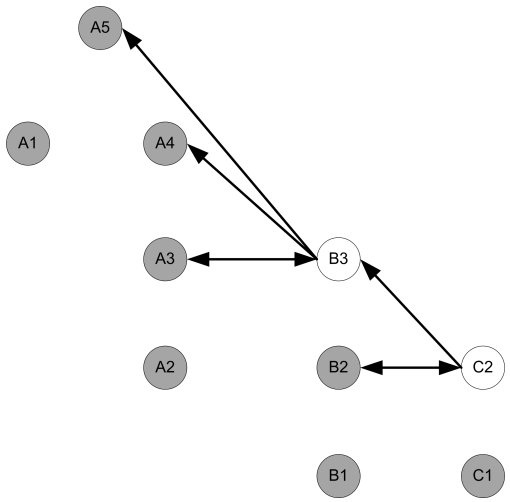


Figure 9.4: Abstract argumentation theory with a second preferred labelling

EXERCISE 4.8.3.

1. The following argument for t can be created.

   A_1: p
   A_2: q
   A_3: $A_1, A_2 \Rightarrow r$
   A_4: $A_3 \rightarrow s$
   A_5: $A_4 \Rightarrow t$

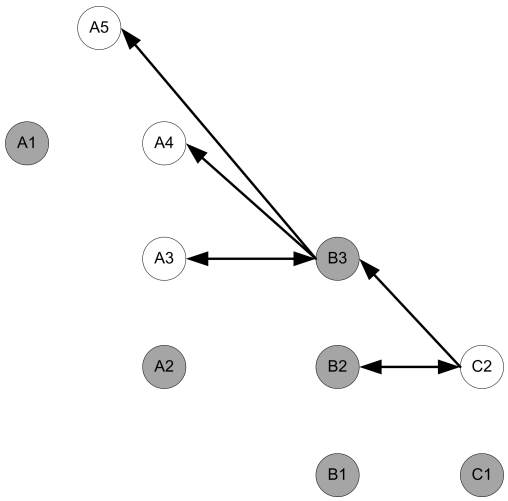


Figure 9.5: Abstract argumentation theory with a third preferred labelling

A_5 has one attacker, namely, the following underminer on A_2:

B_1: u
B_2: $B_1 \Rightarrow v$
B_3: x
B_4: $B_2, B_3 \rightarrow \neg q$

Argument B_4 in turn has one attacker: it is undermined by the following argument for $\neg u$:

C_1: w
C_2: $C_1 \Rightarrow \neg u$

2. We use the G game. Since the argument ordering is empty, all attacks succeed as defeats. So B_4 successfully undermines A_5 on A_2 and thus defeats A_5, while C_2 successfully undermines B_4 on B_1, so C_2 strictly defeats B_4. (Note that this is strict defeat since B_4 does not even attack C_2.) However, we also have that B_1 rebuts C_2 so defeats C_2, so the opponent can reply to C_2 with B_1. Then the game ends with a win by the opponent, since B_1 and C_2 defeat each other, so there is no strict defeater of B_1.

   To see whether A_5 is defensible or overruled, note that the only defeater of A_5 is B_4 but the proponent does not have a winning strategy for B_4: it is defeated by C_2, which has no strict defeater. Hence B_4 is not justified, so A_5 is defensible, which makes t defensible also.

3. t is now justified, since argument B_4 does not defeat argument A_2, so A_5 has no defeaters. To see this, observe that $\texttt{LastDefRules}(A_2) = \varnothing$ while $\texttt{LastDefRules}(B_4) = \{u \Rightarrow v\} \neq \varnothing$, so $\texttt{LastDefRules}(B_4) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(A_2)$, so $B_4 \prec A_2$.

4. Now t is not justified. Note that $\texttt{Prem}_p(A_2) = \{q\}$ while $\texttt{Prem}_p(B_4) = \{u\}$ and since $q <' u$ we have that $\texttt{Prem}_p(A_2) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_4)$. Then despite the fact that $\texttt{DefRules}(B_4) \lhd_{\texttt{Eli}} \texttt{DefRules}(A_2)$ we have that $B_4 \not\prec A_2$, so B_4 defeats A_2 and thus B_4 also defeats A_5.

Next we have to verify whether the attack of C_2 on B_4 succeeds as defeat. We have that $\texttt{Prem}_p(C_2) = \{w\}$ while $\texttt{Prem}_p(B_1) = \{u\}$ and $w <' u$, so we have that $\texttt{Prem}_p(C_2) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_1)$. Moreover, we have that $\texttt{DefRules}(B_1) = \varnothing$ and $\texttt{DefRules}(C_2) = \{w \Rightarrow \neg u\}$ so $\texttt{DefRules}(C_2) \lhd_{\texttt{Eli}} \texttt{DefRules}(B_1)$. So $C_2 \prec B_1$ so C_2 does not defeat B_1. So the opponent has a winning strategy, so A_5 is not justified, so t is not justified since there is no other argument for t.

To see whether A_5 is defensible or overruled, note that the only defeater of A_5 is B_4, which has as the only attacker C_2. Since we have $C_2 \prec B_1$, the opponent has no reply to B_4, so B_4 is justified. Then A_5 is overruled, which makes t overruled also.

5. We must add the following rules to $\mathcal{R}_s$: $\neg s \rightarrow \neg r$ and $v, q \rightarrow \neg x$ and $x, q \rightarrow \neg v$. We first answer question (1) again. There now also is the following rebuttal of B_2:

   D_1: q
   D_2: x
   D_3: $D_1, D_2 \rightarrow \neg v$

   Furthermore, we have that B_4 undermines D_3 on D_1.

   For question (2), note again that with an empty argument ordering all attacks succeed as defeats. In the game for A_5 now B_4 also has a defeater, since D_3 defeats B_4 on B_2. However, B_4 in turn defeats D_3 on D_1, so D_3 does not strictly defeat B_4. So the proponent still has no winning strategy for A_5.

   Next, to see whether A_5 is defensible or overruled, in a game about B_4 the opponent can defeat B_4 with D_3 and since B_4 and D_3 defeat each other and there is no other defeater of D_3, the proponent cannot move so has no winning strategy. So B_4 is not justified either. This makes A_5 and t defensible.

   For question (3) nothing changes, since A_5 has no new attackers.

   For question 4, we must first verify whether D_3's attack on B_2 succeeds. We have that $\texttt{DefRules}(B_2) = \{u \Rightarrow v\}$ and $\texttt{DefRules}(D_3) = \varnothing$, so $\texttt{DefRules}(B_2) \lhd_{\texttt{Eli}} \texttt{DefRules}(D_3)$. However, we also have that $\texttt{Prem}_p(D_3) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. Since D_3 is strict but not firm and B_2 is neither strict nor firm, we have to apply clause (3) of Definition 4.3.24. But then $D_3 \not\prec B_2$ and $B_2 \not\prec D_3$, so D_3 defeats B_2, so D_3's attack on B_2 succeeds. Next we must verify whether B_4's attack on D_1 succeeds. We have to make the same comparisons, so B_4 defeats D_1. But then B_4 also defeats D_3, so D_3 does not strictly defeat B_4. But then the proponent has no legal reply to B_4 in the G-game, so the proponent does not have a winning strategy for A_5. So A_5 is not justified.

   To see whether A_5 is defensible or overruled, we must check whether B_4 is justified. Since D_3 defeats B_4 and B_4 is the only defeater of D_3, we have that the proponent has no legal reply to D_3, so he has no winning strategy for B_4. So A_5 is defensible and so t is defensible.

EXERCISE 4.8.4: see Figure 9.6.

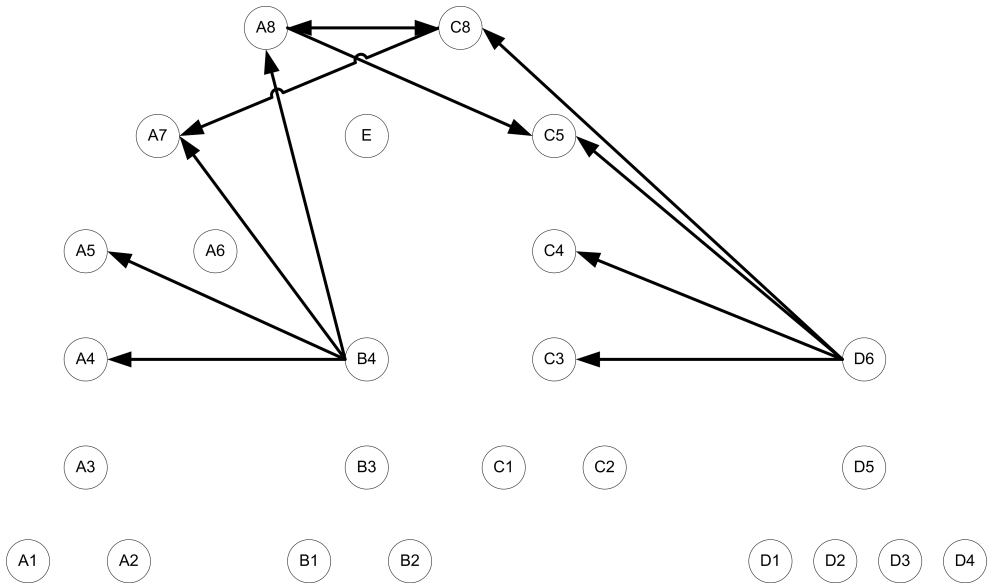


Figure 9.6: Abstract argumentation theory for Figure 4.5

EXERCISE 4.8.5.

1. The following argument for r can be constructed.

   A_1: p
   A_2: $A_1 \Rightarrow q$
   A_3: $A_1, A_2 \rightarrow r$

   We verify the status of r with the P-game. Argument A_3 has one defeater, namely the following undercutter of A_2.

   B_1: s
   B_2: $B_1 \Rightarrow t$
   B_3: $B_2 \rightarrow \neg d_1$

   Argument B_3 has one attacker, rebutting B_3 on B_2:

   C_1: u
   C_2: $C_1 \Rightarrow v$
   C_3: $C_2 \Rightarrow \neg t$

   We are in case (3) of Definition 4.3.24. First, since $u <' s$, we have that $\texttt{Prem}_p(C_3) \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. Next, B_2 uses one defeasible rule, namely, d_2, while C_3 uses two defeasible rules, namely, d_3 and d_4. Since $d_3 < d_2$ we have that $\texttt{DefRules}(C_3) \lhd_{\texttt{Eli}} \texttt{DefRules}(B_2)$. So $C_3 \prec B_2$, so B_2 strictly defeats C_3 and C_3 does not defeat B_3. Moreover, B_3 is *in* in all preferred labellings since it has no defeaters, so A_3 is *out* in all preferred labellings and is therefore overruled. Since there are no other arguments for r, this also makes r overruled.

2. Now r is justified. First, since both arguments are defeasible, the premise ordering is now irrelevant. Next, since arguments B_2 and C_3 are now compared on d_2 and d_4 and since $d_2 < d_4$, we have $\texttt{LastDefRules}(B_2) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(C_3)$.

So $B_2 \prec C_3$ and C_3 strictly defeats both B_2 and B_3. Moreover, there are no defeaters of C_3, so the proponent now has a winning strategy for A_3 in the P-game.

EXERCISE 4.8.6

1. $\mathcal{K}_p$ consists of:

   $\forall x(\texttt{BornInNL}(x) \rightsquigarrow \texttt{Dutch}(x))$
   $\forall x(\texttt{NorwegianName}(x) \rightsquigarrow \texttt{Norwegian}(x))$
   $\forall x((\texttt{Dutch}(x) \vee \texttt{Norwegian}(x)) \rightsquigarrow \texttt{LikesIceSkating}(x))$
   $\texttt{BorninNL}(b)$
   $\texttt{NorwegianName}(b)$
   $\forall x\neg(\texttt{Dutch}(x) \wedge \texttt{Norwegian}(x))$

   The following relevant arguments can be constructed:

   A_1: $\texttt{BorninNL}(b)$
   A_2: $\forall x\ (\texttt{BornInNL}(x) \rightsquigarrow \texttt{Dutch}(x))$
   A_3: $A_2 \rightarrow \texttt{BornInNL}(b) \rightsquigarrow \texttt{Dutch}(b)$
   A_4: $A_1, A_3 \Rightarrow \texttt{Dutch}(b)$
   A_5: $A_4 \rightarrow \texttt{Dutch}(b) \vee \texttt{Norwegian}(b)$
   A_6: $\forall x((\texttt{Dutch}(x) \vee \texttt{Norwegian}(x)) \rightsquigarrow \texttt{LikesIceSkating}(x))$
   A_7: $A_6 \rightarrow (\texttt{Dutch}(b) \vee \texttt{Norwegian}(b)) \rightsquigarrow \texttt{LikesIceSkating}(b)$
   A_8: $A_5, A_7 \Rightarrow \texttt{LikesIceSkating}(b)$

   B_1: $\texttt{BorninNL}(b)$
   B_2: $\forall x\ (\texttt{BornInNL}(x) \rightsquigarrow \texttt{Dutch}(x))$
   B_3: $B_2 \rightarrow \texttt{BornInNL}(b) \rightsquigarrow \texttt{Dutch}(b)$
   B_4: $B_1, B_3 \Rightarrow \texttt{Dutch}(b)$
   B_5: $\forall x\neg(\texttt{Dutch}(x) \wedge \texttt{Norwegian}(x))$
   B_6: $B_4, B_5 \rightarrow \neg\texttt{Norwegian}(b)$

   C_1: $\texttt{NorwegianName}(b)$
   C_2: $\forall x\ (\texttt{NorwegianName}(x) \rightsquigarrow \texttt{Norwegian}(x))$
   C_3: $C_2 \rightarrow \texttt{NorwegianName}(b) \rightsquigarrow \texttt{Norwegian}(b)$
   C_4: $C_1, C_3 \Rightarrow \texttt{Norwegian}(b)$
   C_5: $C_4 \rightarrow \texttt{Dutch}(b) \vee \texttt{Norwegian}(b)$
   C_6: $\forall x((\texttt{Dutch}(x) \vee \texttt{Norwegian}(x)) \rightsquigarrow \texttt{LikesIceSkating}(x))$
   C_7: $C_6 \rightarrow (\texttt{Dutch}(b) \vee \texttt{Norwegian}(b)) \rightsquigarrow \texttt{LikesIceSkating}(b)$
   C_8: $C_5, C_7 \Rightarrow \texttt{LikesIceSkating}(b)$

   D_1: $\texttt{NorwegianName}(b)$
   D_2: $\forall x\ (\texttt{NorwegianName}(x) \rightsquigarrow \texttt{Norwegian}(x))$
   D_3: $D_2 \rightarrow \texttt{NorwegianName}(b) \rightsquigarrow \texttt{Norwegian}(b)$
   D_4: $D_1, D_3 \Rightarrow \texttt{Norwegian}(b)$
   D_5: $\forall x\neg(\texttt{Dutch}(x) \wedge \texttt{Norwegian}(x))$
   D_6: $D_4, D_5 \rightarrow \neg\texttt{Dutch}(b)$

   (If the example is formalised in a propositional language, then the steps A_7 and C_7 must be omitted.)

2. Note first that if no preference relation is specified, it does not hold. Then the relevant defeat relations are as follows:

   - B_6 defeats C_4 and thus also $C_5 - C_8$
   - D_6 defeats B_4 and thus also B_5 and B_6
   - D_6 defeats A_4 and thus also $A_5 - A_8$
   - B_6 defeats D_4 and thus also D_5 and D_6

   Let us first concentrate on B_6 and D_6. Since they defeat each other and have no other defeaters, it is possible to assign no status to them. Then in the grounded status assignments they have no status. But then the same holds for the arguments defeated by one of them. This includes A_8 and C_8. Hence the conclusion $\texttt{LikesIceSkating}(b)$ only has defensible arguments and is therefore itself defensible.

   (The same answer in terms of the fixpoint definition: Since B_6 and D_6 defeat each other and have no other defeaters, they are in no F^i. But then the arguments defeated by one of them also are in no F^i.)

3. Let us again first concentrate on B_6 and D_6. Argument B_6 can be made *in* by making D_6 *out* and vice versa. Then there is a preferred status assignment in which B_6 is *in* and D_6 is *out*. In this status assignment also $C_4 - C_8$ are *out* and $A_1 - A_8$ are *in*. So an argument for the conclusion $\texttt{LikesIceSkating}(b)$ is *in*, namely, A_8. Conversely, there is also a preferred status assignment in which D_6 is *in* and B_6 is *out*. In this status assignment also $A_4 - A_8$ are *out* and $C_1 - C_8$ are *in*. So again an argument for the conclusion $\texttt{LikesIceSkating}(b)$ is *in* but this time it is not A_8 but C_8. So both A_8 and C_8 are defensible, so the conclusion $\texttt{LikesIceSkating}(b)$ is also defensible.

4. Since both preferred extensions contain an argument for the conclusion $\texttt{LikesIceSkating}(b)$, this conclusion is f-justified, even though there is no justified argument for it.

EXERCISE 4.8.7 The following formalisation is based on the intuition that the conclusion that Larry is not rich is justified. The undercutters in the example are based on the principle that statistical defaults about subclasses have priority over statistical defaults about superclasses.

$\mathcal{R}_s$ consists of all valid propositional and first-order inferences.

$\mathcal{R}_d$ consists of:

d_1. $\texttt{Lawyer}(x) \Rightarrow \texttt{Rich}(x)$
d_2. $\texttt{LivesInHollywood}(x) \Rightarrow \texttt{Rich}(x)$
d_3. $\texttt{PublicDefender}(x) \Rightarrow \neg\, \texttt{Rich}(x)$
d_4. $\texttt{RentsinHollywood}(x) \Rightarrow \neg\, \texttt{Rich}(x)$
d_5. $\texttt{PublicDefender}(x) \Rightarrow \neg d_1(x)$
d_6. $\texttt{RentsinHollywood}(x) \Rightarrow \neg d_2(x)$

$\mathcal{K}_p$ consists of

p_1. $\texttt{PublicDefender}(L)$
p_2. $\texttt{RentsInHollywood}(L)$

$\mathcal{K}_n$ consists of

n_1. $\forall x(\texttt{PublicDefender}(x) \supset \texttt{Lawyer}(x))$
n_2. $\forall x(\texttt{RentsInHollywood}(x) \supset \texttt{LivesInHollywood}(x))$

The following relevant arguments can be constructed:

A_1: $\texttt{PublicDefender}(L)$
A_2: $\forall x(\texttt{PublicDefender}(x) \supset \texttt{Lawyer}(x))$
A_3: $A_1, A_2 \rightarrow \texttt{Lawyer}(L)$
A_4: $A_3 \Rightarrow \texttt{Rich}(L)$

B_1: $\texttt{PublicDefender}(L)$
B_2: $B_1 \Rightarrow \neg\texttt{Rich}(L)$

C_1: $\texttt{RentsInHollywood}(L)$
C_2: $\forall x(\texttt{RentsInHollywood}(x) \supset \texttt{LivesInHollywood}(x))$
C_3: $C_1, C_2 \rightarrow \texttt{LivesInHollywood}(L)$
C_4: $C_3 \Rightarrow \texttt{Rich}(L)$

D_1: $\texttt{RentsInHollywood}(L)$
D_2: $B_1 \Rightarrow \neg\texttt{Rich}(L)$

E_1: $\texttt{PublicDefender}(L)$
E_2: $E_1 \Rightarrow \neg d_1(L)$

F_1: $\texttt{RentsInHollywood}(L)$
F_2: $F_1 \Rightarrow \neg d_2(L)$

Let us apply preferred semantics (but in grounded semantics the outcome is the same). Note first that E_2 undercuts A_4 and F_2 undercuts C_4. Moreover, neither E_2 nor F_2 has a defeater, so both of them are in all preferred extensions. But then A_4 and C_4 are not in any preferred extension, so that B_2 and D_2 are in all these extensions. So the conclusion $\neg\texttt{Rich}(L)$ is justified.

EXERCISE 4.8.8.

1. C_2 rebuts D_2 and not vice versa. Since both arguments use defeasible rules and no preference relations hold between them, C_2 successfully rebuts and therefore defeats D_2. Argument C_2 in turn has two defeaters: its subarguments A_2 and B_2 defeat each other and thus also defeat C_2. Since there are no undefeated arguments that defeat A_2 or B_2, none of A_2, B_2, C_2 and D_2 are in the grounded extension. (In terms of status assignments: it is possible to give none of them a status so in the grounded extension, which maximises undecidedness, none of them have a status.) However, none of these arguments are defeated by an argument that is in the grounded extension, so they are all defensible.

2. Note that A_2 can be made *in* if B_2 is made *out* and vice versa. Then at least one preferred status assignment makes A_2 *in* and B_2 *out*, since such assignments minimise undecidedness. But since A_2 defeats C_2, this assignment also makes C_2 *out*. But then it makes D_2 *in*, since its only defeater is C_2. Conversely, a second preferred status assignment makes B_2 *in* and A_2 *out* so it also makes C_2 *out* and D_2 *in*. Since there are no other preferred status assignments, in all such assigments C_2 is *out* and D_2 is *in*. But then C_2 is overruled and D_2 is justified.

EXERCISE 4.8.9.

1. $Cl_{tp}(R_s) = R_s \cup \{-q \to -p;\ -r \to -p;\ p, -s \to -r;\ r, -s \to -p\}$.
2. Yes.
3. No.

EXERCISE 4.8.10. The point of this exercise is that closure under contraposition does not imply closure under transposition.

1. No: $\mathcal{R}_s$ contains $p \to q$ but not $\neg q \to \neg p$.
2. Yes. We have:

   $\{p\} \vdash q$ and $\{\neg q\} \vdash \neg p$
   $\{p\} \vdash \neg r$ and $\{r\} \vdash \neg p$
   $\{\neg r\} \vdash q$ and $\{\neg q\} \vdash r$
   $\{\neg q\} \vdash r$ and $\{\neg r\} \vdash q$

   So an argumentation theory with $\mathcal{R}_s$ satisfies contraposition.

EXERCISE 4.8.11.

1. The following argument for t can be created.

   A_1: s
   A_2: $A_1 \Rightarrow t$

   A_2 is rebutted by the following argument for $\neg t$:

   B_1: p
   B_2: $B_1 \Rightarrow q$
   B_3: $B_1, B_2 \Rightarrow r$
   B_4: $B_2, B_3 \to \neg t$

   (Note that since B_4 is strict, A_2 does not in turn rebut B_4.) We have that $\texttt{LastDefRules}(A_2) = \{d_3\}$ while $\texttt{LastDefRules}(B_4) = \{d_1, d_2\}$. Since $d_2 < d_3$ we have that $\texttt{LastDefRules}(B_4) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(A_2)$, so $B_4 \prec A_2$. Hence B_4 does not defeat A_2. Since A_2 has no other defeaters, we can concude at this point that A_2 will be *in* in all preferred status assignments, which makes it justified. Then t is a justified conclusion.

   It is interesting to verify the status of argument B_4 for $\neg t$. Since the present argumentation theory is well defined, it is to be expected that this conclusion is not justified. This turns out to be indeed the case. First of all, A_2 can be extended to a rebuttal of B_3 by using one of the transpositions of the strict rule:

   $A_3 = B_1$: p
   $A_4 = B_2$: $A_3 \Rightarrow q$
   A_5: $A_2, A_4 \to \neg r$

   We have that $\texttt{LastDefRules}(A_5) = \{d_1, d_3\}$ while $\texttt{LastDefRules}(B_3) = \{d_2\}$. Since $<$ is transitive we have $d_2 < d_1$ so $\{d_2\} \lhd_{\texttt{Eli}} \{d_1, d_3\}$ and $B_3 \prec A_5$.

Hence A_5 successfully rebuts and thus strictly defeats B_3. But then A_5 also defeats B_4.

Yet another relevant argument can be constructed, which starts in the same way as A_5 but uses the other transposition of the strict rule:

$A_3 = B_1$: p
$A_4 = B_2$: $A_3 \Rightarrow q$
$A_6 = B_3$: $A_3, A_4 \Rightarrow r$
A_7: $A_2, A_6 \rightarrow \neg q$

A_7 rebuts B_2 (and not vice versa). We have $\texttt{LastDefRules}(A_7) = \{d_2, d_3\}$ while $\texttt{LastDefRules}(B_2) = \{d_1\}$. Since $d_2 < d_1$ so $A_7 < B_2$ we have that A_7 does not defeat B_2. Since $A_6 = B_2$ we also have that A_7 does not defeat A_6. Finally, A_5 rebuts A_6. Recall that $\texttt{LastDefRules}(A_5) = \{d_1, d_3\}$; moreover, $\texttt{LastDefRules}(A_6) = \{d_2\}$ and we have seen that $\{d_2\} \lhd_{\texttt{Eli}} \{d_1, d_3\}$ so $A_6 \prec A_5$, for which reason A_5 strictly defeats A_6.

Now to evaluate the status of the arguments, A_5 and all its subarguments can be made *in* since they have no defeaters. Since A_5 strictly defeats A_6 and thus also A_7, the latter two arguments can be made *out*. Moreover since A_5 strictly defeats B_3 and thus also B_4, the latter two arguments can also be made *out*. No alternative status assignments are possible, while moreover the present assignment is complete. So B_4 is out in all preferred status assigments, which makes $\neg t$ an overruled conclusion.

2.
- $\texttt{Prem}(A_1) = \{s\}$, $\texttt{Conc}(A_1) = s$, $\texttt{Sub}(A_1) = \{A_1\}$, $\texttt{DefRules}(A_1) = \varnothing$ and $\texttt{TopRule}(A_1) =$ undefined.
- $\texttt{Prem}(A_2) = \{s\}$, $\texttt{Conc}(A_2) = t$, $\texttt{Sub}(A_2) = \{A_1, A_2\}$, $\texttt{DefRules}(A_2) = \{d_3\}$, $\texttt{LastDefRules}(A_2) = \{d_3\}$ and $\texttt{TopRule}(A_2) = d_3$.
- $\texttt{Prem}(A_3) = \{p\}$, $\texttt{Conc}(A_3) = p$, $\texttt{Sub}(A_3) = \{A_3\}$, $\texttt{DefRules}(A_3) = \varnothing$ and $\texttt{TopRule}(A_3) =$ undefined.
- $\texttt{Prem}(A_4) = \{p\}$, $\texttt{Conc}(A_4) = q$, $\texttt{Sub}(A_4) = \{A_3, A_4\}$, $\texttt{DefRules}(A_4) = \{d_1\}$, $\texttt{LastDefRules}(A_4) = \{d_1\}$ and $\texttt{TopRule}(A_4) = d_1$.
- $\texttt{Prem}(A_5) = \{s, p\}$, $\texttt{Conc}(A_5) = \neg r$, $\texttt{Sub}(A_5) = \{A_1, A_2, A_3, A_4, A_5\}$, $\texttt{DefRules}(A_5) = \{d_1, d_3\}$, $\texttt{LastDefRules}(A_5) = \{d_1, d_3\}$ and $\texttt{TopRule}(A_5) = t, q \rightarrow \neg r$.
- $\texttt{Prem}(A_6) = \{p\}$, $\texttt{Conc}(A_6) = r$, $\texttt{Sub}(A_6) = \{A_3, A_4, A_6\}$, $\texttt{DefRules}(A_6) = \{d_1, d_2\}$, $\texttt{LastDefRules}(A_6) = \{d_2\}$ and $\texttt{TopRule}(A_6) = d_2$.
- $\texttt{Prem}(A_7) = \{s, p\}$, $\texttt{Conc}(A_7) = \neg q$, $\texttt{Sub}(A_7) = \{A_1, A_2, A_3, A_4, A_7\}$, $\texttt{DefRules}(A_7) = \{d_1, d_2, d_3\}$, $\texttt{LastDefRules}(A_7) = \{d_2, d_3\}$ and $\texttt{TopRule}(A_7) = t, r \rightarrow \neg q$.
- B_1, B_2, B_3 equal A_3, A_4, A_6.
- $\texttt{Prem}(B_4) = \{p\}$, $\texttt{Conc}(B_4) = \neg t$, $\texttt{Sub}(B_4) = \{B_1, B_2, B_3, B_4\}$, $\texttt{DefRules}(B_4) = \{d_1, d_2\}$, $\texttt{LastDefRules}(B_4) = \{d_1, d_2\}$ and $\texttt{TopRule}(B_4) = q, r \rightarrow \neg t$.

EXERCISE 4.8.12.

(a): $A2 = [[q \Rightarrow t] \Rightarrow s], \neg r \vee \neg t \rightarrow \neg r$.

(b). Overruled. Argument* A_2 for $\neg r$ is rebutted on its subargument* A_1 for t by argument* $B_2 = [p \Rightarrow r], \neg r \vee \neg t \rightarrow \neg t$. We have that $\texttt{DR}(B_2) = \{d_1\}$ while $\texttt{DR}(A_1) = \{d_2, d_3\}$. Since $d_2 < d_1$, we have that $\texttt{DR}(A_1) < \texttt{DR}(B_2)$, so $A_1 \prec B_2$, so B_2 strictly defeats A_2 on B_1. Since B_2 has no defeaters, O has a winning strategy in the G-game, so A_2 is not justified. Moreover, for the same reasons B_2 is justified. So A_2 is overruled. Since there are no other arguments* for $\neg r$, we have that $\neg r$ is overruled as well. (Note that the argument* that can be constructed for r does not attack A_2).

(c) Now we have that $\texttt{LDR}(A_1) = \{d_3\}$ while $\texttt{LDR}(B_2) = \{d_1\}$. Since $d_3 < d_1$, we have that $\texttt{LDR}(B_2) < \texttt{LDR}(A_1)$, so $B_2 \prec A_1$, so A_2 has no defeaters, so A_2 is justified, so $\neg t$ is justified.

EXERCISE 4.8.13.

1. The arguments (shown in Figure 9.7, with their conclusions at the bottom) are:

   A' = a,
   A = $A' \Rightarrow p$,
   B_1 = $\sim s$,
   B_1' = $B_1 \Rightarrow t$,
   B_2 = r,
   B_2' = $B_2 \Rightarrow q$,
   B = B_1',
   $B_2' \rightarrow \neg p$,
   C = $[\neg r]$.

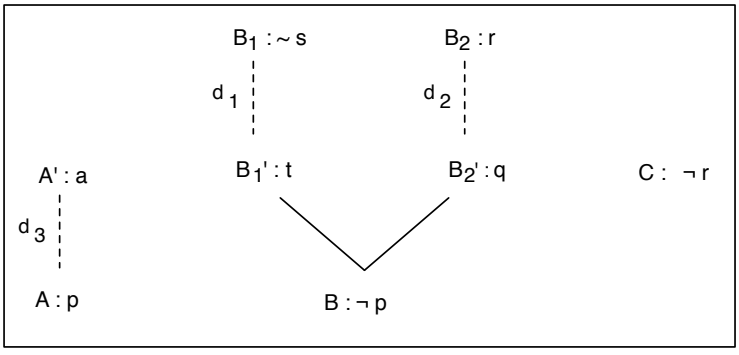


Figure 9.7: *ASPIC*$^+$ arguments and their conclusions, with dashed and solid lines respectively representing application of defeasible and strict inference rules.

2. B rebuts A on A, C undermines B and B_2' on B_2, and C and B_2 undermine each other. Note that A does not rebut B since B has a strict top rule.

3. We have that $\texttt{LastDefRules}(B) = \{d_1, d_2\}$ and $\texttt{LastDefRules}(A) = \{d_3\}$ and since $d_2 < d_3$, we have that $\texttt{LastDefRules}(B) \lhd_{\texttt{Eli}} \texttt{LastDefRules}(A)$.

So B does not defeat A. Moreover, we have that $\texttt{Prem}_p(C) = \{\neg r\}$ while $\texttt{Prem}_p(B_2) = \{r\}$ and $\neg r <' r$, so we also have that $\texttt{Prem}_p() \lhd_{\texttt{Eli}} \texttt{Prem}_p(B_2)$. So $C \prec B_2$ so B_2 strictly defeats C.

4. The transpositions are $p, t \rightarrow \neg q$ and $p, q \rightarrow \neg t$. This yields two new arguments:

   D = $A, B_1' \rightarrow \neg q$,
   E = $A, B_2' \rightarrow \neg t$.

   D rebuts B_2' while E rebuts B_1'.

   We have that $\texttt{LastDefRules}(D) = \{d_1, d_3\}$ and $\texttt{LastDefRules}(B_2') = \{d_2\}$ and since $d_1 \not\lhd d_2$ and $d_2 \not\lhd d_1$, we have that these sets are incomparable in the $\lhd_{\texttt{Eli}}$ ordering. So D defeats B_2'.

   Moreover, we have that $\texttt{LastDefRules}(E) = \{d_2, d_3\}$ and $\texttt{LastDefRules}(B_1') = \{d_1\}$ and since $d_1 \not\lhd d_2$ and $d_2 \not\lhd d_1$, we have that these sets are incomparable in the $\lhd_{\texttt{Eli}}$ ordering. So E defeats B_1'.

EXERCISE 4.8.14.

1. The arguments are

   A_1: $\sim a$
   A_2: $A_1 \rightarrow b$
   A_3: $A_2 \Rightarrow \neg c$

   B_1: $\Rightarrow c$
   B_2: $B_1 \Rightarrow a$

   Argument B_2 contrary-undermines A_1, A_2 and A_3 on A_1. Argument A_3 rebuts B_1 and B_2 on B_1. Finally, B_1 rebuts A_3 on A_3.

2. The attack of B_2 on A_1, A_2 and A_3 succeeds since contrary undermining is a preference-independent form of attack. Moreover, we have that $\texttt{DefRules}(B_1) = \{d_2\}$ and $\texttt{DefRules}(A_3) = \{d_1\}$ and since $d_2 < d_1$, we have that $\texttt{DefRules}(B_1) \lhd_{\texttt{Eli}} \texttt{DefRules}(A_3)$, so $B_1 \prec A_3$. So A_3 strictly defeats B_1 and B_2.

3. The grounded extension is empty, since there are no undefeated arguments.

4. There are two preferred extensions: the first is $\{A_1, A_2, A_3\}$ while the second is $\{B_1, B_2\}$.

EXERCISE 4.8.15.

1. No. We explain this with the G-game. There is an argument* for $guilty$, namely

   $A = murder$, $murder \supset guilty \rightarrow guilty$.

   Argument* A has two strict defeaters, namely:

$B = \neg ab, \neg ab \supset \neg guilty, murder \supset guilty \rightarrow \neg murder$

$C = \neg ab, \neg ab \supset \neg guilty, murder \rightarrow \neg(murder \supset guilty)$

Since $\mathcal{K}_p$ is minimally inconsistent (i.e., taking any element out makes $\mathcal{K}_p$ consistent), both B and C have underminers on any of their premises: these underminers can be formed by replacing the attacked premise with the remaining one. Since the argument ordering is simple, all these undermining attacks succeed as defeats. For example, B is defeated on $\neg ab$ by

$D = murder, \neg ab \supset \neg guilty, murder \supset guilty \rightarrow ab$

In the same way, any further argument* moved in a G-game has defeaters, so the proponent does not have a winning strategy for A.

2. Any argument ordering in which $\neg ab$ is inferior to all other formulas in $\mathcal{K}_p$ will do, since then neither B not C defeats A, so the proponent wins the G-game after moving A.

3. Move all formulas except $\neg ab$ to $\mathcal{K}_n$. Then argument* A has no attackers since all its premises are necessary.

9.5 Exercises Chapter 5

Exercise 5.4.1:

1. We have the same arguments as in Exercise 4.8.5. Argument B_2 attacks C_3 and since $C_3 \prec B_2$ by the last-link ordering, B_2 also defeats C_3. Moreover, B_3 attacks A_2 and A_3 by directly undercutting A_2. However, since $d_2 < d_1$ we have in the last-link ordering that $B_3 \prec A_2$ and $B_3 \prec A_3$. Hence B_3 does not defeat A_2 or A_3. So the only defeat relation in the PAF is that B_2 defeats C_3.

2. A_3 has no defeaters so r is justified.

Exercise 5.4.2:

1. We have the arguments $A_1, A_2, A_3, B_1, B_2, B_3, C_1, C_2, D_3, D_4$ from Example 4.3.6. We have the following attack and defeat relations.

   B_2 and D_4 attack each other. Since $d_2 < d_5$ we have that $B_2 \prec D_4$ so D_4 strictly defeats B_2.
   D_4 attacks B_3 (by directly attacking B_2). Since $d_5 < d_3$ we have that $D_4 \prec B_3$ so D_4 does not defeat B_3.
   B_3 attacks A_2 and A_3 (by undercutting A_2). Since the priority relation between d_1 and d_3 is undefined, the preference relation between these arguments is also undefined. So B_3 defeats A_2 and A_3.

2. The argument A_3 for r has an undefeated defeater, namely, B_3, so r is overruled.

3. In *ASPIC*$^+$ we have that D_4 strictly defeats B_3 since it strictly defeats its subargument B_2. Then A_3's only defeater has an undefeated defeater, so clearly r is justified.

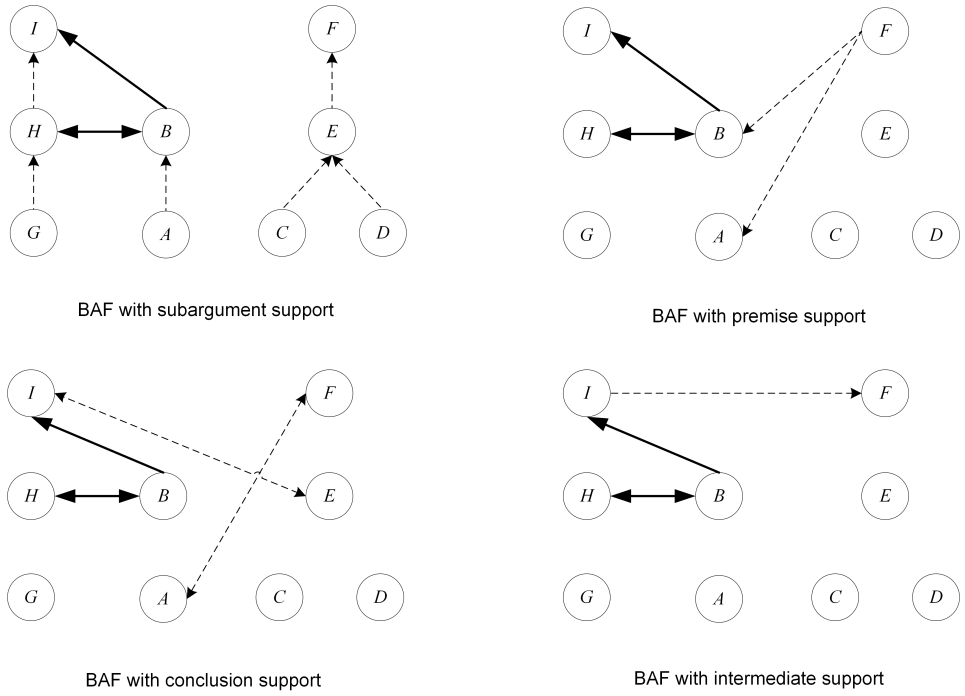


Figure 9.8: *BAFs* for the *ASPIC*$^+$ *SAF*

Exercise 5.4.3:

1. See Figure 9.8, where the support relations are displayed with dashed arrows. The *BAF* with subargument support leaves the transitive closure of support implicit.

2. See the following table for the added defeats.

	Secondary	Extended	Supported	Mediated
Subargument support		$I \to B$	$G \to B$, $A \to H$ $A \to I$	$B \to G$, $H \to A$
Premise support			$F \to H$, $F \to I$	$H \to F$
Conclusion support	$B \to E$			$B \to E$
Intermediate support	$B \to F$			

3. See the following table for the added defeats.

	General	Deductive	Necessary
Subargument support	$G \to B$, $A \to H$ $A \to I$	$G \leftrightarrow B$, $A \leftrightarrow H$, $A \to I$	$B \leftrightarrow I$
Premise support	$F \to H$	$F \leftrightarrow H$	
Conclusion support	$B \to E$		not definable
Intermediate support	$B \to F$		$B \to F$

The reason why for conclusion support the semantics of necessary support is not definable is that in the example the conclusion-support relation cannot be both irreflexive and transitive.

4. See the following table.

	in	*out*	*undecided*
Subargument + general:	A, C, D, E, F, G	B, H, I	
Subargument + deductive:	C, D, E, F		A, B, G, H, I
Subargument + necessary:	A, C, D, E, F, G		B, H, I
Premise + general:	$A, B, C, D, E, F, G,$	H, I	
Premise + deductive:	$A, C, D, E, G,$		B, F, H, I
Premise + necessary:	A, C, D, E, F, G		B, H, I
Conclusion + general:	$A, C, D, E, F, G,$		B, H, I
Conclusion + deductive:	A, C, D, E, F, G		B, H, I
Conclusion + necessary:	$-$	$-$	$-$
Intermediate + general:	A, C, D, E, G		B, F, H, I
Intermediate + deductive:	A, C, D, F, G		B, H, I
Intermediate + necessary:	A, C, D, E, G		B, F, H, I

9.6 Exercises Chapter 6

Exercise 6.6.1:

1. Yes, since there is just one preferred extension, namely, $\{B\}$.

2. No. if the defeat from B to A is deleted, then the preferred extension is empty.

Exercise 6.6.2:

1. In (a) D is justified in all full resolutions. One full resolution deletes the defeat from B to C and another full resolution deletes the defeat from C to B. In both cases the grounded extension is $\{A, D\}$.

   In (b) D is justified in some but not all full resolutions. Any full resolution which deletes the defeat from A to D makes D a member of the grounded extension. But a full resolution that deletes the defeats from D to A and B to A makes instead A a member of the grounded extension.

   In (e) D is also justified in some but not all full resolutions. If the defeat from C to B is deleted, then D is in the grounded extension but if the defeat from B to C is deleted then instead C is in the grounded extension.

2. All answers are the same for preferred semantics.

Exercise 6.6.3:

a:
A defeats A, B defeats A, B defeats C
A defeats A, B defeats A, C defeats B
A defeats A, A defeats B, B defeats C

A defeats A, A defeats B, C defeats B

b: the first one (which has one stable extension) and any of the other three, since these have no stable extensions.

Exercise 6.6.4: Nothing changes. With the elitist last-link ordering the attack of B_3 on A_2 is preference independent. Moreover, no preference can change the preference relation between B_1 and C_3 since $\texttt{LDR}(C_3) \neq \varnothing$ while $\texttt{LDR}(B_1) = \varnothing$. Finally, the preference relation between B_2 and D_4 depends on d_2 and d_5 but it is already given that $d_2 < d_5$ so no further preference can change this. With elitist weakest-link two comparisons are relevant. The first is between $\{d_2\}$ and $\{d_4, d_5\}$ for B_2 versus D_4, where $d_4 < d_2$ is already given and determines the preference between B_2 and D_4, which a preference extension cannot change. The second comparison is between u and s for C_3 versus B_1, where $u <' s$ is already given, so again no further preference can change this.

Exercise 6.6.5:

1. Yes. A is undermined by $B = p, q \rightarrow (p \wedge q)$. We have $\texttt{Prem}_p(A) = \{\neg(p \wedge q)\}$ while $\texttt{Prem}_p(B) = \{p, q\}$ and since $p <' \neg(p \wedge q)$ but the relation between q and $\neg(p \wedge q)$ is undefined, we have that $\texttt{Prem}_p(B) \lhd_{\texttt{Eli}} \texttt{Prem}_p(A)$. So $B \prec A$, so B does not defeat A. Since A has no other defeaters, B is in the grounded extension.

   Note that we have the following orderings between the various premise sets:

   $\{p, q\} \lhd_{\texttt{Eli}} \{\neg(p \wedge q)\}$

   $\{p\} \lhd_{\texttt{Eli}} \{q, \neg(p \wedge q)\}$

   $\{p, \neg(p \wedge q)\} \lhd_{\texttt{Eli}} \{q\}$

2. There are three ways to extend $\leq$: with $q <' \neg(p \wedge q)$, with $\neg(p \wedge q) <' q$ and with $q \approx' \neg(p \wedge q)$. In all three cases the above orderings between the various premise sets does not change. So the set $\mathcal{D}$ of defeat relations does not change, so there exists no full preference-based resolution. So the answer is 'yes'.

Exercise 6.6.6:

For (a) consider a UAT with $\mathcal{L}^u = \{p, \neg p, d_1, \neg d_1, d_2, \neg d_2, d_3, \neg d_3\}$, with $\mathcal{K}^u = \{\neg d_1\}$, with $\mathcal{R}^u_s = \varnothing$ and with $\mathcal{R}^d$ consisting of the rules in the following arguments:

$A:\quad \neg d_1$
$B:\quad \Rightarrow_{d2} p$
$C:\quad \Rightarrow_{d3} \neg d_2$
$D:\quad \Rightarrow_{d1} \neg d_3$

For (b) one of the many examples is with $\mathcal{L}^u = \{p, \neg p, q, \neg q\}$, with $\mathcal{K}^u_n = \{\neg p\}$ and $\mathcal{K}^u_p = \{p, q\}$, and with $\mathcal{R}^u_s = \mathcal{R}^d = \varnothing$. Then

$A:\quad p$
$B:\quad \neg p$
$C:\quad q$

For (c), one example is with the language and rules as specified in the following arguments:

$A: \Rightarrow_{d1} p$ $\quad B: \Rightarrow_{d2} \neg d_1$ $\quad F: \Rightarrow_{d3} \neg d_2$
$C: \Rightarrow_{d4} q$ $\quad D: \Rightarrow_{d5} \neg d_4$ $\quad G: \Rightarrow_{d6} \neg d_5$
$\quad E: \Rightarrow_{d7} \neg d_6$

Exercise 6.6.7:

1. Yes, for instance, adding E as strict defeater of A.

2. No, since any defeater of A also defeats B, So if A is *out* then B is *out*.

Exercise 6.6.8:

1. Yes. AF does not delete arguments and defeat relations, $\mathcal{D} \neq \varnothing$ and the new defeat relation involves one new argument.
   (b): No, since adding C to $\mathcal{A}$ involves adding r to $\mathcal{K}_p$ which induces a new argument $D = r$ (it also involves adding a preference $A \prec C$ but that is irrelevant here). Moreover, an indirect defeat relation from C to B must be added.

Exercise 6.6.9:

1. There are seven arguments:

   A_1: p $\quad$ A_2: $A_1 \Rightarrow_{r1} q$ $\quad$ A_3: $A_2 \Rightarrow_{r2} r$
   B_3: $A_2 \Rightarrow_{r4} t$
   C_1: s $\quad$ C_2: $C_1 \Rightarrow_{r3} \neg r$
   D_1: u $\quad$ D_2: $D_1 \Rightarrow_{r5} \neg t$

   A_3 and C_2 defeat each other;
   B_3 and D_2 defeat each other.

2.

   (a) No, since also an argument $F : \neg u \Rightarrow v$ and an indirect defeat relation (E, D_2) must be added.

   (b) No. Note first that all allowed expansions have to add $\neg u$ to $\mathcal{K}'_p$ since expansions have to contain at least one new argument. Moreover, to change defeat relations also rule preferences must be added. Making the argument for $\neg r$ justified requires $r_1 < r_3$ and making the argument for t justified requires $r_5 < r_1$ and $r_5 < r_4$. These preferences cannot all be added together since $\preceq'$ has to remain a weakest-link ordering and $r_1 < r_3$ and $r_3 < r_5$ together imply $r_1 < r_5$.

9.7 Exercises Chapter 7

EXERCISE 7.6.1

1. For example:

   $P_1 =$ *claim* q
   $O_1 =$ *why* q
   $P_2 = q$ *since* $p, p \supset q$
   $O_2 = \neg p$ *since* $r, r \supset \neg p$
   $P_3 = \neg(r \supset \neg p)$ *since* p, r
   $O_3 =$ *concede* r

   Note that this example illustrates a limitation of the protocol, namely, that O is not allowed to reply to the other premise of P_3.

2. For example:

   $P_1 =$ *claim* q
   $O_1 =$ *why* q
   $P_2 = q$ *since* $p, p \supset q$
   $O_2 = \neg p$ *since* $r, r \supset \neg p$
   $P_3 = \neg(r \supset \neg p)$ *since* p, r
   $O_3 =$ *why* r
   $P_4 =$ *retract* r

EXERCISE 7.6.2

1. O can only reply with *concede* p since, whatever O's acceptance attitude, after P's first move O must reason from $\mathcal{K}_O \cup \{p\}$ so O can construct the trivial argument p *since* p. Here the dialogue terminates.

   This example illustrates that the fact that the players must reason with the commitments of the other player makes that they can learn from each other.

2. Now O can, whatever its acceptance attitude, only reply with *why* p, since O cannot create an allowed argument so cannot concede p or claim $\neg p$.

3. For (1) the answer is the same. This illustrates that the same learning mechanism as in (1) sometimes makes agents learn too easily, since p is not justified by the agents' joint knowledge bases. For (2), O again replies with *why* p.

EXERCISE 7.6.3

1. O can have all all three assertion attitudes but P cannot have any of the three acceptance attitudes since P cannot construct an argument for $\neg q \wedge r$.

   O can have any acceptance attitude. Furthermore, O must move O_1 under any combination of attitudes, since O cannot move an argument for p or for $\neg p$. But O_2 can only be moved if O is confident or careful, since O's argument for $\neg(p \wedge q)$ is not justified given $\mathcal{K}_O$.

2. The protocol is unsound since at termination both players are committed to $\neg q \wedge r$ while on the basis of their joint knowledge bases this conclusion is only defensible.

EXERCISE 7.6.4

1. - p is overruled. It has one argument, viz. $A_1 = (\{q, q \supset p\}, p)$, which has two attackers, viz. $A_2 = (\{\neg p, q \supset p\}, \neg q)$ and $A_3 = (\{q, \neg p\}, \neg(q \supset p))$. We have that A_2 does not defeat A_1 on q since $q \supset p < q$. However, A_3 strictly defeats A_1 on $q \supset p$ since $q \supset p < q$ and $q \supset p < \neg p$. Next, A_3 has two attackers, viz. A_1 and A_2. We have that A_1 does not defeat A_3 on $\neg p$ since $q \supset p < \neg p$. Moreover, A_2 does not defeat A_3 on q since $q \supset p < q$. So neither A_1 nor A_2 defeats A_3. Since A_3 has no other attackers, A_3 is justified and A_1 is overruled. Since p has no other arguments besides A_1, p is also overruled.

   - $\neg p$ is justified. The argument $A_4 = (\{\neg p\}, \neg p)$ has one attacker, viz. A_1, but $A_1 \prec A_4$ since $q \supset p < \neg p$, so A_1 does not defeat A_4. So A_4 has no defeaters and is justified, which makes $\neg p$ justified.

   - $q \supset p$ is defensible. It has one argument, viz. $A_5 = (\{q \supset p\}, q \supset p)$, which has one attacker, viz. A_3. We have that $A_5 \prec A_3$ since $q \subset p < \neg p$, so A_3 strictly defeats A_5. Since we saw under (1) that A_3 is justified, A_5 is overruled so $q \supset p$ is overruled.

2. There is just one legal dialogue, viz.

   W_1: *claim* p $\quad$ B_1: *why* p
   W_2: *claim* $\{q, q \supset p\}$ $\quad$ B_2: *concede* q; B_3: *why* $q \supset p$
   W_4: *claim* $\{q \supset p\}$

Let us explain why. At his first move, W must reason with Σ_W, which contains a justified argument for p (it has no attackers on the basis of Σ_W). Then B at her first move must reason with $\Sigma_B \cup \{p\}$. Then B cannot concede p: although she can construct an argument for p, viz. $A_6 = (\{p\}, p)$, it is not justified: B can construct $A_4 = (\{\neg p\}, \neg p)$, which symmetrically defeats A_6 since $p \approx \neg p$ and is not defeated by other arguments on the basis of $\Sigma_B \cup \{p\}$. Can B claim $\neg p$? No, since her only argument for $\neg p$ is A_4, which is defeated by A_6 and since these arguments have no other attackers on the basis of $\Sigma_B \cup \{p\}$, they are both defensible on the basis of $\Sigma_B \cup \{p\}$. So B must challenge p.

After W's reply with W_2 the information with which B must reason is $\Sigma_B \cup \{p, q, q \supset p\}$. On this basis B's only argument for $\neg q$ is $A_2 = (\{\neg p, q \supset p\}, \neg q)$ but $A_7 = (\{q\}, q)$ is also constructible on this basis and A_7 strictly defeats A_2 since $\neg p < q$ and $q \supset p < q$. So A_7 is justified and B must concede q.

Next, B has an argument against W_2's second premise, viz. A_3, but B can also construct an attacker $A_6 = (\{p\}, p)$ of A_3. Since $p \approx \neg p$, this attack succeeds as defeat, so A_3 is not justified on the basis of $\Sigma_B \cup \{p\}$. So B must challenge. Then W claims $\{q \supset p\}$ and the dialogue terminates without agreement.

At termination, the commitment sets are:

$C_W = \{p, q, q \supset p)\}$, which is consistent;
$C_B = \{q\}$, which is consistent.

On the basis of $\Sigma_W \cup C_W$ we have that p is justified: we have two arguments A_1 and $A_6 = (\{p\}, p)$, for p, which both have no defeater. On the basis of $\Sigma_B \cup C_B$ we have that $\neg p$ is justified since it has a justified argument A_4 (since its only attacker is A_1 and $A_1 \prec A_4$).

In sum, even though on the basis of the players' joint beliefs p is overruled and $\neg p$ is justified, the players do not reach agreement on p.

3. The only legal dialogue now is

W_1: *claim* p $\qquad$ B_1: *claim* $\neg p$

Here the dialogue terminates since W cannot repeat *claim* p. At termination W is committed to p and B to $\neg p$. These sets are both internally consistent and consistent with the agents' own beliefs. Finally, p is justified on the basis of $\Sigma_W \cup C_W$ while $\neg p$ is justified on the basis of $\Sigma_B \cup C_B$.

EXERCISE 7.6.5

1. The only legal dialogue is

W_1: *claim* p $\qquad$ B_1: *why* p
W_2: *claim* $\{q, q \supset p\}$ $\qquad$ B_2: *why* q
W_3: *claim* q

This dialogue terminates without agreement since B is not allowed to repeat *why* q. So B has learned nothing from W.

2. Any player can accept a proposition φ after a *claim* $\{\varphi\}$ move of the other player that was moved after a *why* φ move, provided that the player cannot construct an argument for $\neg\varphi$.

EXERCISE 7.6.6 Assuming the above answer to 7.6.5(2), the only legal dialogue is

W_1: *claim* r	B_1: *why* r
W_2: *claim* $\{p, p \supset q, q \supset r\}$	B_2: *why* p
W_3: *claim* $\{p\}$	B_3: *concede* p, B_4: *claim* $\neg(p \supset q)$
W_4: *claim* $p \supset q$	B_5: *concede* $q \supset r$

This exercise illustrates a number of subtle features of the PWA protocol. Note first that black could make his counterclaim only after first conceding p! Next, at B_5 black could not claim $\neg(p \supset q)$ even though that is allowed by her assertion attitude, since this claim repeats B_4.[1]. Finally, the reason why black must concede $q \supset r$ is that she has a justified argument for it with premises $\{s, s \supset \neg q\}$, which implies not only $\neg q$ but also $q \supset r$ for any r!

EXERCISE 7.6.7

1. A counterexample is Example 7.4.3.

2. A counterexample is $\Sigma_W = \{p, p \supset q, r\}$ and $\Sigma_B = \{r \supset \neg p\}$, with topic q and all formulas of the same preference level. The only legal dialogue on q is:

W_1: *claim* q	B_1: *why* q
W_2: *claim* $\{p, p \supset q\}$	B_2: *why* p
W_3: *claim* $\{p\}$	B_3: *concede* p, B_4: *why* $p \supset q$
W_4: *claim* $\{p \supset q\}$	B_5: *concede* $\{p \supset q\}$

   At termination, we have that $C_W \vdash q$ and $C_B \vdash q$ but q is not justified on the basis of $\Sigma_W \cup \Sigma_B$ because of the counterargument $(\{r, r \supset \neg p\}, \neg p)$.

EXERCISE 7.6.8 No. After $P_1 =$ *claim* q, O cannot construct an argument for or against q, so O must challenge. Then P must reply with *claim* $\{p, p \supset q\}$. Then O has a justified argument against premise p, namely, $\{\neg p\}$, so O moves *claim* $\neg q$. By contrast, O has a justified argument for $p \supset q$, namely, the argument for $\neg p$. Then P moves *claim* p since P has a trivial justified argument for p, after which O cannot repeat *claim* $\{\neg p\}$ and the dialogue terminates without agreement:

$W_1 =$ *claim* q
$B_1 =$ *why* q
$W_2 =$ *claim* $\{p \supset q, p\}$
$B_{2b} =$ *concede* $p \supset q$
$B_{2a} =$ *claim* $\neg p$
$W_{3a} =$ *claim* p

EXERCISE 7.6.9. This follows from result (2) of Section 2.4 of the reader, which implies that finite defeat graphs without cycles have a unique status assignments. (Note

[1] When read literally, PWA's termination condition "when the move required by the procedure cannot be made" implies that the dialogue terminates here, but we read it as meaning that only the 'sub-dialogue' about the first premise of W_2 terminates and the dialogue then continues about the second premise.

that dialogue trees have no such cycles through their reply relations.)

EXERCISE 7.6.10. A surrendered move is *in* by definition regardless of its other replies, so a new reply can never change any dialogical status.

EXERCISE 7.6.11

1. P_1, P_3 and P_7
2. O_2 and O_8
3. P_1, P_9

EXERCISE 7.6.12 For example:

$P_1 =$ *claim* q
$O_1 =$ *why* q
$P_2 = q$ *since* $p, p \supset q$
$O_2 = \neg p$ *since* $r, r \supset \neg p$
$P_3 = \neg(r \supset \neg p)$ *since* p, r
$O_3 =$ *why* r
$P_4 =$ *retract* q
or alternatively
$P_3 =$ *why* r
$O_3 = r$ *since* r
$P_4 =$ *retract* q

There are other examples.

EXERCISE 7.6.13 For example:

$P_1 =$ *claim* q
$O_1 =$ *why* q
$P_2 = q$ *since* $p \wedge q$
$O_2 = \neg(p \wedge q)$ *since* $\neg p$
$P_3 =$ *why* $\neg p$
$O_3 = \neg p$ *since* $\neg p$
$P_4 =$ *retract* q
There are other examples. Note that P cannot attack O_2 or O_3 with p *since* $p \wedge q$, since that argument does not defeat $O's$ arguments.

9.8 Exercises Chapter 8

EXERCISE 8.3.1.

1. The following precedents are citable for the plaintiff: Boeing, Bryce, College-Wat, Den-Tel-Ez, Emery, Space Aero, Televation (not Ferranti, Ecologix, since these were won by defendant)
2. The following are all counterexamples for the defendant against Emery, namely, all precedents citable by the defendant: Arco, Ecologix, Sandlin, Yokana.

3. Arco can be distinguished, since it contains none of the p-factors from the new case and since d-factors 10 en 20 are not in the new case.
   Ecologix can be distinguished on lacking the p-factors 2,4 and 15 in the new case and the d-factors 1 en 19 from the precedent.
   Sandlin can be distinguished since it lacks all p-factors from the new case and since the new case lacks the d-factors 1, 10, 19, 27 from the precedent.
   Yokana can be distinguished since it lacks all p-factors of the new case and since the new case lacks the d-factors 10 en 27 of the precedent.

4. Plaintiff cites Boeing (or Bryce, Den-Tal-Ez, Television), defendant distinguishes since the new case lacks F6, then plaintiff can downplay with F4, since in both cases efforts were made to maintain secrecy (F102).
   Plaintiff cites Boeing (or Bryce, College Watt, Den-Tal-Ez, Emery, Space Aero), defendant distinguishes since the new case contains additional d-factor F23, then plaintiff downplays (type 1) with F4, so in the new case there is still an express confidentiality agreement (F121).

EXERCISE 8.3.2. None. All of them can be distinguished in at least one way:
Boeing on 6, 12,14 and on 16, 23;
Bryce on 6, 18 and on 16, 23;
College Wat on 26 and on 16,23;
Den-Tal-Ez on 6,26 and on 16,23;
Emery on 18 and on 16,23;
Space Aero on 8,18 and on 16,23;
Television on 6,12,18 and on 23.

EXERCISE 8.3.3.

1. The precedent induces $\{T, D\} < \{Not\text{-}H\}$. Deciding W in the new case requires $\{D, H\} > \{Not\text{-}T\}$. This is consistent with the preference, so W is allowed. Deciding *Not-W* in the new case requires $\{D, H\} < \{Not\text{-}T\}$. This is also consistent with the preference, so *Not-W* is also allowed, so W is not forced.

2. The precedent induces $\{T, D\} < \{Not\text{-}H\}$. Deciding W in the new case requires $\{D\} > \{Not\text{-}T, Not\text{-}H\}$. This is inconsistent with the preference, since a fortiori it holds that $\{D\} < \{Not\text{-}T, Not\text{-}H\}$. so W is not allowed. By the same preference derivation we see that *Not-W* is allowed, so *Not-W* is forced.

3. The new case exactly matches the precedent, so the same outcome is forced.

4. Derek's case for watching TV is stronger than Albert's case for watching TV in two ways: Derek is older than Albert, and he is less seriously ill.

EXERCISE 8.3.4

1. The precedent with Albert has $\{T, D\}$ pro W and $\{Not\text{-}H\}$ con W. The new case with Betsy has $\{D, H\}$ pro W and $\{not\text{-}T\}$ con W. Since $\{not\text{-}H\} \not\subseteq \{not\text{-}T\}$, and also since $\{D, H\} \not\subseteq \{T, D\}$, it does not hold that

$$\{T, D, Not\text{-}H\} \leq_{notW} \{D, H, not\text{-}T\}$$

So deciding con W is not forced.

Since there is no precedent for W, deciding for W is not forced either. So both decisions are allowed but not forced.

The relevant differences between Albert and Betsy are $\{H, not\text{-}H\}$.

2. The new case with Carla has $\{D\}$ pro W and $\{Not\text{-}H, not\text{-}T\}$ con W. Since $\{not\text{-}H\} \subseteq \{not\text{-}H, not\text{-}T\}$ and $\{D\} \subseteq \{T, D\}$, it holds that

$$\{T, D, Not\text{-}H\} \leq_{notW} \{D, H, Not\text{-}H, not\text{-}T\}$$

So deciding con W is forced.

There are no relevant differences between Albert and Betsy.

3. The fact situations of the new case and the precedent are the same, so they are obviously equal with respect to $\leq_{notW}$, so $not\text{-}W$ is forced. The cases have no relevant differences.

EXERCISE 8.3.5

(a) The defendant can distinguish on additional pro-defendant factor F1, after which plaintiff can downplay by saying that the additional pro-plaintiff factor F4 shows that F122 is still the case. Defendant can also distinguish on F15, after which plaintiff can downplay with F4, saying that thus the new case still has F101. The defendant can also distinguish on F25, after which plaintiff can downplay with F2, saying that the new case still contains F111.

(b) With Definition 9.2.5: The precedent has $\{F2, F15, F21\}$ pro π and $\{F23\}$ pro δ. The new case has $\{F2, F4, F21\}$ pro π and $\{F1, F25\}$ pro δ. Since $\{F2, F15, F21\} \not\subseteq \{F2, F4, F21\}$, and also since $\{F1, F25\} \not\subseteq \{F23\}$, it does not hold that

$$\{F2, F15, F21, F23\} \leq_{\pi} \{F1, F2, F4, F21 < F25\}$$

So deciding pro π is not forced. Since there is no precedent for δ, deciding for δ is not forced either. So both decisions are allowed but not forced.

With Definition 9.2.8: Clearly, deciding for π is allowed, since there are no decisions for δ in the case base, so adding any decision for π leaves the case base consistent. Moreover, deciding for δ is also allowed, since the preference that could block it, $\{F2, F4, F21\} > \{F1, F25\}$, does not follow from $\{F2, F15, F21\} > \{F23\}$. But note that if the case is decided for one of the sides, then in a next case with the same fact situation, deciding for the same side is forced.

(c) The relevant differences are $\{F1, F15, F25\}$.

EXERCISE 8.3.6

1. Neither is forced. Deciding for the plaintiff is not forced for three reasons:

(1) $v(measures, c_1) >_{\pi} v(measures, F_2)$ since *Entry-By-Visitors-Restricted* $>_{\pi}$ *Access-To-Premises-Controlled*;

(2) $v(\textit{reverse-eng}, c_1) >_\pi v(\textit{reverse-eng}, F_2)$ since $f >_\pi t$;

(3) $v(disclosed, c_1) >_\pi v(disclosed, F_2)$ since $10 >_\pi 15$.

Deciding for the defendant is not forced for two reasons:

$v(deceived, c_2) >_\delta v(deceived, F_2)$ since $f >_\delta t$;

$v(measures, c_2) >_\delta v(measures, F_2)$ since *Minimal* $>_\delta$ *Access-To-Premises-Controlled*.

2. The relevant differences between c_1 and F_2 are $\{(measures$,*Entry-By-Visitors-Restricted*),(*reverse-eng*, $f), (disclosed, 10)\}$. The relevant differences between c_2 and F_2 are $\{(deceived, f), (measures, Minimal)\}$.

EXERCISE 8.3.7

Factors T and $not\text{-}T$ are now combined into a dimension *Age* with V equalling the natural numbers and $x <^W y$ iff $x < y$, so $x \leq_{notW} y$ iff $y \leq x$. Moreover, H now becomes a dimension with at least the values *healthy*, *not-healthy*, *cold* and *flu*, where

not-healthy $<_W$ *healthy*
cold $<_W$ *healthy*
flu $<_W$ *healthy*
flu $<_W$ *cold*

So the value ordering of the health dimension is partial. Finally, D now becomes a two-valued dimension with $V = \{t, f\}$, where $f <_W t$.

1. Deciding $not\text{-}W$ is not forced, since $v(H, Betsy) <_{notW} v(H, Albert)$, since *healthy* $<_{notW}$ *flu*. Since there is no precedent for W, deciding for W is not forced either. So both decisions are allowed but not forced.

   The relevant differences between Albert and Betsy are $\{(H, \textit{flu})\}$.

2. Deciding $not\text{-}W$ is not forced, since $v(H, Albert) \not\leq_{notW} v(H, Carla)$. Since there is no precedent for W, deciding for W is not forced either. So both decisions are allowed but not forced. The relevant differences between Albert and Carla are $\{(H, \textit{flu})\}$.

3. Deciding $not\text{-}W$ is not forced, since $v(Age, Albert) \not<_{notW} v(Age, Derek)$, since $12 >_{notW} 15$. Since there is no precedent for W, deciding for W is not forced either. So both decisions are allowed but not forced. The relevant differences between Albert and Derek are $\{(Age, 12), (H, \textit{flu})\}$.

EXERCISE 8.3.8

Both decisions are allowed.

A relevant difference between c_1 and F is (*employer*, *self-employed*), since *self-employed* and *foreign* are incomparable values of *employer*. So deciding F for *changed* is not forced.

A relevant difference between c_2 and F is (*employer*, *Dutch*), since *foreign* $<_{\textit{not changed}}$ *Dutch*. Another relevant difference is (*duration*,12), since $18 <_{\textit{not changed}} 12$. So deciding F for *not changed* is not forced either.

Bibliography

Aleven, V. (2003). Using background knowledge in case-based legal reasoning: a computational model and an intelligent learning environment, *Artificial Intelligence* **150**: 183–237.

Amgoud, L. and Ben-Naim, J. (2013). Ranking-based semantics for argumentation frameworks, *in* W. Liu, V. Subrahmanian and J. Wijsen (eds), *Scalable Uncertainty Management. SUM 2013*, number 8078 in *Springer Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 134–147.

Amgoud, L. and Besnard, P. (2009). Bridging the gap between abstract argumentation systems and logic, *in* L. Godo and A. Pugliese (eds), *Scalable Uncertainty Management. SUM 2013*, number 5785 in *Springer Lecture Notes in AI*, Springer Verlag, Berlin, pp. 12–27.

Amgoud, L. and Besnard, P. (2013). Logical limits of abstract argumentation frameworks, *Journal of Applied Non-classical Logics* **23**: 229–267.

Amgoud, L., Bodenstaff, L., Caminada, M., McBurney, P., Parsons, S., Prakken, H., van Veenen, J. and Vreeswijk, G. (2006). Final review and report on formal argumentation system, *Deliverable D2.6*, ASPIC IST-FP6-002307.

Amgoud, L. and Cayrol, C. (1998). On the acceptability of arguments in preference-based argumentation, *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pp. 1–7.

Amgoud, L. and Cayrol, C. (2002). A model of reasoning based on the production of acceptable arguments, *Annals of Mathematics and Artificial Intelligence* **34**: 197–215.

Antoniou, G. (1999). A tutorial on default logics, *ACM Computing Surveys* **31**(3): 337–359.

Ashley, K. (1990). *Modeling Legal Argument: Reasoning with Cases and Hypotheticals*, MIT Press, Cambridge, MA.

Baroni, P. and Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics, *Artificial Intelligence* **171**: 675–700.

Baumann, R. (2012). What does it take to enforce an argument? Minimal change in abstract argumentation, *Proceedings of the 20th European Conference on Artificial Intelligence*, pp. 127–132.

Baumann, R. and Brewka, G. (2010). Expanding argumentation frameworks: Enforcing and monotonicity results, *in* P. Baroni, F. Cerutti, M. Giacomin and G. Simari (eds), *Computational Models of Argument. Proceedings of COMMA 2010*, IOS Press, Amsterdam etc, pp. 75–86.

Bench-Capon, T. (2002). The missing link revisited: the role of teleology in representing legal argument, *Artificial Intelligence and Law* **10**: 79–94.

Berman, D. and Hafner, C. (1993). Representing teleological structure in case-based legal reasoning: the missing link, *Proceedings of the Fourth International Conference on Artificial Intelligence and Law*, ACM Press, New York, pp. 50–59.

Besnard, P. and Hunter, A. (2009). Argumentation based on classical logic, *in* I. Rahwan and G. Simari (eds), *Argumentation in Artificial Intelligence*, Springer, Berlin, pp. 133–152.

Bisquert, P., Cayrol, C., Dupin de Saint-Cyr, F. and Lagasquie-Schiex, M.-C. (2013). Goal-driven changes in argumentation: a theoretical framework and a tool, *Proceedings of the 25th International Conference on Tools with Artificial Intelligence (ICTAI 2013)*, pp. 610–617.

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

Caminada, M. (2004). *For the Sake of the Argument. Explorations into Argument-based Reasoning*, Doctoral dissertation Free University Amsterdam.

Caminada, M. (2006). On the issue of reinstatement in argumentation, *in* M. Fischer, W. van der Hoek, B. Konev and A. Lisitsa (eds), *Logics in Artificial Intelligence. Proceedings of JELIA 2006*, number 4160 in *Springer Lecture Notes in AI*, Springer Verlag, Berlin, pp. 111–123.

Caminada, M. and Amgoud, L. (2007). On the evaluation of argumentation formalisms, *Artificial Intelligence* **171**: 286–310.

Caminada, M., Modgil, S. and Oren, N. (2014). Preferences and unrestricted rebut, *in* S. Parsons, N. Oren, C. Reed and F. Cerutti (eds), *Computational Models of Argument. Proceedings of COMMA 2014*, IOS Press, Amsterdam etc, pp. 209–220.

Carlson, L. (1983). *Dialogue Games: an Approach to Discourse Analysis*, Reidel Publishing Company, Dordrecht.

Cayrol, C., Dupin de Saint-Cyr, F. and Lagasquie-Schiex, M.-C. (2010). Change in abstract argumentation frameworks: adding an argument, *Journal of Artificial Intelligence Research* **38**: 49–84.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2005). Graduality in argumentation, *Journal of Artificial Intelligence Research* **23**: 245–297.

Cayrol, C. and Lagasquie-Schiex, M.-C. (2009). Bipolar abstract argumentation systems, *in* I. Rahwan and G. Simari (eds), *Argumentation in Artificial Intelligence*, Springer, Berlin, pp. 65–84.

Cohen, A., Parsons, S., Sklar, E. and McBurney, P. (2018). A characterization of types of support between structured arguments and their relationship with support in abstract argumentation, *International Journal of Approximate Reasoning* **94**: 76–104.

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

Dung, P., Kowalski, R. and Toni, F. (2009). Assumption-based argumentation, *in* I. Rahwan and G. Simari (eds), *Argumentation in Artificial Intelligence*, Springer, Berlin, pp. 199–218.

Dung, P. and Thang, P. (2014). Closure and consistency in logic-associated argumentation, *Journal of Artificial Intelligence Research* **49**: 79–109.

Grooters, D. and Prakken, H. (2016). Two aspects of relevance in structured argumentation: minimality and paraconsistency, *Journal of Artificial Intelligence Research* **56**: 197–245.

Grossi, D. and Modgil, S. (2015). On the graded acceptability of arguments, *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pp. 868–874.

Grossi, D. and Modgil, S. (2019). On the graded acceptability of arguments in abstract and instantiated argumentation, *Artificial Intelligence* **275**: 138–173.

Hamblin, C. (1970). *Fallacies*, Methuen, London.

Hamblin, C. (1971). Mathematical models of dialogue, *Theoria* **37**: 130–155.

Horty, J. (2011). Rules and reasons in the theory of precedent, *Legal Theory* **17**: 1–33.

Horty, J. (2019). Reasoning with dimensions and magnitudes, *Artificial Intelligence and Law* **27**: 309–345.

Kraus, S., Lehmann, D. and Magidor, M. (1990). Nonmonotonic reasoning, preferential models and cumulative logics, *Artificial Intelligence* **44**: 167–207.

Kyburg, H. (1961). *Probability and the Logic of Rational Belief*, Wesleyan University Press, Middletown, CT.

Loui, R. (1987). Defeat among arguments: a system of defeasible inference, *Computational Intelligence* **2**: 100–106.

Modgil, S. (2006). Hierarchical argumentation, *in* M. Fischer, W. van der Hoek, B. Konev and A. Lisitsa (eds), *Logics in Artificial Intelligence. Proceedings of JELIA 2006*, number 4160 in *Springer Lecture Notes in AI*, Springer Verlag, Berlin, pp. 319–332.

Modgil, S. and Prakken, H. (2012). Resolutions in structured argumentation, *in* B. Verheij, S. Woltran and S. Szeider (eds), *Computational Models of Argument. Proceedings of COMMA 2012*, IOS Press, Amsterdam etc, pp. 310–321.

Modgil, S. and Prakken, H. (2013). A general account of argumentation with preferences, *Artificial Intelligence* **195**: 361–397.

Modgil, S. and Prakken, H. (2014). The ASPIC+ framework for structured argumentation: a tutorial, *Argument and Computation* **5**: 31–62.

Modgil, S. and Prakken, H. (2018). Abstract rule-based argumentation, *in* P. Baroni, D. Gabbay, M. Giacomin and L. van der Torre (eds), *Handbook of Formal Argumentation*, Vol. 1, College Publications, London, pp. 286–361.

Parsons, S., Wooldridge, M. and Amgoud, L. (2003). Properties and complexity of some formal inter-agent dialogues, *Journal of Logic and Computation* **13**. 347-376.

Pearl, J. (1992). Epsilon-semantics, *in* S. Shapiro (ed.), *Encyclopedia of Artificial Intelligence*, John Wiley & Sons, New York, pp. 468–475.

Pollock, J. (1974). *Knowledge and Justification*, Princeton University Press, Princeton.

Pollock, J. (1987). Defeasible reasoning, *Cognitive Science* **11**: 481–518.

Pollock, J. (1994). Justification and defeat, *Artificial Intelligence* **67**: 377–408.

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

Pollock, J. (2002). Defeasible reasoning with variable degrees of justification, *Artificial Intelligence* **133**: 233–282.

Pollock, J. (2009). A recursive semantics for defeasible reasoning, *in* I. Rahwan and G. Simari (eds), *Argumentation in Artificial Intelligence*, Springer, Berlin, pp. 173–197.

Prakken, H. (2005). Coherence and flexibility in dialogue games for argumentation, *Journal of Logic and Computation* **15**: 1009–1040.

Prakken, H. (2006). Formal systems for persuasion dialogue, *The Knowledge Engineering Review* **21**: 163–188.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments, *Argument and Computation* **1**: 93–124.

Prakken, H. (2012). Some reflections on two current trends in formal argumentation, *Logic Programs, Norms and Action. Essays in Honour of Marek J. Sergot on the Occasion of his 60th Birthday*, Springer, Berlin/Heidelberg, pp. 249–272.

Prakken, H. (2015). Formalising debates about law-making proposals as practical reasoning, *in* M. Araszkiewicz and K. Płeszka (eds), *Logic in the Theory and Practice of Lawmaking*, Springer, Berlin, pp. 301–321.

Prakken, H. (2016). Rethinking the rationality postulates for argumentation-based inference, *in* P. Baroni, T. Gordon, T. Scheffler and M. Stede (eds), *Computational Models of Argument. Proceedings of COMMA 2016*, IOS Press, Amsterdam etc, pp. 419–430.

Prakken, H. (2020). On validating theories of abstract argumentation frameworks: the case of bipolar argumentation frameworks, *Proceedings of the 20th Workshop on Computational Models of Natural Argument*, Vol. 2669 of *CEUR Workshop Proceedings*, pp. 21–30.

Prakken, H. (2021a). A formal analysis of some factor- and precedent-based accounts of precedential constraint, *Artificial Intelligence and Law* **29**: 559–585.

Prakken, H. (2021b). Philosophical reflections on argument strength and gradual acceptability, *Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 21)*, number 12897 in *Springer Lecture Notes in AI*, Springer Verlag, Berlin, pp. 144–158.

Prakken, H. (2023). Relating abstract and structured accounts of argumentation dynamics: the case of expansions, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning*, IJCAI Organization, pp. 562–571.

Prakken, H. and Sartor, G. (1997). Argument-based extended logic programming with defeasible priorities, *Journal of Applied Non-classical Logics* **7**: 25–75.

Prakken, H. and Sartor, G. (2015). Law and logic: A review from an argumentation perspective, *Artificial Intelligence* **227**: 214–225.

Prakken, H. and Vreeswijk, G. (2002). Logics for defeasible argumentation, *in* D. Gabbay and F. Günthner (eds), *Handbook of Philosophical Logic*, second edn, Vol. 4, Kluwer Academic Publishers, Dordrecht/Boston/London, pp. 219–318.

Rescher, N. and Manor, R. (1970). On inference from inconsistent premises, *Journal of Theory and Decision* **1**: 179–219.

Toni, F. (2014). A tutorial on assumption-based argumentation, *Argument and Computation* **5**: 89–117.

van Eemeren, F., Garssen, B., Krabbe, E., Henkemans, A. S., Verheij, B. and Wagemans, J. (2014). *Handbook of Argumentation Theory*, Springer, Dordrecht.

Vreeswijk, G. (1993a). Defeasible dialectics: a controversy-oriented approach towards defeasible argumentation, *Journal of Logic and Computation* **3**: 317–334.

Vreeswijk, G. (1993b). *Studies in Defeasible Argumentation*, Doctoral dissertation Free University Amsterdam.

Vreeswijk, G. (1997). Abstract argumentation systems, *Artificial Intelligence* **90**: 225–279.

Vreeswijk, G. and Prakken, H. (2000). Credulous and sceptical argument games for preferred semantics, *Proceedings of the 7th European Workshop on Logics in Artificial Intelligence (JELIA'2000)*, number 1919 in *Springer Lecture Notes in AI*, Springer Verlag, Berlin, pp. 239–253.

Walton, D. (1984). *Logical dialogue-games and fallacies*, University Press of America, Inc., Lanham, MD.

Walton, D. (1996). *Argumentation Schemes for Presumptive Reasoning*, Lawrence Erlbaum Associates, Mahwah, NJ.

Walton, D. (2006). *Fundamentals of Critical Argumentation*, Cambridge University Press, Cambridge.

Walton, D. and Krabbe, E. (1995). *Commitment in Dialogue. Basic Concepts of Interpersonal Reasoning*, State University of New York Press, Albany, NY.

Wu, Y. (2012). *Between Argument and Conclusion. Argument-based Approaches to Discussion, Inference and Uncertainty*, Doctoral Dissertation Faculty of Sciences, Technology and Communication, University of Luxemburg.

Zenker, F., Debowska-Kozlowska, K., Godden, D., Selinger, M. and Wells, S. (2020). Five approaches to argument strength: probabilistic, dialectical, structural, empirical, and computational, *Proceedings of the 3rd European Conference on Argumentation*, College Publications, London, pp. 653–674.