Argumentation In Flux (Modelling Change in the Theory of Argumentation)

Tjitze Rienstra

UNIVERSITÉ DU
LUXEMBOURG
PhD-FSTC-2014-36
The Faculty of Science,
Technology and Communication

UNIVERSITÉ MONTPELLIER 2
SCIENCES ET TECHNIQUES
Ecole Doctorale I2S
(Information Structures Systemes)
LIRMM, UMR 5506

DISSERTATION

Defense held on 23/10/2014 in Luxembourg
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN INFORMATIQUE
AND
DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER II
EN INFORMATIQUE

Argumentation in Flux
Modelling Change in the Theory of Argumentation

by

Tjitze Rienstra
Born on 10/10/1981 in Gauw (The Netherlands)

Dissertation defense committee

Dr Leon van der Torre, Supervisor
Professor, Université du Luxembourg

Dr Souhila Kaci, Supervisor
Professor, LIRMM - UMR 5506, Université de Montpellier II

Dr Richard Booth
Université du Luxembourg

Dr Lluís Godo, Chairman
Artificial Intelligence Research Institute (IIIA), CSCI, Bellaterra, Spain

Dr Pietro Baroni, Vice Chairman
Professor, University of Brescia, Italy

Dr Beishui Liao
Professor, Zhejiang University, China

Summary

Dung's theory of abstract argumentation is a widely used formalism in the field of artificial intelligence. It is used to model various types of reasoning, by representing conflicting or defeasible information using an argumentation framework, i.e., a set of arguments and an attack relation. Different so-called semantics have been proposed in the literature to determine, given an argumentation framework, the justifiable points of view on the acceptability of the arguments. The research in this thesis is motivated by the idea that argumentation is not a static process, and that a better understanding of the behaviour and applicability of the theory of abstract argumentation requires a dynamic perspective. We address this issue from three points of view.

First, we identify and investigate two types of change in argumentation. We call them *intervention* and *observation*, due to their similarity to the similarly named types of change in the theory of causal Bayesian networks. While intervention amounts to change due to actions (i.e., bringing new arguments/attacks into play), observation amounts to revision due to new information from the environment. We model these two types of change as two types of inference relations. This allows us to contrast and characterize the behaviour of the two types of change, under a number of different semantics, in terms of properties satisfied by the respective inference relations.

Second, we investigate the relation between abduction in logic programming and change in argumentation. We show that, on the abstract level, changes to an argumentation framework may act as hypotheses to explain an observation. The relation with abduction in logic programming lies in the fact that this abstract model can be instantiated on the basis of an abductive logic program, just like an abstract argumentation framework can be instantiated on the basis of a logic program. We furthermore present dialogical proof theories for the main reasoning problem, i.e., finding hypotheses that explain an observation.

Third, we look at change in preference-based argumentation. Preferences have been introduced in argumentation to encode, for example, relative strength of arguments. An underexposed aspect in these models is change of preferences. We present a dynamic model of preferences in argumentation, based on what we call property-based argumentation frameworks. It is based on Dietrich and List's model of property-based preference and provides an account of how and why preferences in argumentation may change. The idea is that preferences over arguments are derived from preferences over properties of arguments and change as the result of moving to different motivational states. We also provide a dialogical proof theory that establishes whether there exists some motivational state in which an argument is accepted.

Acknowledgements

I would like to thank my supervisors Leon van der Torre, Souhila Kaci and Richard Booth. It is often said that the most important element in obtaining a PhD is to find the right supervisors, and I feel that I could not have been luckier in this regard. They are excellent supervisors without whose guidance this work would have been impossible. I furthermore thank Pietro Baroni, Lluís Godo and Beishui Liao, the external members of my defence committee, for reviewing this thesis and providing invaluable feedback.

I've furthermore had the pleasure to discuss ideas with many colleagues, both inside and outside the University of Luxembourg and University of Montpellier 2. Some of these discussions lead to joint publications, for which I want to thank Matthias Thimm, Nir Oren, Ofer Arieli, Dov Gabbay, Srdjan Vesic, Francois Schwarzentruber, Serena Villata and Alan Perotti.

I want to thank all my colleagues for the great times, both during and after working hours. A special thank you to Martin Caminada, Emil Weydert and Dov Gabbay for all the interesting discussions, many of which influenced the ideas that lead to this thesis.

I would like to thank my parents and my two sisters for their support, as well as my friends in Enschede. Finally, I would like to thank Salome, who has had to pay for this thesis with many lonely weekends, for her love, support and patience.

Contents

Chapter 1

Introduction

1.1 Background

1.1.1 Argumentation in Artificial Intelligence

Argumentation is an activity that lies at the heart of how humans persuade and inform each other, how they make decisions and form beliefs, and how they justify and explain these decisions and beliefs. It is for this reason that argumentation is a central discipline within the field of artificial intelligence. The research in this thesis deals with theoretical applications of argumentation, namely as a model to capture in a general way the notion of *defeasible reasoning*.

Defeasible reasoning is reasoning based on rules of inference that are not necessarily deductively valid. In defeasible reasoning, premises provide support for conclusions, but do not guarantee their truth (it may be the case that the premises are true while the conclusion is false). This means that conclusions arrived at using defeasible reasoning may have to be retracted when additional information is acquired. This is a violation of the *monotony* property, which states that previously drawn conclusions are never retracted when additional information comes into play. In this sense, defeasible reasoning is *non-monotonic*.

Classical monotonic reasoning is suitable for formal or mathematical reasoning. Once we establish the truth of a theorem in mathematics, we do not worry about the truth of this theorem when we acquire new information. For example, Euclid proved more than two centuries ago that the set of prime numbers is infinite, and it is impossible that a new discovery will invalidate this truth. Non-monotonicity is, however, a natural phenomenon in common-sense reasoning, because we often reason using assumptions, general rules with exceptions, rules of thumb, etcetera. This has led to the insight that classical monotonic logics are unsuitable for common-sense reasoning and, starting in the early eighties, much research in the field of Artificial Intelligence has focussed on defeasible, non-monotonic reasoning.

By the early nineties, the number of different defeasible reasoning formalisms was enormous. Some examples are Reiter's default logic [78], Pollock's defea-

sible logic [74], and many variations of logic programming [53, 77, 52]. Some approaches to defeasible reasoning already employed argumentation-theoretic ideas (see Prakken and Vreeswijk [76] for an overview). In 1995, Phan Min Dung showed, however, that many formalisms are based on the same underlying principle, which can be modelled using a simple and elegant theory called *abstract argumentation* [42]. Since 1995, this theory has led to many new insights about the nature of defeasible reasoning and it has been applied in many areas beyond defeasible reasoning, which includes negotiation, decision making and agent communication.

While argumentation is an inherently dynamic activity, the theory of abstract argumentation neglects the notion of change, and takes a static perspective. The research in this thesis is motivated by the idea that a better understanding of the behaviour and applicability of the theory requires a dynamic perspective. This perspective leads to various questions related to change in the theory of abstract argumentation.

The title of this thesis is *Argumentation in Flux*, which is a reference to the book called *Knowledge in Flux* by Peter Gärdenfors's. This book, published in 1988, dealt with the dynamics of epistemic states, and set the stage for the successful and important field of research called *belief change*, which is active to this day. The goal of this thesis is to combine the perspectives taken by Dung in 1995 and by Gärdenfors in 1988.

Before precisely stating the problem that we address, we explain the basics of abstract argumentation, instantiated argumentation (i.e., argumentation as an abstract model of a defeasible reasoning formalism) and a number of extensions of the theory that will be relevant in what follows.

1.1.2 Abstract Argumentation

The starting point in Dung's theory of abstract argumentation is the idea that a debate can be represented using a set of arguments and an attack relation between arguments [42]. Consider, for example, the following exchange of arguments.

- a: Mary will pass her mathematics exam.
- b: No she won't, because the exam is too difficult.
- c: Yes, but Mary is very smart.

A central idea is that we can reason about the acceptance of arguments even if we abstract away from their content. Thus, we only need to identify the arguments a, b and c, and the attacks among them: c attacks b and b attacks a. Formally, a debate can then be represented by what is called an *argumentation framework*. This is a pair $F = (A, \rightsquigarrow)$ where A is a set of abstract arguments and $\rightsquigarrow \subseteq A \times A$ is an attack relation between arguments. Thus, an argumentation framework is essentially a directed graph, in which nodes represent arguments and arrows represent attack between arguments. The debate discussed here can be represented by the argumentation framework shown in figure 1.1.

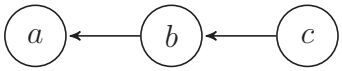


Figure 1.1: An argumentation framework.

Given an argumentation framework, the main reasoning task is to answer the question: which arguments can be accepted? For the argumentation framework shown in figure 1.1, we can reason as follows: The argument c is left unchallenged and is therefore accepted. The argument b is challenged by c which, as we just established, is accepted. This implies that b must be rejected. Finally, the argument a is challenged by b, but since b is rejected, we can accept a.

There is, however, not in general a single answer to the question of which arguments can be accepted. This is due to the fact that argumentation frameworks that contain cycles (such as mutually attacking arguments) may lead to more than one way in which we can choose which arguments to accept. Consider, for example, the following exchange of arguments, which is represented by the argumentation framework shown in figure 1.2.

- a: Germany will win the finals.
- b: No, Argentina is better.
- c: No, Germany is better.
- b: No, Argentina is better.

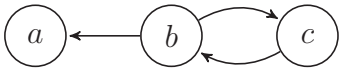


Figure 1.2: An argumentation framework with a cycle.

A method to evaluate an argumentation framework, or to determine the points of view on which arguments to accept, is called an *argumentation semantics*. We will consider in this thesis a number of different semantics, namely the complete, grounded, preferred and stable semantics [42] as well as the semi-stable semantics [89, 27]. Each of these semantics yields, according to some set of criteria, a set of extensions (i.e., sets of acceptable arguments) of a given argumentation framework. These extensions represent the points of view on which arguments are acceptable. An alternative representation for such a point of view is a *labelling*, which assigns to each argument an acceptance status *in* (meaning accepted), *out* (meaning rejected) or *undecided* (neither accepted nor rejected) [26]. A basic condition for a labelling to make sense (and a condition satisfied by all the argumentation semantics we consider) is expressed by the following two rules:

1. An argument is labelled *in* if and only if all attackers are labelled *out*.
2. An argument is labelled *out* if and only if some attacker is labelled *in*.

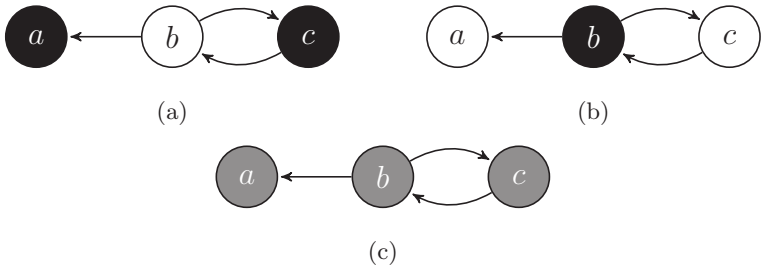


Figure 1.3: Three labellings of the argumentation framework shown in figure 1.2.

We can conveniently represent a labelling by colouring the nodes of the graph. We use the following convention: white nodes are in, black nodes are out, and gray nodes are undecided. The labellings satisfying the rules discussed above for the argumentation framework shown in figure 1.2 are shown in figure 1.3. In figure 1.3a, b is in and a and c are out; in figure 1.3b, a and c are in and b is out; and in figure 1.3c, all arguments are undecided.

To summarize, argumentation frameworks are abstract representations of a debate, in which we identify only the arguments (without specifying their content) and attacks between arguments. The evaluation of an argumentation framework under a given argumentation semantics leads to a set of extensions or labellings, each representing a possible point of view on which arguments to accept. We will formalize these concepts and explain the difference between the different argumentation semantics in chapter 2.

Several extensions of the notion of an argumentation framework have been developed. An extension that we look at in this thesis are *preference-based* argumentation frameworks [2, 86]. Preferences over arguments may be derived from the relative strength of arguments. These preferences suggest that, in some cases, the attack of one argument on another succeeds only on the condition that the latter is not preferred over the former. The notion of a value-based argumentation framework extends this idea [14]. The idea here is that arguments promote certain *values* and that different *audiences* have different preferences over values, from which the preferences over arguments are derived. Thus, a value-based argumentation framework may be evaluated in a different way by different audiences.

1.1.3 Instantiated Argumentation

As we demonstrated, Dung's theory of argumentation is abstract, in the sense that the content of arguments is left unspecified. However, many defeasible reasoning formalisms can be seen as specific forms of *instantiated argumentation*, where arguments are not anymore abstract, but are instantiated on the basis of a knowledge base.

Instantiated argumentation is based on a three-step process, shown in figure 1.4 (this figure also appears in Baroni et al. [4]). In the first step, an argumentation framework is constructed on the basis of a given knowledge base. The arguments

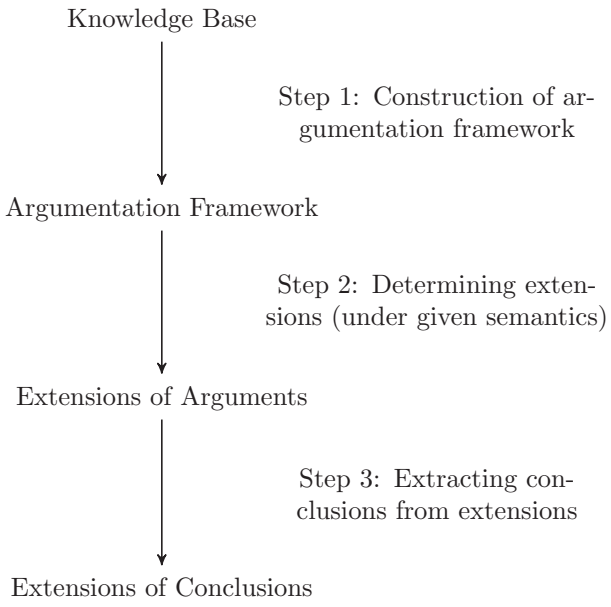


Figure 1.4: The Three-Step Instantiation Process.

in this argumentation framework correspond to defeasible proofs, and attacks are determined by how the different proofs defeat each other. In the second step, an argumentation semantics is applied, resulting in sets of extensions of arguments. In the third step the set of conclusions of the arguments is extracted. The result of this step represents the conclusions of the knowledge base.

Although many non-monotonic reasoning formalisms have been shown to be forms of instantiated argumentation (including Reiter's default logic and Pollock's inductive defeasible logic [42] as well as Nute's defeasible logic [54]), we restrict our attention here to logic programming. When modelling logic programming as argumentation, the knowledge bases that we work with are logic programs. Furthermore, arguments are constructed by constructing proof trees consisting of the rules of the logic program, and the defeasibility of these arguments is due to the use of negation as failure. The outcome of the three-step procedure depends on the argumentation semantics that is applied in step two. Various choices that can be made in this regard have been shown to correspond to various semantics for logic programming. For example, using the stable semantics in step two results in an outcome corresponding to the stable model semantics for LP [42], using the grounded semantics results in an outcome corresponding to the well-founded semantics for LP [42], the complete semantics corresponds in this way to the three-valued stable semantics for LP [91], and the preferred semantics corresponds to the regular model semantics for LP [31].

In addition to defeasible reasoning, it has been shown that the stable marriage problem can be seen as a form of instantiated argumentation [42].

1.2 Change in Argumentation

We now turn to the problem addressed in this thesis. As we pointed out, abstract argumentation theory neglects aspects of change. The theory is agnostic towards change of an argumentation framework, in the sense that it only provides methods to compute a fixed and unchanging evaluation of a given argumentation framework. In real life, however, argumentation is an *activity*, in which the evaluation of arguments evolves as new arguments and attacks come into play. Thus, the evaluation of an argumentation framework may change *due to actions*. Furthermore, there is the general problem of *belief revision*, namely that intelligent agents have to account for changes in their environment and for situations in which beliefs are discovered to be incorrect. This applies to argumentation too: we may arrive, using argumentation, at the conclusion that an argument is accepted, but we have to revise this conclusion if we discover that this is incorrect. Thus, we can distinguish *two types* of change in argumentation. We will now clarify this distinction by giving some examples of these two types of change, and identifying the issues that arise when attempting to model them. We also briefly survey the literature in which aspects of these two types of change have been considered.

1.2.1 Two Types of Change

Change due to Actions

While the notion of an argumentation framework is essentially static, in the sense that the set of arguments and attacks are fixed, argumentation is an inherently dynamic activity. This is because a debate usually evolves due to new arguments and attacks coming into play. These are essentially *actions* performed in a debate. Consider the discussion we used earlier:

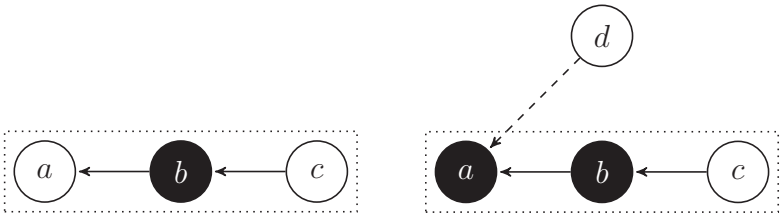
- a: Mary will pass her mathematics exam.
- b: No she won't, because the exam is too difficult.
- c: Yes, but Mary is very smart.

This discussion may continue as follows:

- d: Mary did not pay her tuition fees, so she is barred from taking exams.

The new argument d attacks the argument a and leads to a new evaluation of the argumentation framework. While a was initially in, because its sole attacker b was labelled out, a now becomes out, because d is itself unattacked and therefore in. This is shown in figure 1.5: in the initial situation (figure 1.5a), a and c are *in* and b is *out*. When we add the argument d, a becomes *out* (figure 1.5b).

From a formal point of view, this type of change is easy to deal with, because we can simply add arguments and attacks to an existing argumentation framework and recompute its evaluation. However, it does put some aspects that are



(a) Initial situation. (b) Change due to a new argument.

Figure 1.5: Change due to a new argument.

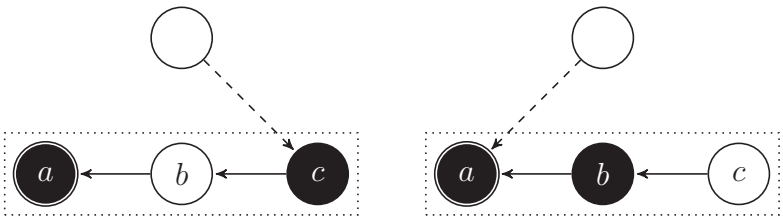
usually studied from a static perspective in a different light. For example, several authors have studied properties that represent natural principles that one may expect an argumentation semantics to satisfy. Many of these properties assume a static setting (examples are **in**-maximialty, (strong) admissibility and (weak) reinstatement [6]). One of our goals is to study properties that apply to the dynamic setting.

Another issue is that of computation. The computational complexity of computing the evaluation of an argumentation framework under various semantics has been studied, e.g., by Dunne [44], while a number of algorithms have been proposed by Modgil and Caminada [68]. In the dynamic setting, however, the question arises whether recomputing the evaluation of an argumentation framework after it is modified can somehow benefit from reusing the evaluation of the initial argumentation framework. Liao et al. [65] investigated the role of *directionality* in this respect. Directionality (a principle first studied by Baroni and Giacomin [6]) ensures, if satisfied by an argumentation semantics, that an argument x has an effect on the status of an argument y only if there is a directed path from x to y. This means that, when an argumentation framework is modified, we can divide the argumentation framework into an affected part and an unaffected part, and only the evaluation of the affected part needs to be recomputed.

Revision of Evaluation

The second type of change is concerned with the revision of the evaluation of an argumentation framework due to new information. The relevance of this type lies in the general problem of belief revision: agents have to account for changes in their environment, or for situations in which their beliefs turn out to be incorrect. This applies to argumentation too. We may, through argumentation, arrive at some conclusion, but we have to revise this conclusion if new information to the contrary becomes known.

We demonstrate the general idea using a simple example. Consider again the argumentation framework shown in figure 1.1, in which we accept a and c and reject b. Suppose we simply learn that Mary did not pass her exam. This means that we must reject a. But how are we supposed to revise the overall evaluation of the argumentation framework? This depends on how we account for this information. We may, for example, give up our belief that Mary is very smart. The new information is then accounted for by adding an attacker to c



(a) Attacking c. (b) Attacking a.

Figure 1.6: Two ways to account for rejection of a.

(figure 1.6a). On the other hand, perhaps we are certain that Mary is very smart, and believe that her failing the exam must have had another reason (perhaps Mary did not pay her tuition fees!). In this situation, the new information can be accounted for by adding an attacker to a (figure 1.6b). This example clearly demonstrates the difference between the two type of change: the *action* of rejecting a by introducing a new attacker simply leads to rejection of a (but not of c) while the *information* that a is rejected may, depending on how we account for this information, also lead to rejection of c.

From a technical point of view, this type of change can be seen as a kind of goal-oriented change of the argumentation framework. Goal-oriented change in argumentation is usually studied from a *multi-agent strategic perspective*. From the multi-agent strategic perspective, the revision of the status of an argument is due not to information from the environment, but represents the goal of an agent in a debate. For example, Kontarinis et al. [61] put this problem as follows: "When several agents are engaged in an argumentation process, they are faced with the problem of deciding how to contribute to the current state of the debate in order to satisfy their own goal, i.e., to make an argument under a given semantics accepted or not." Similar motivations are found in Baumann and Brewka's work on what they call the enforcement problem [11, 12] as well as other work in this direction [17, 20]. Work that takes a multi-agent strategic perspective deals mainly with procedural and economical aspects (e.g., *how* to determine which arguments to attack to satisfy a given goal? And what are the *minimal* contributions to a debate that achieve this?).

In this thesis, however, we take a *single-agent revision perspective*, which has been relatively neglected in the literature. That is, the argumentation framework represents the reasoning of a single agent, and the revision of its evaluation is due to new information that the agent receives from the environment. This leads to different questions than those that are relevant when taking a multi-user strategic perspective. For example, to model revision, we need some mechanism by which a rational agent decides how to change his argumentation framework, in order to revise its evaluation due to an observation. We can also approach this from another angle, and focus on properties that characterize a rational way to revise the evaluation of an argumentation framework.

Outside the area of argumentation theory, this is a very common way of looking at revision. The most notable example is the notion of a *revision operator*, which is widely studied in the area of belief revision [55]. Given some initial knowledge set K, a revision operator $\circledast$ takes as input a piece of new information

ϕ and returns a new knowledge set $K \circledast \phi$ that represents a rational way of revising K by ϕ. This approach permits both constructive characterizations (i.e., the definition of mechanisms that define well-behaved revision operators) as well as postulate-based characterizations (i.e., the definition of properties that characterize a well-behaved revision operator).

In the single-agent revision perspective, the goal-oriented change of the argumentation framework can be seen as a form of *abduction*. That is, the agent needs to find some change to the argumentation framework that explains the new information.

1.2.2 Intervention and Observation

We identified two types of change in argumentation, namely change due to actions in a debate, and revision of the evaluation due to new information. This distinction resembles the distinction between *intervention* and *observation* in causal Bayesian networks. For this reason, we call the two types of change *intervention in argumentation* and *observation in argumentation*. We give a short explanation of intervention and observation in causal Bayesian networks, and then discuss the similarities with the two types of change in argumentation.

Intervention and Observation in Causal Bayesian Nets

Causal Bayesian networks are structures used to represent probabilistic causal relationships between random variables. A causal Bayesian net is a directed acyclic graph, in which nodes represent random variables and edges represent relations of causal influence. Informally speaking, a causal Bayesian net (like a Bayesian net) carries conditional independence relations between different variables of a given probability distribution. That is, once the values of all parent nodes (i.e., all direct causes) of a given node X are known, the value of X is probabilistically independent of all the other ancestor nodes (i.e., all indirect causes). This is called the *Markov assumption*. In addition—and this is what distinguishes *causal* Bayesian nets from regular Bayesian nets—it is assumed that edges actually represent a directed *causal* relation, which is not guaranteed by the Markov assumption alone. We suffice with this informal description, and refer the reader to Pearl [72] for details.

Figure 1.7 (left side) shows an example of a causal Bayesian net (taken from Pearl [72]). It represents the causal relationships between the season ($X_1 \in \{wet, dry\}$), the state of the sprinkler ($X_2 \in \{on, off\}$), the rain ($X_3 \in \{yes, no\}$), wetness of the pavement ($X_4 \in \{yes, no\}$), and slipperiness of the pavement ($X_5 \in \{yes, no\}$). The causal relationship between the season and the state of the sprinkler is because the sprinkler is set in advance according to the season. Furthermore, the season influences the probability of rain and the state of the sprinkler, which both influence wetness of the pavement. In turn, wetness of the pavement influences slipperiness of the pavement.

A causal Bayesian network makes it possible to draw inferences on the basis of two types of events. These are *interventions* and *observations*, which leads to the notion of *intervention-based inference* and *observation-based inference*.

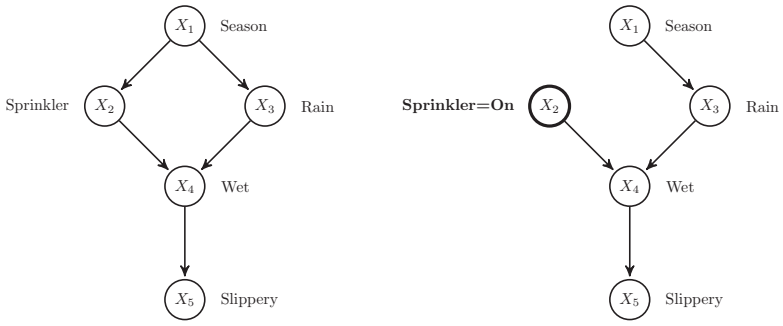


Figure 1.7: A causal Bayesian net (left) and the intervention $X_2 = On$ (right).

First of all, if we observe an event then, a causal Bayesian network informs us about both the possible causal explanations for this event as well as other events caused by it. For example, if we observe that the sprinkler is on, we may infer, as a causal explanation, that the season is dry. This, in turn, may decreases the probability of rain. Other events that may be caused by the sprinkler being on are a wet pavement and, indirectly, a slippery pavement. Thus, observation-based inference involves both backwards abductive reasoning about causes, as well as forward reasoning about effects. Formally, observations in causal Bayesian nets are modelled like in regular Bayesian nets. That is, observing an event amounts to setting the value of the corresponding variable and recomputing the probabilities of the other events *conditional* on the observed event, using the conditional independence relations encoded by the graph.

Interventions, however, represent *actions* where something in the environment is changed. Interventions are modelled by modifying the causal Bayesian network to take the corresponding change in the environment into account. We may, for example, turn on the sprinkler. In the causal Bayesian network, we thus fix the value of X_3 to *on*. Because the influence of the season on the state of the sprinkler is no longer in effect, we furthermore remove the arrow from X_1 to X_3. This is the general idea of how an intervention is interpreted: the original causal Bayesian net is modified by fixing the value of a variable and removing its dependence on its typical causes. The modification corresponding to the intervention of turning on the sprinkler is shown in figure 1.7 on the right.

An important difference between observation and intervention is the way in which they propagate through the causal Bayesian network. An observation propagates to variables that play a role in the causal explanation of the observed event, as well as variables that represent direct and indirect effects of the observed event. An intervention, on the other hand, propagates only to the direct and indirect effects of the variable whose value is fixed. In other words: causal explanation plays no role in determining how an intervention affects other variables.

Intervention and Observation in Argumentation

The distinction between the two types of change in argumentation that we identified resembles the distinction between intervention and observation in causal Bayesian networks. For this reason we refer to the two types of change as *intervention in argumentation* and *observation in argumentation*:

- Given an argumentation framework, an intervention represents a (hypothetical) action in a debate. The action leads to change in the evaluation of the argumentation framework that reflects the effects of the action.
- Given an argumentation framework, an observation represents a new piece of information about the status of one or more arguments. An observation leads to a revised evaluation of the argumentation framework, which accounts both for the explanation of the observation, as well as its effects.

The two types of change in argumentation lead to two types of entailment we can perform on the basis of a given argumentation framework. *Intervention-based entailment* is concerned with the consequences of (hypothetical) actions performed in an argumentation framework, and *observation-based entailment* is concerned with the consequences of new information from the environment about the evaluation of an argumentation framework. This perspective allows us to model the two types of change in a uniform way, and to compare the two in terms of their behaviour, such has how interventions and observations propagate through an argumentation framework.

1.3 Research Questions

Our main goal in chapter 3 and 4 is to formally develop a model to study change due to intervention and observation in argumentation, and to study and compare the behaviour of these two types of change. We start in chapter 3 by looking at intervention. We model change due to intervention as a form of *entailment* between interventions and consequences of interventions. This allows us to study the behaviour of an argumentation framework from a dynamic perspective, by focussing on properties satisfied by these entailment relations. The questions we address are the following.

- How can we model intervention-based entailment in argumentation?
- How does intervention-based entailment behave with respect to some general principles of well-behaved inference?
- What is the role of directionality and noninterference in the behaviour of intervention-based entailment?

In chapter 4 we look at obervation. Like we do for intervention, we model change due to observation as a form of entailment between observations and consequences of observations. The first question we address is the following.

- What is a rational way for an agent to revise the evaluation of an argumentation framework to account for an observation?

This question breaks down into the following two sub-questions.

- By what mechanism does a rational agent decide how to change his argumentation framework in order to revise its evaluation due to an observation?
- What are the conditions that characterize a rational way to revise the evaluation of an argumentation framework due to an observation?

Having modelled the two types of change as two forms of entailment, we are in a position to compare the two.

- What are the main differences between intervention and observation in terms of how interventions and observations propagate through an argumentation framework?

In chapter 5 we focus on abduction in argumentation. The problem of abduction has, in the field of artificial intelligence, been extensively studied in the context of logic programming (see, for example, Denecker and Kakas [39] for an overview). Furthermore, it has been shown that logic programming can be seen as a form of instantiated argumentation. This raises the following question.

- Is there a model of *abduction in abstract argumentation* that can be seen as an abstraction of *abduction in logic programming*, in the same way that *abstract argumentation* has been shown to be an abstraction of *logic programming*?

Proof theories in argumentation answer the question whether or not an argument is accepted under a given semantics. These proof theories are often presented as two-person dialogue games played according to particular sets of rules. As such, they relate different semantics to stereotypical patterns found in real world dialogue. For example, the grounded semantics has been shown to relate to a kind of persuasion dialogue, while the preferred semantics relate to Socratic-style dialogue [30, 32]. Having defined a model for abduction in argumentation, one may wonder whether there are stereotypical dialogue patterns that correspond to the main reasoning tasks associated with this model, namely to find explanations for a given observation. This leads to the following research question.

- Having defined a model of abduction in abstract argumentation, is it possible to define dialogical proof procedures for the problem of finding explanations for a given observation?

In chapter 6 we focus on change in preference-based argumentation frameworks. Preferences are usually assumed to be fixed and no account is provided of how or why they may change. This leads to the final research question.

- How can change in preference-based argumentation be modelled?

1.4 Thesis Overview

- **Chapter 2: Preliminaries**

  We first present the necessary basics of argumentation theory: argumentation frameworks and extension-based semantics [42] and labelling-based semantics [26, 33]. Based on these concepts we define a logical labelling language, which provides us with the flexibility we need to reason about interventions and observations in abstract argumentation.

  We then present the necessary basics of the *KLM* approach to non-monotonic inference due to Kraus, Lehmann and Magidor [62]. This includes both syntactical characterisations (the so called *KLM properties* for well-behaved non-monotonic inference), semantical characterisations, and the results connecting these characterizations.

- **Chapter 3: Intervention-Based Entailment in Argumentation**

  We model change due to intervention (i.e., actions in a debate corresponding to new arguments and attacks being added) as a form of entailment between interventions and consequences of interventions. This allows us to study the behaviour of an argumentation framework from a dynamic perspective, by focussing on properties satisfied by these entailment relations.

  This puts us in a position to study a number of properties of different argumentation semantics, by investigating how these entailment relations behave with respect to some general principles of well-behaved inference. These principles are based on the KLM properties for well-behaved non-monotonic inference.

  We then investigate the role of directionality and noninterference in the behaviour of intervention-based entailment. These properties ensure that the effect of an intervention propagates throughout the argumentation framework in a well-behaved manner.

- **Chapter 4: Observation-Based Entailment in Argumentation**

  We model change due to observation, i.e., new information leading to a revised evaluation of an argumentation framework. Like we do for intervention, we model it as a form of entailment between observations and consequences of observations.

  Having modelled the two types of change as two forms of entailment, we are in a position to compare the two. Specifically, we study the differences between intervention and observation in terms of how interventions and observations propagate through an argumentation framework.

- **Chapter 5: Abduction in Argumentation and Logic Programming**

  We develop a model of abduction in abstract argumentation, based on the idea that changes to an argumentation framework act as hypotheses to explain the support of an observation. We present dialogical proof theories for the main reasoning tasks (i.e., finding hypotheses that explain skeptical/credulous support) and we show that our model can be instantiated on the basis of abductive logic programs.

- **Chapter 6: Change in Preference-Based Argumentation**

  We develop a dynamic model of preference-based argumentation, centring on what we call property-based argumentation frameworks. It is based on Dietrich and List's model of property-based preference and it provides an account of how and why preferences in argumentation may change. The idea is that preferences over arguments are derived from preferences over properties of arguments, and change as the result of moving to different motivational states. We also provide a dialogical proof theory for the task of checking whether there exists some motivational state in which an argument is accepted.

The work presented in chapter 3 and 4 is based in part on joint work with Richard Booth, Souhila Kaci and Leendert van der Torre [25, 23, 24] but contain a number of new ideas and results. The work in chapter 5 is based on joint work with Richard Booth, Dov Gabbay, Souhila Kaci and Leendert van der Torre [21]. The work in chapter 5 is based on joint work with Richard Booth and Souhila Kaci [22].

Chapter 2

Preliminaries

2.1 Abstract Argumentation

The central notion in abstract argumentation theory is that of an *argumentation framework*, which is a directed graph represented by a set A of *arguments* and a binary relation $\rightsquigarrow$ over A called the *attack relation*. [42] To simplify our discussion we assume that A is *finite*, and that it is a subset of an infinite set $\mathcal{U}$ called the *universe of arguments*.

Definition 2.1.1. Let $\mathcal{U}$ be a set whose elements are called *arguments*. An *argumentation framework* is a pair $F = (A, \rightsquigarrow)$ where A is a finite subset of $\mathcal{U}$ and $\rightsquigarrow \subseteq A \times A$ is a relation called the *attack relation*. We denote by $\mathcal{F}$ the set of all argumentation frameworks.

Given an argumentation framework $(A, \rightsquigarrow)$ we say that an argument $a \in A$ *attacks* an argument $b \in A$ if and only if $(a, b) \in \rightsquigarrow$. We will mostly use infix notation and write $a \rightsquigarrow b$ instead of $(a, b) \in \rightsquigarrow$. We extend this notation as follows. Given a set $B \subseteq A$, we say that x attacks B (written $x \rightsquigarrow B$) whenever $x \rightsquigarrow y$ for some $y \in B$ and, conversely, that B attacks x (written $B \rightsquigarrow y$) whenever $x \rightsquigarrow y$ for some $x \in B$. Given two sets $B, B' \subseteq A$ we say that B attacks B (written $B \rightsquigarrow B'$) whenever $x \rightsquigarrow y$ for some $x \in B$ and $y \in B'$. In addition, we write $x \not\rightsquigarrow y$ whenever it is *not* the case that $x \rightsquigarrow y$, and similarly for $x \not\rightsquigarrow B$, $B \not\rightsquigarrow x$ and $B \not\rightsquigarrow B'$. Given an argumentation framework F we furthermore denote by x^- (resp. B^-) the set of arguments attacking x (resp. some $x \in B$) and by x^+ (resp. B^+) the set of arguments attacked by x (resp. some $x \in B$).

An important idea in abstract argumentation is that we can reason about the acceptability of arguments without specifying their content. In chapter 5 we work with argumentation frameworks whose arguments have content that is generated on the basis of a logic program. In this section, however, we do not assign any meaning to abstract arguments.

Figures 2.1a, 2.1b and 2.1c depict examples of argumentation frameworks. We will refer back to these argumentation frameworks throughout this chapter.

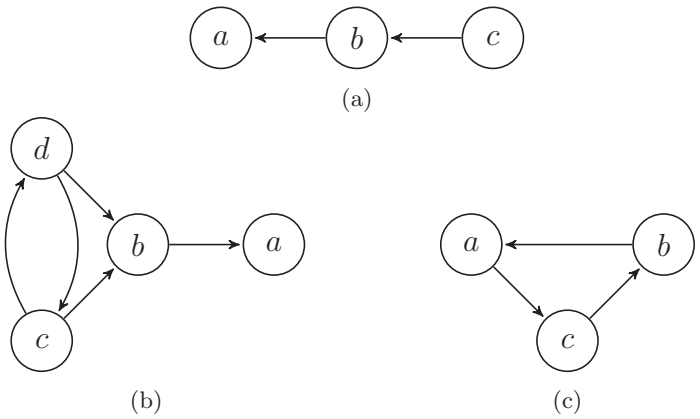


Figure 2.1: Argumentation frameworks used as running examples.

Given an argumentation framework, the main reasoning problem is to determine the positions that a rational agent should take with respect to the acceptability of the arguments. Solutions to this problem can be represented by *extensions*, which are sets of simultaneously acceptable arguments. They can also be represented by *labellings*, which are functions that associate with every argument a label that indicates the argument's *acceptance status*. We use the labelling-based approach in chapters 3 and 4 and the extension-based approach in chapters 5 and 6. The necessary basics of the two approaches are presented in the following two subsections.

2.1.1 Extension-Based Semantics

An extension-based semantics is defined by specifying the conditions that an extension (a set of arguments) must satisfy in order to represent a rational point of view.[1] We represent a semantics by a function $\mathcal{E}_\sigma(F)$ that returns all extensions of F under the semantics σ.

Two basic properties for an extension to represent a rational point of view are *conflict-freeness* and *admissibility*. An extension E is conflict-free if is not self-attacking (i.e., no member of E attacks another member of E). Furthermore, E is admissible if it defends all its members, where defence by E of an argument x is defined as follows.

Definition 2.1.2. [42] Let $F = (A, \rightsquigarrow)$ be a framework. An extension $E \subseteq A$ *defends* an argument $y \in A$ if and only if for all $x \in A$ such that $x \rightsquigarrow y$ it holds that $E \rightsquigarrow x$. We let $\mathcal{D}_F(E)$ denote the set of all arguments $y \in A$ such that E defends y.[2]

Definition 2.1.3. [42] Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An extension $E \subseteq A$ is

[1] The usage of the term semantics in argumentation has its roots in Logic Programming and is somewhat different from its use outside of the area of logic programming and argumentation, such as the preferential model semantics and Kripke semantics.

[2] Defence of x by E was called *acceptability of x w.r.t. E* by Dung [42]. The usage of the term defence is more common in recent literature.

	Co	Gr	Pr	SS	St
$\{a, c\}$	✓		✓	✓	✓
$\{a, d\}$	✓		✓	✓	✓
$\emptyset$	✓	✓			

Table 2.1: Extensions in figure 2.1b.

- *conflict-free* ($E \in \mathcal{E}_{Cf}(F)$) iff $E \not\rightsquigarrow E$.
- *admissible* ($E \in \mathcal{E}_{Ad}(F)$) iff E is conflict-free and $E \subseteq \mathcal{D}_F(E)$.

Using these properties we can define the most widely used semantics, namely the *complete*, *grounded*, *preferred*, *semi-stable* and *stable* semantics.

Definition 2.1.4. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An extension $E \subseteq A$ is

- *complete* ($E \in \mathcal{E}_{Co}(F)$) iff $E \in \mathcal{E}_{Ad}(F)$ and $E = \mathcal{D}_F(E)$. [42]
- *grounded* ($E \in \mathcal{E}_{Gr}(F)$) iff $E \in \mathcal{E}_{Co}(F)$ and there is no $E' \in \mathcal{E}_{Co}(F)$ such that $E' \subset E$. [42]
- *preferred* ($E \in \mathcal{E}_{Pr}(F)$) iff $E \in \mathcal{E}_{Ad}(F)$ and there is no $E' \in \mathcal{E}_{Ad}(F)$ such that $E \subset E'$. [42]
- *semi-stable* ($E \in \mathcal{E}_{SS}(F)$) iff $E \in \mathcal{E}_{Ad}(F)$ and there is no $E' \in \mathcal{E}_{Ad}(F)$ such that $A \setminus (E' \cup E'^{+}) \subset A \setminus (E \cup E^{+})$. [27]
- *stable* ($E \in \mathcal{E}_{St}(F)$) iff E is conflict-free and for all $x \in A \setminus E$, it holds that $E \rightsquigarrow x$. [42]

Table 2.1 shows the extensions under each of these semantics for the argumentation frameworks shown in figure 2.1b. What are the intuitions behind these semantics? First of all, a complete extension represents a point of view that is admissible (i.e., it is conflict-free and defends all its members) and additionally includes all arguments it defends. Note that every argumentation framework has at least one complete (and hence admissible) extension [42]. The grounded extension is a complete extension that is minimal with respect to set-inclusion. It represents the most sceptical point of view and it is always unique. The grounded extension is also characterized by a fix point theory.

Proposition 2.1.1. *[42, Theorem 25] Given an argumentation framework F, the grounded extension of F coincides with the least fixed point of $\mathcal{D}_F$.*

A preferred extension represents one of the maximally credulous positions that one can take. It is an admissible extension that is maximal with respect to set inclusion. Note that every argumentation framework has at least one preferred extension. Table 2.1 demonstrates that not every complete extension is preferred: the extension $\emptyset$, which is complete, is not preferred. The converse, however, does hold.

Proposition 2.1.2. *[42, Theorem 25] For all $F \in \mathcal{F}$, $\mathcal{E}_{Pr}(F) \subseteq \mathcal{E}_{Co}(F)$.*

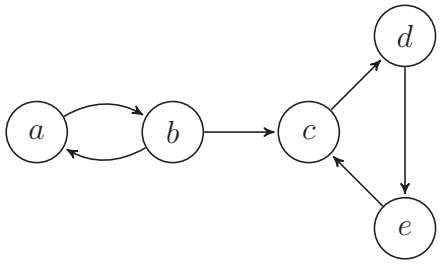


Figure 2.2: Not all preferred extensions are stable.

	Co	*Gr*	*Pr*	*SS*	*St*
$\{a\}$	✓		✓		
$\{b, d\}$	✓		✓	✓	✓
$\emptyset$	✓	✓			

Table 2.2: Extensions in figure 2.2.

A stable extension is a conflict-free extension that makes a decision on *all* arguments. This means that every argument is either a member of the extension or it is attacked by the extension. Note that stable extensions are also preferred.

Proposition 2.1.3. *[42, Theorem 25] For all $F \in \mathcal{F}$, $\mathcal{E}_{St}(F) \subseteq \mathcal{E}_{Pr}(F)$.*

Not every preferred extension is, however, stable. Moreover, a stable extension is not guaranteed to exist. Both facts are demonstrated by the argumentation framework shown in figure 2.1c, which has one extension that is complete, grounded and preferred, namely the empty set, but has no stable extension.

The semi-stable semantics was introduced by Caminada [27], although it is the same as what Verheij called the *admissible stage* semantics [89]. A semi-stable extension is an admissible extension E whose *range*, which is the set $E \cup E^+$, is maximal with respect to set inclusion. The range of E is intuitively the set of arguments about which the extension makes a decision, i.e., those accepted and those rejected. Note that every stable extension is also semi-stable and every semi-stable extension also preferred.

Proposition 2.1.4. *[27, Theorem 1 and 2] For all $F \in \mathcal{F}$, $\mathcal{E}_{St}(F) \subseteq \mathcal{E}_{SS}(F) \subseteq \mathcal{E}_{Pr}(F)$.*

An attractive feature of the semi-stable semantics is that it coincides with the stable semantics *if* a stable extension exists but, unlike a stable extension, the existence of a semi-stable extension is guaranteed. This is not true under the preferred semantics, because not every preferred extension is always stable. This is demonstrated by the argumentation framework shown in figure 2.2, of which the extensions are shown in table 2.2 (this example is due to Caminada [27]). Here, the extension $\{a\}$ is preferred but not semi-stable.

Sceptical and Credulous Acceptance

An argument is said to be *sceptically* accepted under a semantics σ if the argument is a member of *all* σ extensions, and *credulously* accepted if it is a member of *some* σ extension. Intuitively, a credulously accepted argument is an argument that may survive the conflict, depending on the position taken, whereas a sceptically accepted argument survives the conflict no matter what position is taken.

Definition 2.1.5. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and σ a semantics. An argument $x \in A$ is *sceptically* accepted under the σ semantics if and only if $x \in E$ for every $E \in \mathcal{E}_\sigma(F)$. An argument $x \in A$ is *credulously* accepted under the σ semantics if and only if $x \in E$ for some $E \in \mathcal{E}_\sigma(F)$.

Note that the set of arguments sceptically accepted under the complete semantics always coincides with the grounded extension. Furthermore, since the grounded extension is unique, sceptical and credulous acceptance under the grounded semantics coincide.

Figure 2.1b demonstrates that the grounded semantics can be too sceptical in some scenarios. Here, neither c nor d is a member of the grounded extension. As a result, the argument a is also not a member of the grounded extension. But we may still conclude that a is accepted, even if we do not know whether to accept c or d. The preferred semantics correctly captures this, because a is indeed sceptically accepted under the preferred semantics, while b and c are not.

Inclusion Relations Between the Different Semantics

Figure 2.3 gives an overview of the inclusion relations between the extension-based semantics discussed here. It combines proposition 2.1.4, 2.1.2 and definition 2.1.4. This figure also appears in the work of Caminada [26].

2.1.2 Labelling-Based Semantics

An alternative way to represent a position on which arguments to accept is by using labellings. The labelling-based semantics that we present here are the same as the extension-based semantics presented earlier, but their formalization is different. While an extension only captures the arguments that are accepted in a given position, a labelling assigns to each argument an *acceptance status*. This general approach can be traced back to Pollock [75], Jakobovits and Vermeir [59] and Verheij [89].

Three-Valued Labellings

We follow Caminada [26], who defined the semantics described in the previous section using three-valued labellings. The benefit of three-valued labellings over extensions is that these labellings not only distinguish arguments that are accepted and not accepted, but also those that are explicitly rejected and those

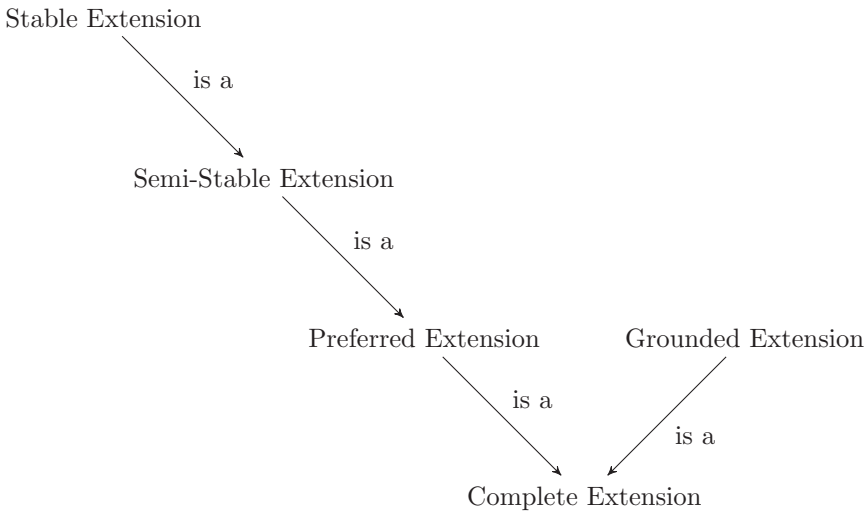


Figure 2.3: Inclusion relations between different semantics.

that are undecided. This means that each argument is assigned one of the following three labels:

- **in** meaning that the argument is accepted,
- **out** meaning that the argument is rejected, and
- **und** (for *undecided*) meaning that the argument is neither rejected nor accepted.

Formally, a labelling of an argumentation framework $F = (A, \rightsquigarrow)$ is a function from A to $\{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$.

Definition 2.1.6. A *labelling* of an argumentation framework $(A, \rightsquigarrow)$ is a function $L : A \to \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$. Given a label $l \in \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$ we define $L^{-1}(l)$ as $\{x \in A \mid L(x) = l\}$. Given an argumentation framework F, we let $\mathcal{L}(F)$ denote the set of all labellings of F.

Given an argumentation framework $F = (\{x_1, \ldots x_n\}, \rightsquigarrow)$ we will also denote a labelling $L \in \mathcal{L}(F)$ by the set of pairs $\{(x_1, L(x_1)), \ldots, (x_n, L(x_n))\}$.

Properties of Labellings

We now define a number of properties that form the basis of the labelling-based semantics that we consider. First of all, an argument is said to be *legally* **in** if all its attackers are labelled **out**, *legally* **out** if some attacker is labelled **in**, and *legally* **und** if no attacker is labelled **in** and some attacker is labelled **und**. [33] Intuitively, an argument is legally labelled if its label is justified on the basis of the labels of the attackers.

Definition 2.1.7. [33] Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $L \in \mathcal{L}(F)$. An argument $x \in A$ is said to be:

1. Legally **in** in L with respect to F iff $L(x) = \textbf{in}$ and for all $y \in x^-$, $L(y) = \textbf{out}$.
2. Legally **out** in L with respect to F iff $L(x) = \textbf{out}$ and for some $y \in x^-$, $L(y) = \textbf{in}$.
3. Legally **und** in L with respect to F iff $L(x) = \textbf{und}$ and for all $y \in x^-$, $L(y) \neq \textbf{in}$ and for some $y \in x^-$, $L(y) = \textbf{und}$.

We say that a labelling L satisfies **in**-legality, **out**-legality or **und**-legality if every argument labelled **in**, **out** or **und** is also legally **in**, **out** or **und**.

Definition 2.1.8. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $L \in \mathcal{L}(F)$.

- L satisfies **in**-*legality* iff for all $x \in A$ s.t. $L(x) = \textbf{in}$, x is legally **in** in L.
- L satisfies **out**-*legality* iff for all $x \in A$ s.t. $L(x) = \textbf{out}$, x is legally **out** in L.
- L satisfies **und**-*legality* iff for all $x \in A$ s.t. $L(x) = \textbf{und}$, x is legally **und** in L.

We furthermore use the properties called *reinstatement* and *rejection* [4]. A labelling satisfies reinstatement if every argument whose attackers are all labelled **out** is labelled **in**, and satisfies rejection if every argument of which some attacker is labelled **in** is labelled **out**.

Definition 2.1.9. [4] Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $L \in \mathcal{L}(F)$.

- L satisfies *reinstatement* iff for all $x \in A$ s.t. for all $y \in x^-$, $L(y) = \textbf{out}$, we have $L(x) = \textbf{in}$.
- L satisfies *rejection* iff for all $x \in A$ s.t. for some $y \in x^-$, $L(y) = \textbf{in}$, we have $L(x) = \textbf{out}$.

Note that every labelling satisfying **out**-legality and **und**-legality also satisfies reinstatement, and every labelling satisfying **in**-legality and **und**-legality also satisfies rejection.

Conflict-Free, Admissible and Complete Labellings

We say that a labelling is *conflict-free* if it satisfies **in**-legality and rejection. This amounts to the condition that every neighbour of every **in**-labelled argument is labelled **out**.[3]

[3]This condition is stronger than the condition of conflict-freeness as it appears, e.g., in [33]. We do this for technical convenience, and it has no consequences for the definitions that follow.

Definition 2.1.10. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A labelling $L \in \mathcal{L}(F)$ is *conflict-free* ($L \in \mathcal{L}_{Cf}(F)$) if and only if L satisfies **in**-legality and rejection.

An admissible labelling is a labelling satisfying **in**-legality and **out**-legality. For a labelling to be complete, it must furthermore satisfy **und**-legality.

Definition 2.1.11. [33] Let $F = (A, \rightsquigarrow)$. A labelling $L \in \mathcal{L}(F)$ is admissible ($L \in \mathcal{L}_{Ad}(F)$) if and only if L satisfies **in**-legality and **out**-legality. A labelling $L \in \mathcal{L}(F)$ is complete ($L \in \mathcal{L}_{Co}(F)$) if and only if L satisfies **in**-legality, **out**-legality and **und**-legality.

The set of arguments labelled **in** in an admissible labelling of an argumentation framework F is an admissible set of F. Moreover, for every admissible set E of F, there is an admissible labelling such that the set of **in**-labelled arguments is exactly E. This admissible labelling is, however, not in general unique, and hence the mapping between admissible labellings and extensions is not one-to-one.

Proposition 2.1.5. *Let F be an argumentation framework.*

- *For all $L \in \mathcal{L}_{Ad}(F)$, $L^{-1}(\mathbf{in}) \in \mathcal{E}_{Ad}(F)$.*
- *For all $E \in \mathcal{E}_{Ad}(F)$, there is an $L \in \mathcal{L}_{Ad}(F)$ such that $E = L^{-1}(\mathbf{in})$.*

Proof. This follows from [33, Theorem 4]. □

Unlike the mapping between admissible labellings and extensions, the mapping between complete labellings and extensions is one-to-one.

Proposition 2.1.6. *Let F be an argumentation framework.*

- *For all $L \in \mathcal{L}_{Co}(F)$, $L^{-1}(\mathbf{in}) \in \mathcal{E}_{Co}(F)$.*
- *For all $E \in \mathcal{E}_{Co}(F)$, there is a unique $L \in \mathcal{L}_{Co}(F)$ such that $E = L^{-1}(\mathbf{in})$.*

Proof. This follows from [33, Theorem 4]. □

Grounded, Preferred, Semi-Stable and Stable Labellings

Grounded, preferred, semi-stable and stable labelling are defined by putting additional restrictions on a complete labelling.

Definition 2.1.12. [33] Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A labelling L is

- *grounded* ($L \in \mathcal{L}_{Gr}(F)$) iff $L \in \mathcal{L}_{Co}(F)$ and there is no $L' \in \mathcal{L}_{Co}(F)$ such that $L'^{-1}(\mathbf{in}) \subset L^{-1}(\mathbf{in})$.
- *preferred* ($L \in \mathcal{L}_{Pr}(F)$) iff $L \in \mathcal{L}_{Co}(F)$ and there is no $L' \in \mathcal{L}_{Co}(F)$ such that $L^{-1}(\mathbf{in}) \subset L'^{-1}(\mathbf{in})$.

Labelling	*Co*	*Gr*	*Pr*	*SS*	*St*
$\{(a, \mathbf{in}), (b, \mathbf{out}), (c, \mathbf{in}), (d, \mathbf{out})\}$	✓		✓	✓	✓
$\{(a, \mathbf{in}), (b, \mathbf{out}), (c, \mathbf{out}), (d, \mathbf{in})\}$	✓		✓	✓	✓
$\{(a, \mathbf{und}), (b, \mathbf{und}), (c, \mathbf{und}), (d, \mathbf{und})\}$	✓	✓			

Table 2.3: Labellings in figure 2.1b.

- *semi-stable* ($L \in \mathcal{L}_{SS}(F)$) iff $L \in \mathcal{L}_{Co}(F)$ and there is no $L' \in \mathcal{L}_{Co}(F)$ such that $L'^{-1}(\mathbf{und}) \subset L^{-1}(\mathbf{und})$.
- *stable* ($L \in \mathcal{L}_{St}(F)$) iff $L \in \mathcal{L}_{Co}(F)$ and $L'(\mathbf{und}) = \emptyset$.

Thus, grounded and preferred labellings are defined by minimizing or maximizing the set of **in**-labelled arguments. Semi-stable extensions are defined by minimizing the set of arguments that is labelled **und**. Finally, a stable labelling is a complete labelling in which no argument is labelled **und**. These definitions correspond to the respective extension-based definitions of these semantics in the following way.

Proposition 2.1.7. *Let F be an argumentation framework and let $\sigma \in \{Gr, Pr, SS, St\}$. For all $L \in \mathcal{L}_{Co}(F)$,*

$$L \in \mathcal{L}_{\sigma}(F) \text{ if and only if } L^{-1}(\mathbf{in}) \in \mathcal{E}_{\sigma}(F).$$

Proof. This follows from proposition 2.1.6 together with definition 2.1.12. □

The benefit of using labellings over extensions is that we can distinguish arguments that are explicitly rejected (labelled **out**) and undecided (labelled **und**). By contrast, an extension only identifies the accepted arguments, and does not explicitly allow a distinction between arguments that are rejected or undecided. Table 2.3 shows the complete labellings of the argumentation framework shown in figure 2.1b. Comparing this with table 2.1, we see that each extension corresponds to a labelling where all accepted arguments are labelled **in** and the non-accepted arguments either **out** or **und**.

For ease of presentation we also represent labellings of argumentation frameworks by colourings of the nodes. We adopt the convention to colour arguments labelled **in** white, arguments labelled **out** black, and arguments labelled **und** gray. Figure 2.4 shows the colourings for three labellings listed in table 2.3.

Properties

The correspondence between labellings and extensions established in proposition 2.1.6 and 2.1.7 means that a number of results carry over from the extension-based setting to the labelling based setting. We state these results here so that we can refer back to them in what follows. First of all, the inclusion relations between the labelling semantics are the same as those between the extension-based semantics.

Proposition 2.1.8. *Let F be an argumentation framework.*

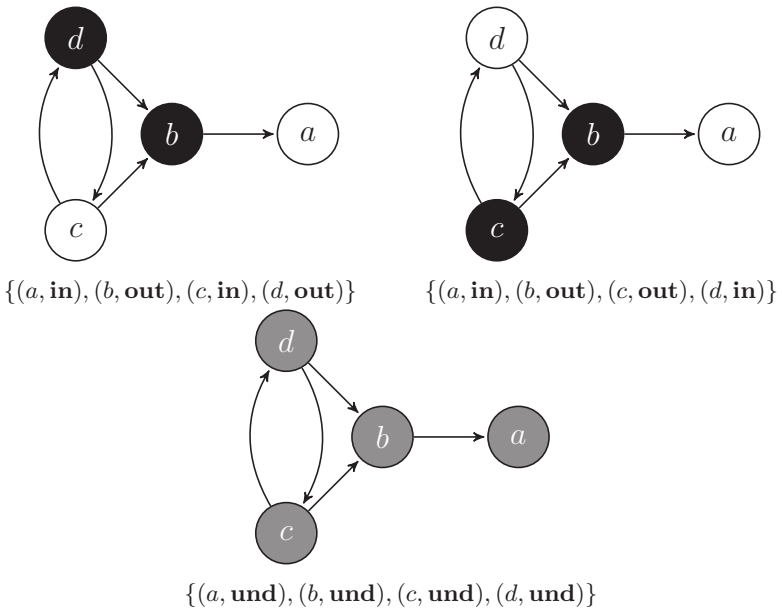


Figure 2.4: Three labellings as colourings of nodes.

1. $\mathcal{L}_{Gr}(F) \subseteq \mathcal{L}_{Co}(F)$.
2. $\mathcal{L}_{St}(F) \subseteq \mathcal{L}_{SS}(F) \subseteq \mathcal{L}_{Pr}(F) \subseteq \mathcal{L}_{Co}(F)$.

Proof. See preceding discussion. □

Like the grounded extension, the grounded labelling is unique.

Proposition 2.1.9. *For all $F \in \mathcal{F}$, $\mid \mathcal{L}_F(Gr) \mid = 1$.*

Proof. Follows from proposition 2.1.7 together with the fact that the grounded extension is unique. □

Like under the extension-based semantics, existence of complete, grounded, preferred and semi-stable labellings is guaranteed, but not of stable labellings.

Proposition 2.1.10. *For all $F \in \mathcal{F}$, $\sigma \in \{Co, Gr, Pr, SS\}$, $\mathcal{L}_F(\sigma) \neq \emptyset$. For some $F \in \mathcal{F}$, $\mathcal{L}_F(St) = \emptyset$.*

Proof. Follows from proposition 2.1.7 together with the fact that the existence of complete, grounded, preferred and semi-stable extensions is guaranteed, but the existence of stable extensions is not. □

Finally, we will later refer later to the property of *language independence*. [4] This property states, if satisfied by a semantics, that isomorphic argumentation frameworks give rise to equivalent (modulo isomorphism) sets of labellings. It is an expression of the idea that arguments are indeed abstract, in the sense that the assigned labels depend only on the topology of the argumentation

framework. To define language independence, we need the following notion of isomorphism.

Definition 2.1.13. Let $F_1 = (A_1, \rightsquigarrow_1)$ and $F_2 = (A_2, \rightsquigarrow_2)$ be two argumentation frameworks. We write $F_1 =_m F_2$ if and only if m is a bijection from A_1 to A_2 such that $x \rightsquigarrow_1 y$ if and only if $m(x) \rightsquigarrow_2 m(y)$. We say that F_1 and F_2 are *isomorphic* if and only if $F_1 =_m F_2$ for some m.

Definition 2.1.14. A labelling-based semantics σ satisfies the language independence principle iff for every $F_1, F_2 \in \mathcal{F}$ such that $F_1 =_m F_2$ it holds that $\mathcal{L}_\sigma(F_1) = \{L_m \mid L \in \mathcal{L}_\sigma(F_2)\}$, where L_m is a labelling of F_1 defined by $L_m(x) = L(m(x))$.

The language independence property is satisfied by all the semantics that we consider.

Proposition 2.1.11. *For all $\sigma \in \{Co, Gr, Pr, SS, St\}$, σ satisfies the language independence principle.*

2.1.3 Labelling-Based Entailment

Instead of identifying the evaluation of an argumentation framework with a set of labellings, we use a more flexible method based on the notion of *labelling-based entailment*. This method allows us to model dynamic modes of evaluation (the subject of chapter 3 and 4) in a more convenient way. The idea is to use a symbolic representation in terms of a logical *labelling language*. The reasoning about argument acceptance under a given semantics can then be captured by an entailment relation between argumentation frameworks and consequences expressed in this language. This language was used before by Booth et al. [23, 24].

Syntax

Given a set A of arguments we denote by $\texttt{lang}(A)$ the language it determines. The atoms of this language are of the form $\mathbf{in}(x)$, $\mathbf{out}(x)$ and $\mathbf{und}(x)$ for some argument $x \in A$. The meaning of these atoms is that the argument x is labelled **in**, **out** or **und**. The language furthermore consists of $\top$ and $\bot$ (representing truth and falsity) and is closed under the usual logical connectives.

Definition 2.1.15. Let $A \subseteq \mathcal{U}$ be a set of arguments

- An *atom* is a symbol of the form $\mathbf{in}(x)$, $\mathbf{out}(x)$ or $\mathbf{und}(x)$, where $x \in A$.
- $\top, \bot$ and every atom is a formula.
- If ϕ, ψ are formulas then so is $(\phi \vee \psi)$.
- If ϕ is a formula then so is $\neg\phi$.
- Nothing else is a formula.

We denote the set of formulas by $\texttt{lang}(A)$. Given an argumentation framework $F = (A, \rightsquigarrow)$ we also write $\texttt{lang}(F)$ instead $\texttt{lang}(A)$.

Other connectives we use are $\wedge$, $\rightarrow$ and $\leftrightarrow$, defined as usual in terms of $\neg$ and $\vee$ (i.e., $(\phi \wedge \psi) = \neg(\neg\phi \vee \neg\psi)$, $(\phi \rightarrow \psi) = (\neg\phi \vee \psi)$ and $(\phi \leftrightarrow \psi) = (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$). We omit parentheses if this does not lead to confusion. We shall furthermore denote literals by α, β or γ and, by slight abuse of notation, denote by $\neg\alpha$ the literal that is the negation of the literal α.

Semantics

We define the relation $\models$, along with a number of related notions, as follows.

Definition 2.1.16. Given an argumentation framework $F = (A, \rightsquigarrow)$, the relation $\models \subseteq \mathcal{L}(F) \times \texttt{lang}(F)$ is defined by:

- $L \models \top$ and $L \not\models \bot$,
- $L \models \mathbf{in}(x)$ iff $L(x) = \mathbf{in}$,
- $L \models \mathbf{out}(x)$ iff $L(x) = \mathbf{out}$,
- $L \models \mathbf{und}(x)$ iff $L(x) = \mathbf{und}$,
- $L \models \neg\phi$ iff $L \not\models \phi$,
- $L \models (\phi \vee \psi)$ iff $L \models \phi$ or $L \models \psi$.

We say that L *satisfies* ϕ whenever $L \models \phi$, and that ϕ *classically entails* ψ (written as $\phi \models \psi$) if every labelling that satisfies ϕ also satisfies ψ.

Labelling-Based Entailment Relations

Posing a query about an argumentation framework F can be interpreted as asking whether some formula $\phi \in \texttt{lang}(F)$ is a consequence of F under the semantics σ. This can be seen as a process of entailment between argumentation frameworks and formulas. Accordingly, each semantics determines such an entailment relation, which we call the *σ-entailment relation* and denote by $\models_\sigma$.

Definition 2.1.17. Given a semantics σ we define the *σ entailment relation* $\models_\sigma$ by:

$$\text{for all } F \in \mathcal{F}, \phi \in \texttt{lang}(F): \quad F \models_\sigma \phi \text{ iff } \forall L \in \mathcal{L}_\sigma(F), L \models \phi.$$

This representation covers not only sceptical acceptance (i.e., x is sceptically accepted iff $F \models_\sigma \mathbf{in}(x)$) but also credulous acceptance (i.e. x is credulously accepted iff $F \not\models_\sigma \neg\mathbf{in}(x)$). It also covers more general queries that are not covered by the simple notions of sceptical or credulous acceptance. Examples are queries involving conjunctions, disjunctions, and arguments being labelled **out** and **und**.

Example 2.1.1. *Let F be the argumentation framework shown in figure 2.1b. We have, for example, $F \models_{Pr} \mathbf{in}(c) \vee \mathbf{in}(d)$, $F \models_{Pr} \mathbf{in}(a)$, $F \not\models_{Co} \mathbf{in}(c) \vee \mathbf{in}(d)$ and $F \models_{Co} (\mathbf{in}(c) \vee \mathbf{in}(d)) \rightarrow \mathbf{in}(a)$.*

2.2 KLM Logics

In this section we present the basics of the influential KLM approach to non-monotonic reasoning. It is due to Kraus, Lehmann and Magidor [62] (hence the name KLM) and was extended by Lehmann [63]. The approach contains elements of work on non-monotonic logics by Gabbay [48] and Shoham [85] and can be traced back to philosophical work on conditional statements by Adams [1]. The starting point in the KLM approach is the notion of a *non-monotonic entailment relation*. This is a relation $\mid\sim$ between formulas of a propositional language where $\phi \mid\sim \psi$ intuitively means that ψ plausibly or defeasibly follows from ϕ. Non-monotonicity means that, unlike classical monotonic entailment relations, these relations generally do not satisfy the property of Monotony:

(Monotony) If $\phi \mid\sim \psi$ then $\phi \wedge \chi \mid\sim \psi$.

A typical example against monotony is the following: suppose we learn that Tweety is a bird and that it plausibly follows that Tweety flies: *bird* $\mid\sim$ *flies*. If we subsequently learn that Tweety is a penguin, then it no longer follows that Tweety flies. However, monotony implies that it *still* follows that Tweety flies: *bird* $\wedge$ *penguin* $\mid\sim$ *flies*. This demonstrates the undesirability of monotony from a common sense reasoning perspective.

One may ask: how should an entailment relation behave in the absence of monotony? This is the question addressed by Kraus, Lehmann and Magidor. They introduced four sets of properties (often called the *KLM properties*) and corresponding semantic models with the aim of characterizing classes of well-behaved non-monotonic entailment relations. These are, from the strongest to the weakest, the class of *rational*, *preferential*, *loop-cumulative* and *cumulative* entailment relations.

In the following two chapters we look at entailment relations for entailment in abstract argumentation on the basis of interventions and observations. These relations are also non-monotonic, and hence we can evaluate and try to characterize their behaviour in terms of the KLM properties. In this section we introduce the definitions and results concerning KLM logics that are relevant to this objective.

2.2.1 Syntactic Characterizations

Syntactically, the classes of *cumulative*, *loop-cumulative*, *preferential* and *rational* inference relations are characterized by sets of properties they satisfy. For the definition we work with an inference relation $\mid\sim$ over a propositional language `lang` that is closed under the usual connectives. The symbols ϕ, ψ and χ used in specifying the properties in the following definition range over

all formulas in `lang`. A relation $\mathrel{|\!\sim}$ is said to satisfy a property if it satisfies all instances.

Definition 2.2.1. Let $\mathrel{|\!\sim}$ be a relation over a propositional language `lang`. The properties *Reflexivity*, *Left Logical Equivalence*, *Right Weakening*, *Cut*, *Cautious Monotony*, *Loop*, *Or* and *Rational Monotony* are defined as follows.

(Reflexivity) $\phi \mathrel{|\!\sim} \phi$.

(Left Logical Equivalence) If $\phi \equiv \psi$ and $\phi \mathrel{|\!\sim} \chi$ then $\psi \mathrel{|\!\sim} \chi$.

(Right Weakening) If $\phi \mathrel{|\!\sim} \psi$ and $\psi \models \chi$ then $\phi \mathrel{|\!\sim} \chi$.

(Cut) If $\phi \mathrel{|\!\sim} \psi$ and $\phi \wedge \psi \mathrel{|\!\sim} \chi$ then $\phi \mathrel{|\!\sim} \chi$.

(Cautious Monotony) If $\phi \mathrel{|\!\sim} \psi$ and $\phi \mathrel{|\!\sim} \chi$ then $\phi \wedge \psi \mathrel{|\!\sim} \chi$.

(Loop) If $\phi_0 \mathrel{|\!\sim} \phi_1$, $\phi_1 \mathrel{|\!\sim} \phi_2$, $\ldots$, $\phi_{k-1} \mathrel{|\!\sim} \phi_k$, $\phi_k \mathrel{|\!\sim} \phi_0$ then $\phi_0 \mathrel{|\!\sim} \phi_k$.

(Or) If $\phi \mathrel{|\!\sim} \chi$ and $\psi \mathrel{|\!\sim} \chi$ then $\phi \vee \psi \mathrel{|\!\sim} \chi$.

(Rational Monotony) If $\phi \not\mathrel{|\!\sim} \neg\psi$ and $\phi \mathrel{|\!\sim} \chi$ then $\phi \wedge \psi \mathrel{|\!\sim} \chi$.

The relation relation $\mathrel{|\!\sim}$ is said to be:

- *cumulative* iff it satisfies Reflexivity, Right Weakening, Left Logical Equivalence, Cut and Cautious Monotony [62].
- *loop-cumulative* iff it is cumulative and satisfies Loop [62].
- *preferential* iff it is loop-cumulative and satisfies Or [62].
- *rational* iff it is preferential and satisfies Rational Monotony [63].

We briefly explain the intuition behind these properties.

Reflexivity is a basic principle of inference, which states that any premise ϕ has, among its consequences, ϕ itself.

Left Logical Equivalence states that the consequences of a premise do not depend on the syntactical representation of the premise. That is, premises that are logically equivalent have the same consequences.

Right Weakening states that if ϕ non-monotonically entails ψ and ψ classically entails χ, then ϕ also non-monotonically entails χ.

Cautious Monotony is a weakening of Monotony. It states that strengthening a premise with something that is already a consequence of this premise does not lead to the retraction of consequences.

Cut is the dual of Cautious Monotony. It states that strengthening a premise with something that is already a consequence of this premise does not lead to new consequences. Note that Cautious Monotony and Cut together imply that if $\phi \mathrel{|\!\sim} \psi$ then the consequences of ϕ and $\phi \wedge \psi$ coincide.

The Or property states that if χ is a consequence of ϕ as well as ψ then it is also a consequence of $\phi \vee \psi$.

Loop states that "if propositions may be arranged in a loop, in a way each one is a plausible consequence of the previous one, then each one of them is a plausible consequence of any one of them" [62].

Rational Monotony is, like Cautious Monotony, a weakening of monotony. It states that strengthening a premise with something the negation of which is not a consequence does not lead to the retraction of consequences. Intuitively, Rational Monotony states that only information that is completely surprising, in the sense that it is believed to be false, leads to the retraction of a consequence.

Note that it is only the Or property that distinguishes the class of preferential and loop-cumulative entailment relations, because the Loop property is satisfied not only by every loop-cumulative entailment relation but also by every preferential entailment relation (i.e., it is derivable from Reflexivity, Left Logical Equivalence, Right Weakening, Cut, Cautious Monotony and Or).

We mention two properties that can be derived using these properties. The first is Equivalence. It states that two propositions that are consequences of each other have are equivalent in the sense that they have the same consequences. Equivalence follows from Cautious Monotony, Left Logical Equivalence and Cut and is therefore satisfied by every cumulative entailment relation [62].

(Equivalence) If $\phi \mathrel{|\!\sim} \psi$, $\psi \mathrel{|\!\sim} \phi$ and $\phi \mathrel{|\!\sim} \chi$ then $\psi \mathrel{|\!\sim} \chi$.

The second property is And. It states that the conjunction of two consequences is also a consequence. This property follows from Cautious Monotony and Cut and is therefore satisfied by every cumulative entailment relation [62].

(And) If $\phi \mathrel{|\!\sim} \psi$ and $\phi \mathrel{|\!\sim} \chi$ then $\phi \mathrel{|\!\sim} \psi \wedge \chi$.

Finally, we mention two properties that do not follow from the properties discussed above.

(Transitivity) If $\phi \mathrel{|\!\sim} \psi$ and $\psi \mathrel{|\!\sim} \chi$ then $\phi \mathrel{|\!\sim} \chi$.

(Contraposition) If $\phi \mathrel{|\!\sim} \psi$ then $\neg\psi \mathrel{|\!\sim} \neg\phi$.

Even though these properties look reasonable at first sight it, turns out that, in the presence of the properties of a cumulative entailment relation, they each imply Monotony [62].

2.2.2 Semantic Characterizations

The semantic characterization of the classes of cumulative, loop-cumulative, preferential and rational inference relations is based on models consisting of a preference relation over states and a mapping from states to valuations. These

models define an inference relation by letting ψ be a consequence of ϕ if ψ is true in the most preferred states in which ϕ is true. The four classes of inference relations are characterized by placing increasingly strong restrictions on these models.

The weakest class of models are the *cumulative models*. The definition presupposes a set V of valuations associated with a propositional language `lang` and a satisfaction relation $\models \subseteq V \times \texttt{lang}$.

Definition 2.2.2. [62] A *cumulative model over* V is a triple $W = (S, \prec, l)$, where

- S is a set containing elements called *states*,
- $\prec$ is a binary relation over S,
- l is a function mapping every state $s \in S$ to a non-empty set $l(s) \subseteq V$,
- $(S, \prec, l)$ satisfies the *smoothness* condition (defined below).

For every formula $\phi \in \texttt{lang}$ we define $\widehat{\phi}$ by $\widehat{\phi} = \{s \in S \mid \forall v \in l(s), v \models \phi\}$. A state s is said to be $\prec$*-minimal* in a set $X \subseteq S$ iff $s \in X$ and there is no $s' \in X$ such that $s' \prec s$. Furthermore, W is called *finite* iff S is finite.

Preferences over states are associated with minimality, meaning that a state s is preferred over a state s' whenever $s \prec s'$. The association of preference with minimality rather than maximality is mainly due to historical reasons, namely that preference was associated with minimizing exceptions.

The smoothness condition is rather technical but not of great importance in the current setting. It ensures that, for every formula ϕ, it is possible to determine the preferred states in $\widehat{\phi}$. In the rest of this chapter we only deal with finite models, in which the smoothness condition is trivially satisfied.

Definition 2.2.3. [62] A triple $(S, \prec, l)$ satisfies the *smoothness condition* iff for all $\phi \in \texttt{lang}$ and $s \in \widehat{\phi}$, either s is $\prec$-minimal in $\widehat{\phi}$, or there is some $s' \in \widehat{\phi}$ such that s' is $\prec$-minimal in $\widehat{\phi}$ and $s' \prec s$.

In a *cumulative ordered model* the preference relation is a strict partial order.[4]

Definition 2.2.4. [62] A *cumulative ordered model over* V is a triple $(S, \prec, l)$ defined like a cumulative model over V except that $\prec$ is a strict partial order.

A *preferential model* is a cumulative model in which every state maps to a single valuation.

Definition 2.2.5. [62] A *preferential model over* V is a triple $(S, \prec, l)$ defined like a cumulative-ordered model over V except that for all $s \in S$, $l(s)$ is a singleton. In this case we also denote by $l(s)$ the member of the set instead of the set.

[4] A binary relation $\prec$ is a strict partial order if it is irreflexive ($s \not\prec s$) and transitive ($s \prec s'$ and $s' \prec s''$ imply $s \prec s''$).

Finally, a *ranked model* is a preferential model in which states can be ordered according to a numerical ranking.

Definition 2.2.6. [63] A *ranked model over* V is a triple $(S, \prec, l)$ defined like a preferential model over V except that there exists some mapping $R : S \to \mathbb{N}$ such that $s \prec s'$ iff $R(s) < R(s')$.

A triple $(S, \prec, l)$ determines an entailment relation as follows.

Definition 2.2.7. [62] A triple $W = (S, \prec, l)$ determines a consequence relation (denoted by $\mid\!\sim^W$) by the following rule:

$$\phi \mid\!\sim^W \psi \text{ iff for all } s \prec\text{-minimal in } \widehat{\phi} \text{ we have } \forall v \in l(s), v \models \psi.$$

The following theorem establishes the characterization of the four classes of entailment relations and cumulative models.

Theorem 2.2.1. *Let* $\mid\!\sim \subseteq \texttt{lang} \times \texttt{lang}$. *It holds that* $\mid\!\sim$ *is cumulative (resp. loop-cumulative, preferential, rational) iff* $\mid\!\sim$ *is defined by a cumulative (resp. cumulative-ordered, preferential, ranked) model. Furthermore, if* `lang` *is logically finite (i.e., contains a finite number of atoms) and* $\mid\!\sim$ *is cumulative (resp. loop-cumulative, preferential, rational) then* $\mid\!\sim$ *is defined by a* finite *cumulative (resp. cumulative-ordered, preferential, ranked) model.*

The proof of this theorem can be found (for the cumulative, loop-cumulative and preferential case) in [62] and (for the rational case) in [63].

Chapter 3

Intervention-Based Entailment in Argumentation

3.1 Introduction

In the introduction we identified two types of change in argumentation. One (called intervention) is change due to actions performed in a debate. These actions correspond to new arguments and attacks that are added to an argumentation framework, which causes the evaluation of the argumentation framework to change. The other (called observation) is the change in argumentation due to observations (information coming from the environment) that requires the revision of the status of one or more arguments. In this section we focus on change in argumentation due intervention in argumentation.

Let us look at an example. Consider the argumentation framework F shown in figure 3.1. The nodes are coloured according to the unique complete (and grounded, preferred, semi-stable and stable) labelling of this argumentation

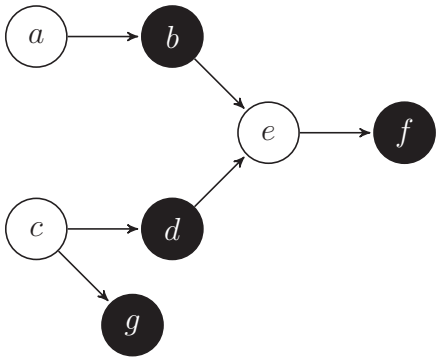


Figure 3.1: An Argumentation Framework.

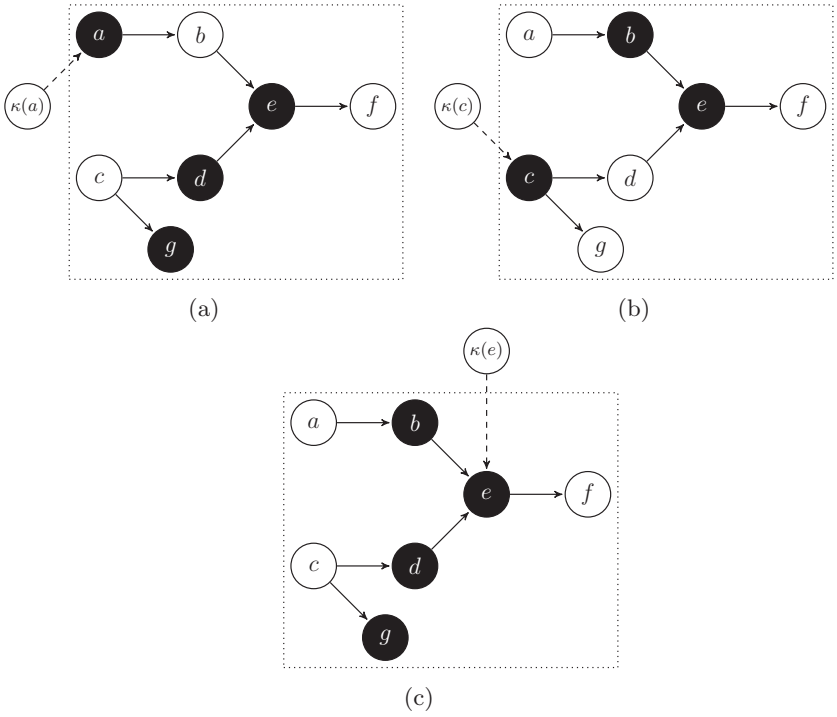


Figure 3.2: Three examples of change to the argumentation framework F in figure 3.1.

framework. There are numerous ways in which this argumentation framework can be extended. Figure 3.2 shows three out of many possibilities. In figure 3.2a, a new argument attacking a is added, which causes a to be **out**, b to be **in**, e to be **out** and f to be **in**, while the labels of c, d and g are not affected. In figure 3.2b and 3.2c, new arguments are added that attack, respectively, c and e. Each of these changes can be thought of as an intervention, or an action in a debate, and each intervention leads to some changes of the evaluation of the argumentation framework with respect to the initial situation.

Our motivation for investigating this type of change is that it allows us to investigate how the evaluation of an argumentation framework under a given semantics *changes* when new arguments and attacks come into play. Unlike the behaviour of argumentation semantics in a static context (that has been studied using properties like **in**-maximialty, (strong) admissibility and reinstatement [6]) this type of behaviour has been relatively neglected.

We model intervention as a form of entailment. This permits us to study this type of change by looking at the properties satisfied by a precisely defined entailment relation. The properties that we focus on are based on the KLM properties, which we discussed in chapter 2. These properties (Cautious Monotony, Cut, Rational Monotony and Loop) represent general principles of well-behaved inference. In the current setting, these properties express principles of how the

evaluation of an argumentation framework changes when new arguments and attacks come into play. For example, Cautious Monotony and Cut together imply that the evaluation of an argumentation framework does not change when a new argument is added that attacks an argument that is already rejected. We present a complete picture of which of these properties are satisfied under the different argumentation semantics that we consider (the complete, grounded, preferred, semi-stable and stable semantics). We furthermore present a number of results concerning argumentation frameworks that are free of odd-length and/or even-length directed cycles.

We also investigate the role of the directionality [6] and noninterference [4] principles in the behaviour of intervention-based entailment. We show that these principles, if satisfied by the argumentation semantics that is used, ensure that the effects of actions propagate through the argumentation framework in a well-behaved manner. On the other hand, using an argumentation semantics not satisfying directionality (the stable and semi-stable semantics) or noninterference (the stable semantics) leads to unintuitive behaviour.

Finally, the results we obtain in this chapter serve as a basis to compare the two types of change we identified. In the next chapter we focus on observation in argumentation. Like we do in this chapter, we investigate the behaviour of change due to observation by modelling it as a form of entailment (observation-based entailment). This approach allows us to compare the behaviour of the two types of change by contrasting them in terms of the properties that satisfy.

We proceed as follows. In section 3.2 we present the basic definitions of intervention-based entailment as well as a number of basic results that will be useful in what follows. In section 3.3 we evaluate the notion of intervention-based entailment using the KLM properties discussed in section 2.2. We then investigate in section 3.4 the role of the directionality and noninterference principles in the behaviour of intervention-based entailment. We discuss in section 3.5 related work and we conclude in section 3.6.

3.2 Intervention-Based Entailment

3.2.1 Defeat and Provisional Defeat

As we explained, an intervention represents an action in a debate that corresponds to the addition of new arguments and attacks to an argumentation framework. Such an action leads to change of the evaluation of the argumentation framework, and this is the type of change that we aim to model. We simplify our model by abstracting away from the actual arguments and attacks that can be added to an argumentation framework. Instead, we focus only on how the addition of a new argument that attacks an existing argument affects the status of the existing argument.

There are two ways in which the status of a new argument y that attacks an existing argument x may affect the status of x. The first is when y is labelled **in**, which causes x to be labelled **out**. The second is when y is labelled **und**, which causes x not to be labelled **in** but still leaves open the possibility for x

to be labelled **und**. Note that the third possibility, namely that y labelled **out**, does not involve any effect on the status of x.

The two effects sketched here are implicit in the definition of a complete labelling. More precisely, the fact that y being labelled **in** causes x to be labelled **out** follows from the *rejection* property of a complete labelling (see definition 2.1.9) and the fact that y being labelled **und** causes x not to be labelled **in** follows from the **in**-*legality* property of a complete labelling (see definition 2.1.8).

Thus, as far as the effect on the label of an argument is concerned, there are two types of actions that can be performed in a debate on a given argument x. We call these two actions *defeat of* x (causing x to be labelled **out**) and *provisional defeat of* x (causing x not to be labelled **in** but still leaving open the possibility for x to be labelled **und**). The distinction between these two types of defeat originates from the work of Pollock [75] and appears in the context of abstract argumentation in the work of Baroni, Giacomin and colleagues (see, e.g., [7, 8]).

Note that other types of change of the label of an argument x, like making x **in**, *not* **out**, **und** or *not* **und**, cannot be achieved in a direct way through (provisional) defeat. Nevertheless, such types of change can still be achieved in an indirect way. For example, given the argumentation framework $a \rightsquigarrow b$, the argument b can be made **in** by defeating a.[1] More generally, we show that any formula that is conflict-free can be made true using only the actions of defeat and provisional defeat on sets of arguments. In this sense, these two actions are functionally complete. This will be proven after we have presented the main definitions.

Formally, we represent an action on an argument x by a literal $\mathbf{out}(x)$ (defeat of x) or $\neg\mathbf{in}(x)$ (provisional defeat of x). An intervention for F is a set Φ of actions such that no two members of Φ refer to the same argument.

Definition 3.2.1. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An *action* for F is a literal of the form $\mathbf{out}(x)$ or $\neg\mathbf{in}(x)$ for some $x \in A$. We use α, β and γ to denote actions and we denote by $Arg(\alpha)$ the argument occurring in α. We denote the set of all actions for F by $\mathtt{Act}(F)$.

Definition 3.2.2. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An *intervention* for F is a set $\Phi \subseteq \mathtt{Act}(F)$ such that for all $\alpha, \beta \in \Phi$, $Arg(\alpha) = Arg(\beta)$ implies $\alpha = \beta$. We denote by $\mathtt{Int}(F)$ the set of interventions for F.

The empty intervention $\emptyset$ will also be called the *vacuous* intervention. In what follows we will see that provisional defeat sometimes leads to undesirable behaviour that does not occur if we focus only on defeat. For this reason we also use the notion of a *stable intervention*, which is an intervention that only consists of defeat.

Definition 3.2.3. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An intervention $\Phi \in \mathtt{Int}(F)$ is *stable* if it consists only of literals of the form $\mathbf{out}(x)$, for some $x \in A$. We denote by $\mathtt{StInt}(F)$ the set of stable interventions for F.

[1] Alternatively, one may consider actions in a debate corresponding to the removal of elements from an argumentation framework, including attacks between existing arguments. Given an argumentation framework $a \rightsquigarrow b$, the change of b to **in** can then be achieved by removing the attack from a to b. We have chosen not to pursue this possibility, the reason being that the addition of elements to an argumentation framework reflects a more natural way of how a debate evolves than the removal of elements.

3.2.2 Intervention-Based Entailment

The (provisional) defeat of an argument x can be achieved adding a new argument to the argumentation framework that attacks x. To achieve defeat of x, we let this attacker itself be unattacked, so that it is labelled **in** in every complete labelling, with the effect that x is labelled **out**. To achieve provisional defeat of x, we let this attacker be self-attacking, so that it is labelled **und** in every complete labelling, with the effect that x is not labelled **in**. To formalize this we need a number of definitions. Given an argumentation framework F, an *F-mapping* is an injective function κ that maps every argument in F to a unique new argument not occurring in F.

Definition 3.2.4. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An *F-mapping* is an injective function $\kappa : A \to \mathcal{U} \setminus A$.

Given an intervention Φ, we denote by $F \oplus^{\kappa} \Phi$ the argumentation framework that represents the effect of Φ on F by the addition of new attacking arguments according to κ.

Definition 3.2.5. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and κ an F-mapping. Given an intervention $\Phi \subseteq \mathtt{Int}(F)$, we define $F \oplus^{\kappa} \Phi$ by $F \oplus^{\kappa} \Phi = (A', \rightsquigarrow')$, where

- $A' = A \cup \{\kappa(x) \mid \mathbf{out}(x) \in \Phi \vee (\neg\mathbf{in}(x)) \in \Phi\}$,
- $\rightsquigarrow' = \rightsquigarrow \cup \{(\kappa(x), x) \mid \mathbf{out}(x) \in \Phi \vee (\neg\mathbf{in}(x)) \in \Phi\} \cup$
  $\{(\kappa(x), \kappa(x)) \mid (\neg\mathbf{in}(x)) \in \Phi\}$,

Example 3.2.1. *Let F be the argumentation framework shown in figure 3.1. The argumentation frameworks $F \oplus^{\kappa} \{\mathbf{out}(a)\}$, $F \oplus^{\kappa} \{\mathbf{out}(c)\}$ and $F \oplus^{\kappa} \{\mathbf{out}(e)\}$ are shown in figure 3.2 (a), (b) and (c), respectively. The argumentation frameworks $F \oplus^{\kappa} \{\neg\mathbf{in}(a)\}$, $F \oplus^{\kappa} \{\neg\mathbf{in}(c)\}$ and $F \oplus^{\kappa} \{\neg\mathbf{in}(e)\}$ are shown in figure 3.3 (a), (b) and (c), respectively.*

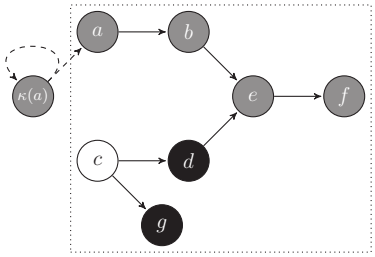
We denote by $\mathcal{L}_{\sigma}(F, \Phi, \kappa)$ the σ labellings of $F \oplus^{\kappa} \Phi$ restricted to the arguments of F.

Definition 3.2.6. Let $F = (A, \rightsquigarrow)$ be an argumentation framework, κ an F-mapping and σ a semantics. We define $\mathcal{L}_{\sigma}(F, \Phi, \kappa)$ by

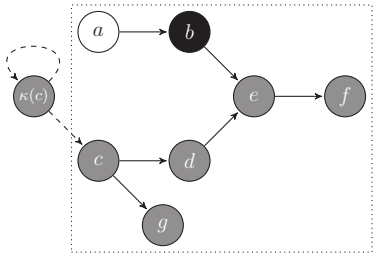
$$\mathcal{L}_{\sigma}(F, \Phi, \kappa) = \mathcal{L}_{\sigma}(F \oplus^{\kappa} \Phi) \downarrow A.$$

We can simplify the notation introduced so far because, for any semantics σ satisfying the language-independence principle, and every argumentation framework F and intervention $\Phi \in \mathtt{Int}(F)$, the set of labellings $\mathcal{L}_{\sigma}(F, \Phi, \kappa)$ is invariant under different choices of κ.

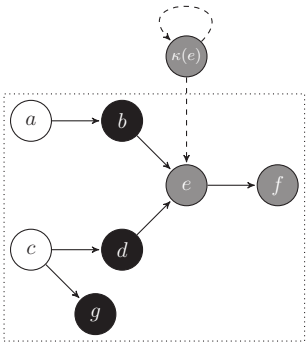
Proposition 3.2.1. *Let σ be a labelling-based semantics and F an argumentation framework. If σ satisfies the language independence principle then for every two F-mappings κ and κ' and every intervention $\Phi \in \mathtt{Int}(F)$ it holds that $\mathcal{L}_{\sigma}(F, \Phi, \kappa) = \mathcal{L}_{\sigma}(F, \Phi, \kappa')$.*



(a) $F \oplus^{\kappa} \{\neg\mathbf{in}(a)\}$.



(b) $F \oplus^{\kappa} \{\neg\mathbf{in}(c)\}$.



(c) $F \oplus^{\kappa} \{\neg\mathbf{in}(e)\}$.

Figure 3.3: Three modifications of the argumentation framework F in figure 3.1.

Proof. Let σ be a labelling-based semantics and $F = (A, \rightsquigarrow)$ an argumentation framework. Suppose σ satisfies the language independence principle. Let κ and κ' be two F-mappings. We then have that $F \oplus^{\kappa} \Phi =_m F \oplus^{\kappa'} \Phi$, where m is defined by $m(x) = x$, if $x \in A$; and $m(x) = \kappa'(\kappa^{-1}(x))$, if $x \notin A$. From definition 2.1.14 it then follows that $\mathcal{L}_\sigma(F \oplus^{\kappa} \Phi) = \{L_m \mid L \in \mathcal{L}_\sigma(F \oplus^{\kappa'} \Phi)\}$, where L_m is a labelling of $F \oplus^{\kappa} \Phi$ defined by $L_m(x) = L(m(x))$. From this it follows that $\mathcal{L}_\sigma(F \oplus^{\kappa} \Phi) \downarrow A = \mathcal{L}_\sigma(F \oplus^{\kappa'} \Phi) \downarrow A$. Via definition 3.2.6 it finally follows that $\mathcal{L}_\sigma(F, \Phi, \kappa) = \mathcal{L}_\sigma(F, \Phi, \kappa')$. □

Due to this invariance (which holds under all semantics that we consider) we can omit the κ argument and simply denote by $\mathcal{L}_\sigma(F, \Phi)$ the set $\mathcal{L}_\sigma(F, \Phi, \kappa)$ for an arbitrary choice of κ.

Definition 3.2.7. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and σ a labelling-based semantics satisfying the language-independence principle. We define $\mathcal{L}_\sigma(F, \Phi)$ by

$$\mathcal{L}_\sigma(F, \Phi) = \mathcal{L}_\sigma(F, \Phi, \kappa), \text{ for an arbitrary choice of } \kappa.$$

Each argumentation framework F and semantics σ determines a relation $\|\!\approx_\sigma^F$ between interventions and formulas of F. We call this an σ *intervention-based entailment relation*.

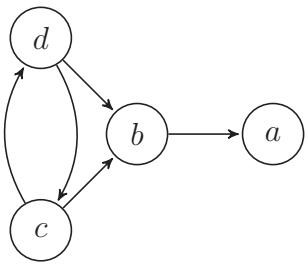


Figure 3.4: An argumentation framework.

Definition 3.2.8. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and σ a semantics. The relation $\|\!\approx_\sigma^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ is defined by the following rule.

$$\text{For all } \Phi \in \mathtt{Int}(F), \phi \in \mathtt{lang}(F): \quad \Phi \|\!\approx_\sigma^F \phi \text{ iff } \forall L \in \mathcal{L}_\sigma(F, \Phi), L \models \phi.$$

Some terminology: Given an argumentation framework F and semantics σ, we say that an intervention $\Phi \in \mathtt{Int}(F)$ skeptically (resp. credulously) entails ϕ if we have $\Phi \|\!\approx_\sigma^F \phi$ (resp. $\Phi \not\|\!\approx_\sigma^F \neg\phi$). If we speak of entailment without specifying whether it is skeptical or credulous, we refer to skeptical entailment.

Example 3.2.2. *Let F be the argumentation framework shown in figure 3.1. We have the following entailments.*

- $\emptyset \|\!\approx_{Co}^F \mathbf{in}(a) \wedge \mathbf{out}(g)$ *(see fig. 3.1)*
- $\{\mathbf{out}(a)\} \|\!\approx_{Co}^F \mathbf{out}(e) \wedge \mathbf{out}(g)$ *(see fig. 3.2a)*
- $\{\neg\mathbf{in}(a)\} \|\!\approx_{Co}^F \mathbf{und}(e) \wedge \mathbf{out}(g)$ *(see fig. 3.3a)*
- $\{\mathbf{out}(c)\} \|\!\approx_{Co}^F \mathbf{out}(e) \wedge \mathbf{in}(g)$ *(see fig. 3.2b)*
- $\{\neg\mathbf{in}(c)\} \|\!\approx_{Co}^F \mathbf{und}(e) \wedge \mathbf{und}(g)$ *(see fig. 3.3b)*
- $\{\mathbf{out}(e)\} \|\!\approx_{Co}^F \mathbf{out}(e) \wedge \mathbf{out}(g)$ *(see fig. 3.2c)*
- $\{\neg\mathbf{in}(e)\} \|\!\approx_{Co}^F \mathbf{und}(e) \wedge \mathbf{out}(g)$ *(see fig. 3.3c)*

The following example involves an argumentation framework with multiple labellings.

Example 3.2.3. *Let F be the argumentation framework shown in figure 3.4.*

- *Under the grounded semantics we have that normally, a is undecided: $\emptyset \|\!\approx_{Gr}^F \mathbf{und}(a)$. However, defeat c or defeat of d implies acceptance of a: $\{\mathbf{out}(c)\} \|\!\approx_{Gr}^F \mathbf{in}(a)$ and $\{\mathbf{out}(d)\} \|\!\approx_{Gr}^F \mathbf{in}(a)$.*
- *Under the complete semantics we have that normally, a is accepted only if c or d is accepted: $\emptyset \|\!\approx_{Co}^F \mathbf{in}(a) \rightarrow (\mathbf{in}(c) \vee \mathbf{in}(d))$. If we defeat b, this no longer holds: $\{\mathbf{out}(b)\} \not\|\!\approx_{Co}^F \mathbf{in}(a) \rightarrow (\mathbf{in}(c) \vee \mathbf{in}(d))$.*

- *Under the preferred semantics we have, for example, that a is accepted:* $\emptyset \mid\!\!=^F_{Pr} \mathbf{in}(a)$. *If we provisionally defeat c and b, this no longer holds:* $\{\neg\mathbf{in}(c), \neg\mathbf{in}(d)\} \mid\!\!\not=^F_{Pr} \mathbf{in}(a)$.

Definition 3.2.8 forms the basis for the further investigations in this chapter. This is because, for each semantics σ and argumentation framework F, we can investigate the behaviour of σ by checking the properties of $\mid\!\!=^F_\sigma$.

Finally, note that the definition of a Φ-modified argumentation framework is very similar to that of a *standard argumentation framework* with respect to an argumentation framework with *input*, which appears in the work of Baroni et al. [3] on decomposability-related properties of argumentation semantics. Similarly, the notion of intervention-based entailment is related to what they call *canonical local functions*. The difference is mainly technical. While their approach is based on a *function* that takes an argumentation framework with input, and returns a set of labellings, in our approach an argumentation framework and semantics determines a *relation* between input (interventions) and consequences of this input.

3.2.3 Basic Properties

Before continuing we prove some basic properties of intervention-based entailment under the semantics that we consider. Some of the properties are rather obvious, but it useful to state them here so that we can refer back to them later on.

Reflexivity

In a Φ-modified argumentation framework, the intended effect of the intervention Φ is achieved only if the semantics that we use satisfies certain conditions. A sufficient condition is that every σ labelling of an argumentation framework is also a complete labelling. In terms of intervention-based entailment this means that the relation $\mid\!\!=^F_\sigma$ satisfies Reflexivity whenever this condition is satisfied.

Definition 3.2.9. Let $F \in \mathcal{F}$. A relation $\mid\!\!=^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Reflexivity* iff for all $\Phi \in \mathtt{Int}(F)$,

$$\Phi \mid\!\!=^F \alpha \text{ for all } \alpha \in \Phi.$$

Proposition 3.2.2. *Let σ be a labelling-based semantics. If for all $F \in \mathcal{F}$, $\mathcal{L}_\sigma(F) \subseteq \mathcal{L}_{Co}(F)$, then for all $F \in \mathcal{F}$, $\mid\!\!=^F_\sigma$ satisfies Reflexivity.*

Proof. Let σ be a semantics and suppose that for $F \in \mathcal{F}$, $\mathcal{L}_\sigma(F) \subseteq \mathcal{L}_{Co}(F)$. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and $\Phi \in \mathtt{Int}(F)$ be an intervention. We prove that $\mid\!\!=^F_\sigma$ satisfies Reflexivity by showing that, for all $\alpha \in \Phi$, $\Phi \mid\!\!=^F_\sigma \alpha$. Let κ be an F-mapping and let $F' = (A', \rightsquigarrow') = F \oplus^\kappa \Phi$. Let L be a σ-labelling of F'. Our assumption implies that L is a complete labelling of F'. We use the fact that L satisfies **out**-legality, **in**-legality, reinstatement and rejection (definition 2.1.8 and 2.1.9).

- Let $x \in A$ be an argument such that $\mathbf{out}(x) \in \Phi$. Definition 3.2.5 then implies that $\kappa(x) \in A'$, $\kappa(x)$ is unattacked in F' and $\kappa(x) \rightsquigarrow' x$. Reinstatement implies that $L(\kappa(x)) = \mathbf{in}$ and hence rejection implies that $L(x) = \mathbf{out}$.
- Let $x \in A$ be an argument such that $(\neg\mathbf{in}(x)) \in \Phi$. Definition 3.2.5 then implies that $\kappa(x) \in A'$, $\kappa(x) \rightsquigarrow' \kappa(x)$ and $\kappa(x) \rightsquigarrow' x$. Then **in**-legality and **out**-legality imply that $L(\kappa(x)) = \mathbf{und}$. Finally, **in**-legality implies that $L(x) \neq \mathbf{in}$.

This implies that for all $\alpha \in \Phi$, $L \models \alpha$. Via definition 3.2.8 it follows that for all $\alpha \in \Phi$, $\Phi \mid\!\models^F_\sigma \alpha$. □

We now obtain the following.

Proposition 3.2.3. *For all $F \in \mathcal{F}$, $\mid\!\models^F_{Co}$, $\mid\!\models^F_{Gr}$, $\mid\!\models^F_{Pr}$, $\mid\!\models^F_{St}$ and $\mid\!\models^F_{SS}$ satisfy Reflexivity.*

Proof. Follows from proposition 3.2.2 and definition 2.1.11 and 2.1.12. □

Note that some labelling-based semantics considered in the literature do not satisfy the property that every labelling is complete. Examples are the stage semantics and CF2 semantics [4]. We leave the treatment of intervention under these semantics for future work.

Vacuous Interventions

A vacuous intervention Φ does not lead to any change of an argumentation framework. This means that the consequences generated by a relation $\mid\!\models^F_\sigma$ given the vacuous intervention coincide with the consequences generated given F by the relation $\models_\sigma$.

Proposition 3.2.4. *For all $F \in \mathcal{F}$, $F \models_\sigma \phi$ iff $\emptyset \mid\!\models^F_\sigma \phi$.*

Proof. Follows directly from definition 3.2.5 and 3.2.8. □

Relative Strength

The following proposition concerns the relative strength of the different intervention-based entailment relations. This is due to the inclusion relations between the corresponding semantics.

Proposition 3.2.5. *For all $F \in \mathcal{F}$,*

1. $\mid\!\models^F_{Co} \subseteq \mid\!\models^F_{Gr}$,
2. $\mid\!\models^F_{Co} \subseteq \mid\!\models^F_{Pr} \subseteq \mid\!\models^F_{SS} \subseteq \mid\!\models^F_{St}$.

Proof. Follows from definition 3.2.8 and proposition 2.1.8. □

Consistency of Interventions

Because every argumentation framework has at least one complete, grounded, preferred and semi-stable labelling, every intervention yields consistent consequences under the complete, grounded, preferred and semi-stable semantics.

Proposition 3.2.6. *For all* $F \in \mathcal{F}$, $\sigma \in \{Co, Gr, Pr, SS\}$ *and* $\Phi \in \texttt{Int}(F)$, $\Phi \not\Vdash^F_\sigma \bot$.

Proof. Follows from definition 3.2.8 and proposition 2.1.10. □

Provisional defeat leads to the addition of a self-attacking argument. This means that the resulting argumentation framework has no stable labellings. Thus, under the stable semantics, only stable interventions generate consistent conclusions.

Proposition 3.2.7. *For all* $F \in \mathcal{F}$ *and* $\Phi \in \texttt{Int}(F)$, *if* Φ *is not stable then* $\Phi \Vdash^F_{St} \bot$.

Proof. See preceding discussion. □

Enforceability Of Formulas

We already mentioned that any constraint on the labels of arguments that is conflict-free can be made true using only the actions of defeat and provisional defeat. We now make this formal. First a definition: a formula is conflict-free with respect to an argumentation framework if it is satisfied by at least one conflict-free labelling of this argumentation framework.

Definition 3.2.10. Let $F \in \mathcal{F}$. A formula $\phi \in \texttt{lang}(F)$ is *conflict-free with respect to* F if and only if there is some $L \in \mathcal{L}_{Cf}(F)$ such that $L \models \phi$.

The following theorem states that, given an argumentation framework F, *every* formula that is conflict-free with respect to F is a consequence of *some* intervention. Conversely, every consequence of every intervention is conflict-free with respect to F. This holds under the complete, grounded, preferred and semi-stable semantics, but not under the stable semantics. Note that, for the sake of readability, we have moved some of the longer proofs, including the proof for the following two theorems, to section 3.7.

Theorem 3.2.8. *Let* F *be an argumentation framework and* $\sigma \in \{Co, Gr, Pr, SS\}$. *The following are equivalent.*

1. *For some* $\Phi \in \texttt{Int}(F)$, $\Phi \Vdash^F_\sigma \phi$.
2. ϕ *is conflict-free with respect to* F.

Proof. See section 3.7 □

Under the stable semantics we have the following. Given an argumentation framework F, *every* formula that is *stable conflict-free* with respect to F is a consequence of *some* intervention Φ that is consistent (i.e., for which we do not have $\Phi \Vdash^F_{St} \bot$). Conversely, every consequence of every intervention that is consistent is stable conflict-free with respect to F.

Definition 3.2.11. A formula $\phi \in \texttt{lang}(F)$ is *stable conflict-free* with respect to F if and only if there is some $L \in \mathcal{L}_{Cf}(F)$ such that $L^{-1}(\mathbf{und}) = \emptyset$ and $L \models \phi$.

Theorem 3.2.9. *Let F be an argumentation framework. The following are equivalent.*

1. *For some $\Phi \in \texttt{Int}(F)$ we have $\Phi \Vdash^F_{St} \phi$ and $\Phi \nVdash^F_{St} \bot$.*
2. *ϕ is a stable conflict-free with respect to F.*

Proof. See section 3.7 □

These results show that, given an argumentation framework F and a semantics σ, any constraint ϕ, as long as it is (stable) conflict-free, can be translated into an intervention Φ that makes ϕ true. As the following example demonstrates, however, there may be more than one intervention that makes a given constraint true.

Example 3.2.4. *Let F be the argumentation framework shown in figure 3.1*

- *Let $\sigma \in \{Co, Gr, Pr, SS, St\}$. The formula $\mathbf{in}(f)$ is (stable) conflict-free with respect to F. We have $\{\mathbf{out}(a)\} \Vdash^F_\sigma \mathbf{in}(f)$, $\{\mathbf{out}(c)\} \Vdash^F_\sigma \mathbf{in}(f)$ and $\{\mathbf{out}(e)\} \Vdash^F_\sigma \mathbf{in}(f)$ (see figure 3.2). That is: $\{\mathbf{out}(a)\}$, $\{\mathbf{out}(c)\}$ and $\{\mathbf{out}(e)\}$ all make $\mathbf{in}(f)$ true.*
- *Let $\sigma \in \{Co, Gr, Pr, SS\}$. The formula $\mathbf{und}(f)$ is conflict-free with respect to F. We have $\{\neg\mathbf{in}(a)\} \Vdash^F_\sigma \mathbf{und}(f)$, $\{\neg\mathbf{in}(c)\} \Vdash^F_\sigma \mathbf{und}(f)$ and $\{\neg\mathbf{in}(e)\} \Vdash^F_\sigma \mathbf{und}(f)$ (see figure 3.3). That is: $\{\neg\mathbf{in}(a)\}$, $\{\neg\mathbf{in}(c)\}$ and $\{\neg\mathbf{in}(e)\}$ all make f undecided.*

This result shows that the actions of defeat and provisional defeat are sufficient to make an argumentation framework satisfy any constraint on the status of the arguments, as long as this constraint is (stable) conflict-free. In the next chapter we look at observation-based entailment, and the selection of (minimal) interventions that make a constraint (now taken to be an observation) true will play a central role. In that setting, these interventions play the role of explanations for the observation, and their selection can be seen as a process of abduction.

Note that Baumann and Brewka have proven a result for extension-based semantics that is related to what we prove in theorem 3.2.8 and 3.2.9, namely that every conflict-free set of an argumentation framework can be turned into a (unique) complete extension by adding a new argument attacking existing arguments [12].

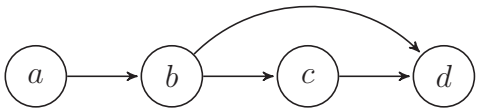


Figure 3.5: Failure of Transitivity and Contraposition.

3.3 KLM Properties

Intervention-based entailment allows us to evaluate the different semantics in terms of their behaviour in a dynamic context, where we not only focus on the evaluation of an argumentation framework in isolation, but instead focus on the evaluation of an argumentation framework given the possible interventions. We characterize the behaviour of a semantics σ in terms of the properties satisfied by a relation $\|\!\approx^F_\sigma$ for every possible argumentation framework F, as well as for a number of special classes (argumentation frameworks that are even-cycle-free, odd-cycle-free or completely cycle-free). A property that is not satisfied in general is monotony, which we define for intervention-based entailment as follows.

Definition 3.3.1. Let $F \in \mathcal{F}$. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Monotony* iff for all $\Phi, \Psi \in \texttt{Int}(F)$ and $\phi \in \texttt{lang}(F)$,

$$\text{if } \Phi \|\!\approx^F \phi \text{ then } \Phi \cup \Psi \|\!\approx^F \phi.$$

An example of the failure of Monotony is easy to come up with. For example, let F be the argumentation framework shown in figure 3.1 and let $\sigma \in \{Co, Gr, Pr, SS, St\}$. The rejection of b given the vacuous intervention ($\emptyset \|\!\approx^F_\sigma \mathbf{out}(b)$) no longer follows if we defeat a ($\{\mathbf{out}(a)\} \not\|\!\approx^F_\sigma \mathbf{out}(b)$), which is a violation of Monotony. The Transitivity and Contraposition properties, which can be defined for intervention-based entailment as follows, also fail.

Definition 3.3.2. Let $F \in \mathcal{F}$. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Transitivity* iff for all $\alpha, \beta, \gamma \in \texttt{Act}(F)$,

$$\text{if } \{\alpha\} \|\!\approx^F \beta \text{ and } \{\beta\} \|\!\approx^F \gamma \text{ then } \{\alpha\} \|\!\approx^F \gamma$$

Definition 3.3.3. Let $F \in \mathcal{F}$. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Contraposition* iff for all $\alpha, \beta \in \texttt{Act}(F)$,

$$\{\alpha\} \|\!\approx^F \beta \text{ then } \{\neg\beta\} \|\!\approx^F \neg\alpha.$$

The following two examples demonstrate their failure.

Example 3.3.1. *Let F be the argumentation framework shown in figure 3.5.*

- *We have $\{\mathbf{out}(a)\} \|\!\approx^F_{Pr} \mathbf{out}(c)$ and $\{\mathbf{out}(c)\} \|\!\approx^F_{Pr} \mathbf{in}(d)$. Transitivity would imply that we have $\{\mathbf{out}(a)\} \|\!\approx^F_{Pr} \mathbf{in}(d)$ but this is not the case. The reason is that defeat of c justifies acceptance of d as long as b is not accepted. Defeating a not only justifies rejection of c, but also acceptance of b, and hence acceptance of d is not justified if we defeat a.*

- *We have* $\{\mathbf{out}(b)\} \mid\!\!\approx^F_{Pr} \mathbf{in}(c)$*. According to contraposition it then follows that* $\{\neg\mathbf{in}(c)\} \mid\!\!\approx^F_{Pr} \neg\mathbf{out}(b)$*, but this is not the case. The reason is that, while defeating* b *indeed justifies acceptance of* c*, provisionally defeating* c *does not justify non-rejection of* b*.*

These examples can be adapted to apply to the other semantics.

The KLM properties that we discussed in section 2.2.1 are desirable properties for well-behaved non-monotonic inference. Thus, we can ask: how well-behaved are the different argumentation semantics with respect to these properties? This is the question we address in this section.

We already proved that Reflexivity is satisfied under all semantics that we consider. It is furthermore clear from definition 3.2.8 that every relation $\mid\!\!\approx^F_\sigma$ satisfies Right Weakening (if $\phi \models \psi$ and $\Phi \mid\!\!\approx^F_\sigma \phi$ then $\Phi \mid\!\!\approx^F_\sigma \psi$), and that it trivially satisfies a form of Left Logical Equivalence, because logical equivalence of two interventions implies syntactical equivalence ($\Phi \models \Psi$ and $\Psi \models \Phi$ implies $\Phi = \Psi$). In the rest of this section we consider variations of the remaining KLM properties, namely Cautious Monotony, Cut, Rational Monotony and Loop. For each property, we formulate an analogue that is appropriate for intervention-based entailment. We then determine the conditions under which each property is satisfied, and we demonstrate the failure if they are not satisfied. We omit discussion of the Or property, as it involves disjunction in the premise, which cannot be expressed by an intervention.

3.3.1 Cautious Monotony

Cautious Monotony states that we do not lose consequences if we strengthen a premise, if what we add is already a consequence of this premise. We express this for intervention-based entailment as follows.

Definition 3.3.4. Let $F \in \mathcal{F}$. A relation $\mid\!\!\approx^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Cautious Monotony* iff for all $\Phi \in \mathtt{Int}(F)$, $\alpha \in \mathtt{Act}(F)$ and $\phi \in \mathtt{lang}(F)$,

$$\text{if } \Phi \mid\!\!\approx^F \alpha \text{ and } \Phi \mid\!\!\approx^F \phi \text{ then } \Phi \cup \{\alpha\} \mid\!\!\approx^F \phi.$$

In the setting of intervention-based entailment, Cautious Monotony states that if an argument x is already rejected (resp. not accepted) then defeating (resp. provisionally defeating) x does not lead to loss of consequences. We believe that this is an intuitive property for intervention-based entailment. It corresponds to a natural principle: if a new argument comes into play that defeats an argument that we already reject, we do not change our mind about the status of any of the other arguments.

In what follows we will see that some semantics behave better when we focus only on stable interventions. For this reason we also define the following weakening of Cautious Monotony, which we call *Stable Cautious Monotony*. While Cautious Monotony applies to arbitrary interventions, Stable Cautious Monotony applies only to stable interventions. It is easy to see that Cautious Monotony implies Stable Cautious Monotony, but that the converse does not hold.

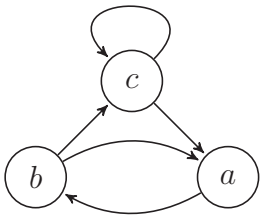


Figure 3.6: Failure of Cautious Monotony under the preferred, semi-stable and stable semantics.

Definition 3.3.5. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\|\!\approx^F \subseteq \texttt{StInt}(F) \times \texttt{lang}(F)$ satisfies *Stable Cautious Monotony* iff for all $\Phi \in \texttt{StInt}(F)$, $\phi \in \texttt{lang}(F)$ and $x \in A$,

$$\text{if } \Phi \|\!\approx^F \mathbf{out}(x) \text{ and } \Phi \|\!\approx^F \phi \text{ then } \Phi \cup \{\mathbf{out}(x)\} \|\!\approx^F \phi.$$

Our first result is that Cautious Monotony is satisfied under the complete and grounded semantics.

Theorem 3.3.1. *For all $F \in \mathcal{F}$, $\|\!\approx^F_{Gr}$ satisfies Cautious Monotony.*

Proof. See section 3.7 □

Theorem 3.3.2. *For all $F \in \mathcal{F}$, $\|\!\approx^F_{Co}$ satisfies Cautious Monotony.*

Proof. See section 3.7 □

However, (Stable) Cautious Monotony may fail under the preferred, semi-stable and stable semantics. This is demonstrated by the following example.

Example 3.3.2. *Let F be the argumentation framework shown in figure 3.6. Note that acceptance is maximized only if b is accepted. This implies, under the preferred semantics, rejection of c. However, if we defeat c, acceptance is maximized not only if we accept b but also if we accept c. This causes failure of Cautious Monotony. We have $\emptyset \|\!\approx^F_{Pr} \mathbf{out}(c)$ and $\emptyset \|\!\approx^F_{Pr} \mathbf{in}(b)$. (Stable) Cautious Monotony implies that we have $\{\mathbf{out}(c)\} \|\!\approx^F_{Pr} \mathbf{in}(b)$ but this is not the case. This example of the failure of Cautious Monotony also applies to $\|\!\approx^F_{SS}$ and $\|\!\approx^F_{St}$.*

Note that the counterexample above does not rely on the fact that c is self-attacking. Cautious Monotony fails in a similar way if we let c be part of any odd-length directed cycle.

In section 3.3.6 we show that the absence of odd-length cycles is a sufficient condition for the satisfaction of Stable Cautious Monotony (but not of Cautious Monotony) under the preferred, semi-stable semantics and stable semantics.

3.3.2 Cut

The Cut property can be seen as the converse of Cautious Monotony. It states that we do not *gain* consequences if we strengthen a premise, if what we add is already a consequence of this premise. We express this for intervention-based entailment as follows.

Definition 3.3.6. Let $F \in \mathcal{F}$. A relation $\mid\!\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Cut* iff for all $\Phi \in \texttt{Int}(F)$, $\alpha \in \texttt{Act}(F)$ and $\phi \in \texttt{lang}(F)$,

$$\text{if } \Phi \mid\!\!\approx^F \alpha \text{ and } \Phi \cup \{\alpha\} \mid\!\!\approx^F \phi \text{ then } \Phi \mid\!\!\approx^F \phi.$$

We believe that this is an intuitive property for intervention-based entailment. The motivation is the same as for Cautious Monotony: if a new argument comes into play that defeats an argument that we already reject, we do not change our mind about the status of any of the other arguments. Together, Cautious Monotony and Cut state that if $\Phi \mid\!\!\approx^F \alpha$ then the consequences given the interventions $\Phi \cup \{\alpha\}$ and Φ coincide.

Like we did for Cautious Monotony, we also consider a weakening of Cut, which we call *Stable Cut*. While Cut applies to arbitrary interventions, Stable Cut applies only to stable interventions. It is easy to see that Cut implies Stable Cut, but that the converse does not hold.

Definition 3.3.7. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\mid\!\!\approx^F \subseteq \texttt{StInt}(F) \times \texttt{lang}(F)$ satisfies *Stable Cut* iff for all $\Phi \in \texttt{StInt}(F)$, $\phi \in \texttt{lang}(F)$ and $x \in A$,

$$\text{if } \Phi \mid\!\!\approx^F \mathbf{out}(x) \text{ and } \Phi \cup \{\mathbf{out}(x)\} \mid\!\!\approx^F \phi \text{ then } \Phi \mid\!\!\approx^F \phi.$$

Cut is, like Cautious Monotony, satisfied under the complete and grounded semantics. But unlike Cautious Monotony, Cut is also satisfied under the preferred semantics.

Theorem 3.3.3. *For all $F \in \mathcal{F}$, $\mid\!\!\approx^F_{Gr}$ satisfies Cut.*

Proof. See section 3.7 □

Theorem 3.3.4. *For all $F \in \mathcal{F}$, $\mid\!\!\approx^F_{Co}$ satisfies Cut.*

Proof. See section 3.7 □

Theorem 3.3.5. *For all $F \in \mathcal{F}$, $\mid\!\!\approx^F_{Pr}$ satisfies Cut.*

Proof. See section 3.7 □

Cut fails under the stable semantics due to the fact that non-stable interventions cause inconsistency. This is demonstrated by the following example.

Example 3.3.3. *Let F be the argumentation framework shown in figure 3.5. We have $\emptyset \mid\!\!\approx^F_{St} \neg\mathbf{in}(c)$ and $\{\neg\mathbf{in}(c)\} \mid\!\!\approx^F_{St} \bot$. Cut implies that we have $\emptyset \mid\!\!\approx^F_{St} \bot$, but this is not the case.*

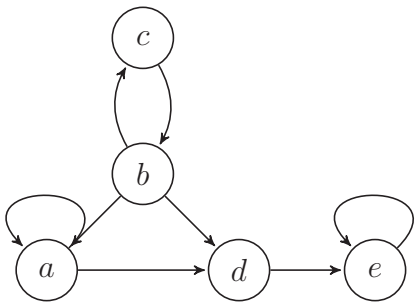


Figure 3.7: Failure of Cut under the semi-stable semantics.

Nevertheless, Stable Cut is satisfied under the stable semantics.

Theorem 3.3.6. *For all $F \in \mathcal{F}$, $\mid\!\!\sim^F_{St}$ satisfies Stable Cut.*

Proof. See section 3.7 □

Cut, as well as Stable Cut, may fail under the semi-stable semantics. This is demonstrated by the following example.

Example 3.3.4. *Let F be the argumentation framework shown in figure 3.7. Note that acceptance of b is necessary to minimize undecidedness. This implies rejection of a. If we defeat a, however, then it is not the acceptance of b but instead the acceptance of c that minimizes undecidedness. This causes failure of Cut: We have $\emptyset \mid\!\!\sim^F_{SS} \mathbf{out}(a)$ and $\{\mathbf{out}(a)\} \mid\!\!\sim^F_{SS} \mathbf{in}(c)$. Cut implies that we have $\emptyset \mid\!\!\sim^F_{SS} \mathbf{in}(c)$, but this is not the case.*

Note that, like the counterexample for the failure of Cautious Monotony under the preferred, semi-stable and stable semantics, the counterexample above does not rely on the fact that a and e are self-attacking. Cut fails under the semi-stable semantics in a similar way if we let a or e be part of any odd-length directed cycle. In section 3.3.6 we show that the absence of odd-length cycles is a sufficient condition for the satisfaction of Stable Cut under the semi-stable semantics.

It is worth noting that, in the KLM framework, Cautious Monotony and Cut imply *Equivalence*, which expresses that two propositions that are each other's consequence are equivalent in terms of the consequences they generate [62]. For intervention-based entailment we express this as follows.

Definition 3.3.8. Let $F \in \mathcal{F}$. A relation $\mid\!\!\sim^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Equivalence* iff for all $\alpha, \beta \in \mathtt{Act}(F)$ and $\phi \in \mathtt{lang}(F)$,

$$\text{if } \{\alpha\} \mid\!\!\sim^F \beta, \{\beta\} \mid\!\!\sim^F \alpha \text{ and } \{\alpha\} \mid\!\!\sim^F \phi \text{ then } \{\beta\} \mid\!\!\sim^F \phi.$$

In the current setting Cautious Monotony and Cut also imply Equivalence.

Proposition 3.3.7. *Let $F \in \mathcal{F}$ and $\mid\!\!\sim^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$. If $\mid\!\!\sim^F$ satisfies Cautious Monotony and Cut then it satisfies Equivalence.*

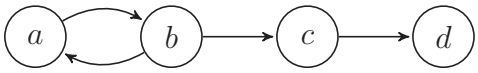


Figure 3.8: Failure of Rational Monotony under all but the grounded semantics.

Proof. Let $F \in \mathcal{F}$ and $\mid\!\!\approx^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$. Suppose $\mid\!\!\approx^F$ satisfies Cautious Monotony and Cut. Assume we have (1) $\{\alpha\} \mid\!\!\approx^F \beta$, (2) $\{\beta\} \mid\!\!\approx^F \alpha$ and (3) $\{\alpha\} \mid\!\!\approx^F \phi$. Cautious Monotony, (1) and (3) imply $\{\alpha, \beta\} \mid\!\!\approx^F \phi$. This implies, together with Cut and (2), $\{\beta\} \mid\!\!\approx^F \phi$. □

Under the complete and grounded semantics, both Cautious Monotony and Cut are satisfied and thus Equivalence as well. It can be verified that Equivalence is not generally satisfied under the preferred, semi-stable and stable semantics.

3.3.3 Rational Monotony

Rational Monotony is, like Cautious Monotony, a weakening of Monotony. It states that strengthening a premise with something the negation of which is not a consequence does not lead to the retraction of consequences. In the intervention-based setting, this means that strengthening an intervention by adding something that is already a credulous consequence (as opposed to a skeptical consequence, as in Cautious Monotony) does not lead to the retraction of consequences. For intervention-based entailment we express this as follows.

Definition 3.3.9. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\mid\!\!\approx^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Rational Monotony* iff for all $\Phi \in \mathtt{Int}(F)$, $\alpha \in \mathtt{Act}(F)$ and $\phi \in \mathtt{lang}(F)$,

$$\text{if } \Phi \not\mid\!\!\approx^F \neg\alpha \text{ and } \Phi \mid\!\!\approx^F \phi \text{ then } \Phi \cup \{\alpha\} \mid\!\!\approx^F \phi.$$

Rational Monotony has been argued for as an attractive property for non-monotonic inference because it expresses the principle that only information that is completely surprising, in the sense that it believed to be false, should force one to withdraw conclusions [62]. Others have argued, however, that this principle is too strong (see, for example, [87]).

Rational Monotony may fail under the complete, preferred, semi-stable and stable semantics. This is demonstrated by the following example.

Example 3.3.5. *Let F be the argumentation framework shown in figure 3.8. Normally, d is accepted in a complete labelling only if b is accepted. Hence we have $\emptyset \mid\!\!\approx^F_{Co} \mathbf{in}(d) \rightarrow \mathbf{in}(b)$. Furthermore, there is a complete labelling in which c is rejected, thus we have $\emptyset \not\mid\!\!\approx^F_{Co} \neg\mathbf{out}(c)$. But defeating c leads to a new complete labelling, in which d is accepted but b is not. This causes failure of Rational Monotony. We have $\emptyset \mid\!\!\approx^F_{Co} \mathbf{in}(d) \rightarrow \mathbf{in}(b)$ and $\emptyset \not\mid\!\!\approx^F_{Co} \neg\mathbf{out}(c)$. Rational Monotony would imply that we have $\{\mathbf{out}(c)\} \mid\!\!\approx^F_{Co} \mathbf{in}(d) \rightarrow \mathbf{in}(b)$, but this is not the case. Note that this counterexample also applies to the preferred, semi-stable and stable semantics.*

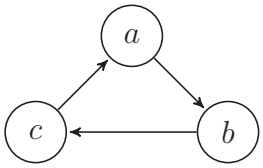


Figure 3.9: Failure of Loop under all semantics.

Speaking in loose terms, the failure of Rational Monotony is due to the fact that an intervention may 'override' the relationship between the status of different arguments, which is accompanied by loss of consequences. In the example above, it is the defeat of c that overrides the relation that exists between the status of d and b expressed by $\mathbf{in}(d) \rightarrow \mathbf{in}(b)$. Indeed, we have $\emptyset \mid\!\!\approx^F_{Co} \mathbf{in}(d) \rightarrow \mathbf{in}(b)$ but not $\{\mathbf{out}(c)\} \mid\!\!\approx^F_{Co} \mathbf{in}(d) \rightarrow \mathbf{in}(b)$.

The only semantics under which intervention-based entailment satisfies Rational Monotony is the grounded semantics. This follows from the fact that the grounded labelling is unique, which implies the following.

Proposition 3.3.8. *For all $F \in \mathcal{F}$, $\Phi \in \mathtt{Int}(F)$ and $\phi \in \mathtt{lang}(F)$, either $\Phi \mid\!\!\approx^F_{Gr} \phi$ or $\Phi \mid\!\!\approx^F_{Gr} \neg\phi$.*

Proof. Follows from definition 3.2.8 and proposition 2.1.9. □

This means that for all F, $\Phi \in \mathtt{Int}(F)$ and $\alpha \in \mathtt{Act}(F)$ we have $\Phi \not\mid\!\!\approx^F_{Gr} \neg\alpha$ if and only if $\Phi \mid\!\!\approx^F_{Gr} \alpha$. This makes Cautious Monotony and Rational Monotony equivalent. Because $\mid\!\!\approx^F_{Gr}$ satisfies Cautious Monotony, it also satisfies Rational Monotony.

Proposition 3.3.9. *For all $F \in \mathcal{F}$, $\mid\!\!\approx^F_{Gr}$ satisfies Rational Monotony.*

Proof. See preceding discussion. □

3.3.4 Loop

Loop states that "if propositions may be arranged in a loop, in a way each one is a plausible consequence of the previous one, then each one of them is a plausible consequence of any one of them" [62]. According to Kraus, Lehmann and Magidor, this is a desirable property for non-monotonic inference. In the current setting we express it as follows.

Definition 3.3.10. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\mid\!\!\approx^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Loop* iff for all $\alpha_1, \ldots, \alpha_k \in \mathtt{Act}(F)$,

$$\text{if } \{\alpha_0\} \mid\!\!\approx^F \alpha_1, \{\alpha_1\} \mid\!\!\approx^F \alpha_2, \ldots, \{\alpha_{k-1}\} \mid\!\!\approx^F \alpha_k, \{\alpha_k\} \mid\!\!\approx^F \alpha_0 \text{ then } \{\alpha_0\} \mid\!\!\approx^F \alpha_k.$$

The following example demonstrates the failure of Loop under all semantics.

Example 3.3.6. *Let F be the argumentation framework shown in figure 3.9. Defeating a (resp. b, c) results in acceptance of b (resp. c, a), which in*

turn results in rejection of c (resp. a, b). Thus we have $\{\mathbf{out}(a)\} \not\models^{F}_{Co} \mathbf{out}(c)$, $\{\mathbf{out}(c)\} \not\models^{F}_{Co} \mathbf{out}(b)$ *and* $\{\mathbf{out}(b)\} \not\models^{F}_{Co} \mathbf{out}(a)$. *Loop would imply that we have* $\{\mathbf{out}(a)\} \not\models^{F}_{Co} \mathbf{out}(b)$ *but this is not the case. Instead, we have* $\{\mathbf{out}(a)\} \not\models^{F}_{Co} \mathbf{in}(b)$. *This counterexample also applies to the preferred, grounded, semi-stable and stable semantics.*

It is easy to check that the construction of this counterexamples applies to every unattacked odd-length directed cycle (i.e., not attacked by arguments outside the cycle) of length more than one, and hence that the existence of such a cycle is a sufficient condition for the failure of Loop. It is an open question whether the absence of odd-length directed cycles is a sufficient condition for the satisfaction of Loop. In section 3.3.8 we do show, however, that the absence of any directed cycle (both odd and even) ensures satisfaction of Loop.

3.3.5 Summary

Let us summarize the results obtained so far. Table 3.1 shows whether the KLM property referred to in each row is satisfied by the relation $\not\models^{F}_{\sigma}$ for every argumentation framework F. A check mark indicates that the property is satisfied for all $F \in \mathcal{F}$, and a cross mark indicates that the property fails for some $F \in \mathcal{F}$. This table summarizes the positive results obtained in theorem 3.3.1, 3.3.2, 3.3.3, 3.3.4, 3.3.5, 3.3.6 and 3.3.9 and the negative results demonstrated in example 3.3.2, 3.3.4, 3.3.5 and 3.3.6.

	Co	*Gr*	*Pr*	*SS*	*St*
(Stable) Cautious Monotony	✓	✓	✗	✗	✗
(Stable) Cut	✓	✓	✓	✗	✓
Rational Monotony	✗	✓	✗	✗	✗
Loop	✗	✗	✗	✗	✗

Table 3.1: Properties satisfied by $\not\models^{F}_{\sigma}$ for all $F \in \mathcal{F}$.

3.3.6 Odd-Cycle-Free Argumentation Frameworks

It has been argued (e.g. by Bench-Capon [13]) that odd-length directed cycles in argumentation frameworks have the nature of a paradox, in the sense that nothing can be accepted. This suggests that argumentation frameworks that contain no odd-length directed cycles may exhibit better behaviour than those that do, as these argumentation frameworks can be thought of as being free of paradoxes. In this section we determine whether the properties discussed in the preceding sections that fail under some of the semantics we consider, nevertheless hold if the argumentation framework contains no odd-length directed cycles. We call these argumentation frameworks *odd-cycle-free*.

Definition 3.3.11. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A sequence $x_0, \ldots, x_n$ of arguments is called an *odd cycle* if and only if n is odd; $x_0 = x_n$; and $x_i \rightsquigarrow x_{i+1}$ for all $0 \leq i < n$. We say that F is *odd-cycle-free* if it contains no odd cycles.

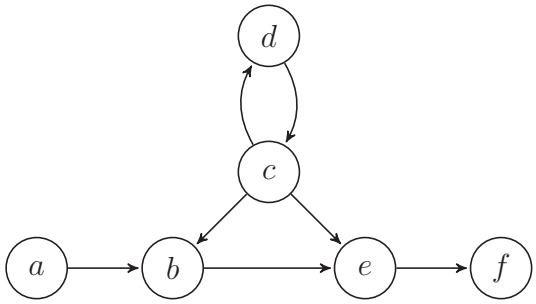


Figure 3.10: Failure of Cut and Cautious Monotony in the odd-cycle-free case.

Dung has shown that odd-cycle-freeness implies *coherence*, which is the condition that the set of preferred and stable extensions coincide [42].[2] Because the existence of at least one stable extension implies that the stable and semi-stable extensions coincide, this result implies that odd-cycle-freeness implies that the preferred and stable extensions also coincide with the semi-stable extensions. We state this result for labelling-based semantics.

Proposition 3.3.10. *If F is odd-cycle-free then $\mathcal{L}_{Pr}(F) = \mathcal{L}_{SS}(F) = \mathcal{L}_{St}(F)$.*

We have shown that Cautious Monotony may fail under the preferred, semi-stable and stable semantics, and that Cut may fail under the semi-stable semantics. Odd-cycle-freeness is, however, a sufficient condition for the satisfaction of Stable Cautious Monotony and Stable Cut under the preferred, semi-stable and stable semantics.

Theorem 3.3.11. *For all $F \in \mathcal{F}$, if F is odd-cycle-free then $\|\!\approx^F_{Pr}$, $\|\!\approx^F_{SS}$ and $\|\!\approx^F_{St}$ satisfy Stable Cautious Monotony.*

Proof. See section 3.7. □

Theorem 3.3.12. *For all $F \in \mathcal{F}$, if F is odd-cycle-free then $\|\!\approx^F_{SS}$ satisfies Stable Cut.*

Proof. See section 3.7 □

Why are the non-stable versions of Cautious Monotony and Cut not satisfied? This is because provisional defeat of an argument is interpreted by adding a self-attacking attacker to this argument. This means that provisional defeat introduces odd cycles. Under the semi-stable semantics, this causes failure of Cautious Monotony and Cut. This is demonstrated by the following example.

Example 3.3.7. *Let F be the argumentation framework shown in figure 3.10. Note that F is odd-cycle-free. We have $\{\neg\mathbf{in}(a), \neg\mathbf{in}(f)\} \|\!\approx^F_{SS} \mathbf{out}(b)$. The failure of Cautious Monotony is due to the fact that we have $\{\neg\mathbf{in}(a), \neg\mathbf{in}(f)\} \|\!\approx^F_{SS} \mathbf{in}(c)$ but not $\{\neg\mathbf{in}(a), \neg\mathbf{in}(f), \mathbf{out}(b)\} \|\!\approx^F_{SS} \mathbf{out}(b)$. The failure of Cut is*

[2] Dung actually proved something stronger, namely that every *limited controversial* argumentation framework is coherent. However, in the finite case, limited controversiality is the same as odd-cycle-freeness.

due to the fact that we have $\{\neg\mathbf{in}(a), \neg\mathbf{in}(f), \mathbf{out}(b)\} \mid\!\approx^F_{SS} \mathbf{out}(c)$ *but not* $\{\neg\mathbf{in}(a), \neg\mathbf{in}(f)\} \mid\!\approx^F_{SS} \mathbf{out}(c)$.

The results obtained in this section show that, indeed, odd-cycle-free argumentation frameworks behave better in the sense that they ensure satisfaction of Stable Cautious Monotony and Stable Cut under semantics that do not satisfy these properties in general.

Odd-cycle-freeness does not ensure satisfaction of Rational Monotony. To see why, consider again example 3.3.5, which demonstrates the failure of Rational Monotony. This example relies on an argumentation framework that is odd-cycle-free.

Table 3.2 summarizes the results obtained concerning odd-cycle-free argumentation frameworks. Note that the question of whether Loop is satisfied remains open.

	Co	*Gr*	*Pr*	*SS*	*St*
(Stable) Cautious Monotony	✓	✓	✓	✓	✓
(Stable) Cut	✓	✓	✓	✓	✓
Rational Monotony	✗	✓	✗	✗	✗
Loop	?	?	?	?	?

Table 3.2: Properties satisfied by $\mid\!\approx^F_\sigma$ for every odd-cycle-free $F \in \mathcal{F}$.

3.3.7 Even-Cycle-Free Argumentation Frameworks

The next class of argumentation frameworks are those containing no even-length directed cycles. Whereas odd-length directed cycles have the nature of a paradox, even-length directed cycles have the nature of a dilemma, in the sense that they force one to select one of two possibilities. In this section we look at the behaviour of argumentation frameworks that contain no odd-length directed cycles. We call these argumentation frameworks *even-cycle-free*.

Definition 3.3.12. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A sequence $x_0, \ldots, x_n$ of arguments is called an *even cycle* if and only if n is even; $x_0 = x_n$; $x_i \rightsquigarrow x_{i+1}$ for all $0 \leq i < n$; and for all $0 \leq i, j < n$ s.t. $i \neq j$, $x_i \neq x_j$.[3] We say that F is *even-cycle-free* if it contains no even cycles.

It was proven by Dvorak [47] (strengthening a result obtained by Dunne and Bench-Capon [14]) that these argumentation frameworks have a unique complete extension, and hence a unique complete labelling.

Proposition 3.3.13. *If F is even-cycle-free then $|\mathcal{L}_{Co}(F)| = 1$.*

Proof. (Adapted from [47]) We prove it by contraposition. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. Suppose $|\mathcal{L}_{Co}(F)| \neq 1$. Because $|\mathcal{L}_{Co}(F)| > 1$ it follows that there are two labellings $L, L' \in \mathcal{L}_{Co}(F)$ such that $L \neq L'$. Suppose furthermore that L is the grounded labelling of F. Then there is some $x_0 \in A$

[3] This condition ensures that odd cycles do not 'double count' as even cycles.

such that $L'(x_0) = \mathbf{in}$ and $L(x_0) \neq \mathbf{in}$. Because $L(x_0) \neq \mathbf{in}$, there is a $x_1 \in A$ such that $x_1 \rightsquigarrow x_0$ and $L(x_1) \neq \mathbf{out}$. Furthermore, we have $L'(x_0) = \mathbf{in}$ and thus $L'(x_1) = \mathbf{out}$ and hence there is a $x_2 \in A$ such that $x_2 \rightsquigarrow x_1$ and $L'(x_2) = \mathbf{in}$. Because $L(x_1) \neq \mathbf{out}$, we furthermore have $L(x_2) \neq \mathbf{in}$. Inductively, we obtain a sequence $x_0, x_1, x_2, \ldots$ such that $x_0 \rightsquigarrow x_1 \rightsquigarrow x_2 \ldots$; for each i that is even, $L'(i) = \mathbf{in}$; and for each i that is odd, $L'(i) = \mathbf{out}$. Now let n be the smallest integer such that $x_0 = x_n$ (finiteness of A ensures existence of n). We then have that n is even, and that, for all $0 \leq i, j < n$ s.t. $i \neq j$, $x_i \neq x_j$. It then follows that $x_0, \ldots, x_n$ is an even cycle and hence F is not even-cycle-free. □

We now state two immediate consequences of this fact. The first is that even-cycle-freeness ensures that intervention-based entailment under the grounded, complete, preferred and semi-stable semantics coincides. The second is that intervention-based entailment under the grounded and stable semantics coincide as far as interventions are concerned that do not yield inconsistent conclusions under the stable semantics.

Proposition 3.3.14. *If F is even-cycle-free then $\|\!\approx^F_{Gr} = \|\!\approx^F_{Co} = \|\!\approx^F_{Pr} = \|\!\approx^F_{SS}$.*

Proposition 3.3.15. *If F is even-cycle-free then for all $\Phi \in \mathtt{Int}(F)$, if $\Phi \not\|\!\approx^F_{St} \bot$ then for all $\phi \in \mathtt{lang}(F)$, $\Phi \|\!\approx^F_{St} \phi$ iff $\Phi \|\!\approx^F_{Gr} \phi$.*

Let us start with the complete semantics. We have seen that Rational Monotony fails under the complete semantics. But because rational monotony is satisfied under the grounded semantics, proposition 3.3.14 implies that rational monotony is satisfied in the even-cycle-free case under the complete semantics.

Theorem 3.3.16. *For all $F \in \mathcal{F}$, if F is even-cycle-free then $\|\!\approx^F_{Co}$ satisfies Rational Monotony.*

What about the preferred and semi-stable semantics? Recall that, in the general case, Cautious Monotony fails under the preferred and semi-stable semantics; Cut is satisfied under the preferred semantics but not under the semi-stable semantics; and the preferred and semi-stable semantics both fail Rational Monotony. Proposition 3.3.14 implies that, in the even-cycle-free case, the preferred and semi-stable semantics satisfy Cautious Monotony, Cut and Rational Monotony.

Theorem 3.3.17. *For all $F \in \mathcal{F}$, if F is even-cycle-free then $\|\!\approx^F_{Pr}$ and $\|\!\approx^F_{SS}$ satisfy Cautious Monotony, Cut and Rational Monotony.*

Proof. This follows immediately from proposition 3.3.14 together with the fact that for all $F \in \mathcal{F}$, $\|\!\approx^F_{Gr}$ satisfies Cautious Monotony, Cut and Rational Monotony (theorem 3.3.1 and 3.3.3 and proposition 3.3.9). □

Under the stable semantics, even-cycle-freeness does not ensure satisfaction of (Stable) Cautious Monotony. Here, (Stable) Cautious Monotony still fails if an initial premise entails inconsistency while a strengthening of this premise does not. This is demonstrated by the following example.

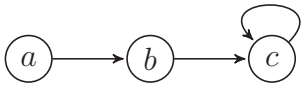


Figure 3.11: Failure of Cautious Monotony due to inconsistency under the stable semantics.

Example 3.3.8. *Let F be the argumentation framework shown in figure 3.11. Note that F is even-cycle-free. Because F has no stable labelling we have $\emptyset \mid\!\!\approx^F_{St} \bot$ and hence, trivially, $\emptyset \mid\!\!\approx^F_{St} \mathbf{out}(a)$. Cautious Monotony implies that we have $\{\mathbf{out}(a)\} \mid\!\!\approx^F_{St} \bot$ but this is not the case, because $F \oplus^{\kappa} \{\mathbf{out}(a)\}$ does have a stable labelling, and hence $\{\mathbf{out}(a)\} \not\mid\!\!\approx^F_{St} \bot$.*

Finally, proposition 3.3.15 implies that even-cycle-freeness ensures satisfaction of Rational Monotony under the stable semantics. Note that the stable semantics fails Rational Monotony in the general case (see example 3.3.5).

Theorem 3.3.18. *For all $F \in \mathcal{F}$, if F is even-cycle-free then $\mid\!\!\approx^F_{St}$ satisfies Rational Monotony.*

Proof. Let $F \in \mathcal{F}$ and assume that F is even-cycle-free. Suppose $\Phi \mid\!\!\approx^F_{St} \phi$ and $\Phi \not\mid\!\!\approx^F_{St} \neg\alpha$. Then $\Phi \not\mid\!\!\approx^F_{St} \bot$ and hence proposition 3.3.15 implies $\Phi \mid\!\!\approx^F_{Gr} \phi$ and $\Phi \not\mid\!\!\approx^F_{Gr} \neg\alpha$. Because $\mid\!\!\approx^F_{Gr}$ satisfies Rational Monotony (theorem 3.3.9) it follows that $\Phi \cup \{\alpha\} \mid\!\!\approx^F_{Gr} \phi$. If $\Phi \cup \{\alpha\} \mid\!\!\approx^F_{St} \bot$, it trivially follows that $\Phi \cup \{\alpha\} \mid\!\!\approx^F_{St} \phi$, and we are done. If we have $\Phi \cup \{\alpha\} \not\mid\!\!\approx^F_{St} \bot$ then the fact that $\Phi \cup \{\alpha\} \mid\!\!\approx^F_{Gr} \phi$ implies, via proposition 3.3.15, that $\Phi \cup \{\alpha\} \mid\!\!\approx^F_{St} \phi$. □

Even-cycle-freeness does not ensure satisfaction of Loop under any of the semantics that we consider, as the counterexample that we used to demonstrate the failure of Loop relies on an argumentation framework that is even-cycle-free.

Table 3.3 summarizes the results obtained here concerning even-cycle-free argumentation frameworks (i.e., theorem 3.3.16, 3.3.17 and 3.3.18).

	Co	*Gr*	*Pr*	*SS*	*St*
(Stable) Cautious Monotony	✓	✓	✓	✓	✗
(Stable) Cut	✓	✓	✓	✓	✓
Rational Monotony	✓	✓	✓	✓	✓
Loop	✗	✗	✗	✗	✗

Table 3.3: Properties satisfied by $\mid\!\!\approx^F_{\sigma}$ for every even-cycle-free $F \in \mathcal{F}$.

3.3.8 Acyclic Argumentation Frameworks

We now consider the acyclic case. Acyclic argumentation frameworks were called *well-founded* by Dung [42].

Definition 3.3.13. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. We say that F is *acyclic* if there is no sequence $x_0, \ldots, x_n$ of arguments such that $x_0 = x_n$ and $x_i \rightsquigarrow x_{i+1}$ for all $0 \leq i < n$.

Obviously, an acyclic argumentation framework is also odd-cycle-free and even-cycle-free. Hence, all the properties that are satisfied in the odd-cycle-free and even-cycle-free case are satisfied in the acyclic case.

Proposition 3.3.19. *For all $F \in \mathcal{F}$, if F is acyclic then $\models^F_{Co}$, $\models^F_{Pr}$, $\models^F_{SS}$ and $\models^F_{St}$ all satisfy (Stable) Cautious Monotony, (Stable) Cut and Rational Monotony.*

Proof. Satisfaction of Cautious Monotony, Cut and Rational Monotony under the complete, preferred and semi-stable semantics follows from the even-cycle-free case (theorem 3.3.17). Satisfaction of Stable Cautious Monotony under the stable semantics follows from the odd-cycle-free case (theorem 3.3.11). Satisfaction of Stable Cut under the stable semantics follows from the general case (theorem 3.3.6). Satisfaction of Rational Monotony under the stable semantics follows from the even-cycle-free case (theorem 3.3.18). □

While this result is easily obtained, the following theorem is non-trivial. It states that, in the acyclic case, Loop is satisfied under the complete, grounded, preferred and semi-stable semantics.

Theorem 3.3.20. *For all $F \in \mathcal{F}$, if F is acyclic then $\models^F_{Co}$, $\models^F_{Gr}$, $\models^F_{Pr}$ and $\models^F_{SS}$ satisfy Loop.*

Proof. Let $F = (A, \rightsquigarrow)$ be an acyclic argumentation framework. We first show that $\models^F_{Gr}$ satisfies Loop. Assume that $\{\alpha_0\} \models^F_{Gr} \alpha_1$, $\{\alpha_1\} \models^F_{Gr} \alpha_2$, …, $\{\alpha_{k-1}\} \models^F_{Gr} \alpha_k$, $\{\alpha_k\} \models^F_{Gr} \alpha_0$. We prove that, for some $i \in \{0, \dots k\}$, it holds that $\emptyset \models^F_{Gr} \alpha_i$. Suppose the contrary: for all $i \in \{0, \dots k\}$ we have $\emptyset \not\models^F_{Gr} \alpha_i$. Because $\models^F_{Gr}$ satisfies conditional directionality (see definition 3.4.7 and theorem 3.4.3) we then have $\alpha_0 \rightsquigarrow^* \alpha_1$, $\alpha_1 \rightsquigarrow^* \alpha_2$, …, $\alpha_{k-1} \rightsquigarrow^* \alpha_k$, $\alpha_k \rightsquigarrow^* \alpha_0$. This is a contradiction because F is acyclic. Hence for some $i \in \{0, \dots k\}$ we have $\emptyset \models^F_{Gr} \alpha_i$. But we also have $\alpha_i \models^F_{Gr} \alpha_{i+1}$ (addition is understood modulo $k+1$). Via Cut we then get $\emptyset \models^F_{Gr} \alpha_{i+1}$. By repeatedly applying Cut like this we get $\emptyset \models^F_{Gr} \alpha_0$ and $\emptyset \models^F_{Gr} \alpha_k$. Via Cautious Monotony we then get $\{\alpha_0\} \models^F_{Gr} \alpha_k$. Hence $\models^F_{Gr}$ satisfies Loop. Via proposition 3.3.14 we get that $\models^F_{\sigma}$ satisfies Loop, for all $\sigma \in \{Co, Gr, Pr, SS\}$. □

Under the stable semantics, the following weakening of Loop is satisfied in the acyclic case. This weakening is similar in nature to the weakening of Cautious Monotony and Cut that we called Stable Cautious Monotony and Stable Cut. We call it *Stable Loop*.

Definition 3.3.14. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\models^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Stable Loop* iff for all $x_1, \dots, x_k \in A$,

$$\text{if } \{\mathbf{out}(x_0)\} \models^F \mathbf{out}(x_1), \{\mathbf{out}(x_1)\} \models^F \mathbf{out}(x_2), \dots, \{\mathbf{out}(x_{k-1})\} \models^F \mathbf{out}(x_k),$$

$$\{\mathbf{out}(x_k)\} \models^F \mathbf{out}(x_0) \text{ then } \{\mathbf{out}(x_0)\} \models^F \mathbf{out}(x_k).$$

Theorem 3.3.21. *If F is acyclic then $\models^F_{St}$ satisfies Stable Loop.*

Proof. Let $F = (A, \rightsquigarrow)$ be an acyclic argumentation framework. Theorem 3.3.20 implies that $\|\!\approx^F_{Gr}$ satisfies Loop and thus also Stable Loop. Because F is odd-cycle-free, we have for all $x \in A$, $\{\mathbf{out}(x)\} \not\|\!\approx^F_{St} \bot$. Because F is even-cycle-free, proposition 3.3.15 then implies that for all $x, y \in A$, $\{\mathbf{out}(x)\} \|\!\approx^F_{St} \mathbf{out}(y)$ iff $\{\mathbf{out}(x)\} \|\!\approx^F_{Gr} \mathbf{out}(y)$. From this it follows that $\|\!\approx^F_{St}$ satisfies Stable Loop. □

Table 3.4 summarizes the results obtained concerning acyclic argumentation frameworks (i.e., proposition 3.3.19 and theorem 3.3.20 and 3.3.21).

	Co	*Gr*	*Pr*	*SS*	*St*
(Stable) Cautious Monotony	✓	✓	✓	✓	✓
(Stable) Cut	✓	✓	✓	✓	✓
Rational Monotony	✓	✓	✓	✓	✓
(Stable) Loop	✓	✓	✓	✓	✓

Table 3.4: Properties satisfied by $\|\!\approx^F_\sigma$ for every acyclic $F \in \mathcal{F}$.

3.4 Directionality and Noninterference

In this section we study the role of *directionality* and *noninterference* in the behaviour of intervention-based entailment. The notion of directionality was introduced by Baroni and Giacomin [6] for extension-based semantics and adapted by Baroni et al. [4] for labelling-based semantics. Intuitively, directionality expresses the idea that the notion of attack is directional: an argument x has an effect on the label of an argument y only if there is a directed path from x to y. Noninterference was introduced by Caminada [27] and adapted by Baroni et al. [4] for labelling-based semantics. It is a weakening of directionality: it states that an argument x has an effect on the label of an argument y only if there is an undirected path between x and y.

Formally, these properties are defined by the condition that certain sets of arguments of an argumentation framework can be evaluated independently of the arguments outside this set. In the case of directionality, these are the *unattacked* sets of an argumentation framework, which are sets of arguments not attacked by an argument outside this set. In the case of noninterference, these are the *isolated* sets, which are sets of arguments not attacked by arguments outside this set and not attacking arguments outside this set.

In this section we show that, in the dynamic setting of intervention-based entailment, directionality and noninterference imply that the effect of an intervention propagates through an argumentation framework in a well-behaved manner. That is, if we use a semantics satisfying directionality, then an intervention involving the argument x only affects consequences about arguments attacked by x, attacked by arguments attacked by x, and so forth. We call this *Conditional Directionality*. Similarly, noninterference implies that an intervention involving the argument x only affects consequences about arguments connected to x, in either direction, by attacks. We call this *Conditional Noninterference*.

3.4.1 Directionality

The Directionality Principle

The *directionality principle* is due to Baroni and Giacomin [6] and was later adapted for labelling-based semantics by Baroni et al. [4]. Intuitively, it expresses idea that the notion of attack is directional: an argument x has an effect on an argument y only if x attacks y. This is formalized by the condition that the *unattacked* sets of an argumentation framework (sets of arguments not attacked by an argument outside this set) can be evaluated independently of the rest of the argumentation framework. First some definitions.

Definition 3.4.1. [4] Given an argumentation framework $F = (A, \rightsquigarrow)$ a labelling $L \in \mathcal{L}(F)$ and a set $B \subseteq A$ we denote by $L \downarrow B$ the *restriction of L by B*, which is defined to be the labelling $(L \downarrow B) : B \to \{\mathbf{in}, \mathbf{out}, \mathbf{und}\}$ such that $(L \downarrow B)(x) = L(x)$, for all $x \in B$. Given a set $M \subseteq \mathcal{L}(F)$, we denote by $M \downarrow B$ the set $\{L \downarrow B \mid L \in M\}$.

Definition 3.4.2. [6] Given an argumentation framework $F = (A, \rightsquigarrow)$ and a set $B \subseteq A$ we denote by $F \downarrow B$ the *restriction of F by B*, which is defined to be the argumentation framework $(B, \rightsquigarrow \cap (B \times B))$.

Definition 3.4.3. [6] Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $B \subseteq A$. We say that B is *unattacked* iff there is no $x \in B$ and $y \in A \setminus B$ such that $y \rightsquigarrow x$. We let $U(F)$ denote the set of unattacked sets of F.

Formally, the condition that unattacked sets can be evaluated independently of the rest of the argumentation framework means that, given an argumentation framework F and unattacked set $B \in U(F)$, the labellings of the restriction of F by B coincide with the labellings of F, restricted by B.

Definition 3.4.4. [4] A labelling-based semantics σ satisfies the directionality principle iff for all $F \in \mathcal{F}$ and $B \in U(F)$,

$$\mathcal{L}_\sigma(F) \downarrow B = \mathcal{L}_\sigma(F \downarrow B).$$

We can characterize the directionality principle in terms of a labelling-based entailment relation as follows.

Proposition 3.4.1. *A labelling-based semantics σ satisfies the directionality principle if and only if*

$$\forall B \in U(F), \phi \in \mathtt{lang}(B), (F \downarrow B) \models_\sigma \phi \textit{ iff } F \models_\sigma \phi.$$

Proof. This follows from definition 2.1.17. □

Not all semantics satisfy directionality. While the complete, grounded and preferred semantics do, the semi-stable and stable semantics do not. This was shown by Baroni et al. [6].

Proposition 3.4.2. *[6] The Co, Gr and Pr semantics satisfy the directionality principle but the SS and St semantics do not.*

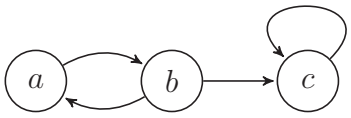


Figure 3.12: Failure of directionality.

The failure of the directionality principle under the semi-stable and stable semantics is demonstrated by the following example.

Example 3.4.1. *Let F be the argumentation framework shown in figure 3.12. Let $\sigma \in \{St, SS\}$. We have*

$$(\mathcal{L}_\sigma(F) \downarrow \{a, b\}) = \{\{(a, \mathbf{out}), (b, \mathbf{in})\}\}.$$

But we also have $\{a, b\} \in U(F)$ and

$$\mathcal{L}_\sigma(F \downarrow \{a, b\}) = \{\{(a, \mathbf{in}), (b, \mathbf{out})\}, \{(a, \mathbf{out}), (b, \mathbf{in})\}\}.$$

In terms of labelling-based entailment, this means that we have, for example, $F \models_\sigma \mathbf{in}(b)$ but not $F \downarrow \{a, b\} \models_\sigma \mathbf{in}(b)$. This is a violation of directionality.

The Conditional Directionality Property

We now introduce the *Conditional Directionality* property for intervention-based entailment. Informally speaking, Conditional Directionality states that an intervention that applies to an argument x only changes the label of an argument y if there is a directed path from x to y. We make this formal using the relation of *structural relevance*[4].

Definition 3.4.5. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. We say that x is *structurally relevant* to y (written $x \rightsquigarrow^* y$) if $x = y$ or if there is a directed path in F from x to y. We similarly say that $B \subseteq A$ is structurally relevant to $B' \subseteq A$ (written $B \rightsquigarrow^* B'$) iff for some $x \in B$ and $y \in B'$ it holds that $x \rightsquigarrow^* y$.

Note that this definition implies that every argument is structurally relevant to itself. We also apply the relation of structural relevance to interventions and formulas. For example, an intervention Φ is structurally relevant to a formula ϕ whenever the set of arguments occurring in Φ is structurally relevant to the set of arguments occurring in ϕ.

Definition 3.4.6. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. We denote by $\mathit{Args}(\Phi)$ (resp. $\mathit{Args}(\phi)$) the set of all arguments occurring in Φ (resp. ϕ). We slightly abuse notation and write $\phi \rightsquigarrow^* \psi$, $\phi \rightsquigarrow^* \Psi$ and $\Phi \rightsquigarrow^* \psi$ to mean $\mathit{Args}(\phi) \rightsquigarrow^* \mathit{Args}(\psi)$, $\mathit{Args}(\phi) \rightsquigarrow^* \mathit{Args}(\Psi)$ and $\mathit{Args}(\Phi) \rightsquigarrow^* \mathit{Args}(\psi)$, respectively.

[4]Not to be confused with the relation of relevance used by Caminada in [27], which is what we call *structural connectedness* in definition 3.4.10.

The Conditional Directionality property states that an intervention only changes the consequences to which the intervention is structurally relevant. In other words, whether or not a formula is a consequence is independent of any intervention not structurally relevant to this formula. Formally, we express this by saying that the consequences of two interventions $\Phi \cup \Psi$ and Φ are the same as far as consequences to which Ψ is not structurally relevant are concerned.

Definition 3.4.7. Let $F \in \mathcal{F}$. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Conditional Directionality* iff for all $\Phi, \Psi \in \texttt{Int}(F)$ and $\phi \in \texttt{lang}(F)$,

$$\text{if } \Psi \not\rightsquigarrow^* \phi \text{ then } \Phi \cup \Psi \|\!\approx^F \phi \text{ iff } \Phi \|\!\approx^F \phi.$$

The principle of directionality, if satisfied by a labelling-based semantics σ, ensures that $\|\!\approx^F_\sigma$ satisfies Conditional Directionality.

Theorem 3.4.3. *If σ satisfies the directionality principle then for all $F \in \mathcal{F}$, $\|\!\approx^F_\sigma$ satisfies Conditional Directionality.*

Proof. See section 3.7. □

The complete, grounded and preferred semantics all satisfy the directionality principle and hence for all $F \in \mathcal{F}$, the relations $\|\!\approx^F_{Co}$, $\|\!\approx^F_{Gr}$ and $\|\!\approx^F_{Pr}$ satisfy Conditional Directionality. This does not hold for the semi-stable and stable semantics, which do not satisfy the directionality principle. The following example demonstrates unintuitive behaviour as a result of this.

Example 3.4.2. *Let F be the argumentation framework shown in figure 3.12. Note that acceptance of b is necessary to minimize undecidedness. Hence we have $\emptyset \|\!\approx^F_{SS} \mathbf{in}(b)$. Furthermore we have $\{\mathbf{out}(c)\} \not\rightsquigarrow^* \mathbf{in}(b)$, and hence the intervention $\{\mathbf{out}(c)\}$ should not affect whether or not $\mathbf{in}(b)$ is a consequence. However, if we defeat c then acceptance of b is no longer necessary to minimize undecidedness. Thus we have $\{\mathbf{out}(c)\} \not\|\!\approx^F_{SS} \mathbf{in}(b)$, which is a violation of Conditional Directionality. This example shows that under the semi-stable semantics an intervention might affect the labels of arguments to which the intervention is not structurally relevant. This example also applies to the stable semantics.*

3.4.2 Noninterference

The Noninterference Principle

Noninterference is a weakening of directionality: it states that an argument x has an effect on an argument y only if x and y are connected (in either direction) by an attack. Noninterference was introduced by Caminada [27] and adapted by Baroni et al. [4] for labelling-based semantics. Formally, noninterference is defined by the condition that the *isolated* sets (sets of arguments not attacked by arguments outside this set and not receiving attacks from outside this set) can be evaluated independently of the rest of the argumentation framework.

Definition 3.4.8. [4] Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $B \subseteq A$. We say that B is *isolated* iff there is no $x \in B$ and $y \in A \setminus B$ such that $x \rightsquigarrow y$ or $y \rightsquigarrow x$. We let $I(F)$ denote the set of isolated sets of F.

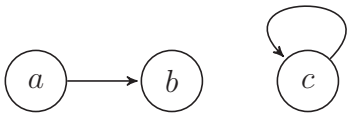


Figure 3.13: A counterexample for noninterference under the stable semantics.

Formally, the condition that isolated sets can be evaluated independently of the rest of the argumentation framework means that, given an argumentation framework F and isolated set $B \in I(F)$, the labellings of the restriction of F by B coincide with the labellings of F, restricted by B. Note that, since an isolated set is also an unattacked set, the directionality principle implies the noninterference principle.

Definition 3.4.9. [4] A labelling-based semantics σ satisfies the noninterference principle if and only if for all $F \in \mathcal{F}$ and $B \in I(F)$, $\mathcal{L}_\sigma(F) \downarrow B = \mathcal{L}_\sigma(F \downarrow B)$.

The noninterference principle is characterized in terms of a labelling-based entailment relation as follows.

Proposition 3.4.4. *A labelling-based semantics σ satisfies noninterference if and only if*

$$\forall B \in I(F), \phi \in \mathtt{lang}(B), (F \downarrow B) \models_\sigma \phi \text{ iff } F \models_\sigma \phi.$$

Proof. This follows from definition 2.1.17. □

The complete, grounded and preferred semantics satisfy the directionality principle and therefore they also satisfy the noninterference principle. Caminada has shown that the semi-stable semantics also satisfies noninterference, but that the stable semantics does not [27].

Proposition 3.4.5. *The Co, Gr, Pr and SS semantics satisfy the noninterference principle but the St semantics does not.*

The failure of the noninterference principle under the stable semantics is demonstrated by the following example.

Example 3.4.3. *Let F be the argumentation framework shown in figure 3.13. Note that $\{a, b\} \in I(F)$. We have $\mathcal{L}_{St}(F) = \emptyset$ and hence*

$$(\mathcal{L}_{St}(F) \downarrow \{a, b\}) = \emptyset.$$

But we also have

$$\mathcal{L}_{St}(F \downarrow \{a, b\}) = \{\{(a, \mathbf{in}), (b, \mathbf{out})\}\}.$$

In terms of labelling-based entailment, this means that we have, for example, $F \models_{St} \bot$ but not $F \downarrow \{a, b\} \models_{St} \bot$. This is a violation of noninterference.

The Conditional Noninterference Property

The *Conditional Noninterference* property for intervention-based entailment is related to the noninterference property of a semantics. Informally speaking, Conditional Noninterference states that an intervention that applies to an argument x only changes the label of an argument y if there is an *undirected* path from x to y. We make this formal using the relation of *structural connectedness*. We say that an argument x is *structurally connected* to an argument y if there is an undirected path between x and y.

Definition 3.4.10. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. We say that x is *structurally connected* to y (written $x \rightsquigarrow^* y$) if $x = y$ or if there is an undirected path in F from x to y. We similarly say that $B \subseteq A$ is structurally connected to $B' \subseteq A$ (written $B \leftrightsquigarrow^* B'$) iff for some $x \in B$ and $y \in B'$ it holds that $x \leftrightsquigarrow^* y$.

It is easy to check that the structural connectedness relation is an equivalence relation, whose classes are exactly the isolated sets of the argumentation framework. Like we did for the relation of structural relevance, we extend the relation of structural connectedness to apply to interventions and formulas. If, e.g., it holds that $\mathit{Args}(\Phi) \leftrightsquigarrow^* \mathit{Args}(\phi)$ then we also write $\Phi \leftrightsquigarrow^* \phi$.

Conditional Noninterference states that an intervention only changes consequences to which the intervention is structurally connected. In other words, whether or not a formula is a consequence is independent of any intervention not structurally connected to this formula.

Definition 3.4.11. Let $F \in \mathcal{F}$. A relation $\|\!\approx^F \subseteq \mathtt{Int}(F) \times \mathtt{lang}(F)$ satisfies *Conditional Noninterference* iff for all $\Phi, \Psi \in \mathtt{Int}(F)$ and $\phi \in \mathtt{lang}(F)$,

$$\text{if } \Psi \not\leftrightsquigarrow^* \phi \text{ then } \Phi \cup \Psi \|\!\approx^F \phi \text{ iff } \Phi \|\!\approx^F \phi.$$

It is easy to see that structural relevance implies structural connectedness, and hence that we have the following.

Proposition 3.4.6. *If $\|\!\approx^F$ satisfies Conditional Directionality then $\|\!\approx^F$ satisfies Conditional Noninterference.*

Proof. Suppose $\|\!\approx^F$ satisfies Conditional Directionality and assume $\Psi \not\leftrightsquigarrow^* \phi$. It then follows that $\Psi \not\rightsquigarrow^* \phi$ and hence, by Conditional Directionality, $\Phi \cup \Psi \|\!\approx^F \phi$ iff $\Phi \|\!\approx^F \phi$. □

Furthermore, if a labelling-based semantics σ satisfies the noninterference principle then for every argumentation framework F, the relation $\|\!\approx^F_\sigma$ satisfies Conditional Noninterference.

Theorem 3.4.7. *For any labelling-based semantics σ that satisfies the noninterference principle it holds that for all $F \in \mathcal{F}$, $\|\!\approx^F_\sigma$ satisfies Conditional Noninterference.*

Proof. See section 3.7. □

The complete, grounded, preferred and semi-stable semantics all satisfy the noninterference principle and hence for all $F \in \mathcal{F}$, the relations $\|\!\sim^F_{Co}$, $\|\!\sim^F_{Gr}$, $\|\!\sim^F_{Pr}$ and $\|\!\sim^F_{SS}$ satisfy Conditional Noninterference. This does not hold for the stable semantics, which does not satisfy the noninterference principle. The following example demonstrates unintuitive behaviour as a result of this.

Example 3.4.4. *Let F be the argumentation framework shown in figure 3.13. We have $\{\mathbf{out}(c)\} \not\|\!\sim^F_{St} \mathbf{in}(b)$ and $\{\mathbf{out}(c)\} \not\rightsquigarrow^* \mathbf{in}(b)$. Conditional Noninterference would imply that we have $\emptyset \not\|\!\sim^F_{St} \mathbf{in}(b)$ but this is not the case. We have instead $\emptyset \|\!\sim^F_{St} \bot$ and hence $\emptyset \|\!\sim^F_{St} \mathbf{in}(b)$.*

3.4.3 Summary and Discussion of Results

We have shown that, whenever a semantics σ satisfies the directionality and noninterference principle, it holds that for all $F \in \mathcal{F}$, the relation $\|\!\sim^F_\sigma$ satisfies Conditional Directionality and Conditional Noninterference, respectively. The results with respect to the semantics that we consider are shown in table 3.5.

	$\|\!\sim^F_{Co}$	$\|\!\sim^F_{Gr}$	$\|\!\sim^F_{Pr}$	$\|\!\sim^F_{SS}$	$\|\!\sim^F_{St}$
Conditional Directionality	✓	✓	✓	✗	✗
Conditional Noninterference	✓	✓	✓	✓	✗

Table 3.5: Properties satisfied by the relation $\|\!\sim^F_\sigma$, for all $F \in \mathcal{F}$.

It is interesting to contrast the properties of Conditional Directionality and Conditional Noninterference on the one hand, and Cautious Monotony and Cut on the other. While Cautious Monotony and Cut together state that two interventions Φ and $\Phi \cup \{\alpha\}$ are equivalent in terms of their consequences if α is already a consequence of Φ, Conditional Directionality and Conditional Noninterference implies that Φ and $\Phi \cup \{\alpha\}$ are the same as far as the consequences to which α is not structurally relevant or connected are concerned. The relation between these properties is that they all express conditions under which a consequence (resp. non-consequence) of an intervention is still a consequence (resp. non-consequence) if we change the intervention.

Although the technical differences between the theories of abstract argumentation and causal Bayesian networks complicates their comparison, the property of Conditional Directionality expresses a principle that also applies to intervention in causal Bayesian networks. This is the principle that in a causal Bayesian network, the effect of an intervened variable propagates only to the direct and indirect effects of this variable. This is because an intervened variable in a causal Bayesian network is made independent of its typical causes.

3.5 Related Work

We have developed the notion of intervention-based entailment as a formal tool to study the idea that an argumentation framework is a system whose evaluation changes as new arguments and attacks are added. A similar perspective is

taken by Baroni et al. [3], who study the so called input/output behaviour of argumentation frameworks. Roughly speaking, the idea is that an argumentation framework is seen as a sort of black box, with both input and output, formed by interactions with arguments outside the argumentation framework. They study two types of properties. The first type of properties concern the *decomposability* of a semantics, or the question of whether putting together the labellings of different subframeworks under a given semantics (while taking into account the interactions between the subframeworks) results in a set of labellings that coincides with the labellings of the whole argumentation framework under the same semantics. The first type of properties concern the *replacement* properties: the question of whether different argumentation frameworks that behave the same with respect to a given set of input and output arguments are interchangeable in the context of a larger argumentation framework, without affecting the labellings of this larger argumentation framework.

A specific type of decomposability, called *SCC-recusiveness*, was studied earlier by Baroni et al. [8]. The SCC-recursive scheme exploits the property that an argumentation framework can be decomposed into a set of strongly connected components (SCCs) and that the graph obtained considering SCCs as single nodes is acyclic. This means that this decomposition yields a partial order over SCCs. This partial order can be used for the incremental computation of the extensions of an argumentation framework, by first computing the extensions of the initial SCCs (i.e., those not attacked by other SCCs), and using these results to recursively compute the extensions of the whole argumentation framework. A semantics is called SCC-recursive if its extensions can be computed using this recursive scheme.

We did not, like the work mentioned here, focus on properties related to decomposability or replacement. Nevertheless, the results that we have obtained contribute, like these properties, to a deeper understanding of the behaviour of an argumentation framework when external input is considered.

Liao et al. [65] addressed the question of whether, after modifying an argumentation framework, it is possible to partially reuse extensions computed before the modification, when computing the extensions of the modified argumentation framework. They show that, roughly speaking, directionality ensures the evaluation of the set of arguments not reachable from some argument affected by the modification remains the same. They call this set of unreachable arguments the *unaffected part*. This is very similar to what the Conditional Directionality property states (that is, consequences referring to arguments not reachable from an argument affected by an intervention remain the same). In a follow-up work, Baroni et al. [9] show that the SCC-recursiveness property enables one to reuse the (unchanged) evaluation of the unaffected part when computing the evaluation of the modified argumentation framework.

Sakama [84] studied properties of counterfactual conditionals about the status of arguments in an argumentation framework. He defines a semantics for counterfactual conditionals in abstract argumentation of the form $\alpha \ \Box\!\!\rightarrow \beta$ or $\alpha \ \Diamond\!\!\rightarrow \beta$, where α and β are single literals of the form $\mathbf{in}(x)$ or $\mathbf{out}(x)$. The intended meaning of $\alpha \ \Box\!\!\rightarrow \beta$ (resp. $\alpha \ \Diamond\!\!\rightarrow \beta$) is that "if it were the case that α, then it would (resp. might) be the case that β." The interpretation is similar to ours, i.e., the premise is translated into a modification of the argumentation

framework within which the consequent is evaluated. In particular, acceptance in the premise is translated by removing all attackers pointing towards the accepted argument. A number of properties of these conditionals are studied, and some of them are similar to the properties we considered. These include Reflexivity and a restricted form of Cautious Monotony, as well as properties such as Transitivity and Contraposition. We have extended these results by, for example, demonstrating the failure of Cautious Monotony under the preferred, semi-stable and stable semantics, and proving results with respect to argumentation frameworks containing no (odd-length or even-length) directed cycles.

Cayrol et al. [34] study the impact of modifying an argumentation framework by adding a single new argument that interacts with old arguments. They define a number of properties that describe the impact of such a modification on the extensions of the argumentation framework, such as *restrictiveness*, *conservativeness* and *questioning* (i.e., modifications that result, respectively, in a smaller, equivalent, and larger set of extensions). They then define necessary and sufficient conditions under which a modification satisfies these properties. While the properties that we studied in this chapter are, like those studied by Cayrol et al., also about the impact of modifying an argumentation framework, the nature of the type of properties used to describe this impact are of a different nature. For example, all properties used by Cayrol et al. refer to extensions instead of labellings, and hence no distinction is made between **out** and **und** labelled arguments. Furthermore, the majority of these properties refer explicitly to the number of extensions before and after the modification, while this number does not appear in the definition of the properties that we consider.

3.6 Conclusion and Future Work

We modelled change due to actions performed in a debate by introducing the notion of intervention-based entailment. We simplified our model by abstracting away from the actual new arguments and attacks that may be introduced in a debate, and instead we focussed on two actions: defeat of an argument and provisional defeat of an argument, which are modelled by the addition of a (self-attacking) attacker. We have shown that this is sufficient to obtain a number of fundamental results concerning the behaviour of this type of change. Nevertheless, there are several directions in which our model can be generalized, which is a subject for future research. For example, we may look at changes to an argumentation framework where elements are removed, or where more complex additions (e.g. addition of arguments that are in turn attacked by existing arguments) are performed.

We investigated how intervention-based entailment behaves with respect to the KLM properties, and obtained a number of surprising results, like the failure of Cautious Monotony and Cut under a number of semantics, except in cases where the argumentation framework contains no odd-length or even-length directed cycles.

Finally, we investigated the role of directionality and noninterference in the behaviour of intervention-based entailment. We showed that these properties

ensure that the effect of an action propagates in a well-behaved manner. In particular, directionality of a semantics σ ensures that the for all $F \in \mathcal{F}$, the relation $\|\!\approx_\sigma^F$ satisfies Conditional Directionality, which expresses a principle that also applies to intervention in causal Bayesian networks, namely that the effect of an intervened variable propagates only to the direct and indirect effects of this variable. This raises the question whether notions such as the *Markov assumption* and the *d-separation* criterion [72] have analogues in the theory of abstract argumentation. This is a question we plan to address in future work.

A question that we have not addressed is: what is the meaning of intervention in instantiated argumentation? The main obstacle is that the instantiated setting imposes certain restrictions on how an argumentation framework can be modified. If these restrictions are not respected, the labellings of the argumentation framework do not any more correspond to consistent sets of conclusions on the instantiated level. If, for example, an argument x contains no defeasible rules, then it is impossible to construct an instantiated argument that attacks x and hence the rejection of x cannot be enforced. In the instantiated setting, it may furthermore be more appropriate to look at interventions that are expressed in the object language of the instantiation itself.

Another direction for future work is to relate the notion of intervention-based entailment with the relation of strong equivalence of argumentation frameworks [71]. Strong equivalence between argumentation frameworks means that they generate equivalent sets of extensions, and that this equivalence is robust with respect to the addition of new arguments and attacks. The main issue addressed is the characterization of strong equivalence in terms of syntactical (i.e., topology related) conditions. In the setting of intervention-based entailment we can address the question: given two argumentation frameworks F_1 and F_2, what are the conditions under which it holds that $\|\!\approx_\sigma^{F_1} = \|\!\approx_\sigma^{F_2}$?

3.7 Proofs

Theorem 3.2.8. *Let F be an argumentation framework and $\sigma \in \{Co, Gr, Pr, SS\}$. The following are equivalent.*

1. *For some $\Phi \in \mathtt{Int}(F)$, $\Phi \|\!\approx_\sigma^F \phi$.*
2. *ϕ is conflict-free with respect to F.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and $\sigma \in \{Co, Gr, Pr, SS\}$. (1) to (2): Let $\Phi \in \mathtt{Int}(F)$ and $\phi \in \mathtt{lang}(F)$ be a formula such that $\Phi \|\!\approx_\sigma^F \phi$. Suppose, toward a contradiction, that ϕ is not conflict-free w.r.t. F. Because we have $\mathcal{L}_\sigma(F, \Phi) \neq \emptyset$, this implies, via definition 3.2.10, that there is some $L \in \mathcal{L}_\sigma(F, \Phi)$ such that $L \notin \mathcal{L}_{Cf}(F)$. But this implies that there is an $L \in \mathcal{L}_\sigma(F \oplus^\kappa \Phi)$ for some F-mapping κ, and $L \notin \mathcal{L}_{Cf}(F \oplus^\kappa \Phi)$. This is a contradiction. Hence ϕ is conflict-free w.r.t. F. (2) to (1): Let $\phi \in \mathtt{lang}(F)$ be a conflict-free formula w.r.t. F. Let $L \in \mathcal{L}_{Cf}(F)$ be a labelling such that $L \models \phi$ (definition 3.2.10 ensures existence of L). We define Φ by $\Phi = \{\mathbf{out}(x) \mid L(x) = \mathbf{out}\} \cup \{\neg\mathbf{in}(x) \mid L(x) = \mathbf{und}\}$. Let κ be an F-mapping,

let $(A', \rightsquigarrow') = F \oplus^{\kappa} \Phi$ and let $L' \in \mathcal{L}_{\sigma}((A', \rightsquigarrow'))$. We prove that $L = L' \downarrow A$. Let $x \in A$. Three cases:

1. $L(x) = \mathbf{out}$. Then $\mathbf{out}(x) \in \Phi$. From definition 3.2.5 it follows that there is an $y \in A'$ such that $y \rightsquigarrow' x$ and y is unattacked in F'. Condition 2 and 4 in definition 2.1.11 imply that $L'(x) = \mathbf{out}$.

2. $L(x) = \mathbf{in}$. Conflict-freeness of L then implies that for all $y \in A$ s.t. $y \rightsquigarrow x$, $L(y) = \mathbf{out}$ and hence $\mathbf{out}(y) \in \Phi$. From (1) and the fact that $\mathbf{out}(x) \notin \Phi$ and $\neg\mathbf{in}(x) \notin \Phi$ it then follows that for all $y \in A'$ s.t. $y \rightsquigarrow x$, $L'(y) = \mathbf{out}$. Condition 4 in definition 2.1.11 finally implies that $L'(x) = \mathbf{in}$.

3. $L(x) = \mathbf{und}$. Then $(\neg\mathbf{in}(x)) \in \Phi$. Conflict-freeness of L then implies that for all $y \in A$ s.t. $y \rightsquigarrow x$, $L(y) \neq \mathbf{in}$. From (1) and (2) it follows that for all $y \in A$ s.t. $y \rightsquigarrow x$, $L'(y) \neq \mathbf{in}$. Furthermore we have one more attacker $y \in A' \setminus y$ such that $y \rightsquigarrow' x$, and $L'(y) = \mathbf{und}$. Condition 1 and 2 in definition 2.1.11 imply that $L'(x) = \mathbf{und}$.

Hence $L = L' \downarrow A$. Definition 3.2.8 now implies that $\Phi \mid\!\!\approx^{F}_{Co} \phi$. □

Theorem 3.2.9. *Let F be an argumentation framework. The following are equivalent.*

1. *For some $\Phi \in \mathtt{Int}(F)$ we have $\Phi \mid\!\!\approx^{F}_{St} \phi$ and $\Phi \not\mid\!\!\approx^{F}_{St} \bot$.*

2. *ϕ is a stable conflict-free with respect to F.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. (1) to (2): Let $\Phi \in \mathtt{Int}(F)$ and $\phi \in \mathtt{lang}(F)$ a formula such that $\Phi \mid\!\!\approx^{F}_{St} \phi$ and $\Phi \not\mid\!\!\approx^{F}_{St} \bot$. Suppose, toward a contradiction, that ϕ is not stable conflict-free w.r.t. F. Because we have $\Phi \not\mid\!\!\approx^{F}_{St} \bot$, we have $\mathcal{L}_{St}(F, \Phi) \neq \emptyset$. This implies, via definition 3.2.11, that there is some $L \in \mathcal{L}_{St}(F, \Phi)$ such that $L \notin \mathcal{L}_{Cf}(F)$ or $L^{-1}(\mathbf{und}) \neq \emptyset$. But this implies that there is an $L \in \mathcal{L}_{St}(F \oplus^{\kappa} \Phi)$ for some F-mapping κ, and $L \notin \mathcal{L}_{Cf}(F \oplus^{\kappa} \Phi)$ or $L^{-1}(\mathbf{und}) \neq \emptyset$. This is a contradiction. Hence ϕ is conflict-free w.r.t. F. (2) to (1): Let $\phi \in \mathtt{lang}(F)$ be stable conflict-free w.r.t. F. We let $L \in \mathcal{L}_{Cf}(F)$ be a labelling such that $L^{-1}(\mathbf{und}) = \emptyset$ and $L \models \phi$ (definition 3.2.11 ensures existence of L) and define Φ by $\Phi = \{\mathbf{out}(x) \mid L(x) = \mathbf{out}\}$. Now let κ be an F-mapping and define $L' \in \mathcal{L}(F \oplus^{\kappa} \Phi)$ by $L'(x) = L(x)$, if $x \in A$ and $L'(x) = \mathbf{in}$, otherwise. It then follows that every **out**-labelled argument in L' is attacked by an argument that is labelled **in** and every **in** labelled argument is attacked only by arguments that are labelled **out** and hence L' is a stable labelling of $F \oplus^{\kappa} \Phi$. Furthermore it is the *only* stable labelling of $F \oplus^{\kappa} \Phi$, because every **out**-labelled argument is attacked by an argument that is labelled **in** in *every* stable labelling of $F \oplus^{\kappa} \Phi$. Via definition 3.2.8 it follows that $\Phi \mid\!\!\approx^{F}_{St} \phi$ and $\Phi \not\mid\!\!\approx^{F}_{St} \bot$. □

For the proof of theorem 3.3.1 and 3.3.3 we use the following lemmas.

Lemma 3.7.1. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $a \in A$. Let $\{E\} = \mathcal{E}_{Gr}(F)$ and $\{E'\} = \mathcal{E}_{Gr}(F \oplus^{\kappa} \{\mathbf{out}(a)\})$ for some F-mapping κ. If $E \rightsquigarrow a$ then $E = E' \cap A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework, let $a \in A$ and let $F' = (A', \rightsquigarrow') = F \oplus^{\kappa} \{\mathbf{out}(a)\}$ for some F-mapping κ. Let $\{E\} = \mathcal{E}_{Gr}(F)$ and $\{E'\} = \mathcal{E}_{Gr}(F')$. Note that we have $\kappa(a) \rightsquigarrow' a$. Assume $E \rightsquigarrow a$. We prove $E = E' \cap A$. We prove the two inclusions separately.

1. $E \subseteq E' \cap A$: We prove this by showing that, for all $n \in \mathbb{Z}$, $\mathcal{D}^n_F(\emptyset) \subseteq \mathcal{D}^n_{F'}(\emptyset)$ (where $\mathcal{D}^n_F(\emptyset)$ is defined by $\mathcal{D}^1_F(\emptyset) = \mathcal{D}_F(\emptyset)$ and for all $i > 1$, $\mathcal{D}^i_F(\emptyset) = \mathcal{D}_F(\mathcal{D}^{i-1}_F(\emptyset))$). Assume towards contradiction that for some $n \in \mathbb{Z}$, there is an $x \in A$ such that $x \in \mathcal{D}^n_F(\emptyset)$ and $x \notin \mathcal{D}^n_{F'}(\emptyset)$. Let $y \in A'$ be the argument such that $y \rightsquigarrow' x$ while for all $z \in A'$ such that $z \rightsquigarrow' y$, we have $z \notin \mathcal{D}^{n-1}_{F'}(\emptyset)$ (definition 2.1.2 guarantees existence of y). We then also have $y \in A$ (if not, then $y = \kappa(a)$ and hence $x = a$, which is a contradiction, because we have $E \rightsquigarrow a$ and thus $a \notin E$, and therefore $a \notin \mathcal{D}^n_F(\emptyset)$). Because we have $x \in \mathcal{D}^n_F(\emptyset)$, there is some $z \in A$ such that $z \rightsquigarrow y$ and $z \in \mathcal{D}^{n-1}_F(\emptyset)$, while $z \notin \mathcal{D}^{n-1}_{F'}(\emptyset)$. By continuing this argument we arrive at $\mathcal{D}_F(\emptyset) \not\subseteq \mathcal{D}_{F'}(\emptyset)$, which is false. Hence $\mathcal{D}^n_F(\emptyset) \subseteq \mathcal{D}^n_{F'}(\emptyset)$ and, by proposition 2.1.1, $E \subseteq E' \cap A$.

2. $E' \cap A \subseteq E$: We prove this by showing that, for all $n \in \mathbb{Z}$, $\mathcal{D}^n_{F'}(\emptyset) \cap A \subseteq \mathcal{D}^\infty_F(\emptyset)$. Assume towards contradiction that for some $n \in \mathbb{Z}$, there is an $x \in A$ such that $x \in \mathcal{D}^n_{F'}(\emptyset) \cap A$ and $x \notin \mathcal{D}^\infty_F(\emptyset)$. Let $y \in A$ be the argument such that $y \rightsquigarrow x$ while for all $z \in A$ such that $z \rightsquigarrow y$, we have $z \notin \mathcal{D}^\infty_F(\emptyset)$ (definition 2.1.2 guarantees existence of y). Let $z \in A'$ be the argument such that $z \rightsquigarrow' y$ and $z \in \mathcal{D}^{n-1}_{F'}(\emptyset)$ (definition 2.1.2, together with the fact that $y \in A'$ and $y \rightsquigarrow' x$, guarantees existence of z). We then also have $z \in A$ (if not, then $z = \kappa(a)$ and hence $y = a$, which is a contradiction, because we then have $E \rightsquigarrow y$, contradicting that for all $z \in A$ s.t. $z \rightsquigarrow y$, $z \notin \mathcal{D}^\infty_F(\emptyset)$). Thus we have $z \in \mathcal{D}^{n-1}_{F'}(\emptyset) \cap A$ and $z \notin \mathcal{D}^{n-1}_F(\emptyset)$. By continuing this argument we arrive at $\mathcal{D}_{F'}(\emptyset) \cap A \not\subseteq \mathcal{D}_F(\emptyset)$, which is false. Hence $\mathcal{D}^n_{F'}(\emptyset) \cap A \subseteq \mathcal{D}^\infty_F(\emptyset)$ and, by proposition 2.1.1, $E' \cap A \subseteq E$.

□

Lemma 3.7.2. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $a \in A$. Let $\{E\} = \mathcal{E}_{Gr}(F)$ and $\{E'\} = \mathcal{E}_{Gr}(F \oplus^{\kappa} \{\neg\mathbf{in}(a)\})$ for some F-mapping κ. If $a \notin E$ then $E = E' \cap A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework, let $a \in A$ and let $F' = (A', \rightsquigarrow') = F \oplus^{\kappa} \{\neg\mathbf{in}(a)\}$ for some F-mapping κ. Let $\{E\} = \mathcal{E}_{Gr}(F)$ and $\{E'\} = \mathcal{E}_{Gr}(F')$. Note that we have $\kappa(a) \rightsquigarrow' a$ and $\kappa(a) \rightsquigarrow' \kappa(a)$. Assume $a \notin E$. We prove $E = E' \cap A$. We prove the two inclusions separately.

1. $E \subseteq E' \cap A$: We prove this by showing that, for all $n \in \mathbb{Z}$, $\mathcal{D}^n_F(\emptyset) \subseteq \mathcal{D}^n_{F'}(\emptyset)$ (where $\mathcal{D}^n_F(\emptyset)$ is defined by $\mathcal{D}^1_F(\emptyset) = \mathcal{D}_F(\emptyset)$ and for all $i > 1$, $\mathcal{D}^i_F(\emptyset) = \mathcal{D}_F(\mathcal{D}^{i-1}_F(\emptyset))$). Assume towards contradiction that for some $n \in \mathbb{Z}$, there is an $x \in A$ such that $x \in \mathcal{D}^n_F(\emptyset)$ and $x \notin \mathcal{D}^n_{F'}(\emptyset)$. Let $y \in A'$ be the argument such that $y \rightsquigarrow' x$ while for all $z \in A'$ such that $z \rightsquigarrow' y$, we have $z \notin \mathcal{D}^{n-1}_{F'}(\emptyset)$ (definition 2.1.2 guarantees existence of y). We then also have $y \in A$ (if not, then $y = \kappa(a)$ and hence $x = a$,

which is a contradiction, because we have $a \notin E$ and hence $a \notin \mathcal{D}_F^n(\emptyset)$). Because we have $x \in \mathcal{D}_F^n(\emptyset)$, there is some $z \in A$ such that $z \rightsquigarrow y$ and $z \in \mathcal{D}_F^{n-1}(\emptyset)$, while $z \notin \mathcal{D}_{F'}^{n-1}(\emptyset)$. By continuing this argument we arrive at $\mathcal{D}_F(\emptyset) \not\subseteq \mathcal{D}_{F'}(\emptyset)$, which is false. Hence $\mathcal{D}_F^n(\emptyset) \subseteq \mathcal{D}_{F'}^n(\emptyset)$ and, by proposition 2.1.1, $E \subseteq E' \cap A$.

2. $E' \cap A \subseteq E$: We prove this by showing that, for all $n \in \mathbb{Z}$, $\mathcal{D}_{F'}^n(\emptyset) \cap A \subseteq \mathcal{D}_F^\infty(\emptyset)$. Assume towards contradiction that for some $n \in \mathbb{Z}$, there is an $x \in A$ such that $x \in \mathcal{D}_{F'}^n(\emptyset) \cap A$ and $x \notin \mathcal{D}_F^\infty(\emptyset)$. Let $y \in A$ be the argument such that $y \rightsquigarrow x$ while for all $z \in A$ such that $z \rightsquigarrow y$, we have $z \notin \mathcal{D}_F^\infty(\emptyset)$ (definition 2.1.2 guarantees existence of y). Let $z \in A'$ be the argument such that $z \rightsquigarrow' y$ and $z \in \mathcal{D}_{F'}^{n-1}(\emptyset)$ (definition 2.1.2, together with the fact that $y \in A'$ and $y \rightsquigarrow' x$, guarantees existence of z). We then also have $z \in A$ (if not, then $z = \kappa(a)$, which is a contradiction, because we then have $z \rightsquigarrow' z$, contradicting $z \in \mathcal{D}_{F'}^{n-1}(\emptyset)$). Thus we have $z \in \mathcal{D}_{F'}^{n-1}(\emptyset) \cap A$ and $z \notin \mathcal{D}_F^{n-1}(\emptyset)$. By continuing this argument we arrive at $\mathcal{D}_{F'}(\emptyset) \cap A \not\subseteq \mathcal{D}_F(\emptyset)$, which is false. Hence $\mathcal{D}_{F'}^n(\emptyset) \cap A \subseteq \mathcal{D}_F^\infty(\emptyset)$ and, by proposition 2.1.1, $E' \cap A \subseteq E$.

□

Lemma 3.7.3. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Let $\{L\} = \mathcal{L}_{Gr}(F)$ and $\{L'\} = \mathcal{L}_{Gr}(F \oplus^\kappa \{\mathbf{out}(x)\})$ for some F-mapping κ. If $L \models \mathbf{out}(x)$ then $L = L' \downarrow A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework, let $x \in A$ and let $F' = (A', \rightsquigarrow') = F \oplus^\kappa \mathbf{out}(x)$ for some F-mapping κ. Let $\{L\} = \mathcal{L}_{Gr}(F)$ and $\{L'\} = \mathcal{L}_{Gr}(F')$ and assume $L \models \mathbf{out}(x)$. We then have $E \rightsquigarrow x$, where $\{E\} = \mathcal{E}_{Gr}(F)$. Lemma 3.7.1 implies that $E = E' \cap A$, where $\{E'\} = \mathcal{E}_{Gr}(F')$. We now prove that $L = L' \downarrow A$. Proposition 2.1.7 implies that $L^{-1}(\mathbf{in}) = L'^{-1}(\mathbf{in}) \cap A$. We split the remainder of the proof in two parts.

- $L^{-1}(\mathbf{out}) \subseteq L'^{-1}(\mathbf{out}) \cap A$. Suppose $y \in L^{-1}(\mathbf{out})$. We immediately have that $y \in A$. Completeness of L implies that $\exists z \in A$ s.t. $z \rightsquigarrow y$, $L(z) = \mathbf{in}$ and hence $z \in E$. Because $E \subseteq E'$, it follows that $z \in E'$ and hence $L'(z) = \mathbf{in}$. Completeness of L' implies that $y \in L'^{-1}(\mathbf{out})$.

- $L'^{-1}(\mathbf{out}) \cap A \subseteq L^{-1}(\mathbf{out})$. Suppose $y \in L'^{-1}(\mathbf{out}) \cap A$. If $x = y$ then we immediately have $y \in L^{-1}(\mathbf{out})$. In the remainder we assume $x \neq y$. Because L' is complete there is a $z \in A'$ s.t. $z \rightsquigarrow' y$ and $L'(z) = \mathbf{in}$ and, since $x \neq y$, we have $z \in A$ and $z \rightsquigarrow y$. Thus, there is a $z \in A$ s.t. $z \rightsquigarrow y$, $z \in E$ and hence $L(z) = \mathbf{in}$. Completeness of L implies that $L(y) = \mathbf{out}$.

Thus we have $L^{-1}(\mathbf{in}) = L'^{-1}(\mathbf{in}) \cap A$, $L^{-1}(\mathbf{out}) = L'^{-1}(\mathbf{out}) \cap A$ and consequently $L^{-1}(\mathbf{und}) = L'^{-1}(\mathbf{und}) \cap A$. It follows that $L = L' \downarrow A$. □

Lemma 3.7.4. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Let $\{L\} = \mathcal{L}_{Gr}(F)$ and $\{L'\} = \mathcal{L}_{Gr}(F \oplus^\kappa \{\neg\mathbf{in}(x)\})$ for some F-mapping κ. If $L \models \neg\mathbf{in}(x)$ then $L = L' \downarrow A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework, let $x \in A$ and let $F' = (A', \rightsquigarrow') = F \oplus^{\kappa} \neg\mathbf{in}(x)$ for some F-mapping κ. Let $\{L\} = \mathcal{L}_{Gr}(F)$ and $\{L'\} = \mathcal{L}_{Gr}(F')$ and assume $L \models \neg\mathbf{in}(x)$. We then have $x \notin E$, where $\{E\} = \mathcal{E}_{Gr}(F)$. Lemma 3.7.2 implies that $E = E' \cap A$, where $\{E'\} = \mathcal{E}_{Gr}(F')$. We now prove that $L = L' \downarrow A$. Proposition 2.1.7 implies that $L^{-1}(\mathbf{in}) = L'^{-1}(\mathbf{in}) \cap A$. We split the remainder of the proof in two parts.

- $L^{-1}(\mathbf{out}) \subseteq L'^{-1}(\mathbf{out}) \cap A$. Suppose $y \in L^{-1}(\mathbf{out})$. We immediately have that $y \in A$. Completeness of L implies that $\exists z \in A$ s.t. $z \rightsquigarrow y$, $L(z) = \mathbf{in}$ and hence $z \in E$. Because $E \subseteq E'$, it follows that $z \in E'$ and hence $L'(z) = \mathbf{in}$. Completeness of L' implies that $y \in L'^{-1}(\mathbf{out})$.
- $L'^{-1}(\mathbf{out}) \cap A \subseteq L^{-1}(\mathbf{out})$. Suppose $y \in L'^{-1}(\mathbf{out}) \cap A$. Completeness of L' implies that there is a $z \in A'$ s.t. $z \rightsquigarrow' y$ and $L'(z) = \mathbf{in}$. Because $z \notin A$ implies $z \rightsquigarrow' z$, which contradicts $L'(z) = \mathbf{in}$, we have that $z \in A$. Hence there is a $z \in A$ s.t. $z \rightsquigarrow y$ and $z \in E'$. It then follows that $z \in E$ and hence $L(z) = \mathbf{in}$. Completeness of L implies that $L(y) = \mathbf{out}$.

Thus we have $L^{-1}(\mathbf{in}) = L'^{-1}(\mathbf{in}) \cap A$, $L^{-1}(\mathbf{out}) = L'^{-1}(\mathbf{out}) \cap A$ and consequently $L^{-1}(\mathbf{und}) = L'^{-1}(\mathbf{und}) \cap A$. It follows that $L = L' \downarrow A$. □

We are now ready to prove theorem 3.3.1 and 3.3.3.

Theorem 3.3.1. *For all $F \in \mathcal{F}$, $\|\!\approx^{F}_{Gr}$ satisfies Cautious Monotony.*

Proof. Let $F \in \mathcal{F}$, $\Phi \in \mathtt{Int}(F)$ and let κ be an F-mapping. Suppose $\Phi \|\!\approx^{F}_{Gr} \alpha$. Lemma 3.7.16 implies that $\emptyset \|\!\approx^{F \oplus^{\kappa} \Phi}_{Gr} \alpha$. Lemma 3.7.3 and 3.7.4 together imply, via definition 3.2.8, that $\emptyset \|\!\approx^{F \oplus^{\kappa} \Phi}_{Gr} \phi$ iff $\{\alpha\} \|\!\approx^{F \oplus^{\kappa} \Phi}_{Gr} \phi$. Lemma 3.7.16 implies that $\Phi \|\!\approx^{F}_{Gr} \phi$ iff $\Phi \cup \{\alpha\} \|\!\approx^{F}_{Gr} \phi$. The only-if direction implies that $\|\!\approx^{F}_{Gr}$ satisfies Cautious Monotony. (The if direction implies that $\|\!\approx^{F}_{Gr}$ satisfies Cut.) □

Theorem 3.3.3. *For all $F \in \mathcal{F}$, $\|\!\approx^{F}_{Gr}$ satisfies Cut.*

Proof. See proof of theorem 3.3.1. □

For the proof of theorem 3.3.2 and 3.3.4 we use the following lemmas.

Lemma 3.7.5. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. If for all $L \in \mathcal{L}_{Co}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{Co}(F, \{\mathbf{out}(x)\}) \downarrow A = \mathcal{L}_{Co}(F)$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that for all $L \in \mathcal{L}_{Co}(F)$, $L(x) = \mathbf{out}$. We prove that $\mathcal{L}_{Co}(F, \{\mathbf{out}(x)\}) \downarrow A = \mathcal{L}_{Co}(F)$. The $\supseteq$ direction is trivial. We prove the $\subseteq$ direction. Suppose $L \in \mathcal{L}_{Co}(F, \{\mathbf{out}(x)\}) \downarrow A$. We immediately have that every $y \in A \setminus \{x\}$ is legally labelled in L w.r.t. F. We prove that x, too, is legally labelled in L w.r.t. F. Let $\{L'\} = \mathcal{L}_{Gr}(F)$ and $\{L''\} = \mathcal{L}_{Gr}(F, \{\mathbf{out}(x)\})$. Because $L' \in \mathcal{L}_{Co}(F)$ we have $L' \models \mathbf{out}(x)$. Lemma 3.7.3 then implies that $L'' \downarrow A = L'$. Thus, for some $y \in A$ s.t. $y \rightsquigarrow x$, we have $L''(y) = \mathbf{in}$ and hence $L(y) = \mathbf{in}$. This implies that x is legally **out** in L w.r.t. F, and hence that $L \in \mathcal{L}_{Co}(F)$. □

To prove lemma 3.7.7 below we use the following definition and proposition.

Definition 3.7.1. [29] Let $F = (A, \rightsquigarrow)$ be an argumentation framework. The *comittedness* relation $\sqsubseteq \subseteq \mathcal{L}(F) \times \mathcal{L}(F)$ is defined by by $L \sqsubseteq L'$ iff $L^{-1}(\mathbf{in}) \subseteq L'^{-1}(\mathbf{in})$ and $L^{-1}(\mathbf{out}) \subseteq L'^{-1}(\mathbf{out})$.

Proposition 3.7.6. *[29, Theorem 11] Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $L \in \mathcal{L}(F)$ be an admissible labelling of F. There exists a labelling $L' \in \mathcal{L}_F(Co)$ such that $L \sqsubseteq L'$.*

Lemma 3.7.7. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. If for all $L \in \mathcal{L}_{Co}(F)$, $L(x) \neq \mathbf{in}$ then $\mathcal{L}_{Co}(F, \{\neg\mathbf{in}(x)\}) \downarrow A = \mathcal{L}_{Co}(F)$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that for all $L \in \mathcal{L}_{Co}(F)$, $L(x) \neq \mathbf{in}$. We prove that $\mathcal{L}_{Co}(F, \{\neg\mathbf{in}(x)\}) \downarrow A = \mathcal{L}_{Co}(F)$. The $\supseteq$ direction is trivial. We prove the $\subseteq$ direction. Suppose $L \in \mathcal{L}_{Co}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. Assume towards a contradiction that $L \notin \mathcal{L}_{Co}(F)$. It then holds that x is illegally **und** in L w.r.t. F, because all other arguments are legally labelled in L w.r.t. F. More precisely we have $L(y) = \mathbf{out}$ for all $y \in A$ s.t. $y \rightsquigarrow x$. But then we still have that L is an admissible labelling of F. Proposition 3.7.6 then implies that there is an $L' \in \mathcal{L}_F(Co)$ such that $L \sqsubseteq L'$. But we then have $L'(x) = \mathbf{in}$, which contradicts our assumption. Hence $L \in \mathcal{L}_{Co}(F)$. □

We are now ready to prove theorem 3.3.2 and 3.3.4.

Theorem 3.3.2. *For all $F \in \mathcal{F}$, $\|\!\approx^F_{Co}$ satisfies Cautious Monotony.*

Proof. The proof is similar to the proof of theorem 3.3.1 (using lemma 3.7.5 and 3.7.7). □

Theorem 3.3.4. *For all $F \in \mathcal{F}$, $\|\!\approx^F_{Co}$ satisfies Cut.*

Proof. The proof is similar to the proof of theorem 3.3.1 (using lemma 3.7.5 and 3.7.7). □

For the proof of theorem 3.3.5 we use the following lemmas.

Lemma 3.7.8. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. If for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{Pr}(F) \subseteq \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) = \mathbf{out}$. We prove $\mathcal{L}_{Pr}(F) \subseteq \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$. Suppose that $L \in \mathcal{L}_{Pr}(F)$ and assume towards a contradiction that $L \notin \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$. Because $L \in \mathcal{L}_{Co}(F)$ and $L \models \mathbf{out}(x)$ it follows that $L \in \mathcal{L}_{Co}(F, \{\mathbf{out}(x)\}) \downarrow A$. Thus there must be an $L' \in \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$ such that $L^{-1}(\mathbf{in}) \subset L'^{-1}(\mathbf{in})$. From this it follows that there is some $y \in A$ such that $y \rightsquigarrow x$ and $L'(y) = \mathbf{in}$. This means that all members of A are legally labelled in L' w.r.t. F, and hence that $L' \in \mathcal{L}_{Co}(F)$. But then we have $L \notin \mathcal{L}_{Pr}(F)$, which is a contradiction. Hence $L \in \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$. □

Lemma 3.7.9. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. If for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) \neq \mathbf{in}$ then $\mathcal{L}_{Pr}(F) \subseteq \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) \neq \mathbf{in}$. We prove that $\mathcal{L}_{Pr}(F) \subseteq \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. Suppose $L \in \mathcal{L}_{Pr}(F)$. If $L(x) = \mathbf{out}$ it follows by argument similar to the one used in the proof of lemma 3.7.8 that $L \in \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. In the remainder we assume $L(x) = \mathbf{und}$. Assume towards a contradiction that $L \notin \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. Because $L \in \mathcal{L}_{Co}(F)$ and $L \models \mathbf{und}(x)$ it follows that $L \in \mathcal{L}_{Co}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. Thus there must be an $L' \in \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$ such that $L^{-1}(\mathbf{in}) \subset L'^{-1}(\mathbf{in})$. We know that $L'(x) \neq \mathbf{in}$, Thus, there are two cases left:

1. $L'(x) = \mathbf{out}$. Because we have $L^{-1}(\mathbf{in}) \subset L'^{-1}(\mathbf{in})$, there is an $y \in A$ such that $y \rightsquigarrow x$ and $L'(y) = \mathbf{in}$. This means that all members of A are legally labelled in L' w.r.t. F, and hence that $L' \in \mathcal{L}_{Co}(F)$. But then we have $L \notin \mathcal{L}_{Pr}(F)$, which is a contradiction. Thus, this case is impossible.

2. $L'(x) = \mathbf{und}$. Two sub-cases:

   (a) x is legally **und** in L' w.r.t. F. It then follows that $L' \in \mathcal{L}_{Co}(F)$. But then we have $L \notin \mathcal{L}_{Pr}(F)$, which is a contradiction. Thus, this case is impossible.

   (b) x is illegally **und** in L' w.r.t. F. It then holds that $L' \in \mathcal{L}_{Ad}(F)$. Proposition 3.7.6 then implies that there is an $L' \in \mathcal{L}_F(Co)$ such that $L' \sqsubseteq L''$. But we then have $L''(x) = \mathbf{in}$, and hence there is an $L''' \in \mathcal{L}_F(Pr)$ such that $L'''(x) = \mathbf{in}$, which contradicts our assumption. Thus, this case is impossible.

Thus we have that $L \in \mathcal{L}_{Pr}(F, \{\neg\mathbf{in}(x)\}) \downarrow A$. □

Theorem 3.3.5. *For all $F \in \mathcal{F}$, $\models^F_{Pr}$ satisfies Cut.*

Proof. The proof is similar to the proof of theorem 3.3.1 (using lemma 3.7.8 and 3.7.9). □

For the proof of theorem 3.3.6 we use the following lemma.

Lemma 3.7.10. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. If for all $L \in \mathcal{L}_{St}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{St}(F) \subseteq \mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) \downarrow A$.*

Proof. Follows easily. □

Theorem 3.3.6. *For all $F \in \mathcal{F}$, $\models^F_{St}$ satisfies Stable Cut.*

Proof. Similar to the proof of theorem 3.3.1 (using lemma 3.7.10). □

For the proof of theorem 3.3.11 we use the following lemmas.

Lemma 3.7.11. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. Assume that F contains no odd cycles. If for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{Pr}(F)$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that F contains no odd cycles. Assume that for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) = \mathbf{out}$. We prove that $\mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{Pr}(F)$. Let κ be an F-mapping and let $L \in \mathcal{L}_{Pr}(F \oplus^{\kappa} \{\mathbf{out}(x)\})$. Assume towards contradiction that $L \downarrow A \notin \mathcal{L}_{Pr}(F)$.

We show that this implies that there is an $L' \in \mathcal{L}_{Pr}(F)$ such that $L'(x) = \mathbf{in}$. We start by constructing an admissible set E of F. Let κ be an F-mapping and let $F' = (A', \rightsquigarrow') = F \oplus^{\kappa} \{\mathbf{out}(x)\}$. We now define E by

$$E = \{y \in A \mid L(y) = \mathbf{in} \text{ and there is no odd path in } F' \text{ from } x \text{ to } y\}.$$

We first prove that E is an admissible set of F. It is easy to see that E is conflict-free with respect to F. What remains is to show that $E \subseteq \mathcal{D}_F(E)$. Let $y \in E$ and $z \in A$ be an argument such that $z \rightsquigarrow y$. Then $L(y) = \mathbf{in}$ and $z \rightsquigarrow' y$ and hence admissibility of L w.r.t. F' implies that there is a $z' \in L^{-1}(\mathbf{in})$ s.t. $z' \rightsquigarrow' z$. We prove that (1) $z' \rightsquigarrow z$ and (2) $z' \in E$.

1. Because there is no odd path in F' from x to y and because $z \rightsquigarrow' y$ it follows that $z \neq x$ and hence $z' \in A$. This implies $z' \rightsquigarrow z$.
2. Because $y \in E$, there is no odd path in F' from x to y and hence no odd path in F' from x to z'. Furthermore we have $z' \in A$ (see (1)) and hence $z' \in E$.

Hence $E \subseteq \mathcal{D}_F(E)$, meaning that E is an admissible set of F.

We now prove that for all $y \in A$ such that $y \rightsquigarrow x$, there is a $z \in E$ s.t. $z \rightsquigarrow y$. Suppose $y \in A$ and $y \rightsquigarrow x$. Because F and F' contain no odd cycles we have $\mathcal{L}_{Pr}(F) = \mathcal{L}_{St}(F)$ and $\mathcal{L}_{Pr}(F') = \mathcal{L}_{St}(F')$ (proposition 3.3.10). Hence we have $L \in \mathcal{L}_{St}(F')$ and $L \downarrow A \notin \mathcal{L}_{St}(F)$. This implies that x is illegally **out** in $L \downarrow A$ w.r.t. F. Hence $L(y) = \mathbf{out}$, and hence there is a $z \in E'$ s.t. $z \rightsquigarrow' y$. We prove that (1) $z \in A$ and (2) there is no odd path in F' from x to z.

1. If $z \notin A$ then $x = y$, but we then have $x \rightsquigarrow x$, which is a contradiction because F contains no odd cycles. hence $z \in A$.
2. If there is an odd path in F' from x to z then there is an odd path in F' from x to x, because $z \rightsquigarrow' y \rightsquigarrow' x$. But F' contains no odd cycles, and hence there is no odd path in F' from x to z.

From this it follows that $z \in E$. Hence for all $y \in A$ such that $y \rightsquigarrow x$, there is a $z \in E$ s.t. $z \rightsquigarrow y$. This in turn implies that there is a preferred extension E' of F such that $E \subseteq E'$ and $x \in E'$. Hence there is a preferred labelling $L' \in \mathcal{L}_{Pr}(F)$ such that $L'(x) = \mathbf{in}$. This is a contradiction. Hence $L \downarrow A \in \mathcal{L}_{Pr}(F)$. □

Lemma 3.7.12. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. Assume that F contains no odd cycles. If for all $L \in \mathcal{L}_{SS}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{SS}(F)$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that F contains no odd cycles. Assume that for all $L \in \mathcal{L}_{SS}(F)$, $L(x) = \mathbf{out}$. We prove that $\mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{SS}(F)$. Let $L \in \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A$. Because every semistable labelling is also preferred, it follows that $L \in$

$\mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A$. Because F contains no odd cycles, proposition 3.3.10 implies $\mathcal{L}_{SS}(F) = \mathcal{L}_{Pr}(F)$. Hence for all $L' \in \mathcal{L}_{Pr}(F)$, $L'(x) = \mathbf{out}$. Lemma 3.7.11 then implies that $L \in \mathcal{L}_{Pr}(F)$. Applying proposition 3.3.10 again yields $L \in \mathcal{L}_{SS}(F)$. □

Lemma 3.7.13. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. Assume that F contains no odd cycles. If for all $L \in \mathcal{L}_{St}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{St}(F)$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that F contains no odd cycles. Assume that for all $L \in \mathcal{L}_{St}(F)$, $L(x) = \mathbf{out}$. Because F contains no odd cycles, proposition 3.3.10 implies $\mathcal{L}_{St}(F) = \mathcal{L}_{Pr}(F)$. Hence for all $L \in \mathcal{L}_{Pr}(F)$, $L(x) = \mathbf{out}$. Lemma 3.7.11 then implies that $\mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{Pr}(F)$. Furthermore, $F \oplus^{\kappa} \mathbf{out}(x)$ contains no odd cycles, thus proposition 3.3.10 implies $\mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) = \mathcal{L}_{Pr}(F, \{\mathbf{out}(x)\})$. It follows that $\mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) \downarrow A \subseteq \mathcal{L}_{St}(F)$. □

We are now ready to prove theorem 3.3.11.

Theorem 3.3.11. *For all $F \in \mathcal{F}$, if F is odd-cycle-free then $\|\!\approx^F_{Pr}$, $\|\!\approx^F_{SS}$ and $\|\!\approx^F_{St}$ satisfy Stable Cautious Monotony.*

Proof. We prove the preferred case. Let $F \in \mathcal{F}$, $\Phi \in \mathtt{StInt}(F)$ and let κ be an F-mapping. Assume that F is odd-cycle-free. Suppose $\Phi \|\!\approx^F_{Pr} \mathbf{out}(x)$. Lemma 3.7.16 implies that $\emptyset \|\!\approx^{F \oplus^{\kappa} \Phi}_{Pr} \mathbf{out}(x)$. Because Φ is stable, $F \oplus^{\kappa} \Phi$ is also odd-cycle-free, and hence lemma 3.7.11 implies that if $\emptyset \|\!\approx^{F \oplus^{\kappa} \Phi}_{Pr} \phi$ then $\{\mathbf{out}(x)\} \|\!\approx^{F \oplus^{\kappa} \Phi}_{Pr} \phi$. By applying lemma 3.7.16 again we get that if $\Phi \|\!\approx^F_{Pr} \phi$ then $\Phi \cup \{\mathbf{out}(x)\} \|\!\approx^F_{Pr} \phi$. Hence $\|\!\approx^F_{Pr}$ satisfies Stable Cautious Monotony. The semi-stable and stable case follow similarly, using lemma 3.7.12 for the semi-stable case and lemma 3.7.13 for the stable case. □

For the proof of theorem 3.3.12 we use the following lemma.

Lemma 3.7.14. *Let $F = (A, \rightsquigarrow)$ and $x \in A$. Assume that F is odd-cycle-free. If for all $L \in \mathcal{L}_{SS}(F)$, $L(x) = \mathbf{out}$ then $\mathcal{L}_{SS}(F) \subseteq \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A$.*

Proof. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $x \in A$. Assume that F is odd-cycle-free. Assume that for all $L \in \mathcal{L}_{SS}(F)$, $L(x) = \mathbf{out}$. We prove that $\mathcal{L}_{SS}(F) \subseteq \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A$. Let $L \in \mathcal{L}_{SS}(F)$. Because $L(x) = \mathbf{out}$ we immediately have that $L \in \mathcal{L}_{Co}(F, \{\mathbf{out}(x)\}) \downarrow A$. We now show that $L \in \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A$. Because F contains no odd cycles, proposition 3.3.10 implies $\mathcal{L}_{SS}(F) = \mathcal{L}_{St}(F)$ and hence for all $y \in A$, $L(y) \neq \mathbf{und}$. This implies that $L \in \mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) \downarrow A$ and, because $\mathcal{L}_{St}(F, \{\mathbf{out}(x)\}) \subseteq \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\})$, $L \in \mathcal{L}_{SS}(F, \{\mathbf{out}(x)\}) \downarrow A$. □

We are now ready to prove theorem 3.3.12.

Theorem 3.3.12. *For all $F \in \mathcal{F}$, if F is odd-cycle-free then $\|\!\approx^F_{SS}$ satisfies Stable Cut.*

Proof. Similar to the proof of theorem 3.3.11 (using lemma 3.7.14). □

For the proof of theorem 3.4.3 we use lemma 3.7.15 and 3.7.16.

Lemma 3.7.15. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework. If σ satisfies the directionality principle then $\Phi \not\rightsquigarrow^* \phi$ implies ($\Phi \mid\!\!\approx^F_\sigma \phi$ iff $\emptyset \mid\!\!\approx^F_\sigma \phi$).*

Proof. Suppose σ satisfies the directionality principle. Proposition 3.4.1 implies that for all $F \in \mathcal{F}$,

$$\forall B \in U(F), \phi \in \mathtt{lang}(B), (F \downarrow B) \models_\sigma \phi \text{ iff } F \models_\sigma \phi. \tag{3.1}$$

Let $F = (A, \rightsquigarrow)$ be an argumentation framework, $\Phi \in \mathtt{Int}(F)$ be an intervention, and $\phi \in \mathtt{lang}(F)$ a formula such that $\Phi \not\rightsquigarrow^* \phi$. Let κ be an F-mapping. Define B by $B = \{x \in A \mid x \rightsquigarrow^* \phi\}$. We then have $\phi \in \mathtt{lang}(B)$ and $B \in U(F)$ and hence $B \in U(F \oplus^\kappa \Phi)$. We now prove that $\emptyset \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^F_\sigma \phi$: From proposition 3.2.4 it follows that $\emptyset \mid\!\!\approx^F_\sigma \phi$ iff $F \models_\sigma \phi$. From 3.1, together with the fact that $B \in U(F)$ and $\phi \in \mathtt{lang}(B)$, it follows that $F \models_\sigma \phi$ iff $(F \downarrow B) \models_\sigma \phi$. Because $(F \downarrow B) = ((F \oplus^\kappa \Phi) \downarrow B)$ it follows that $(F \downarrow B) \models_\sigma \phi$ iff $((F \oplus^\kappa \Phi) \downarrow B) \models_\sigma \phi$. From 3.1 together with the fact that $B \in U(F \oplus^\kappa \Phi)$ and $\phi \in \mathtt{lang}(B)$, it follows that $((F \oplus^\kappa \Phi) \downarrow B) \models_\sigma \phi$ iff $(F \oplus^\kappa \Phi) \models_\sigma \phi$. From proposition 3.2.4 it finally follows that $(F \oplus^\kappa \Phi) \models_\sigma \phi$ iff $\Phi \mid\!\!\approx^F_\sigma \phi$. Hence we have $\emptyset \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^F_\sigma \phi$. □

Lemma 3.7.16. *Let $\sigma \in \{Gr, Co, Pr, SS, St\}$, $F \in \mathcal{F}$, $\Phi, \Psi \in \mathtt{Int}(F)$ and let κ be an F-mapping. It holds that $\Phi \cup \Psi \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^{(F \oplus^\kappa \Psi)}_\sigma \phi$.*

Proof. Let $\sigma \in \{Gr, Co, Pr, SS, St\}$, $F = (A, \rightsquigarrow)$ an argumentation framework, $\Phi, \Psi \in \mathtt{Int}(F)$ and let κ be an F-mapping. Because $F \oplus^\kappa (\Phi \cup \Psi)$ and $(F \oplus^\kappa \Phi) \oplus^{\kappa'} \Psi)$ are isomorphic and σ statisfies language independence, we have $\mathcal{L}_\sigma(F \oplus^\kappa (\Phi \cup \Psi)) \downarrow A = \mathcal{L}_\sigma((F \oplus^\kappa \Phi) \oplus^{\kappa'} \Psi)) \downarrow A$. Via definition 3.2.8 it follows that $\Phi \cup \Psi \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^{(F \oplus^\kappa \Psi)}_\sigma \phi$. □

We are now ready to prove theorem 3.4.3.

Theorem 3.4.3. *If σ satisfies the directionality principle then for all $F \in \mathcal{F}$, $\mid\!\!\approx^F_\sigma$ satisfies Conditional Directionality.*

Proof of theorem 3.4.3. Suppose σ satisfies the directionality principle. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. Let $\Phi, \Psi \in \mathtt{Int}(F)$ and $\phi \in \mathtt{lang}(F)$. Suppose $\Psi \not\rightsquigarrow^* \phi$. We need to prove that $\Phi \cup \Psi \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^{F'}_\sigma \phi$. Let κ be an F-mapping and let $F' = (A', \rightsquigarrow') = F \oplus^\kappa \Phi$. It then follows that $\Psi \not\rightsquigarrow^{*\,\prime} \phi$ and hence (via lemma 3.7.15) $\Psi \mid\!\!\approx^{F'}_\sigma \phi$ iff $\emptyset \mid\!\!\approx^{F'}_\sigma \phi$. Lemma 3.7.16 then implies that $\Phi \cup \Psi \mid\!\!\approx^F_\sigma \phi$ iff $\Phi \mid\!\!\approx^{F'}_\sigma \phi$. □

Theorem 3.4.7. *If σ satisfies the noninterference principle then for all $F \in \mathcal{F}$, $\mid\!\!\approx^F_\sigma$ satisfies Conditional Noninterference.*

Proof. The proof is exactly the same as the proof of theorem 3.4.3, except that every occurrence of $U(F)$ (i.e., the unattacked sets of F) is replaced with $I(F)$ (i.e., the isolated sets of F). □

Chapter 4

Observation-Based Entailment in Argumentation

4.1 Introduction

In this section we focus on change in argumentation due to observations. What we mean by an observation is some piece of information from the environment that requires the revision of the status of one or more arguments. The question we address is: what is a rational way for an agent to revise the evaluation of an argumentation framework to account for an observation?

Let us look at an example. Let F be the argumentation framework shown in figure 4.1. The nodes are coloured according to the unique complete (and grounded, preferred, semi-stable and stable) labelling of this argumentation framework. Even though e is accepted, an agent may, through observation, want to revise the status of e to rejected. The rejection of e can be achieved in various ways by changing F. Three of them are shown in figure 4.2: adding an

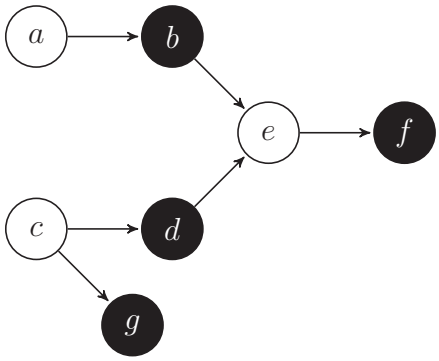


Figure 4.1: An Argumentation Framework.

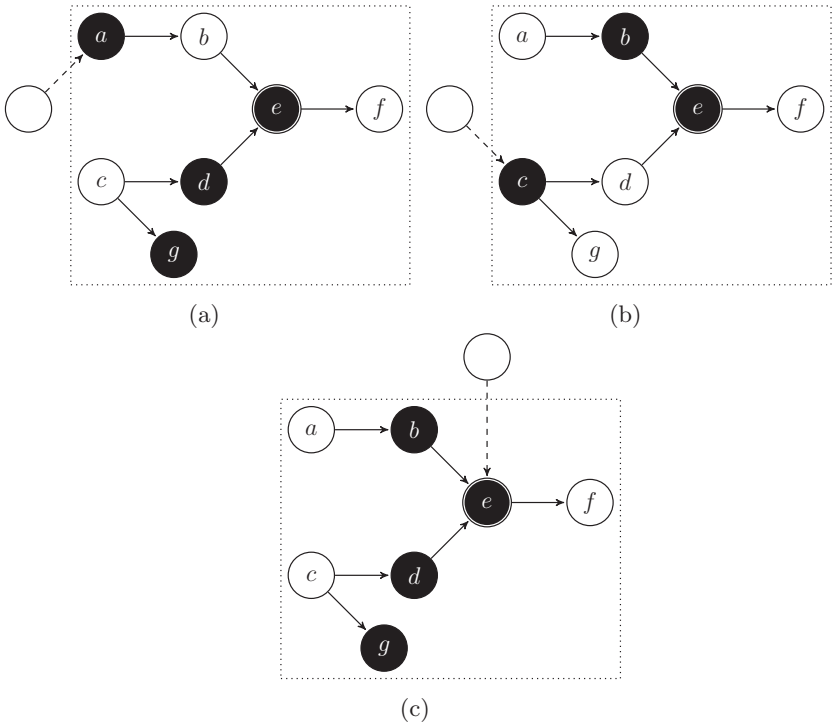


Figure 4.2: Three ways to account for the rejection of e.

attacker to a, c and e all account for the rejection of e. This example clearly demonstrates the difference between intervention and observation: while the *action* of defeating e only affects the label of e and f, the *observed* rejection of e may (depending on how it is accounted for) also affect the labels of the other arguments.

This example shows that what we are dealing with is a kind of goal-oriented change of the argumentation framework. That is, the goal is to revise the evaluation of the argumentation framework to satisfy a constraint (e.g., rejection of e) and this goal can be achieved by changing the argumentation framework (e.g., by attacking a, c or e). Goal-oriented change in argumentation is usually studied from a *multi-agent strategic perspective*. From the multi-agent strategic perspective, the revision of the status of an argument is due not to information from the environment, but represents the goal of an agent in a debate. For example, Kontarinis et al. [61] put this problem as follows: "When several agents are engaged in an argumentation process, they are faced with the problem of deciding how to contribute to the current state of the debate in order to satisfy their own goal, i.e., to make an argument under a given semantics accepted or not." Similar motivations are found in Baumann and Brewka's work on what they call the enforcement problem [11, 12] as well as other work in this direction [17, 20]. Considerations that dominate in the multi-agent strategic perspective are mostly of procedural and economical nature. These works fo-

cus, for example, on how to determine which arguments to attack to satisfy a given goal, or what the minimal contributions to a debate are to achieve this. By contrast, we take a *single-agent revision perspective*: the argumentation framework represents the reasoning of a single agent, and the revision of its evaluation is due to new information that the agent receives from the environment. This leads to the following two questions:

1. By what mechanism does a rational agent decide how to change his argumentation framework, in order to revise its evaluation due to an observation?
2. What are the conditions that characterize a rational way to revise the evaluation of an argumentation framework due to an observation?

We address the first question by proposing a model for how a rational agent performs revision, based on an abductive principle. Different hypotheses (which are changes to an argumentation framework) are considered. A hypothesis is an explanation for an observation if it accounts for its truth. An agent accommodates an observation by finding the most plausible explanations for the observation. We simplify this model by abstracting away from the changes that can be made to an argumentation framework, and instead we focus on change represented by interventions. Thus, roughly speaking, an abductive model for a given argumentation framework consists of a set of interventions along with a preference relation encoding their relative plausibility.

Analogous to the notion of intervention-based entailment, we introduce the notion of observation-based entailment. An observation-based entailment for a given argumentation framework F is a relation between observations about the status of the arguments in F and consequences of observations. An observation-based entailment relation is determined by an abductive model by letting something be a consequence of an observation whenever it is a consequence of the most preferred explanations for the observation. This scheme can be applied credulously (i.e., by focusing on explanations for credulous truth of the observation) and sceptically (i.e., by focusing on explanations for sceptical truth of the observation).

We address the second question by showing that the class of observation-based entailment relations for a given argumentation framework F (i.e., the observation-based entailment relations determined by some abductive model for F) is characterized by a strengthening of the class of preferential (in the credulous case) and loop-cumulative (in the sceptical case) entailment relations. This characterization is complete in the credulous case, but not in the sceptical case. This result gives a handle on how credulous and sceptical observation-based entailment differ: while credulous entailment satisfies the Or rule, sceptical entailment generally does not.

Having introduced the notion of observation-based entailment, we are in a position to compare it with intervention-based entailment. We investigate the main difference between these two types of entailment, which is related to how the effects of interventions and observations propagate through an argumentation framework. The results we obtain also demonstrate the role of the principles of directionality and noninterference in observation-based entailment. Namely,

these principles ensure (if we make some reasonable further assumptions) that the effect of an observation propagates through the argumentation framework in a well-behaved manner.

The overview of this chapter is as follows. In section 4.2 we present the basic definitions of the notion of an abductive model, which represents a mechanism by which a rational agent decides how to change his argumentation framework, in order to revise its evaluation due to an observation. This leads to the notion of credulous and sceptical observation-based entailment. In section 4.3 we show that credulous (resp. sceptical) observation-based entailment relations are characterized by a strengthening of the class of preferential (in the credulous case) and loop-cumulative (in the sceptical case) entailment relations. In section 4.4 and 4.5 we investigate the role of directionality and noninterference in the behaviour of observation-based entailment and the difference between intervention and observation in terms of how the effects of interventions and observations propagate through an argumentation framework. We discuss related work in section 4.6 and we conclude and discuss some directions for future work in section 4.7.

4.2 Observation-Based Entailment

In this section we present a model of how an agent revises the evaluation of an argumentation framework due to observations. Intuitively, the idea is to regard an observation as something that can be explained within the agent's *abductive model*. The basic definitions concerning abductive models will be presented in section 4.2.1. An abductive model determines an entailment relation by letting ψ be a consequence of ϕ if all most preferred explanations for ϕ also entail ψ. Thus, intuitively, the most preferred explanations for ϕ are used to predict the effect of observing ϕ on the overall evaluation of the argumentation framework. We present the definitions for this scheme in section 4.2.2 for the credulous case and in section 4.2.3 for the sceptical case.

4.2.1 Abductive Models

Let us first describe abductive models informally. Given an argumentation framework F and semantics σ, an abductive model consists essentially of a set of *hypotheses* (each of which maps to some intervention for F) and a preference relation over hypotheses. This relation represents the relative plausibility that the agent attributes to the different hypotheses. For example, given two arguments a and b, the agent may deem it more plausible that a is defeated than that b is defeated. To keep things simple, we focus from here on only on semantics under which the existence of labellings is guaranteed. This means that we restrict our attention to the complete, grounded, preferred and semi-stable semantics, and leave the stable semantics out of consideration.

Definition 4.2.1. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. An *abductive model based on F under semantics σ* is a tuple $M = (F, H, m, <, \sigma)$ where:

- H is a finite set containing elements called *hypotheses*,
- $<$ is a strict partial order over H,
- $m : H \to \mathtt{Int}(F)$ is a function mapping hypotheses to interventions,
- $\sigma \in \{Co, Gr, Pr, SS\}$.

A relation $<$ represents the agent's preference among hypotheses. To be consistent with the KLM framework, we associate preference with minimality. Thus, $h < h'$ holds whenever h is preferred over h'.

Note that it suffices in most practical cases to identify hypotheses directly with interventions, meaning that H is a subset of $\mathtt{Int}(F)$ and m is the identity function. The more general setting is, however, necessary to prove the characterization result in section 4.3.

We consider two properties that, as we show later, result in good behaviour. The first is *closure under weakening*. It states that, if an intervention Φ is considered possible then every subset of Φ is also considered possible. Intuitively, it represents a kind of independence between different arguments. It means that, for example, if the agent considers the intervention $\{\mathbf{out}(x), \mathbf{out}(y)\}$ possible, then he also considers $\{\mathbf{out}(x)\}$ and $\{\mathbf{out}(y)\}$ individually possible, as well as the vacuous intervention $\emptyset$.

Definition 4.2.2. An abductive model $M = (F, H, m, <, \sigma)$ is *closed under weakening* if and only if

$$\text{for all } h \in H \text{ and } \Phi \subseteq m(h), \text{ there is a } h' \in H \text{ such that } m(h') = \Phi.$$

The second property is the *minimality assumption*. It states that interventions that are strictly logically weaker are always preferred by the agent. Intuitively, it reflects an assumption of minimal explanation: logically weaker hypotheses are always more plausible.

Definition 4.2.3. An abductive model $M = (F, H, m, <, \sigma)$ satisfies the *minimality assumption* if and only if

$$\text{for all } h, h' \in H, \text{ if } m(h) \models m(h') \text{ and } m(h') \not\models m(h) \text{ then } m(h') < m(h).$$

Let us illustrate these definitions with an example.

Example 4.2.1. *Let F be the argumentation framework shown in figure 4.1. Let $M = (F, H, m, <, Pr)$ be the abductive model based on F under the preferred semantics, defined by:*

- $H = \{\emptyset, \{\mathbf{out}(c)\}, \{\mathbf{out}(a)\}, \{\mathbf{out}(e)\}\}$.
- $\emptyset < \{\mathbf{out}(c)\} < \{\mathbf{out}(a)\} < \{\mathbf{out}(e)\}$.
- *m is the identity function.*

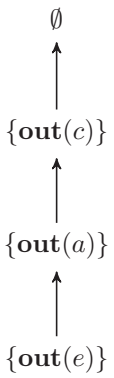


Figure 4.3: An example of an abductive model for the AF shown in figure 4.1.

This abductive model is shown in figure 4.3. All hypotheses are depicted and an arrow from a hypothesis h to a hypothesis h' means that $h' < h$ (transitive arrows are omitted). Thus, the non-vacuous hypotheses considered here are the defeat of a, c and e. The vacuous intervention is the most preferred one. Defeat of c is preferred over defeat of a and defeat of a is preferred over defeat of e. This abductive model is closed under weakening because every subset of every intervention is also an intervention. It also satisfies the minimality assumption: $\emptyset$ is logically weaker than all other interventions and is also the most preferred, while all the other interventions are incomparable with respect to logical strength.

In the next two sections we define the notion of a credulous and a sceptical observation-based entailment relation. Each abductive model determines a credulous observation-based entailment relation and a sceptical observation-based entailment relation. This means that, unlike in the setting of intervention-based entailment, there is no unique definition of a credulous or sceptical observation-based entailment relation for a given argumentation framework and semantics. Instead, we work with classes of credulous and sceptical observation-based entailment relations, each defined by some abductive model for a given argumentation framework. This set-up is similar to that of most theories of belief revision, where a revision operator is determined by an epistemic state, which is a representation of the beliefs and revision strategy of an agent.

4.2.2 Credulous Observation-Based Entailment

We focus first on explanations for observations that must become *credulous* consequences: if M is based on F under the semantics σ then a hypothesis h is a *credulous* explanation for ϕ if ϕ is a credulous consequence of the intervention $m(h)$ under semantics σ. We furthermore say that h is a most preferred credulous explanation if there is no h' that is a credulous explanation for ϕ and $h' < h$.

Definition 4.2.4. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $M = (F, H, m, <, \sigma)$: Given a formula $\phi \in \mathtt{lang}(F)$ we say that a hypothesis $h \in H$ is a *credulous explanation for* ϕ if and only if $m(h) \not\models^F_\sigma \neg\phi$. We say that

h is a *most preferred credulous explanation for* ϕ if and only if h is a credulous explanation for ϕ and there is no credulous explanation h' for ϕ such that $h' < h$.

Finding explanations for an observation is not the main goal. Instead, the goal is to use the explanations for an observation to predict how the observation affects the overall evaluation of the argumentation framework. In the credulous case this amounts to the following: If, given an abductive model M, we want to know whether ψ is a consequence of the observation ϕ, we first determine the most preferred credulous explanations for ϕ. We then check, for every most preferred credulous explanation h of ϕ, whether $m(h) \mid\!\!\approx^F_\sigma \phi \rightarrow \psi$ holds (that is, whether the intervention $m(h)$ entails ψ, presupposing the truth of the observation ϕ). If it does, then ψ is a consequence of the observation ϕ. Using this scheme, every abductive model determines a *credulous observation-based entailment relation*:

Definition 4.2.5. Given an abductive model $M = (F, H, m, <, \sigma)$ the *credulous observation-based entailment relation defined by M* will be denoted by $\mid\!\sim^M_{Cr}$ and is defined by: $\phi \mid\!\sim^M_{Cr} \psi$ iff for every most preferred credulous explanation h for ϕ we have $m(h) \mid\!\!\approx^F_\sigma \phi \rightarrow \psi$.

Note that credulous observation-based entailment puts a relatively weak burden on what constitutes an explanation, because what is observed only needs to be credulously true given the explanation.

We now illustrate the definition of credulous observation-based entailment with an example.

Example 4.2.2. *Let M be the abductive model that we considered in example 4.2.1.*

- *Consider the observation* $\mathbf{in}(f)$*. The unique most preferred credulous explanation is* $\{\mathbf{out}(c)\}$*, because we have* $\{\mathbf{out}(c)\} \not\mid\!\!\approx^F_{Pr} \neg\mathbf{in}(f)$*. Thus, we also have acceptance of g, because we have* $\{\mathbf{out}(c)\} \mid\!\!\approx^F_{Pr} \mathbf{in}(f) \rightarrow \mathbf{in}(g)$*. On the other hand, we have rejection of b (*$\{\mathbf{out}(c)\} \mid\!\!\approx^F_{Pr} \mathbf{in}(f) \rightarrow \mathbf{out}(b)$*). Thus we have*

$$\mathbf{in}(f) \mid\!\sim^M_{Cr} \mathbf{in}(g) \text{ and } \mathbf{in}(f) \mid\!\sim^M_{Cr} \mathbf{out}(b).$$

- *Consider the observation* $\mathbf{in}(b)$*. The unique most preferred credulous explanation is* $\{\mathbf{out}(a)\}$*, because we have* $\{\mathbf{out}(a)\} \not\mid\!\!\approx^F_{Pr} \neg\mathbf{in}(b)$*. This implies, e.g., rejection of e, because we have* $\{\mathbf{out}(a)\} \mid\!\!\approx^F_{Pr} \mathbf{in}(b) \rightarrow \mathbf{out}(e)$*. It also implies rejection of g, because we have* $\{\mathbf{out}(a)\} \mid\!\!\approx^F_{Pr} \mathbf{in}(b) \rightarrow \mathbf{out}(g)$*. Then*

$$\mathbf{in}(b) \mid\!\sim^M_{Cr} \mathbf{out}(e) \text{ and } \mathbf{in}(b) \mid\!\sim^M_{Cr} \mathbf{out}(g).$$

The following example involves an argumentation framework with (potentially) multiple labellings.

Example 4.2.3. *Let F be the argumentation framework shown in figure 4.4. Let $M = (F, H, m, <, Pr)$ be the abductive model based on F under the preferred semantics, defined by:*

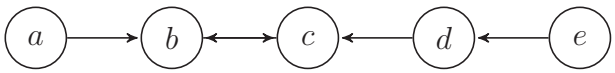


Figure 4.4: An Argumentation Framework.

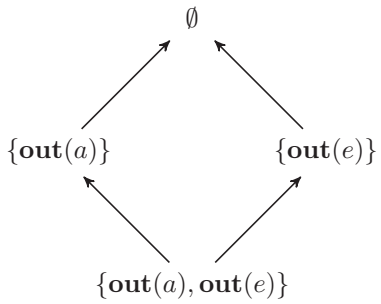


Figure 4.5: An example of an abductive model for the AF shown in figure 4.4.

- $H = \{\emptyset, \{\mathbf{out}(a)\}, \{\mathbf{out}(e)\}, \{\mathbf{out}(a), \mathbf{out}(e)\}\}$.
- $\emptyset < \{\mathbf{out}(a)\}$, $\emptyset < \{\mathbf{out}(e)\}$, $\{\mathbf{out}(a)\} < \{\mathbf{out}(a), \mathbf{out}(e)\}$ *and* $\{\mathbf{out}(e)\} < \{\mathbf{out}(a), \mathbf{out}(e)\}$.
- m *is the identity function.*

This abductive model is shown in figure 4.5. Note that this abductive model is both closed under weakening and satisfies the minimality assumption. We have that defeat of a is enough to make b credulously accepted under the preferred semantics, because we have $\{\mathbf{out}(a)\} \not\models^F_{Pr} \neg\mathbf{in}(b)$. *We do not, however, have* $\{\mathbf{out}(e)\} \not\models^F_{Pr} \neg\mathbf{in}(b)$. *Hence* $\{\mathbf{out}(a)\}$ *is the unique preferred credulous explanation for the observation* $\mathbf{in}(b)$. *Thus we have* $\mathbf{in}(b) \mid\sim^M_{Cr} \mathbf{out}(a)$ *and* $\mathbf{in}(b) \mid\sim^M_{Cr} \mathbf{in}(e)$.

4.2.3 Sceptical Observation-Based Entailment

Apart from determining the *credulous* explanations for an observation, an abductive model allows us to determine *sceptical* explanations. Analogous to the notion of (most preferred) credulous explanation we define here the notion of (most preferred) sceptical explanation.

Definition 4.2.6. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $M = (F, H, m, <, \sigma)$: Given a formula $\phi \in \mathtt{lang}(F)$ we say that a hypothesis $h \in H$ is a *sceptical explanation for* ϕ if and only if $m(h) \mid\!\models^F_\sigma \phi$. We say that h is a *most preferred sceptical explanation for* ϕ if and only if h is a sceptical explanation for ϕ and there is no sceptical explanation h' for ϕ such that $h' < h$.

Every abductive model M determines a sceptical observation-based entailment relation $\mid\sim^M_{Sk}$ by setting $\phi \mid\sim^M_{Sk} \psi$ iff the most preferred sceptical explanations for ϕ sceptically entail ψ.

Definition 4.2.7. Given an abductive model $M = (F, H, m, <, \sigma)$ the *sceptical observation-based entailment relation defined by M* will be denoted by $\mid\!\sim^M_{Sk}$ and is defined by $\phi \mid\!\sim^M_{Sk} \psi$ iff for every most preferred sceptical explanation h for ϕ we have $m(h) \parallel\!\!\models^F_\sigma \psi$.

Thus, an abductive model determines two observation-based entailment relations, one credulous and one sceptical. Because there is generally no relation between sets of most preferred sceptical and most preferred credulous explanations for a given observation, these two entailment relations usually behave differently. The grounded case forms an exception:

Proposition 4.2.1. *For every abductive model M based on an argumentation framework F under the grounded semantics it holds that* $\mid\!\sim^M_{Cr} = \mid\!\sim^M_{Sk}$.

In the general case, credulous and sceptical observation-based entailment is different because in the sceptical case a higher burden is placed on what constitutes an explanation than in the credulous case. This is demonstrated by the following example.

Example 4.2.4. *Let F be the argumentation framework shown in figure 4.4. Let* $M = (F, H, m, <, Pr)$ *be the abductive model based on F used in example 4.2.3 (shown in figure 4.5). While defeat of a is enough to make b* credulously *accepted under the preferred semantics it is not enough to make b* sceptically *accepted under the preferred semantics. That is, we have* $\{\mathbf{out}(a)\} \not\parallel\!\!\models^F_{Pr} \neg\mathbf{in}(b)$ *but not* $\{\mathbf{out}(a)\} \parallel\!\!\models^F_{Pr} \mathbf{in}(b)$. *To make b sceptically accepted we must use the intervention* $\{\mathbf{out}(a), \mathbf{out}(e)\}$, *because we have* $\{\mathbf{out}(a), \mathbf{out}(e)\} \parallel\!\!\models^F_{Pr} \mathbf{in}(b)$ *but not* $\{\mathbf{out}(a)\} \parallel\!\!\models^F_{Pr} \mathbf{in}(b)$ *or* $\{\mathbf{out}(e)\} \parallel\!\!\models^F_{Pr} \mathbf{in}(b)$. *Hence* $\{\mathbf{out}(a), \mathbf{out}(e)\}$ *is the preferred sceptical explanation for the observation* $\mathbf{in}(b)$. *This demonstrates the difference between credulous and sceptical observation-based entailment: while on the one hand we have* $\mathbf{in}(b) \mid\!\sim^M_{Cr} \mathbf{in}(e)$, *we have on the other hand* $\mathbf{in}(b) \mid\!\sim^M_{Sk} \mathbf{out}(e)$.

In the following section we prove a result that allows us to make a more precise statement about the difference between credulous and sceptical observation-based entailment.

4.3 A Syntactic Characterization

In the previous section we presented a semantic or constructive definition of credulous and sceptical observation-based entailment. It turns out that, by strengthening the classes of preferential (in the credulous case) and loop-cumulative (in the sceptical case) entailment relations (defined in section 2.2) we obtain a syntactic characterization. More precisely, given an argumentation framework F, the class of credulous (resp. sceptical) observation-based entailment relations based on F is characterized by a restricted class of preferential (resp. loop-cumulative) entailment relations. This characterization is complete in the credulous case, but not in the sceptical case. The details are presented in the following two subsections.

4.3.1 The Credulous Case

Given an argumentation framework F, the class of credulous observation-based entailment relations based on F coincides with a restriction of the class of preferential entailment relations (i.e., the class of entailment relations satisfying Reflexivity, Left Logical Equivalence, Right Weakening, Cut, Cautious Monotony, Loop and Or) over $\texttt{lang}(F)$ that furthermore satisfy what we call the property of *conflict-freeness with respect to* F. The property of conflict-freeness with respect to F is defined by requiring that, for every formula that is not conflict-free with respect to F, the negation of this formula is a consequence of every premise.

Definition 4.3.1. Let F be an argumentation framework and $\mid\!\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$. We say that $\mid\!\sim$ is *conflict-free w.r.t.* F iff for all $\phi, \psi \in \texttt{lang}(F)$,

$$\text{if } \psi \text{ is not conflict-free w.r.t. } F \text{ then } \phi \mid\!\sim \neg\psi.$$

The following lemma establishes the correspondence between preferential entailment relations over $\texttt{lang}(F)$ that are conflict-free with respect to F and preferential models defined over conflict-free labellings of F.

Lemma 4.3.1. *Let F be an argumentation framework and let $\mid\!\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$. The following are equivalent:*

1. *$\mid\!\sim$ is preferential and conflict-free w.r.t. F.*
2. *$\mid\!\sim = \mid\!\sim^W$ for a preferential model W over $\mathcal{L}_{Cf}(F)$.*

Proof. Let F be an argumentation framework and let $\mid\!\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$.

(1 implies 2): Suppose $\mid\!\sim$ is preferential and conflict-free w.r.t. F. Theorem 2.2.1 implies that there is a preferential model $W = (S, \prec, l)$ over $\mathcal{L}(F)$ such that $\mid\!\sim = \mid\!\sim^W$. We prove that W is defined over $\mathcal{L}_{Cf}(F)$ (i.e. for all $s \in S$, $l(s) \in \mathcal{L}_{Cf}(F)$). Suppose the contrary: there is an $s \in S$, and $l(s) \notin \mathcal{L}_{Cf}(F)$. Then $Form(l(s))$ is not conflict-free w.r.t. F but we do not have have $Form(l(s)) \mid\!\sim^W \neg Form(l(s))$. This means that $\mid\!\sim$ is not conflict-free w.r.t. F, which is a contradiction.

(2 implies 1): Let $W = (S, \prec, l)$ be a preferential model over $\mathcal{L}_{Cf}(F)$. Theorem 2.2.1 implies that $\mid\!\sim^W$ is preferential. We prove that $\mid\!\sim^W$ is conflict-free w.r.t. F. Let $\phi, \psi \in \texttt{lang}(F)$ and assume that ψ is not conflict-free w.r.t. F. If there is no $s \in S$ such that $l(s) \models \phi$, it follows trivially that $\phi \mid\!\sim^W \neg\psi$ and we are done. Now let $s \in S$ be a state such that $l(s) \models \phi$. Then $l(s) \in \mathcal{L}_{Cf}(F)$, while there is no $L \in \mathcal{L}_{Cf}(F)$ such that $L \models \psi$. Hence $l(s) \not\models \psi$. This implies $\phi \mid\!\sim^W \neg\psi$. Hence $\mid\!\sim^W$ is preferential and conflict-free w.r.t. F. □

The following lemma states that, for every abductive model M based on F, we can construct a preferential model W over $\mathcal{L}_{Cf}(F)$ such that $\mid\!\sim^W = \mid\!\sim^M_{Cr}$. Conversely, for every preferential model W over $\mathcal{L}_{Cf}(F)$, we can construct an abductive model M based on F such that $\mid\!\sim^M_{Cr} = \mid\!\sim^W$.

Note that, for the sake of readability, we have moved some of the longer proofs, including the proof for the following lemma, to section 4.8.

Lemma 4.3.2. *Let F be an argumentation framework and let $\mid\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$. The following are equivalent:*

1. $\mid\sim = \mid\sim^W$ *for a preferential model W over $\mathcal{L}_{Cf}(F)$.*
2. $\mid\sim = \mid\sim^M_{Cr}$ *for an abductive model M based on F.*

Proof. See section 4.8. □

Lemma 4.3.1 and 4.3.2 lead to the following characterization result.

Theorem 4.3.3. *Let F be an argumentation framework and let $\mid\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$. The following are equivalent:*

- $\mid\sim$ *is preferential and conflict-free with respect to F.*
- $\mid\sim = \mid\sim^M_{Cr}$ *for an abductive model M based on F.*

Proof. Follows directly from lemma 4.3.1 and lemma 4.3.2. □

Note that this result implies that preferential models over conflict-free labellings of an argumentation framework can be considered as an alternative but equivalent semantics for credulous observation-based entailment. Preferential models and preferential entailment relations for the evaluation of an argumentation framework were considered before by Booth et al. [24] while ranked models appear in [23]. Preferential models over extensions of an argumentation framework have furthermore been considered by Roos [80].

4.3.2 The Sceptical Case

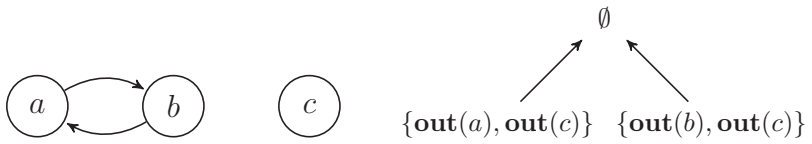
We now show that, given an argumentation framework F, every sceptical observation-based entailment relation based on F is a loop-cumulative entailment relation that is conflict-free w.r.t. F. That is, it satisfies the properties of Reflexivity, Left Logical Equivalence, Right Weakening, Cut, Cautious Monotony and Loop, as described in section 2.2. We prove this by showing that, for every abductive model M based on F, we can construct a cumulative-ordered model W over $\mathcal{L}_{Cf}(F)$ such that $\mid\sim^W = \mid\sim^M_{Sk}$.

Lemma 4.3.4. *Let F be an argumentation framework and let $\mid\sim \subseteq \texttt{lang}(F) \times \texttt{lang}(F)$. It holds that if $\mid\sim = \mid\sim^M_{Sk}$ for an abductive model M based on F then $\mid\sim = \mid\sim^W$ for a cumulative-ordered model W over $\mathcal{L}_{Cf}(F)$.*

Proof. See section 4.8. □

We now obtain the following result.

Theorem 4.3.5. *Let F be an argumentation framework. For every abductive model M based on F it holds that $\mid\sim^M_{Sk}$ is loop-cumulative and conflict-free w.r.t. F.*



(a) The Argumentation Framework. (b) The Abductive Model.

Figure 4.6: Sceptical entailment fails Or (example 4.3.1).

Proof. Follows directly from lemma 4.3.1 and lemma 4.3.4. □

The following proposition shows that, unlike in the credulous case, the characterization in the sceptical case is not complete. This result is due to the fact that, given an abductive model M based on F under semantics σ, $\mid\!\sim^W = \mid\!\sim^M_{Sk}$ holds only if every state in W satisfies the condition that it maps to a set of conflict-free labellings that is *realizable*, in the sense that it coincides with the set $\mathcal{L}_\sigma(F, \Phi)$ for some intervention Φ. The problem is, however, that not every set of conflict-free labellings is realizable. The notion of realizability and its limits has been studied, for extension-based semantics, by Dunne et al. [45].

Proposition 4.3.6. *It is not the case that for every argumentation framework F and every entailment relation $\mid\!\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ that is loop-cumulative and conflict-free w.r.t. F, there is an abductive model M based on F such that $\mid\!\sim = \mid\!\sim^M_{Sk}$.*

Proof of proposition 4.3.6. Let $F = (\{a\}, \emptyset)$. Let $W = (\{s\}, \emptyset, l)$ be the cumulative-ordered model over $\mathcal{L}_{Cf}(F)$ where $l(s) = \{\{(a, \mathbf{in})\}, \{(a, \mathbf{out})\}\}$. It holds that $\mid\!\sim^W$ is a loop-cumulative entailment relation that is conflict-free w.r.t. F. Now assume that $\mid\!\sim^W = \mid\!\sim^M_{Sk}$ for the abductive model $M = (F, H, m, <, \sigma)$. It can be verified that we have $\mathbf{out}(a) \vee \mathbf{in}(a) \not\mid\!\sim^M_{Sk} \bot$, $\mathbf{out}(a) \mid\!\sim^M_{Sk} \bot$ and $\mathbf{in}(a) \mid\!\sim^M_{Sk} \bot$. This implies that there is some $h \in H$ s.t. $m(h) \models^F_\sigma \mathbf{out}(a) \vee \mathbf{in}(a)$, $m(h) \not\models^F_\sigma \mathbf{out}(a)$ and $m(h) \not\models^F_\sigma \mathbf{in}(a)$ and hence $(\mathcal{L}_\sigma(F, m(h)) \downarrow \{a\}) = l(s)$. But then we have $\{(a, \mathbf{out})\} \in (\mathcal{L}_\sigma(F, m(h)) \downarrow \{a\})$, which implies $\mathbf{out}(a) \in m(h)$ and hence $\{(a, \mathbf{in})\} \notin (\mathcal{L}_\sigma(F, m(h)) \downarrow \{a\})$. This is a contradiction. Hence there is no abductive model M such that $\mid\!\sim^W = \mid\!\sim^M_{Sk}$. □

Sceptical observation-based entailment relations do not in general satisfy Or, and hence they are not in general preferential. On the semantic side, this is due to the fact that, in the credulous case, we have that a most preferred credulous explanation for an observation $\phi \vee \psi$ is also a most preferred credulous explanation for the observations ϕ and for ψ, but this does not hold in the sceptical case. The following example demonstrates the failure of Or in the sceptical case.

Example 4.3.1. *Let F be the argumentation framework shown in figure 4.6a. Let $M = (F, H, m, <, Pr)$ be the abductive model based on F under the preferred semantics, defined by:*

- $H = \{\emptyset, \{\mathbf{out}(a), \mathbf{out}(c)\}, \{\mathbf{out}(b), \mathbf{out}(c)\}\}$.

- $\emptyset < \{\mathbf{out}(a), \mathbf{out}(c)\}$ *and* $\emptyset < \{\mathbf{out}(a), \mathbf{out}(c)\}$.
- m *is the identity function.*

This abductive model is shown in figure 4.6b. Note that this abductive model satisfies the minimality assumption but is not closed under weakening, because it contains the interventions $\{\mathbf{out}(a), \mathbf{out}(c)\}$ *and* $\{\mathbf{out}(b), \mathbf{out}(c)\}$ *but not the interventions* $\{\mathbf{out}(a)\}$, $\{\mathbf{out}(c)\}$ *and* $\{\mathbf{out}(b)\}$.

We have $\mathbf{in}(a) \mid\!\sim^{M}_{Sk} \mathbf{out}(c)$ *and* $\mathbf{in}(b) \mid\!\sim^{M}_{Sk} \mathbf{out}(c)$ *but not* $\mathbf{in}(a) \vee \mathbf{in}(b) \mid\!\sim^{M}_{Sk} \mathbf{out}(c)$. *This is a violation of the Or rule.*

4.3.3 Summary and Discussion of Results

The results obtained in this section show that observation-based entailment relations are preferential entailment relations (in the credulous case) and loop-cumulative entailment relations (in the sceptical case). This means that they satisfy the properties Reflexivity, Left Logical Equivalence, Right Weakening, Cut, Cautious Monotony, Loop and (in the credulous case) Or. In the credulous case, we have furthermore obtained a complete characterization: given an argumentation framework F, the class of credulous observation-based entailment relations defined by some abductive model based on F coincides with the class of preferential entailment relations over $\texttt{lang}(F)$ that are, in addition, conflict-free with respect to F.

We have shown in the previous chapter that intervention-based entailment relations fail a number of analogues of the KLM properties. For example, Cautious Monotony fails under the preferred, semi-stable and stable semantics, Cut fails under the semi-stable semantics, and Loop fails under all semantics. But note that, unlike in the intervention-based case, the question of whether an observation-based entailment relation satisfies these properties does not depend on the argumentation semantics that is used. The satisfaction of these properties is due purely to the correspondence between abductive models and preferential/cumulative-ordered models that we established in lemma 4.3.2 and 4.3.4. Thus, unlike in the intervention-based case, the results obtained in this section do not say anything about the behaviour of the complete, grounded, preferred or semi-stable semantics. The results only say something about the mechanism that we used to define observation-based entailment, i.e., using abductive models.

4.4 Directionality in Observation-Based Entailment

In this section we investigate the role of directionality in the behaviour of observation-based entailment. In the setting of intervention-based entailment, we proved that there is a relation between the directionality principle and the property of Conditional Directionality, which expresses that an intervention only affects the status of an argument if the intervention is structurally relevant to

this argument. We proved that the directionality property, if satisfied by a semantics σ, ensures that for all $F \in \mathcal{F}$, the relation $\|\!\approx_\sigma^F$ satisfies Conditional Directionality.

However, the idea behind the property of Conditional Directionality does not apply to observation-based entailment, because observations involve abductive reasoning, which may result in the change of the status of an argument, even if the observation is not structurally relevant to this argument. This means that the analogue of Conditional Directionality, which we reformulate for observation-based entailment as follows, is generally not satisfied, even if the semantics we use satisfies directionality. This analogue states that two observations ϕ and $\phi \wedge \psi$ are the same as far as consequences to which ψ is not structurally relevant are concerned. Note that, compared to Conditional Directionality for intervention-based entailment, we introduce the additional consistency constraint $\psi \not\sim \bot$ to rule out cases where the the consequences of ϕ and $\phi \wedge \psi$ differ due to inconsistency.

Definition 4.4.1. Let $(A, \rightsquigarrow) \in \mathcal{F}$. A relation $\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ satisfies *Conditional Directionality* iff for all $\phi, \psi, \chi \in \mathtt{lang}(F)$,

$$\text{if } \psi \not\rightsquigarrow^* \chi \text{ and } \psi \not\sim \bot \text{ then } \phi \wedge \psi \sim \chi \text{ iff } \phi \sim \chi.$$

The following example demonstrates the failure of Conditional Directionality in the setting of observation-based entailment.

Example 4.4.1. *Let F be the argumentation framework shown in figure 4.1. Let $M = (F, H, m, <, \sigma)$ be the abductive model based on F defined by:*

- $H = 2^{\{\mathbf{out}(a), \mathbf{out}(b), \mathbf{out}(c), \mathbf{out}(d), \mathbf{out}(e), \mathbf{out}(f)\}}$.
- *m is the identity function.*
- *$<=\subset$.*
- *$\sigma \in \{Co, Gr, Pr, SS\}$.*

Note that this abductive model is closed under weakening and satisfies the minimality assumption.

We have $\mathbf{out}(e) \not\sim_{Sk}^M \bot$ and $\mathbf{out}(e) \not\rightsquigarrow^* \mathbf{in}(a)$. Thus, Conditional Directionality would imply that we have $\mathbf{out}(e) \sim_{Sk}^M \mathbf{in}(a)$ iff $\top \sim_{Sk}^M \mathbf{in}(a)$. However, we have:

- *$\mathbf{out}(e) \not\sim_{Sk}^M \mathbf{in}(a)$, because among the most preferred sceptical explanations for $\mathbf{out}(e)$ we have the intervention $\{\mathbf{out}(a)\}$, which does not entail acceptance of a: $\{\mathbf{out}(a)\} \not\|\!\approx_\sigma^F \mathbf{in}(a)$.*
- *$\top \sim_{Sk}^M \mathbf{in}(a)$, because the unique most preferred sceptical explanation for $\top$ is the vacuous intervention, which entails acceptance of a: $\emptyset \|\!\approx_\sigma^F \mathbf{in}(a)$.*

This is a violation of Conditional Directionality. This counterexample also applies to $\sim_{Cr}^M$.

If Conditional Directionality does not apply, then what is the role of directionality in the setting of observation-based entailment? We address this question by looking at two properties that demonstrate this role: Directional **out**-legality and Directional Reinstatement. Before we turn to their definition, we must look at two properties of intervention-based entailment, namely *Conditional* **out**-*legality* and *Conditional Reinstatement*.

4.4.1 Conditional out-legality and Reinstatement

Before we can introduce the properties of Directional **out**-legality and Directional Reinstatement, we must introduce two properties that we call *Conditional* **out**-*legality* and *Conditional Reinstatement*. They are properties for intervention-based entailment. Intuitively, they reflect the assumption that, given an intervention, we *only* unjustifiably reject an argument x if x is defeated by the intervention, and we *only* unjustifiably refrain from accepting an argument x if x is provisionally defeated by the intervention.

Definition 4.4.2. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Conditional* **out**-*legality* iff for all $x \in A$,

$$\text{if } \Phi \not\models \mathbf{out}(x) \text{ then } \Phi \|\!\approx^F \mathbf{out}(x) \rightarrow \vee_{y \in x^-} \mathbf{in}(y).$$

Definition 4.4.3. Let $F = (A, \rightsquigarrow)$ be an argumentation framework. A relation $\|\!\approx^F \subseteq \texttt{Int}(F) \times \texttt{lang}(F)$ satisfies *Conditional Reinstatement* iff for all $x \in A$,

$$\text{if } \Phi \not\models \neg\mathbf{in}(x) \text{ then } \Phi \|\!\approx^F (\wedge_{y \in x^-} \mathbf{out}(y)) \rightarrow \mathbf{in}(x).$$

The **out**-legality (definition 2.1.8) and reinstatement (definition 2.1.9) properties, if satisfied by all labellings under a semantics σ, ensure that for all $F \in \mathcal{F}$, the relation $\|\!\approx^F_\sigma$ satisfies Conditional **out**-legality and Conditional Reinstatement, respectively. Because under the complete, grounded, preferred and semi-stable semantics, all labellings satisfy **out**-legality and reinstatement, this implies that the relations $\|\!\approx^F_{Co}$, $\|\!\approx^F_{Gr}$, $\|\!\approx^F_{Pr}$ and $\|\!\approx^F_{SS}$ satisfy Conditional **out**-legality and Conditional Reinstatement.

Proposition 4.4.1. *Let $F \in \mathcal{F}$ and let σ be a labelling-based semantics.*

1. *If all σ labellings of F satisfy* **out**-*legality then $\|\!\approx^F_\sigma$ satisfies Conditional* **out**-*legality.*
2. *If all σ labellings of F satisfy Reinstatement then $\|\!\approx^F_\sigma$ satisfies Conditional Reinstatement.*

Proof. Let $F = (A, \rightsquigarrow)$ and let σ be a labelling-based semantics.

(1): Suppose for all $L \in \mathcal{L}_\sigma(F)$, L satisfies **out**-legality. Let $\Phi \in \texttt{Int}(F)$, let κ be an F-mapping and let $L' \in \mathcal{L}_\sigma(F \oplus^\kappa \Phi)$. We prove that, for all $x \in A$ s.t. $\Phi \not\models \mathbf{out}(x)$, $L' \models \mathbf{out}(x) \rightarrow \vee_{y \in x^-} \mathbf{in}(y)$ (from here on x^- refers to the attackers of x in F). Let $x \in A$ and suppose $\Phi \not\models \mathbf{out}(x)$. If $L' \not\models \mathbf{out}(x)$ we are done. In the remainder we assume $L' \models \mathbf{out}(x)$. Because L' satisfies **out**-legality there is an y such that $L'(y) = \mathbf{in}$ and y attacks x in $F \oplus \Phi$.

Because $\Phi \not\models \mathbf{out}(x)$, definition 3.2.5 implies that if $y \notin A$ then y is self-attacking in $F \oplus^{\kappa} \Phi$, which is impossible. Hence $y \in A$ and thus $y \in x^{-}$. Thus $L' \models \mathbf{out}(x) \rightarrow \vee_{y \in x^{-}} \mathbf{in}(y)$. It follows that $\Phi \mid\!\!\approx^{F}_{\sigma} \mathbf{out}(x) \rightarrow \vee_{y \in x^{-}} \mathbf{in}(y)$. Hence $\mid\!\!\approx^{F}_{\sigma}$ satisfies Conditional **out**-legality.

(2): Suppose for all $L \in \mathcal{L}_{\sigma}(F)$, L satisfies reinstatement. Let $\Phi \in \mathtt{Int}(F)$, let κ be an F-mapping and let $L' \in \mathcal{L}_{\sigma}(F \oplus^{\kappa} \Phi)$. We prove that, for all $x \in A$ s.t. $\Phi \not\models \neg\mathbf{in}(x)$, $L' \models (\wedge_{y \in x^{-}} \mathbf{out}(y)) \rightarrow \mathbf{in}(x)$ (from here on x^{-} refers to the attackers of x in F). Let $x \in A$ and suppose $\Phi \not\models \neg\mathbf{in}(x)$. If $L' \not\models (\wedge_{y \in x^{-}} \mathbf{out}(y))$ we are done. In the remainder we assume $L' \models (\wedge_{y \in x^{-}} \mathbf{out}(y))$. Because $\Phi \not\models \neg\mathbf{in}(x)$, definition 3.2.5 implies that every attacker of x in $F \oplus^{\kappa} \Phi$ is a member of x^{-}. Because L' satisfies reinstatement it then follows that $L'(x) = \mathbf{in}$. Hence $L' \models (\wedge_{y \in x^{-}} \mathbf{out}(y)) \rightarrow \mathbf{in}(x)$. It follows that $\Phi \mid\!\!\approx^{F}_{\sigma} (\wedge_{y \in x^{-}} \mathbf{out}(y)) \rightarrow \mathbf{in}(x)$. Hence $\mid\!\!\approx^{F}_{\sigma}$ satisfies Conditional Reinstatement. □

4.4.2 Directional out-legality and Reinstatement

The idea behind the properties of Conditional **out**-legality and Conditional Reinstatement does not apply to observation-based entailment. The reason is similar to the reason why Conditional Directionality does not apply: observations involve abductive reasoning, which may result in (provisional) defeat of an argument that is not itself observed to be rejected or not accepted. Keeping the directionality principle in mind, however, it stands to reason that, if we observe ϕ, we (provisionally) defeat an argument x *only* if x is also structurally relevant to ϕ. Intuitively, this is because an argument that is not structurally relevant to an observation should play no role in explaining it. This is expressed by the following two properties, which we call *Directional* **out**-*legality* and *Directional Reinstatement*

Definition 4.4.4. Let $F = (A, \rightsquigarrow)$. A relation $\mid\!\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ satisfies *Directional* **out**-*legality* iff for all $\phi \in \mathtt{lang}(F)$ and $x \in A$,

$$\text{if } x \not\rightsquigarrow^{*} \phi \text{ then } \phi \mid\!\sim \mathbf{out}(x) \rightarrow \vee_{y \in x^{-}} \mathbf{in}(y).$$

Definition 4.4.5. Let $F = (A, \rightsquigarrow)$. A relation $\mid\!\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ satisfies *Directional Reinstatement* iff for all $\phi \in \mathtt{lang}(F)$ and $x \in A$,

$$\text{if } x \not\rightsquigarrow^{*} \phi \text{ then } \phi \mid\!\sim (\wedge_{y \in x^{-}} \mathbf{out}(y)) \rightarrow \mathbf{in}(x).$$

These two properties demonstrate the role of directionality in the setting of observation-based. More precisely, given an abductive model M based on F under the semantics σ that is both closed under weakening and satisfies the minimality assumption, the relations $\mid\!\sim^{M}_{Cr}$ and $\mid\!\sim^{M}_{Sk}$ satisfy Directional **out**-legality (resp. Directional Reinstatment) whenever the corresponding relation $\mid\!\!\approx^{F}_{\sigma}$ satisfies Conditional **out**-legality (resp. Conditional Reinstatement) as well as Conditional Directionality.

Theorem 4.4.2. *Let $F \in \mathcal{F}$ and let σ be a labelling-based semantics. Let $M = (F, H, m, <, \sigma)$ be an abductive model that is closed under weakening and satisfies the minimality assumption.*

1. *If* $\|\!\approx^F_\sigma$ *satisfies Conditional Directionality and Conditional* **out***-legality then both* $\mid\!\sim^M_{Cr}$ *and* $\mid\!\sim^M_{Sk}$ *satisfy Directional* **out***-legality.*

2. *If* $\|\!\approx^F_\sigma$ *satisfies Conditional Directionality and Conditional Reinstatement then both* $\mid\!\sim^M_{Cr}$ *and* $\mid\!\sim^M_{Sk}$ *satisfy Directional Reinstatement.*

For the proof we use the following lemma.

Lemma 4.4.3. *Let* $F = (A, \rightsquigarrow)$ *and let* σ *be a labelling-based semantics. Let* $M = (F, H, m, <, \sigma)$ *be an abductive model that is closed under weakening and satisfies the minimality assumption. Suppose* $\|\!\approx^F_\sigma$ *satisfies Conditional Directionality. If* $h \in H$ *is a most preferred sceptical or credulous explanation for* ϕ *then* $x \in \mathit{Args}(m(h))$ *implies* $x \rightsquigarrow^* \phi$.

Proof. Let $F = (A, \rightsquigarrow)$, σ be a labelling-based semantics and let $M = (F, H, m, <, \sigma)$ be an abductive model that is closed under weakening and satisfies the minimality assumption. Suppose $\|\!\approx^F_\sigma$ satisfies Conditional Directionality. We prove the credulous case (the sceptical case is similar): Let $h \in H$ be a most preferred credulous explanation for ϕ. We then have $m(h) \not\|\!\approx^F_\sigma \neg\phi$ and there is no $h' \in H$ such that $h' < h$ and $m(h') \not\|\!\approx^F_\sigma \neg\phi$. Now suppose the contrary: $x \in \mathit{Args}(m(h))$ and $x \not\rightsquigarrow^* \phi$. We then have $\{\mathbf{out}(x)\} \in m(h)$ or $\{\neg\mathbf{in}(x)\} \in m(h)$. Assume (w.l.o.g.) that $\{\mathbf{out}(x)\} \in m(h)$. Because $\|\!\approx^F_\sigma$ satisfies Conditional Directionality and $\{\mathbf{out}(x)\} \not\rightsquigarrow^* \phi$ it follows that $m(h) \|\!\approx^F_\sigma \phi$ iff $m(h) \setminus \{\mathbf{out}(x)\} \|\!\approx^F_\sigma \phi$ and hence we have $m(h) \setminus \{\mathbf{out}(x)\} \not\|\!\approx^F_\sigma \neg\phi$. But because M is closed under weakening, there is some $h' \in H$ such that $m(h') = m(h) \setminus \{\mathbf{out}(x)\}$. The minimality assumption furthermore implies that $h' < h$. But this is a contradiction, because it means that h is not a most preferred credulous explanation for ϕ. Thus it must hold that $x \in \mathit{Args}(m(h))$ and $x \rightsquigarrow^* \phi$. □

We now prove theorem 4.4.2.

Proof of theorem 4.4.2. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let σ be a labelling-based semantics. Let M be an abductive model based on F under semantics σ that is closed under weakening and satisfies the minimality assumption.

(1): Suppose $\|\!\approx^F_\sigma$ satisfies Conditional Directionality and Conditional **out**-legality. We prove that $\mid\!\sim^M_{Cr}$ satisfies Directional **out**-legality (the proof for $\mid\!\sim^M_{Sk}$ is similar): Let $x \in A$ and suppose $x \not\rightsquigarrow^* \phi$. Let $h \in H$ be a most preferred credulous explanation for ϕ. We must prove that $m(h) \|\!\approx^F_\sigma \mathbf{out}(x) \rightarrow \vee_{y \in x^-} \mathbf{in}(y)$. Via lemma 4.4.3 it follows that for all $z \in A$, $z \in \mathit{Args}(m(h))$ implies $z \rightsquigarrow^* \phi$. Thus we have $x \notin \mathit{Args}(m(h))$. Because $\|\!\approx^F_\sigma$ satisfies Conditional **out**-legality it then follows that $m(h) \|\!\approx^F_\sigma \mathbf{out}(x) \rightarrow \vee_{y \in x^-} \mathbf{in}(y)$. Hence $\mid\!\sim^M_{Cr}$ satisfies Directional **out**-legality.

(2): The proof is similar to (1), except that here we have to show that if $h \in H$ is a most preferred credulous explanation for ϕ then $m(h) \|\!\approx^F_\sigma (\wedge_{y \in x^-} \mathbf{out}(y)) \rightarrow \mathbf{in}(x)$, which follows via lemma 4.4.3 from the fact that $\|\!\approx^F_\sigma$ satisfies Conditional Reinstatement. □

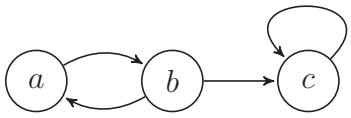


Figure 4.7: Failure of Directional **out**-legality and Reinstatement (example 4.4.2).

Because for all $F \in \mathcal{F}$, $\Vdash^F_{Co}$, $\Vdash^F_{Gr}$ and $\Vdash^F_{Pr}$ all satisfy Conditional Directionality, Conditional **out**-legality and Conditional Reinstatement, this result implies that for all abductive models that are based on F under the complete, grounded and preferred semantics, and are closed under weakening and satisfy the minimality assumption, the relations $\mid\!\sim^M_{Cr}$ and $\mid\!\sim^M_{Sk}$ satisfy Directional **out**-legality and Directional Reinstatement.

In the following example we show that the failure of Directional **out**-legality and Directional Reinstatement leads to unintuitive behaviour. This failure occurs in some abductive models under the semi-stable semantics, which does not satisfy directionality.

Example 4.4.2. *Let F be the argumentation framework shown in figure 4.7. Let $M = (F, H, m, <, SS)$ be the abductive model based on F under the semi-stable semantics, defined by:*

- $H = 2^{\{\mathbf{out}(a), \mathbf{out}(b), \mathbf{out}(c)\}}$.
- $<=\subset$.
- *m is the identity function.*

Note that this abductive model is closed under weakening and satisfies the minimality assumption.

Normally it follows that a is rejected: $\top \mid\!\sim^M_{Cr} \mathbf{out}(a)$. We furthermore have that c is not structurally relevant to the argument a and hence that $c \not\rightsquigarrow^* \mathbf{in}(a)$. Intuitively, c should play no role in explaining the observation $\mathbf{in}(a)$. Thus, Directional **out***-legality implies that observing $\mathbf{in}(a)$ does not lead to the unjustified rejection of c: $\mathbf{in}(a) \mid\!\sim^M_{Cr} \mathbf{out}(c) \rightarrow (\mathbf{in}(b) \vee \mathbf{in}(c))$. Furthermore, Directional Reinstatement implies that observing $\mathbf{in}(a)$ leads to acceptance of c whenever its attackers are rejected: $\mathbf{in}(a) \mid\!\sim^M_{Cr} (\mathbf{out}(b) \wedge \mathbf{out}(c)) \rightarrow \mathbf{in}(c)$. However, neither of these hold. This is due to the fact that, among the most preferred credulous explanations for $\mathbf{in}(a)$ under the semi-stable semantics, we have the intervention $\{\mathbf{out}(c)\}$, which entails neither $\mathbf{out}(c) \rightarrow (\mathbf{in}(b) \vee \mathbf{in}(c))$ nor $(\mathbf{out}(b) \wedge \mathbf{out}(c)) \rightarrow \mathbf{in}(c)$. Thus, we have $\mathbf{in}(a) \not\mid\!\sim^M_{Cr} \mathbf{out}(c) \rightarrow (\mathbf{in}(b) \vee \mathbf{in}(c))$, which is a failure of Directional* **out***-legality, and $\mathbf{in}(a) \not\mid\!\sim^M_{Cr} (\mathbf{out}(b) \wedge \mathbf{out}(c)) \rightarrow \mathbf{in}(c)$ which is a failure of Directional Reinstatement. Thus, we have demonstrated that, if we use the semi-stable semantics (which does not satisfy directionality) then arguments that are not structurally relevant to an observation may play a role in explaining it, and may be unjustifiably rejected.*

Note that theorem 4.4.2 not only establishes the role of directionality but also the role of closure under weakening and the assumption of minimality. Roughly

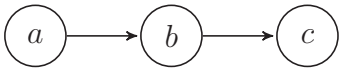
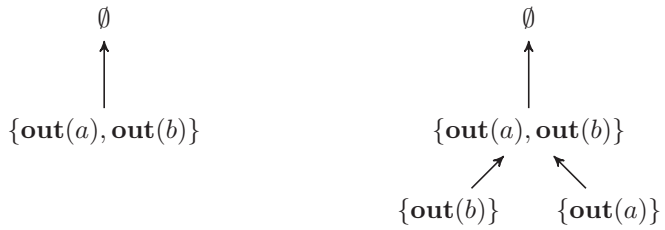


Figure 4.8: The argumentation framework for example 4.4.3.



(a) No closure under weakening. (b) No minimality assumption.

Figure 4.9: Two abductive models (example 4.4.3).

speaking, these properties ensure that an argument is (provisionally) defeated when making an observation only if this (provisional) defeat contributes to explaining the observation. The failure of Directional **out**-legality and Directional Reinstatement due to the violation of closure under weakening or the assumption of minimality is demonstrated by the following example.

Example 4.4.3. *Let F be the argumentation framework shown in figure 4.8.*

1. *Let $M_1 = (F, H, m, <, Gr)$ be the abductive model based on F under the grounded semantics, defined by:*

   - $H = \{\emptyset, \{\mathbf{out}(a), \mathbf{out}(b)\}\}$.
   - $\emptyset < \{\mathbf{out}(a), \mathbf{out}(b)\}$.
   - *m is the identity function.*

   This abductive model is shown in figure 4.9a. Note that this abductive model satisfies the minimality assumption but is not closed under weakening: we have $\{\mathbf{out}(a), \mathbf{out}(b)\} \in H$ but not $\{\mathbf{out}(a)\} \in H$ and $\{\mathbf{out}(b)\} \in H$.

2. *Let $M_2 = (F, H, m, <, Gr)$ be the abductive model based on F under the grounded semantics, defined by:*

   - $H = \{\emptyset, \{\mathbf{out}(a), \mathbf{out}(b)\}, \{\mathbf{out}(a)\}, \{\mathbf{out}(b)\}\}$.
   - $\emptyset < \{\mathbf{out}(a), \mathbf{out}(b)\}$, $\{\mathbf{out}(a), \mathbf{out}(b)\} < \{\mathbf{out}(a)\}$, $\{\mathbf{out}(a), \mathbf{out}(b)\} < \{\mathbf{out}(b)\}$.
   - *m is the identity function.*

   This abductive model is shown in figure 4.9b. Note that we now have closure under weakening, but the minimality assumption is violated: we have, e.g., $\{\mathbf{out}(a), \mathbf{out}(b)\} \models \{\mathbf{out}(a)\}$ and $\{\mathbf{out}(a)\} \not\models \{\mathbf{out}(a), \mathbf{out}(b)\}$ but not $\{\mathbf{out}(a)\} < \{\mathbf{out}(a), \mathbf{out}(b)\}$.

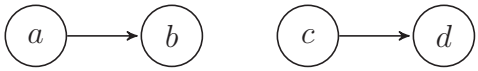


Figure 4.10: Violation of Conditional Noninterference (example 4.5.1).

While $\models^{F}_{Gr}$ *satisfies Conditional Directionality and Conditional* **out***-legality, neither* $\sim^{M_1}_{Sk}$ *nor* $\sim^{M_1}_{Sk}$ *satisfies Directional* **out***-legality: we have* $b \not\rightsquigarrow^{*} \{\mathbf{out}(a)\}$ *but* $\mathbf{out}(a) \not\sim^{M_1}_{Sk} \mathbf{out}(b) \rightarrow \mathbf{in}(a)$ *and* $\mathbf{out}(a) \not\sim^{M_2}_{Sk} \mathbf{out}(b) \rightarrow \mathbf{in}(a)$*. Directional Reinstatement fails in a similar way. The reason for this failure is the fact that in both* M_1 *and* M_2*, the intervention* $\{\mathbf{out}(a), \mathbf{out}(b)\}$ *is the most preferred explanation for the observation* $\mathbf{out}(a)$*, leading to the defeat of* b*, even though this is not necessary to explain* $\mathbf{out}(a)$*.*

4.5 Noninterference in Observation-Based Entailment

In the previous section we explained that the idea behind the property of Conditional Directionality does not apply to observation-based entailment. However, the idea behind the Conditional Noninterference property looks at first sight reasonable in the setting of observation-based. That is, an observation ϕ should not change the status of an argument if the argument is not structurally connected to ϕ. This can be considered reasonable because an argument x that is not structurally connected to an observation ϕ is not structurally relevant to ϕ, and should therefore play no role in explaining ϕ and, conversely, ϕ is not structurally relevant to x, and hence the status of x should not change as a result of observing ϕ.

Let us define the analogue of Conditional Noninterference for observation-based entailment. This property states that two observations ϕ and $\phi \wedge \psi$ are the same as far as consequences that are not structurally connected to ψ are concerned. Note that, compared to Conditional Noninterference for intervention-based entailment, we introduce the additional consistency constraint $\psi \not\sim \bot$ to rule out cases where the the consequences of ϕ and $\phi \wedge \psi$ differ due to inconsistency.

Definition 4.5.1. Let $(A, \rightsquigarrow) \in \mathcal{F}$. A relation $\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ satisfies *Conditional Noninterference* iff for all $\phi, \psi \in \mathtt{lang}(F)$,

$$\text{if } \psi \not\rightsquigarrow^{*} \chi \text{ and } \psi \not\sim \bot \text{ then } \phi \wedge \psi \sim \chi \text{ iff } \phi \sim \chi.$$

It turns out that this property is generally not satisfied. The reason is that, if ϕ involves arguments of two distinct isolated sets then it effectively makes the labels of the arguments in these two sets dependent. We may, for example, observe $\mathbf{in}(x) \rightarrow \mathbf{out}(y)$, where x and y are members of distinct isolated sets. This amounts to an "observed attack", which makes the status of the arguments in the two isolated sets dependent, causing a violation of Conditional Noninterference if the observation is strengthened. This is demonstrated by the following example.

Example 4.5.1. *Let* F *be the argumentation framework shown in figure 4.10. Let* $M = (F, H, m, <, \sigma)$ *be the abductive model defined by:*

- $H = 2^{\{\mathbf{out}(a),\mathbf{out}(b),\mathbf{out}(c),\mathbf{out}(d)\}}$.
- $<=\subset$.
- *m is the identity function.*
- $\sigma \in \{Co, Gr, Pr, SS\}$.

Note that this abductive model is closed under weakening and satisfies the minimality assumption.

Now, we have that $\mathbf{out}(a) \not\sim^{M}_{Cr} \bot$ *and* $\mathbf{out}(a) \not\leftrightsquigarrow^{*} \mathbf{out}(d)$. *Conditional Noninterference would imply that we have, for all* $\phi \in \mathtt{lang}(F)$,

$$\phi \wedge \mathbf{out}(a) \mathrel{|\!\sim}^{M}_{Cr} \mathbf{out}(d) \text{ iff } \phi \mathrel{|\!\sim}^{M}_{Cr} \mathbf{out}(d).$$

Let us see whether this holds if we set ϕ *to* $\mathbf{in}(b) \rightarrow \mathbf{out}(c)$. *Note that this observation involves both isolated sets in* F.

- *On the one hand we have* $\mathbf{in}(b) \rightarrow \mathbf{out}(c) \mathrel{|\!\sim}^{M}_{Cr} \mathbf{out}(d)$. *This is because the unique most preferred credulous explanation for* $\mathbf{in}(b) \rightarrow \mathbf{out}(c)$ *is the vacuous intervention, which entails rejection of* d: $\emptyset \mid\!\models^{F}_{\sigma} \mathbf{out}(d)$.
- *On the other hand we have* $(\mathbf{in}(b) \rightarrow \mathbf{out}(c)) \wedge \mathbf{out}(a) \not\sim^{M}_{Cr} \mathbf{out}(d)$. *This is because among the most preferred credulous explanations for* $(\mathbf{in}(b) \rightarrow \mathbf{out}(c)) \wedge \mathbf{out}(a)$ *we have the intervention* $\{\mathbf{out}(a), \mathbf{out}(c)\}$, *which does not entail rejection of* d: $\{\mathbf{out}(a), \mathbf{out}(c)\} \not\mid\!\models^{F}_{\sigma} \mathbf{out}(d)$.

This is a violation of Conditional Noninterference. This counterexample also applies to $\mathrel{|\!\sim}^{M}_{Sk}$.

The following weaker variant of Conditional Noninterference is still desirable. We call it *Weak Conditional Noninterference*. It states that what we believe after observing ϕ coincides with what we initially believe (i.e., given the observation $\top$) as far as all consequences that are not structurally connected to ϕ are concerned.

Definition 4.5.2. Let $(A, \rightsquigarrow) \in \mathcal{F}$. A relation $\mathrel{|\!\sim} \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$ satisfies *Weak Conditional Noninterference* iff for all $\phi, \psi \in \mathtt{lang}(F)$,

$$\text{if } \phi \not\leftrightsquigarrow^{*} \psi \text{ and } \phi \not\sim \bot \text{ then } \top \mathrel{|\!\sim} \psi \text{ iff } \phi \mathrel{|\!\sim} \psi.$$

We have that, given an abductive model M based on F under the semantics σ that is both closed under weakening and satisfies the minimality assumption, the relations $\mathrel{|\!\sim}^{M}_{Cr}$ and $\mathrel{|\!\sim}^{M}_{Sk}$ satisfy Weak Conditional Noninterference whenever the relation $\mathrel{|\!\models}^{F}_{\sigma}$ satisfies Conditional Noninterference.

Theorem 4.5.1. *Let* $F \in \mathcal{F}$ *and let* σ *be a labelling-based semantics. Let* $M = (F, H, m, <, \sigma)$ *be an abductive model that is closed under weakening and satisfies the minimality assumption. If* $\mathrel{|\!\models}^{F}_{\sigma}$ *satisfies Conditional Noninterference then both* $\mathrel{|\!\sim}^{M}_{Cr}$ *and* $\mathrel{|\!\sim}^{M}_{Sk}$ *satisfy Weak Conditional Noninterference.*

For the proof we use the following lemma.

Lemma 4.5.2. *Let $F = (A, \rightsquigarrow)$ and let σ be a labelling-based semantics. Let $M = (F, H, m, <, \sigma)$ be an abductive model that is closed under weakening and satisfies the minimality assumption. Suppose $\Vdash^F_\sigma$ satisfies Conditional Noninterference. If $h \in H$ is a most preferred sceptical or credulous explanation for ϕ then $x \in Args(m(h))$ implies $x \leftsquigarrow^* \phi$.*

Proof. Let $F = (A, \rightsquigarrow)$, σ be a labelling-based semantics and let $M = (F, H, m, <, \sigma)$ be an abductive model that is closed under weakening and satisfies the minimality assumption. Suppose $\Vdash^F_\sigma$ satisfies Conditional Directionality. We prove the credulous case (the sceptical case is similar): Let $h \in H$ be a most preferred credulous explanation for ϕ. We then have $m(h) \not\Vdash^F_\sigma \neg\phi$ and there is no $h' \in H$ such that $h' < h$ and $m(h') \not\Vdash^F_\sigma \neg\phi$. Now suppose the contrary: $x \in Args(m(h))$ and $x \not\leftsquigarrow^* \phi$. We then have $\{\mathbf{out}(x)\} \in m(h)$ or $\{\neg\mathbf{in}(x)\} \in m(h)$. Assume (w.l.o.g.) that $\{\mathbf{out}(x)\} \in m(h)$. Because $\Vdash^F_\sigma$ satisfies Conditional Noninterference and $\{\mathbf{out}(x)\} \not\leftsquigarrow^* \phi$ it follows that $m(h) \Vdash^F_\sigma \phi$ iff $m(h) \setminus \{\mathbf{out}(x)\} \Vdash^F_\sigma \phi$ and hence we have $m(h) \setminus \{\mathbf{out}(x)\} \not\Vdash^F_\sigma \neg\phi$. But because M is closed under weakening, there is some $h' \in H$ such that $m(h') = m(h) \setminus \{\mathbf{out}(x)\}$. The minimality assumption furthermore implies that $h' < h$. But this is a contradiction, because it means that h is not a most preferred credulous explanation for ϕ. Thus it must hold that $x \in Args(m(h))$ and $x \leftsquigarrow^* \phi$. □

We now prove theorem 4.5.1.

Proof of theorem 4.5.1. Let $F \in \mathcal{F}$ and let σ be a labelling-based semantics. Let $M = (F, H, m, <, \sigma)$ be an abductive model that is closed under weakening and satisfies the minimality assumption. Assume $\Vdash^F_\sigma$ satisfies Conditional Noninterference. We prove that $\mid\!\sim^M_{Cr}$ satisfies Weak Conditional Noninterference (the proof that $\mid\!\sim^M_{Sk}$ satisfies Weak Conditional Noninterference follows similarly.) Suppose $\phi \not\leftsquigarrow^* \psi$ and assume $\phi \not\mid\!\sim^M_{Cr} \bot$.

(Only if:) Suppose $\top \mid\!\sim^M_{Cr} \psi$. We denote by $h \in H$ a hypothesis such that $m(h) = \emptyset$ (existence of h is guaranteed because M is closed under weakening). Because M satisfies the minimality assumption, h is a most preferred credulous explanation for $\top$. Thus we have $\emptyset \Vdash^F_\sigma \psi$. Now let $h' \in H$ be a most preferred credulous explanation for ϕ (existence of h' is guaranteed because we have $\phi \not\mid\!\sim^M_{Cr} \bot$). Lemma 4.5.2 implies that if $x \in Args(m(h'))$ then $x \leftsquigarrow^* \phi$. Because $\phi \not\leftsquigarrow^* \psi$ it follows that $m(h') \not\leftsquigarrow^* \psi$. Because $\Vdash^F_\sigma$ satisfies Conditional Noninterference it follows that $m(h') \Vdash^F_\sigma \psi$. Hence $\phi \mid\!\sim^M_{Cr} \psi$.

(If:) Suppose $\phi \mid\!\sim^M_{Cr} \psi$. Let $h \in H$ be a most preferred credulous explanation for ϕ (existence of h' is guaranteed because we have $\phi \not\mid\!\sim^M_{Cr} \bot$). We then have $m(h) \Vdash^F_\sigma \psi$. Lemma 4.5.2 implies that if $x \in Args(m(h))$ then $x \leftsquigarrow^* \phi$. Because $\phi \not\leftsquigarrow^* \psi$ it follows that $m(h) \not\leftsquigarrow^* \psi$. Because $\Vdash^F_\sigma$ satisfies Conditional Noninterference it follows that $\emptyset \Vdash^F_\sigma \psi$. Hence $\top \mid\!\sim^M_{Cr} \psi$. □

Because, for all $\sigma \in \{Co, Gr, Pr, SS\}$, σ satisfies the noninterference principle, we have that for all $F \in \mathcal{F}$, $\Vdash^F_\sigma$ satisfies Conditional Noninterference. This implies that, given any abductive model M based on F under the complete, grounded, preferred or semi-stable semantics that is closed under weakening and satisfies the minimality assumption, the relations $\mid\!\sim^M_{Cr}$ and $\mid\!\sim^M_{Sk}$ satisfy Weak Conditional Noninterference.

As we explained in the previous section, closure under weakening and the minimality assumption imply, roughly speaking, that an argument is (provisionally) defeated only if this contributes to explaining the observation. For this reason, the violation of closure under weakening or the minimality assumption may lead to failure of Weak Conditional Noninterference. For a demonstration we can consider again the argumentation framework and abductive model used in example 4.3.1. This abductive model is not closed under weakening. As a result, the sceptical observation $\mathbf{in}(a)$ leads to rejection of c, even though a and c are members of two distinct isolated sets.

4.6 Related Work

In describing the problem of revising the evaluation of an argumentation framework we made a distinction between the single-agent revision perspective and the multi-agent strategic perspective. While we have taken the single-agent revision perspective, the multi-agent strategic perspective has received most attention in the literature. This includes the work of Baumann and Brewka, who call this the *enforcing* problem [12]. The precise question they address is: can some set of arguments be enforced (i.e., made credulously accepted) by modifying the argumentation framework? They distinguish different types of modifications, one type being a *normal expansion*, where new arguments/attacks are added but no attacks between existing arguments. For each type they determine the conditions under which a set of arguments is enforceable under a given semantics. Baumann also looked at the problem of minimal change, that is: if some set of arguments is enforceable, then how many modifications (i.e., how many added or removed attacks) are necessary to do so? [11]

Kontarinis et al. [61] also take the multi-agent strategic perspective. Their approach is based on the notion of a *game board*, which is an argumentation framework together with a specification of the set of possible changes (considering only the addition/removal of attacks between existing arguments). They then study the problem of changing the argumentation framework with the goal of making an argument accepted or not accepted under a given semantics. They propose a method of computation, based on term rewriting logic, where the initial goal is rewritten into sub-goals which, if successful, leads to a so called *target set*, or a minimal set of attacks that must be added/removed to satisfy the goal.

Boella et al.[20] study goal-oriented change under the grounded semantics. Their approach centres on the concept of a conditional labelling, which associates each argument with three formulas. These formulas describe which arguments must be attacked to make the argument labelled **in**, **out** or **und**.

Bisquert et al. [17] takes a slightly different perspective. They look at change of the evaluation of an argumentation framework as a process that is part of persuasion. That is, an agent wants to revise an audience's evaluation of a debate, and the revision operator tells the agent how to do this. They introduce a language to describe sets of argumentation frameworks as well as a language to describe goals concerning the acceptance of arguments. Generalized enforcement is then a process that takes as input two formulas, one representing a set of initial argumentation frameworks and one describing a set of arguments to accept.

The result is a revised set of argumentation frameworks in which the desired arguments are accepted. They prove a representation theorem that relates a particular class of generalized enforcement operators with distance measures defined over argumentation frameworks.

A clear single-agent revision perspective is taken by Coste-Marquis et al. [37]. They focus on revising the evaluation of an argumentation framework by adding and removing attacks between existing arguments. The motivating example they use is that an agent has to revise the evaluation of its argumentation framework when confronted with trustworthy information (for practical purposes, this can be considered to be the same as what we call observation). They focus on revision operators that take as input an argumentation framework and a revision formula, and return a set of revised argumentation frameworks. Specifically, they focus on revision operators that are characterized by a set of AGM-style postulates for revision with minimal change. That is, the set of extensions of the revised argumentation frameworks taken together should differ minimally from the set of extensions of the initial argumentation framework. They then look at a number of distance measures over extensions and argumentation frameworks, that can be used to define concrete revision operators.

Rotstein et al. [81, 82] and Moguillansky et al. [69] study revision in a more structured setting. They assume that there is, apart from a set of arguments (and a distinct set of *active* arguments), a *subargument* relation among arguments. This is called a *dynamic argumentation framework*. This additional structure allows the study of operators that add an argument and afterwards make the necessary additional change to ensure that the added argument is accepted.

It is well known that non monotonic inference relations can be used to define belief revision operators and vice versa. This means that our theory of observation-based entailment can alternatively be seen as a theory about belief revision operators. We could formalize this as follows: an an observation-based entailment relation $\mid\!\sim^M$ defined by an abductive model M based on F defines a revision operator $*$ for F defined by $\psi \in (F * \phi)$ iff $\phi \mid\!\sim^M \psi$. Results due to Gardenfors and Makinson [51] imply that, if $\mid\!\sim^M$ is rational, this operator behaves much like a belief revision operator that satisfies the AGM postulates. A weaker set of postulates for belief revision, that applies in the case where $\mid\!\sim^M$ is not rational but still preferential, has been studied by Benferhat et al. [16].

We made a distinction between credulous and sceptical observation-based entailment, which differs in how explanations are selected. That is, credulous (resp. sceptical) entailment is based on selecting the most preferred explanations that make an observation credulously (resp. sceptically) true. Bochman investigated credulous and sceptical non-monotonic inference based on a different distinction [18]. Roughly speaking, sceptical entailment in Bochman's investigation is based on the rule that $\phi \mid\!\sim \psi$ holds if *all* preferred states consistent with ϕ satisfy $\phi \rightarrow \psi$, while credulous entailment is based on the rule that $\phi \mid\!\sim \psi$ holds if *some* preferred state consistent with ϕ satisfies $\phi \rightarrow \psi$. Whereas in our setting, the distinction applies to how the premise is interpreted, it is, in Bochman's setting, applied to how the conclusion is interpreted.

Observation-based entailment relations are closely related to *abductive conse-*

quence relations, as studied by Pino Pérez and Uzcátegui [73]. These are consequence relations that are defined via an underlying principle of abduction. The idea is very close to the one that we applied: Given a background theory Σ they associate with each *explanatory relation* $\triangleright$ between formulas (where $\phi \triangleright \psi$ means that ψ is a preferred explanation of ϕ) a consequence relation $\mid\!\sim^{ab}$ by setting

$$\phi \mid\!\sim^{ab} \psi \text{ iff } \Sigma \cup \{\chi\} \vdash \psi \text{ for every } \chi \text{ such that } \phi \triangleright \chi.$$

They then isolate postulates for explanatory relations based on the interplay between $\triangleright$ and $\mid\!\sim^{ab}$. In particular, they isolate postulates for $\triangleright$ when $\mid\!\sim^{ab}$ is assumed to be an entailment relation that satisfies the properties studied by Kraus et al. [62]. Lobo and Uzcátegui [66] studied abductive consequence relations defined by cumulative ordered models that capture preferences among explanations. This is similar to our notion of sceptical observation-based entailment which, as we saw in section 4.3, is also closely related to cumulative ordered models. They observe that the resulting notion of consequence does not satisfy the Or-rule. We proved that the situation is the same in the sceptical case, but not in the credulous case.

Finally, Roos [80] studied the relationship between the preferential model and argumentation semantics. He first defines a preferential model as as consisting of a preference relation over states, where each state maps to a conflict-free extension of a given argumentation framework. He considers preferential models in which a sufficient (but not necessary) condition for one extension E to be preferred over another extension E' is when E is a superset of E'. He then considers various preference relations where the most preferred states of these preferential models coincide with the complete, grounded, preferred and stable extensions.

4.7 Conclusion and Future Work

We identified two types of change: intervention (representing actions in a debate) and observation (information from the environment that requires the revision of the status of one or more arguments). In this chapter we addressed the question of how a rational agent should revise the evaluation of an argumentation framework to account for an observation. We proposed a model, based on an abductive principle, that determines an observation-based entailment relation. This entailment relation captures how an argumentation framework is evaluated given an observation.

We proved that the class of observation-based entailment relations for a given argumentation framework F (i.e., the observation-based entailment relations determined by some abductive model for F) is characterized by a strengthening of the class of preferential (in the credulous case) and loop-cumulative (in the sceptical case) entailment relations. This characterization is complete in the credulous case, but not in the sceptical case. This result gives a handle on how credulous and sceptical observation-based entailment differ, namely that the former satisfies the Or-rule, but the latter generally does not.

Finally, we investigated the difference between interventions and observations, in terms of how their effect propagate through an argumentation framework.

We also investigated the role of the principles of directionality and noninterference in the behaviour of an observation-based entailment relation. Namely, these principles ensure (if we make some reasonable further assumptions) that the effect of an observation propagates through the argumentation framework in a well-behaved manner. If we use a semantics that satisfies the directionality, the difference between intervention and observation is that the effect of an intervention only propagates to arguments that are directly or indirectly attacked by an argument that is (provisionally) defeated, whereas the effect of an observation also propagates to arguments that play a role in explaining the observation (i.e., arguments that directly or indirectly attack arguments of which the status is observed).

We simplified our model by abstracting away from the actual changes that can be made to an argumentation framework in order to account for an observation, and instead we focussed on change represented by interventions. This is a limitation in the sense that not every change that can be made to an argumentation framework is represented by some intervention. This includes changes where attacks between existing arguments are removed, and changes where multiple arguments and attacks are added at once. Considering more general forms of change will lead to more general forms of revision. For example, the observation that two arguments a and b where a attacks b are both accepted, can be explained by removing the attack from a to b. In the current setting, observations that are not conflict-free cannot be consistently dealt with. These considerations suggest more general forms of revision in argumentation, which is a subject for future research.

Finally, we have been working in the setting of abstract argumentation, without taking into account possible instantiations of abstract argumentation frameworks. This raises the question: can observation-based entailment be applied if we use instantiated forms of argumentation? We partially address this question in the following chapter, where we present a model of abduction in abstract argumentation that is an abstraction of abduction in logic programming. However, we focus in the following chapter on *finding* explanations for a given observation, and we do not, like we did in this chapter, look at the specific problem of revision due to observation.

4.8 Proofs

Lemma 4.3.2. *Let F be an argumentation framework and let $\mid\!\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$. The following are equivalent:*

1. *$\mid\!\sim = \mid\!\sim^{M}_{Cr}$ for an abductive model M based on F.*
2. *$\mid\!\sim = \mid\!\sim^{W}$ for a preferential model W over $\mathcal{L}_{Cf}(F)$.*

For the (1) to (2) direction of lemma 4.3.2 we use definition 4.8.1 and lemma 4.8.1 and 4.8.2.

Definition 4.8.1. Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $M = (F, H, m, <, \sigma)$ be an abductive model. We define W^{M}_{Cr} by $W^{M}_{Cr} = (S, \prec, l)$, where

- $S = \{(h, L) \mid h \in H, L \in \mathcal{L}_\sigma(F, m(h)) \downarrow A\}$,
- $(h, L) \prec (h', L')$ iff $h < h'$,
- $l((h, L)) = L$.

Lemma 4.8.1. *Let F be an argumentation framework, $M = (F, H, m, <, \sigma)$ an abductive model and let $W^M_{Cr} = (S, \prec, l)$. For all $(h, L) \in S$ it holds that (h, L) is $\prec$-minimal in $\widehat{\phi}$ iff h is a most preferred credulous explanation for ϕ and $L \models \phi$.*

Proof. Let F be an argumentation framework, $M = (F, H, m, <, \sigma)$ a abductive model and let $W^M_{Cr} = (S, \prec, l)$. Let $(h, L) \in S$.

(Only if) Suppose (h, L) is $\prec$-minimal in $\widehat{\phi}$. It follows immediately that $L \models \phi$. We prove by contradiction that h is a most preferred credulous explanation for ϕ. Suppose the contrary. Then there is a h' that is a credulous explanation for ϕ and $h' < h$. But then there is an $L' \in \mathcal{L}_\sigma(F, m(h')) \downarrow A$ and $L' \models \phi$, and hence a state $(h', L') \in \widehat{\phi}$ such that $(h', L') \prec (h, L)$ (i.e., (h, L) is not $\prec$-minimal in $\widehat{\phi}$). This is a contradiction. Hence h is a most preferred credulous explanation for ϕ.

(If) Suppose h is a most preferred credulous explanation for ϕ and $L \models \phi$. It immediately follows that $(h, L) \in \widehat{\phi}$. Suppose (h, L) is not $\prec$-minimal in $\widehat{\phi}$. Then there is a state $(h', L') \in \widehat{\phi}$ and $(h', L') \prec (h, L)$. But then h is not a most preferred credulous explanation for ϕ. This is a contradiction. Hence (h, L) is $\prec$-minimal in $\widehat{\phi}$. □

Lemma 4.8.2. *Let F be an argumentation framework and let $M = (F, H, m, <, \sigma)$ an abductive model. It holds that W^M_{Cr} is a preferential model over $\mathcal{L}_{Cf}(F)$.*

Proof. Let F be an argumentation framework and let $M = (F, H, m, <, \sigma)$ an abductive model. Let $W^M_{Cr} = (S, \prec, l)$. From definition 4.8.1 together with the fact that $<$ is a strict partial order it follows that $\prec$ is a strict partial order. Furthermore finiteness of H implies finiteness of S and hence the smoothness condition is satisfied. Theorem 3.2.8 furthermore implies that for all $s \in S$, $l(s) \subseteq \mathcal{L}_{Cf}(F)$, while proposition 2.1.10 implies that for all $s \in S$, $l(s)$ is non-empty. By definition 2.2.5 it follows that W^M_{Cr} is a preferential model over $\mathcal{L}_{Cf}(F)$. □

For the (2) to (1) direction of lemma 4.3.2 we use definitions 4.8.3 and 4.8.2 and lemmas 4.8.3, 4.8.4 and 4.8.5.

Definition 4.8.2. Let F be an argumentation framework and let $L \in \mathcal{L}_{Cf}(F)$. We define the *enforcing intervention* $Enf(L)$ of L by

$$Enf(L) = \{\mathbf{out}(x) \mid L(x) = \mathbf{out}\} \cup \{\neg\mathbf{in}(x) \mid L(x) = \mathbf{und}\}.$$

Lemma 4.8.3. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework. Let $\sigma \in \{Co, Gr, Pr, SS\}$. For all $L \in \mathcal{L}_{Cf}(F)$ it holds that $\mathcal{L}_\sigma(F, Enf(L)) \downarrow A = \{L\}$.*

Proof. The case $\sigma = Co$ follows from the principles of conflict-freeness, admissibility and reinstatement. The other cases follow directly. □

Definition 4.8.3. Let F be an argumentation framework and let $W = (S, \prec, l)$ be a finite preferential model over $\mathcal{L}_{Cf}(F)$. Let $\sigma \in \{Co, Gr, Pr, SS\}$. We define M_W^σ by $M_W^\sigma = (F, H, m, <, \sigma)$, where

- $H = S$.
- m is defined by $m(s) = Enf(l(s))$ (See definition 4.8.2).
- $<=\prec$.

Lemma 4.8.4. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and let $W = (S, \prec, l)$ be a finite preferential model over $\mathcal{L}_{Cf}(F)$. Let $\sigma \in \{Co, Gr, Pr, SS\}$. Let $(F, H, m, <, \sigma) = M_W^\sigma$. For all $s \in S$ it holds that s is a credulous explanation for ϕ iff $s \in \widehat{\phi}$.*

Proof. Let $s \in S$. Definition 4.2.4 implies that s is a credulous explanation for ϕ iff $m(s) \not\models_\sigma^F \neg\phi$. Because $m(s) = Enf(l(s))$, lemma 4.8.3 implies that $\{l(s)\} = \mathcal{L}_\sigma(F, m(s)) \downarrow A$ and hence that $m(s) \not\models_\sigma^F \neg\phi$ iff $l(s) \models \phi$. Hence s is a credulous explanation for ϕ iff $s \in \widehat{\phi}$. □

Lemma 4.8.5. *Let F be an argumentation framework. Let $W = (S, \prec, l)$ be a finite preferential model over $\mathcal{L}_{Cf}(F)$. Let $\sigma \in \{Co, Gr, Pr, SS\}$. It holds that M_W^σ is an abductive model based on F.*

Proof. From the fact that $\prec$ is a strict partial order it follows that $<$ is a strict partial order. Finiteness of S implies finiteness of H. Lemma 4.8.3 implies that for no $h \in H$, $m(h) \models_\sigma^F \bot$. Hence M_W^σ is an abductive model based on F. □

We are now ready to prove lemma 4.3.2.

Proof of lemma 4.3.2. We first prove that (1) implies (2): Let M be an abductive model based on F. Lemma 4.8.2 implies that $W_{Cr}^M = (S, \prec, l)$ is a finite preferential model over $\mathcal{L}_{Cf}(F)$. The following equivalences prove that $\mid\!\sim_{Cr}^M = \mid\!\sim^{W_{Cr}^M}$:

1. $\phi \mid\!\sim^{W_{Cr}^M} \psi$
2. For every $(h, L) \in S$ that is $\prec$-minimal in $\widehat{\phi}$ we have $l((h, L)) \models \psi$.
3. For every most preferred credulous explanation h for ϕ and every $L \in \mathcal{L}_\sigma(F, m(h))$ s.t. $L \models \phi$ we have $L \models \psi$.
4. For every most preferred credulous explanation h for ϕ we have $m(h) \models_\sigma^F \phi \rightarrow \psi$.
5. $\phi \mid\!\sim_{Cr}^M \psi$

Equivalence of 1/2 follows from definition 2.2.7. Equivalence of 2/3 follows from lemma 4.8.1. Equivalence of 3/4 follows from definitions 2.1.16 and 2.1.17. Equivalence of 4/5 follows from definition 4.2.5. Hence $\mid\!\sim = \mid\!\sim_{Cr}^M = \mid\!\sim^W$ for a finite preferential model W over $\mathcal{L}_{Cf}(F)$.

We now prove that (2) implies (1): Let W be a finite preferential model over $\mathcal{L}_{Cf}(F)$. Let $\sigma \in \{Co, Gr, Pr, SS\}$. Lemma 4.8.5 implies that $M^{\sigma}_{W} = (F, H, m, <, \sigma)$ is an abductive model based on F. The following equivalences prove that $\mid\!\sim^{W} = \mid\!\sim^{M^{\sigma}_{W}}_{Cr}$:

1. $\phi \mid\!\sim^{W} \psi$.
2. For every $s \in S$ that is $\prec$-minimal in $\widehat{\phi}$ we have $l(s) \models \psi$.
3. For every $h \in H$ that is a most preferred credulous explanation for ϕ we have $m(h) \parallel\!\!\approx^{F}_{\sigma} \phi \rightarrow \psi$.
4. $\phi \mid\!\sim^{M^{\sigma}_{W}}_{Cr} \psi$.

Equivalence of 1/2 follows from definition 2.2.7. Equivalence of 2/3 follows from lemma 4.8.4. Equivalence of 3/4 follows from definition 4.2.5. Hence $\mid\!\sim = \mid\!\sim^{W} = \mid\!\sim^{M}_{Cr}$ for an abductive model based on F. □

Lemma 4.3.4. *Let F be an argumentation framework and let $\mid\!\sim \subseteq \mathtt{lang}(F) \times \mathtt{lang}(F)$. It holds that if $\mid\!\sim = \mid\!\sim^{M}_{Sk}$ for an abductive model M based on F then $\mid\!\sim = \mid\!\sim^{W}$ for a cumulative-ordered model W over $\mathcal{L}_{Cf}(F)$.*

To prove lemma 4.3.4 we use definition 4.8.4 and lemma 4.8.6 and 4.8.7.

Definition 4.8.4. Let $F = (A, \rightsquigarrow)$ and let $M = (F, H, m, <, \sigma)$ be an abductive model. We define W^{M}_{Sk} by $W^{M}_{Sk} = (S, \prec, l)$, where

- $S = H$.
- $< = \prec$.
- $l(s) = \mathcal{L}_{\sigma}(F, m(s)) \downarrow A$.

Lemma 4.8.6. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework, $M = (F, H, m, <, \sigma)$ an abductive model and $W^{M}_{Sk} = (S, \prec, l)$. For all $s \in S$ it holds that s is a sceptical explanation for ϕ iff $s \in \widehat{\phi}$.*

Proof. Follows directly from definitions 4.2.6, 3.2.8, 2.1.17, 4.8.4 and 2.2.7. □

Lemma 4.8.7. *Let F be an argumentation framework and $M = (F, H, m, <, \sigma)$ an abductive model based on F. It holds that W^{M}_{Sk} is a finite cumulative ordered model over $\mathcal{L}_{Cf}(F)$.*

Proof. Let $M = (F, H, m, <, \sigma)$ an abductive model based on F. Let $W^{M}_{Sk} = (S, \prec, l)$. The fact that $<$ is a strict partial order implies that $\prec$ is too. Theorem 3.2.8 furthermore implies that for all $s \in S$, $l(s) \subseteq \mathcal{L}_{Cf}(F)$, while proposition 2.1.10 implies that for all $s \in S$, $l(s)$ is non-empty. Because $H = S$ it holds that S is finite. By definition 2.2.4 it follows that W^{M}_{Sk} is a cumulative ordered model over $\mathcal{L}_{Cf}(F)$. □

We are now ready to prove lemma 4.3.4.

Proof of lemma 4.3.4. Suppose $\mid\!\sim = \mid\!\sim^{M}_{Sk}$ for an abductive model $M = (F, H, m, <, \sigma)$. Lemma 4.8.7 implies that $W^{M}_{Sk} = (S, \prec, l)$ is a finite cumulative ordered model over $\mathcal{L}_{Cf}(F)$. We show that $\mid\!\sim^{W} = \mid\!\sim^{W^{M}_{Sk}}_{Sk}$. This follows from the following equivalences:

1. $\phi \mid\!\sim^{W} \psi$.
2. For all s $\prec$-minimal in $\widehat{\phi}$ we have $\forall v \in l(s), v \models \psi$.
3. For every most preferred sceptical explanation h for ϕ we have $m(h) \parallel\!\models^{F}_{\sigma} \psi$.
4. $\phi \mid\!\sim^{W^{M}_{Sk}}_{Sk} \psi$.

Equivalence of 1/2 follows from definition 2.2.7. Equivalence of 2/3 follows from lemma 4.8.6. Equivalence of 3/4 follows from definition 4.2.7. Hence $\mid\!\sim = \mid\!\sim^{W} = \mid\!\sim^{W}_{Sk}$ for some cumulative ordered model W over $\mathcal{L}_{Cf}(F)$. □

Chapter 5

Abduction in Argumentation and Logic Programming

5.1 Introduction

In the context of abstract argumentation, abduction can be seen as the problem of finding *changes* to an argumentation framework with the goal of explaining observations about the evaluation of the argumentation framework. This was the mechanism by which we defined the notion of observation-based entailment in chapter 4. In this chapter we look at two further aspects of combining argumentation and abduction.

Firstly, proof theories in argumentation are often formulated as *dialogical* proof theories, which aim at relating the problem they address with stereotypical patterns found in real world dialogue. For example, proof theories for sceptical/credulous acceptance have been modelled as dialogues in which a proponent persuades an opponent to accept the necessity/possibility of an argument [68], while credulous acceptance has also been related to Socratic style dialogue [32]. This raises the question of whether proof procedures for abduction in argumentation can similarly be modelled as dialogues. Secondly, abstract argumentation can be seen as an abstraction of logic programming. That is, an *instantiated* argumentation framework can be generated on the basis of a logic program, and the consequences of the logic program be computed by looking at the extensions of the instantiated argumentation framework [42]. In the context of abduction, one may ask whether a model of abduction in argumentation can similarly be seen as an abstraction of *abductive* logic programming. These are the two questions we address in this chapter.

We first present a model of abduction in abstract argumentation, based on the notion of an abductive argumentation framework that encodes different possible changes to an argumentation framework, each of which may act as a hypothesis to explain an observation that can be justified by making an argument accepted.

We then do two things:

1. We present sound and complete dialogical proof procedures for the main reasoning tasks, i.e., finding hypotheses that explain sceptical/credulous acceptance of arguments in support of an observation. These proof procedures show that the problem of abduction is related to an extended form of persuasion, where the proponent uses *hypothetical* moves to persuade the opponent.

2. We show that abductive argumentation frameworks can be instantiated by ALPs (abductive logic programs) in such a way that the hypotheses generated for an observation by the ALP can be computed by translating the ALP into an abductive argumentation framework. The type of ALPs we focus on are based on Sakama and Inoue's model of *extended* abduction [57, 58], in which hypotheses have a positive as well as a negative element (i.e., facts added to the logic program as well as facts removed from it).

In sum, our contribution is a model of abduction in argumentation which can be seen as an abstraction of abduction in logic programming, and we present dialogical proof procedures for determining what the explanations for a given observation are.

The overview of this chapter is as follows. After introducing the necessary preliminaries we present in section 5.2 a model of abduction in abstract argumentation. In section 5.3 we present dialogical proof procedures for two of the main problems (explaining sceptical/credulous acceptance). In section 5.4 we show that our model of abduction can be used to instantiate abduction in logic programming. We conclude by discussing related work in section 5.5 we conclude in section 5.6.

This chapter is based on joint work with Richard Booth, Dov Gabbay, Souhila Kaci and Leendert van der Torre [21].

5.2 Abductive Argumentation Frameworks

To simplify our definitions, we work in this chapter with extension-based semantics instead of labelling-based semantics. We assume that an observation translates into a set $B \subseteq A$. Intuitively, B is a set of arguments that each individually support the observation. If at least one argument $x \in B$ is sceptically (resp. credulously) accepted w.r.t. the complete semantics, we say that the observation B is sceptically (resp. credulously) *supported*.

Definition 5.2.1. Given an argumentation framework $F = (A, \rightsquigarrow)$, an observation $B \subseteq A$ is sceptically (resp. credulously) supported iff for all (resp. some) $E \in \mathcal{E}_{Co}(F)$ it holds that $x \in E$ for some $x \in B$.

The following proposition implies that checking whether an observation B is sceptically supported can be done by checking whether an individual argument $x \in B$ is in the grounded extension.

Proposition 5.2.1. *Let $F = (A, \rightsquigarrow)$, $B \subseteq A$. It holds that F sceptically supports B iff some $x \in B$ is a member of the grounded extension of F.*

It may be that an argumentation framework F does not sceptically or credulously support an observation B. Abduction then amounts to finding a change to F so that B is supported. We use the following definition of an *abductive argumentation framework* to capture the changes with respect to F (each change represented by an argumentation framework G called an *abducible* argumentation framework) that an agent considers. We assume that F itself is also an abducible argumentation framework, namely one that captures the case where no change is necessary. Other abducible argumentation frameworks may be formed by addition of arguments and attacks to F, removal of arguments and attacks from F, or a combination of both.

Definition 5.2.2. An *abductive argumentation framework* is a pair $M = (F, I)$ where F is an argumentation framework and $I \subseteq \mathcal{F}$ a set of argumentation frameworks called *abducible* such that $F \in I$.

Given an abductive argumentation framework (F, I) and observation B, sceptical/credulous support for B can be explained by the change from F to some $G \in I$ that sceptically/credulously supports B. In this case we say that G *explains* sceptical/credulous support for B. The arguments/attacks added to and absent from G can be seen as the actual explanation.

Definition 5.2.3. Let $M = (F, I)$ be an abductive argumentation framework. An abducible argumentation framework $G \in I$ *explains* sceptical (resp. credulous) support for an observation B iff G sceptically (resp. credulously) supports B.

One can focus on explanations satisfying additional criteria, such as minimality w.r.t. the added or removed arguments/attacks. We leave the formal treatment of such criteria for future work.

Example 5.2.1. *Let $M = (F, \{F, G_1, G_2, G_3\})$, where F, G_1, G_2 and G_3 are defined as shown in figure 5.1. Let $B = \{b\}$ be an observation. It holds that G_1 and G_3 both explain sceptical support for B, while G_2 only explains credulous support for B.*

5.3 Explanation Dialogues

In this section we present methods to determine, *given an abductive argumentation framework $M = (F, I)$ (for $F = (A, \rightsquigarrow)$) whether an abducible argumentation framework $G \in I$ explains credulous or sceptical support for an observation $B \subseteq A$.* We build on ideas behind the *grounded* and *preferred games*, which are dialogical procedures that determine sceptical or credulous acceptance of an argument [68]. To sketch the idea behind these games (for a detailed discussion cf. [68]): two imaginary players (PRO and OPP) take alternating turns in putting forward arguments according to a set of rules, PRO either as an initial claim or in defence against OPP's attacks, while OPP initiates different disputes by attacking the arguments put forward by PRO. Sceptical or credulous

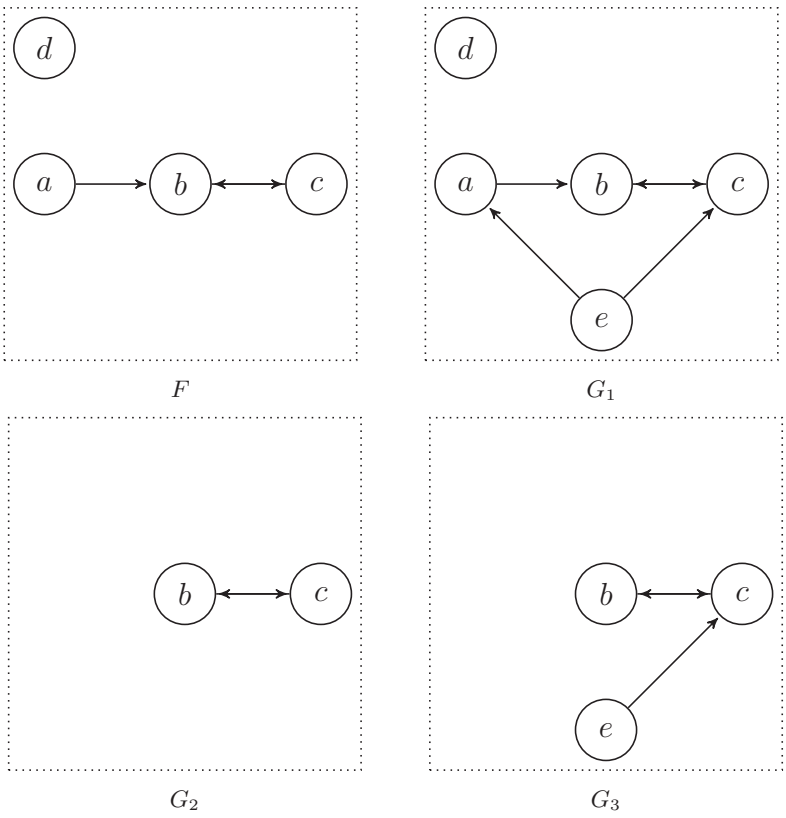


Figure 5.1: The argumentation frameworks of the abductive argumentation framework $(F, \{F, G_1, G_2, G_3\})$.

acceptance is proven if PRO can win the game by ending every dispute in its favour according to a "last-word" principle.

Our method adapts this idea so that the moves made by PRO are essentially *hypothetical* moves. That is, to defend the initial claim (i.e., to explain an observation) PRO can put forward, by way of hypothesis, any attack $x \rightsquigarrow y$ present in some $G \in I$. This marks a choice of PRO to focus only on those abducible argumentation frameworks in which the attack $x \rightsquigarrow y$ is present. Similarly, PRO can reply to an attack $x \rightsquigarrow y$, put forward by OPP, with the claim that this attack is invalid, marking the choice of PRO to focus only on the abducible argumentation frameworks in which the attack $x \rightsquigarrow y$ is *not* present. Thus, each move by PRO narrows down the set of abducible argumentation frameworks in which all of PRO's moves are valid. The objective is to end the dialogue with a non-empty set of abducible argumentation frameworks. Such a dialogue represents a proof that these abducible argumentation frameworks explain sceptical or credulous support for the observation.

Alternatively, such dialogues can be seen as games that determine sceptical or credulous support of an observation by an argumentation framework that are played simultaneously over all abducible argumentation frameworks in the abductive argumentation framework. In this view, the objective is to end the dialogue in such a way that it represents a proof for at least one abducible argumentation framework. Indeed, in the case where $M = (F, \{F\})$, our method reduces simply to a proof theory for sceptical or credulous support of an observation by F.

Before we move on we need to introduce some notation.

Definition 5.3.1. Given a set I of argumentation frameworks we define:

- $A_I = \cup\{A \mid (A, \rightsquigarrow) \in I\}$,
- $\rightsquigarrow_I = \cup\{\rightsquigarrow \mid (A, \rightsquigarrow) \in I\}$,
- $I_{x \rightsquigarrow y} = \{(A, \rightsquigarrow) \in I \mid x, y \in A, x \rightsquigarrow y\}$,
- $I_B = \{(A, \rightsquigarrow) \in I \mid B \subseteq A\}$.

We model dialogues as sequences of *moves*, each move being of a certain type, and made either by PRO or OPP.

Definition 5.3.2. Let $M = (F, I)$ be an abductive argumentation framework. A *dialogue based on* M is a sequence $S = (m_1, \ldots, m_n)$, where each m_i is either:

- an *OPP attack* "**OPP:** $x \rightsquigarrow y$", where $x \rightsquigarrow_I y$,
- a *hypothetical PRO defence* "**PRO:** $y \rightsquigarrow^+ x$", where $y \rightsquigarrow_I x$,
- a *hypothetical PRO negation* "**PRO:** $y \rightsquigarrow^- x$", where $y \rightsquigarrow_I x$,
- a *conceding move* "**OPP: ok**",
- a *success claim move* "**PRO: win**".

We denote by $S \cdot S'$ the concatenation of S and S'.

Intuitively, a move **OPP:** $y \rightsquigarrow x$ represents an attack by OPP on the argument x by putting forward the attacker y. A hypothetical PRO defence **PRO:** $y \rightsquigarrow^+ x$ represents a defence by PRO who puts forward y to attack the argument x put forward by OPP. A hypothetical PRO negation **PRO:** $y \rightsquigarrow^- x$, on the other hand, represents a claim by PRO that the attack $y \rightsquigarrow x$ is *not* a valid attack. The conceding move **OPP: ok** is made whenever OPP runs out of possibilities to attack a given argument, while the move **PRO: win** is made when PRO is able to claim success.

In the following sections we specify how dialogues are structured. Before doing so, we introduce some notation that we use to keep track of the abducible argumentation frameworks on which PRO chooses to focus in a dialogue D. We call this set the *information state* of D after a given move. While it initially contains all abducible argumentation frameworks in M, it is restricted when PRO makes a move **PRO:** $x \rightsquigarrow^+ y$ or **PRO:** $x \rightsquigarrow^- y$.

Definition 5.3.3. Let $M = (F, I)$ be an abductive argumentation framework. Let $D = (m_1, \ldots, m_n)$ be a dialogue based on M. We denote the *information state in D after move i* by $J(D, i)$, which is defined recursively by:

$$J(D,i) = \begin{cases} I & \text{if } i = 0, \\ J(D, i-1) \cap I_{x \rightsquigarrow y} & \text{if } m_i = \textbf{PRO:}\ x \rightsquigarrow^+ y, \\ J(D, i-1) \setminus I_{x \rightsquigarrow y} & \text{if } m_i = \textbf{PRO:}\ x \rightsquigarrow^- y, \\ J(D, i-1) & \text{otherwise.} \end{cases}$$

We denote by $J(D)$ the information state $J(D, n)$.

5.3.1 Sceptical Explanation Dialogues

We define the rules of a dialogue using a set of production rules that recursively define the set of sequences constituting dialogues. In a sceptical explanation dialogue for an observation B, an initial argument $x \in B$ is challenged by the opponent, who puts forward all possible attacks **OPP:** $y \rightsquigarrow x$ present in any of the abducible argumentation frameworks, followed by **OPP: ok**. We call this a *sceptical OPP reply* to x. For each move **OPP:** $y \rightsquigarrow x$, PRO responds with a *sceptical PRO reply to* $y \rightsquigarrow x$, which is either a hypothetical defence **PRO:** $z \rightsquigarrow^+ y$ (in turn followed by a sceptical OPP reply to z) or a hypothetical negation **PRO:** $y \rightsquigarrow^- x$. Formally:

Definition 5.3.4 (Sceptical explanation dialogue)**.** Let $F = (A, \rightsquigarrow)$, $M = (F, I)$ and $x \in A$.

- A *sceptical OPP reply to x* is a finite sequence (**OPP:** $y_1 \rightsquigarrow x) \cdot S_1 \cdot \ldots \cdot$ (**OPP:** $y_n \rightsquigarrow x) \cdot S_n \cdot$ (**OPP: ok**) where $\{y_1, \ldots, y_n\} = \{y \mid y \rightsquigarrow_I x\}$ and each S_i is a sceptical PRO reply to $y_i \rightsquigarrow x$.
- A *sceptical PRO reply to $y \rightsquigarrow x$* is either: (1) A sequence (**PRO:** $z \rightsquigarrow^+ y) \cdot S$ where $z \rightsquigarrow_I y$ and where S is a sceptical OPP reply to z, or (2) The sequence (**PRO:** $y \rightsquigarrow^- x$).

Given an observation $B \subseteq A$ we say that M generates the *sceptical explanation*

dialogue D for B iff $D = S \cdot (\textbf{PRO: win})$, where S is a sceptical OPP reply to some $x \in B$.

The following theorem establishes soundness and completeness.

Theorem 5.3.1. *Let $M = (F, I)$ be an abductive argumentation framework where $F = (A, \rightsquigarrow)$. Let $B \subseteq A$ and $G \in I$. It holds that G explains sceptical support for B iff M generates a sceptical explanation dialogue D for B such that $G \in J(D)$.*

The proof requires the following definitions and results. We define the *degree* of an argument x that is a member of the grounded extension to be the number of times that the characteristic function must be applied in order to obtain x.

Definition 5.3.5. Given an argumentation framework $F = (A, \rightsquigarrow)$ we define the *degree* $Deg_F(x)$ of an argument x that is a member of the grounded extension of F to be the smallest positive integer n s.t. $x \in \mathcal{D}_F^n(\emptyset)$.

The following lemma establishes an important relationship between the degree of an argument and the degrees of its defenders.

Lemma 5.3.2. *Let $F = (A, \rightsquigarrow)$ be an argumentation framework and $x \in A$ an argument that is a member of the grounded extension of F. For every $y \in A$ s.t. $y \rightsquigarrow x$ there is a $z \in A$ that is a member of the grounded extension of F such that $z \rightsquigarrow y$ and $Deg_F(z) < Deg_F(x)$.*

Proof of lemma 5.3.2. Let $F = (A, \rightsquigarrow)$, $x \in A$ a member of the grounded extension of F and $y \in A$ an argument s.t. $y \rightsquigarrow x$. Definition 2.1.4 implies that there is a $z \in A$ that is a member of the grounded extension of F s.t. $z \rightsquigarrow y$. Definition 2.1.2 furthermore implies that for every $B \subseteq A$, if $x \in \mathcal{D}_F(B)$ then $z \in B$. Definition 5.3.5 now implies that $Deg_F(x) > Deg_F(z)$. □

Proof of theorem 5.3.1. Let $M = (F, I)$ be an abductive argumentation framework where $F = (A, \rightsquigarrow)$. Let $B \subseteq A$ and $G \in I$.

Only if: Assume that G explains sceptical support for B. Proposition 5.2.1 implies that there is an $x \in B$ such that x is a member of the grounded extension of G. We prove that M generates a sceptical OPP reply D to x such that $G \in J(D)$. We prove this by strong induction on $Deg_G(x)$.

Let the induction hypothesis $H(i)$ stand for: *If x is a member of the grounded extension of G and $Deg_G(x) = i$ then there is a sceptical OPP reply D to x s.t. $G \in J(D)$.*

Base case ($H(0)$): this follows immediately, because this means that x has no attackers, and hence (**OPP: ok**) is a sceptical OPP reply to x.

Induction step: Assume $H(i)$ holds for all $0 < i < k$. We prove $H(k)$. Assume x is a member of the grounded extension of G and $Deg_G(x) = k$. We construct an OPP reply D to x such that $G \in J(D)$. Given an argument $y \in A_G$ s.t. $y \rightsquigarrow_G x$ we define $Z(y)$ by

$$Z(y) = \{z \mid z \rightsquigarrow_G y, z \text{ is a member of the grounded extension of } G\}.$$

i	m_i	$J(D,i)$
1	**OPP:** $c \rightsquigarrow b$	$\{F, G_1, G_2, G_3\}$
2	**PRO:** $e \rightsquigarrow^+ c$	$\{G_1, G_3\}$
3	**OPP: ok**	$\{G_1, G_3\}$
4	**OPP:** $a \rightsquigarrow b$	$\{G_1, G_3\}$
5	**PRO:** $e \rightsquigarrow^+ a$	$\{G_1\}$
6	**OPP: ok**	$\{G_1\}$
7	**OPP: ok**	$\{G_1\}$
8	**PRO: win**	$\{G_1\}$

Table 5.1: A sceptical explanation dialogue for the observation $\{b\}$.

Definition 2.1.4 implies that for every $y \in A_G$ s.t. $y \rightsquigarrow x$, $Z(y) \neq \emptyset$. Furthermore lemma 5.3.2 implies that for every $y \in A_G$ s.t. $y \rightsquigarrow x$ and for every $z \in Z(y)$ it holds that $Deg_G(z) < k$. We can now define D by $D = D_1 \cdot D_2 \cdot$ (**OPP: ok**) where: $D_1 =$ (**OPP:** $y_1 \rightsquigarrow x$) $\cdot$ (**PRO:** $y_1 \rightsquigarrow^- x$) $\cdot \ldots \cdot$ (**OPP:** $y_n \rightsquigarrow x$) $\cdot$ (**PRO:** $y_n \rightsquigarrow^- x$) where $\{y_1, \ldots, y_n\} = \{y \in A_I \mid y \rightsquigarrow_I x, y \not\rightsquigarrow_G x\}$, and $D_2 =$ (**OPP:** $y'_1 \rightsquigarrow x$) $\cdot$ (**PRO:** $z_1 \rightsquigarrow^+ y'_1$) $\cdot D_{z_1} \cdot \ldots \cdot$ (**OPP:** $y'_m \rightsquigarrow x$) $\cdot$ (**PRO:** $z_m \rightsquigarrow^+ y'_m$) $\cdot D_{z_m}$ where $\{y'_1, \ldots, y'_m\} = \{y \in A_I \mid y \rightsquigarrow_G x\}$, for each $j \in \{1, \ldots, m\}$, $z_j \in Z(y_j)$ and D_{z_j} is a sceptical OPP reply to z_j (because $Deg_G(z_j) < k$ and $H(i)$ holds for all $0 < i < k$, this sceptical OPP reply exists). It holds that D is a sceptical OPP reply to x. Furthermore it holds that $G \in J(D_1)$ and $G \in J(D_2)$ and hence $G \in J(D)$.

By the principle of strong induction it follows that there exists a sceptical OPP reply D to x such that $G \in J(D)$. Hence M generates a sceptical explanation dialogue $D \cdot$ (**PRO: win**) for B such that $G \in J(D \cdot$ (**PRO: win**)).

If: We prove that if D is a sceptical OPP reply to some $x \in B$ such that $G \in J(D)$ then x is a member of the grounded extension of G. We prove this by induction on the structure of D.

Assume that for every proper subsequence D' of D that is a sceptical OPP reply to an argument z it holds that z is a member of the grounded extension of G and $G \in J(D)$. (The base case is the special case where no proper subsequence of D is a sceptical OPP reply.) We prove that x is a member of the grounded extension of G. We write D as (**OPP:** $y_1 \rightsquigarrow x$) $\cdot D_1 \cdot \ldots \cdot$ (**OPP:** $y_n \rightsquigarrow x$) $\cdot D_n \cdot$ (**OPP: ok**). Then every D_i (for $1 \leq i \leq n$) is either of the form **PRO:** $y_i \rightsquigarrow^- x$ or of the form **PRO:** $z \rightsquigarrow^+ y_i \cdot D'$, where D' is a proper subsequence of D that is a sceptical OPP reply to some argument z and $G \in J(D')$. Thus, for every $y \in A_I$ s.t. $y \rightsquigarrow_I x$ it holds that either $y \not\rightsquigarrow_G x$, or y is attacked by some z that is a member of the grounded extension of G. It follows that x is a member of the grounded extension of G.

By the principle of induction it follows that if D is a sceptical OPP reply to some $x \in B$ such that $G \in J(D)$ then x is a member of the grounded extension of G. Thus, if M generates a sceptical explanation dialogue $D \cdot$ (**PRO: win**) for B such that $G \in J(D \cdot$ (**PRO: win**)) then D is a sceptical OPP reply to some $x \in B$ and therefore it holds that x is a member of the grounded extension of G and finally that G explains sceptical support for B. □

i	m_i	$J(D,i)$
1	**OPP:** $c \rightsquigarrow b$	$\{F, G_1, G_2, G_3\}$
2	**PRO:** $e \rightsquigarrow^+ c$	$\{G_1, G_3\}$
3	**OPP: ok**	$\{G_1, G_3\}$
4	**OPP:** $a \rightsquigarrow b$	$\{G_1, G_3\}$
5	**PRO:** $a \rightsquigarrow^- b$	$\{G_3\}$
6	**OPP: ok**	$\{G_3\}$
7	**PRO: win**	$\{G_3\}$

Table 5.2: A sceptical explanation dialogue for the observation $\{b\}$.

Example 5.3.1. *Table 5.1 shows an example of a sceptical explanation dialogue* $D = \{m_1, \ldots, m_8\}$ *for the observation* $\{b\}$ *that is generated by the abductive argumentation framework defined in example 5.2.1. The sequence* $(m_1, \ldots, m_7)$ *is a sceptical OPP reply to b, in which OPP puts forward the two attacks* $c \rightsquigarrow b$ *and* $a \rightsquigarrow b$*. PRO defends b from both c and a by putting forward the attacker e (move 2 and 5). This leads to the focus first on the abducible argumentation frameworks* G_1, G_3 *(in which the attack* $e \rightsquigarrow c$ *exists) and then on* G_1 *(in which the attack* $e \rightsquigarrow a$ *exists). This proves that* G_1 *explains sceptical support for the observation* $\{b\}$*.*

Another dialogue is shown in table 5.2. Here, PRO defends b from c by using the argument e, but defends b from a by claiming that the attack $a \rightsquigarrow b$ *is invalid. This leads to the focus first on the abducible argumentation frameworks* G_1, G_3 *(in which the attack* $e \rightsquigarrow c$ *exists) and then on* G_3 *(in which the attack* $a \rightsquigarrow b$ *does not exist). This dialogue proves that* G_3 *explains sceptical support for* $\{b\}$*.*

5.3.2 Credulous Explanation Dialogues

The definition of a credulous explanation dialogue is similar to that of a sceptical one. The difference lies in what constitutes an acceptable defence. To show that an argument x is sceptically accepted, x must be defended from its attackers by arguments other than x itself. For credulous acceptance, however, it suffices to show that x is a member of an admissible set, and hence x may be defended from its attackers by any argument, including x itself. To achieve this we need to keep track of the arguments that are, according to the moves made by PRO, accepted. Once an argument x is accepted, PRO does not need to defend x again, if this argument is put forward a second time.

Formally a *credulous OPP reply* to (x, Z) (for some $x \in A_I$ and set $Z \subseteq A_I$ used to keep track of accepted arguments) consists of all possible attacks **OPP:** $y \rightsquigarrow x$ on x, followed by **OPP: ok** when all attacks have been put forward. For each move **OPP:** $y \rightsquigarrow x$, PRO responds either by putting forward a hypothetical defence **PRO:** $z \rightsquigarrow^+ y$ which (this time *only if* $z \notin Z$) is followed by a credulous OPP reply to $(z, Z \cup \{z\})$, or by putting forward a hypothetical negation **PRO:** $y \rightsquigarrow^- x$. We call this response a *credulous PRO reply to* $(y \rightsquigarrow x, Z)$. A credulous explanation dialogue for a set B consists of a credulous OPP reply to $(x, \{x\})$ for some $x \in B$, followed by a success claim **PRO: win**.

In addition, arguments put forward by PRO in defence of the observation may

not conflict. Such a conflict occurs when OPP puts forward **OPP:** $x \rightsquigarrow y$ and **OPP:** $y \rightsquigarrow z$ (indicating that both y and z are accepted) while PRO does not put forward **PRO:** $y \rightsquigarrow^{-} z$. If this situation does not occur we say that the dialogue is *conflict-free*.

Definition 5.3.6 (Credulous explanation dialogue)**.** Let $F = (A, \rightsquigarrow)$, $M = (F, I)$, $x \in A$ and $Z \subseteq A$.

- A *credulous OPP reply to* (x, Z) is a finite sequence (**OPP:** $y_1 \rightsquigarrow x) \cdot S_1 \cdot \ldots \cdot$ (**OPP:** $y_n \rightsquigarrow x) \cdot S_n \cdot$ (**OPP: ok**) where $\{y_1, \ldots, y_n\} = \{y \mid y \rightsquigarrow_I x\}$ and each S_i is a credulous PRO reply to $(y_i \rightsquigarrow x, Z)$.
- A *credulous PRO reply to* $(y \rightsquigarrow x, Z)$ is either:
  1. a sequence (**PRO:** $z \rightsquigarrow^{+} y) \cdot S$ such that $z \rightsquigarrow_I y$, $z \notin Z$ and S is a credulous OPP reply to $(z, Z \cup \{z\})$,
  2. a sequence (**PRO:** $z \rightsquigarrow^{+} y$) such that $z \rightsquigarrow_I y$ and $z \in Z$, or
  3. the sequence (**PRO:** $y \rightsquigarrow^{-} x$).

Given a set $B \subseteq A$ we say that M generates the *credulous explanation dialogue* D for B iff $D = S \cdot$ (**PRO: win**), where S is a credulous OPP reply to $(x, \{x\})$ for some $x \in B$. We say that D is *conflict-free* iff for all $x, y, z \in A_I$ it holds that if D contains the moves **OPP:** $x \rightsquigarrow y$ and **OPP:** $y \rightsquigarrow z$ then it contains the move **PRO:** $y \rightsquigarrow^{-} z$.

The following theorem establishes soundness and completeness.

Theorem 5.3.3. *Let $M = (F, I)$ be an abductive argumentation framework where $F = (A, \rightsquigarrow)$. Let $B \subseteq A$ and $G \in I$. It holds that G explains credulous support for B iff M generates a conflict-free credulous explanation dialogue D for B such that $G \in J(D)$.*

Proof of theorem 5.3.3. Let $M = (F, I)$ be an abductive argumentation framework where $F = (A, \rightsquigarrow)$. Let $B \subseteq A$ and $G \in I$.

Only if: Assume that G explains credulous support for B. Then there is an admissible set E of G such that $a \in E$ for some $a \in B$. Based on E and a we construct a conflict-free credulous explanation dialogue D for B such that $G \in J(D)$. Given an argument $x \in E$ we define the credulous OPP reply $D(x, Z)$ recursively by $D(x, Z) =$ (**OPP:** $y_1 \rightsquigarrow x) \cdot S_1 \cdot \ldots \cdot$ (**OPP:** $y_n \rightsquigarrow x) \cdot S_n \cdot$ (**OPP: ok**)

where $\{y_1, \ldots, y_n\} = \{y \mid y \rightsquigarrow_I x\}$ and each S_i is a credulous PRO reply defined by the following cases:

- Case 1: $y_i \rightsquigarrow_G x$. Let z be an argument such that $z \in E$ and $z \rightsquigarrow_G y_i$. (Admissibility of E guarantees the existence of z.)
  - Case 1.1: $z \notin Z$: Then $S_i =$ **PRO:** $z \rightsquigarrow^{+} y_i \cdot D(z, Z \cup \{z\})$.
  - Case 1.2: $z \in Z$: Then $S_i =$ **PRO:** $z \rightsquigarrow^{+} y_i$.
- Case 2: $y_i \not\rightsquigarrow_G x$: Then $S_i =$ **PRO:** $y_i \rightsquigarrow^{-} x$.

Let $D = (m_1, \ldots, m_n) = D(a, \{a\}) \cdot (\textbf{PRO: win})$. It can be checked that D is a credulous explanation dialogue for $\{a\}$. We need to prove that:

- $G \in J(D)$. This follows from the fact that for all $i \in \{1, \ldots, n\}$, $m_i = \textbf{PRO:}\, x \rightsquigarrow^- y$ only if $x \not\rightsquigarrow_G y$ and $m_i = \textbf{PRO:}\, x \rightsquigarrow^+ y$ only if $x \rightsquigarrow_G y$.
- D is finite. This follows from the fact that for every credulous OPP reply $D(x, Z)$ that is a subsequence of a credulous OPP reply $D(y, Z')$ it holds that Z is a strict superset of Z', together with the fact that $Z \subseteq A_I$ and A_I is finite.
- D is conflict-free. We prove this by contradiction. Thus we assume that for some x, y, z there are moves $\textbf{OPP:}\, x \rightsquigarrow y$ and $\textbf{OPP:}\, y \rightsquigarrow z$ and no move $\textbf{PRO:}\, y \rightsquigarrow^- z$. By the construction of D it follows that $y, z \in E$. Furthermore if $y \not\rightsquigarrow_G z$ then by the construction of D, the move $\textbf{OPP:}\, y \rightsquigarrow z$ is followed by $\textbf{PRO:}\, y \rightsquigarrow^- z$, which is a contradiction. Hence $y \rightsquigarrow_G z$. Thus E is not a conflict-free set of G, contradicting our assumption that E is an admissible set of G. Hence D is conflict-free.

Hence there is a conflict-free credulous explanation dialogue D for B such that $G \in J(D)$.

If: Let D be a conflict-free credulous explanation dialogue for an observation B such that $G \in J(D)$. We prove that there is an admissible set E of G s.t. $a \in E$ for some $a \in B$. We define E by $E = \{a\} \cup \{x \mid \textbf{PRO:}\, x \rightsquigarrow^+ z \in D\}$. To prove that E is an admissible set of G we show that (1) for every $x \in E$ and every $y \in A$ such that $y \rightsquigarrow_G x$, there is a $z \in E$ such that $y \rightsquigarrow_G z$ and (2) that E is a conflict-free set of G.

1. Let $x \in E$. Then either $x = a$ or there is a move $m_i = \textbf{PRO:}\, x \rightsquigarrow^+ y$ in D. It follows either that m_{i+1} is a credulous OPP reply to (x, Z) or not, in which case there is a move m_j (for $j < i$) that is a credulous OPP reply to (x, Z). Hence for some $Z \subseteq A_I$ there is an OPP reply to (x, Z) in D. For m_{i+1} there are two cases:
   - $m_{i+1} = \textbf{PRO:}\, z \rightsquigarrow^+ y$. Then $z \in E$ and, because $G \in J(D)$, $z \rightsquigarrow_G y$.
   - $m_{i+1} = \textbf{PRO:}\, y \rightsquigarrow^- x$. But $y \rightsquigarrow_G x$, hence $G \notin J(D)$, which is a contradiction. Thus, this case is not possible.

   Thus for every $x \in E$ and every $y \in A$ s.t. $y \rightsquigarrow_G x$, there is a $z \in E$ such that $z \rightsquigarrow_G y$.
2. Assume the contrary, i.e., E is not conflict-free. Then for some $y, z \in E$ it holds that $y \rightsquigarrow_G z$. From (1) it follows that there is also an $x \in E$ such that $x \rightsquigarrow_G y$. By the construction of E it follows that either $y = a$ or for some x' there is a move $\textbf{PRO:}\, y \rightsquigarrow^+ x'$ in D, and similarly either $z = a$ or for some x' there is a move $\textbf{PRO:}\, z \rightsquigarrow^+ x'$ in D. Hence there are moves $\textbf{OPP:}\, x \rightsquigarrow y$ and $\textbf{OPP:}\, y \rightsquigarrow z$ in D. From the fact that $G \in J(D)$ and $y \rightsquigarrow_G z$ it follows that there is no move $\textbf{PRO:}\, y \rightsquigarrow^- z$ in D. Hence D is not conflict-free, which is a contradiction. It follows that E is a conflict-free set of G.

i	m_i	$J(D,i)$
1	**OPP:** $c \rightsquigarrow b$	$\{F, G_1, G_2, G_3\}$
2	**PRO:** $b \rightsquigarrow^+ c$	$\{F, G_1, G_2, G_3\}$
3	**OPP:** $a \rightsquigarrow b$	$\{F, G_1, G_2, G_3\}$
4	**PRO:** $a \rightsquigarrow^- b$	$\{G_2, G_3\}$
5	**OPP: ok**	$\{G_2, G_3\}$
6	**PRO: win**	$\{G_2, G_3\}$

Table 5.3: A credulous explanation dialogue for the observation $\{b\}$.

It finally follows that E is an admissible set of G and $a \in E$ and hence G explains credulous support for B.

□

Example 5.3.2. *Table 5.3 shows a conflict-free credulous explanation dialogue $D = (m_1, \ldots, m_6)$ for the observation $\{b\}$ generated by the abductive argumentation framework defined in example 5.2.1. Here, the sequence $(m_1, \ldots, m_5)$ is a credulous OPP reply to $(b, \{b\})$. PRO defends b from OPP's attack $c \rightsquigarrow b$ by putting forward the attack $b \rightsquigarrow c$. Since b was already assumed to be accepted, this suffices. At move m_4, PRO defends itself from the attack $a \rightsquigarrow b$ by negating it. This restricts the focus on the abducible argumentation frameworks G_2 and G_3. The dialogue proves that these two abducible argumentation frameworks explain credulous support for the observation $\{b\}$. Finally, the sceptical explanation dialogues from example 5.3.1 are also credulous explanation dialogues.*

5.4 Abduction in Logic Programming

In this section we show that abductive argumentation frameworks can be instantiated with abductive logic programs, in the same way that regular argumentation frameworks can be instantiated with regular logic programs. In sections 5.4.1 and 5.4.2 we recall the necessary basics of logic programming and the relevant results regarding logic programming as instantiated argumentation. In section 5.4.3 we present a model of abductive logic programming based on Sakama and Inoue's model of extended abduction [57, 58], and in section 5.4.2 we show how this model can be instantiated using abductive argumentation frameworks.

5.4.1 The Partial Stable Semantics of Logic Programs

A logic program P is a finite set of rules, each rule being of the form $C \leftarrow A_1, \ldots, A_n, \sim B_1, \ldots, \sim B_m$ where $C, A_1, \ldots, A_n, B_1, \ldots, B_m$ are *atoms*. If $m = 0$ then the rule is called *definite*. If both $n = 0$ and $m = 0$ then the rule is called a *fact* and we identify it with the atom C. We assume that logic programs are ground. Alternatively, P can be regarded as the set of ground instances of a set of non-ground rules. We denote by At_P the set of all (ground) atoms occurring in P. The logic programming semantics we focus on can be defined using *3-valued interpretations* [77]:

Definition 5.4.1. A 3-valued interpretation I of a logic program P is a pair $I = (T, F)$ where $T, F \subseteq At_P$ and $T \cap F = \emptyset$. An atom $A \in At(P)$ is *true* (resp. *false*, *undecided*) in I iff $A \in T$ (resp. $A \in F$, $A \in At_P \setminus (T \cup F)$).

The following definition of a *partial stable model* is due to Przymusinski [77]. Given a logic program P and 3-valued interpretation I of P, the *GL-transformation* $\frac{P}{I}$ is a logic program obtained by replacing in every rule in P every premise $\sim B$ such that B is true (resp. undecided, false) in I by the atoms 0 (resp. $\frac{1}{2}$, 1), where 0 (resp. $\frac{1}{2}$, 1) are defined to be false (resp. undecided, true) in every interpretation. It holds that for all 3-valued interpretations I of P, $\frac{P}{I}$ is definite (i.e., consists only of definite rules). This means that $\frac{P}{I}$ has a unique *least* 3-valued interpretation (T, F) with minimal T and maximal F that satisfies all rules. That is, for all rules $C \leftarrow A_1, \ldots, A_n$, in $\frac{P}{I}$, C is true (resp. *not* false) in (T, F) if for all $i \in \{1, \ldots, n\}$, A_i is true (resp. *not* false) in (T, F). Given a 3-valued interpretation I, the least 3-valued interpretation of $\frac{P}{I}$ is denoted by $\Gamma(I)$. This leads to the following definition of a *partial stable model* of a logic program, along with the associated notions of consequence.

Definition 5.4.2. [77] Let P be a logic program. A 3-valued interpretation I is a *partial stable model* of P iff $I = \Gamma(I)$. We say that an atom C is a sceptical (resp. credulous) consequence of P iff C is true in all (resp. some) partial stable models of P.

It has been shown that the above defined notion of sceptical consequence coincides with the *well-founded* semantics [77].

5.4.2 Logic Programming as Argumentation

Wu et al. [91] have shown that a logic program P can be transformed into an argumentation framework F in such a way that the consequences of P under the partial stable semantics can be computed by looking at the complete extensions of F. The idea is that an argument consists of a conclusion $C \in At_P$, a set of rules $R \subseteq P$ used to derive C and a set $N \subseteq At_P$ of atoms that must be underivable in order for the argument to be acceptable. The argument is attacked by another argument with a conclusion C' iff $C' \in N$. The following definition, apart from notation, is due to Wu et al. [91].

Definition 5.4.3. Let P be a logic program. An instantiated argument is a triple (C, R, N), where $C \in At_P$, $R \subseteq P$ and $N \subseteq At_P$. We say that P generates (C, R, N) iff either:

- $r = C \leftarrow \sim B_1, \ldots, \sim B_m$ is a rule in P, $R = \{r\}$ and $N = \{B_1, \ldots, B_m\}$.
- (1) $r = C \leftarrow A_1, \ldots, A_n, \sim B_1, \ldots, \sim B_m$ is a rule in P, (2) P generates, for each $i \in \{1, \ldots, n\}$ an argument (A_i, R_i, N_i) such that $r \notin R_i$, and (3) $R = \{r\} \cup R_1 \cup \ldots \cup R_n$ and $N = \{B_1, \ldots, B_m\} \cup N_1 \cup \ldots \cup N_n$.

We denote the set of arguments generated by P by A_P. Furthermore, the attack relation generated by P is denoted by $\rightsquigarrow_P$ and is defined by $(C, R, N) \rightsquigarrow_P (C', R', N')$ iff $C \in N'$.

The following theorem states that sceptical (resp. credulous) acceptance in $(A_P, \rightsquigarrow_P)$ corresponds with sceptical (resp. credulous) consequences in P as defined in definition 5.4.2. It follows from theorems 15 and 16 due to Wu et al. [91].

Theorem 5.4.1. *Let P be a logic program. An atom $C \in At_P$ is a sceptical (resp. credulous) consequence of P iff some $(C, R, N) \in A_P$ is sceptically (resp. credulously) accepted in $(A_P, \rightsquigarrow_P)$.*

5.4.3 Abduction in Logic Programming

The model of abduction in logic programming that we use is based on the model of *extended* abduction studied by Inoue and Sakama [57, 58]. They define an abductive logic program (ALP) to consist of a logic program and a set of atoms called *abducibles*.

Definition 5.4.4. An abductive logic program is a pair (P, U) where P is a logic program and $U \subseteq At_P$ a set of facts called abducibles.

Note that, as before, the set U consists of ground facts of the form $C \leftarrow$ (identified with the atom C) and can alternatively be regarded as the set of ground instances of a set of non-ground facts. A hypothesis, according to Inoue and Sakama's model, consists of both a positive element (i.e., abducibles added to P) and a negative element (i.e., abducibles removed from P).

Definition 5.4.5. Let $ALP = (P, U)$ be an abductive logic program. A hypothesis is a pair (Δ^+, Δ^-) such that $\Delta^+, \Delta^- \subseteq U$ and $\Delta^+ \cap \Delta^- = \emptyset$. A hypothesis (Δ^+, Δ^-) sceptically (resp. credulously) explains a query $Q \in At_P$ if and only if Q is a sceptical (resp. credulous) consequence of $(P \cup \Delta^+) \setminus \Delta^-$.

Note that Sakama and Inoue focus on computation of explanations under the stable model semantics of P, and require P to be acyclic to ensure that a stable model of P exists and is unique [58]. We, however, define explanation in terms of the consequences according to the partial stable models of P, which always exist even if P is not acyclic [77], so that we do not need this requirement.

The following example demonstrates the previous two definitions.

Example 5.4.1. *Let $ALP = (P, U)$ where $P = \{(p \leftarrow \sim s, r),\ (p \leftarrow \sim s, \sim q),\ (q \leftarrow \sim p), r\}$ and $U = \{r, s\}$. The hypothesis $(\{s\}, \emptyset)$ sceptically explains q, witnessed by the unique model $I = (\{r, s, q\}, \{p\})$ satisfying $I = \Gamma(I)$. Similarly, $(\{s\}, \{r\}))$ sceptically explains q and $(\emptyset, \{r\}))$ credulously explains q.*

5.4.4 Instantiated Abduction in Argumentation

In this section we show that an abductive argumentation framework (F, I) can be instantiated on the basis of an abductive logic program (P, U). The idea is that every possible hypothesis (Δ^+, Δ^-) maps to an abducible argumentation framework generated by the logic program $(P \cup \Delta^+) \setminus \Delta^-$. The hypotheses for a query Q then correspond to the abducible argumentation frameworks that explain the observation B consisting of all arguments with conclusion Q. The construction of (F, I) on the basis of (P, U) is defined as follows.

Definition 5.4.6. Let $ALP = (P, U)$ be an abductive logic program. Given a hypothesis (Δ^+, Δ^-), we denote by $F_{(\Delta^+,\Delta^-)}$ the argumentation framework $(A_{(P\cup\Delta^+)\setminus\Delta^-}, \rightsquigarrow_{(P\cup\Delta^+)\setminus\Delta^-})$. The abductive argumentation framework *generated by ALP* is denoted by M_{ALP} and defined by $M_{ALP} = (F_P, I_{ALP})$, where $I_{ALP} = \{F_{(\Delta^+,\Delta^-)} \mid \Delta^+, \Delta^- \subseteq U, \Delta^+ \cap \Delta^- = \emptyset\}$.

The following theorem states the correspondence between the explanations of a query Q in an abductive logic program ALP and the explanations of an observation B in the abductive argumentation framework M_{ALP}.

Theorem 5.4.2. *Let $ALP = (P, U)$ be an abductive logic program, $Q \in At_P$ a query and (Δ^+, Δ^-) a hypothesis. Let $M_{ALP} = (F_{ALP}, I_{ALP})$. We denote by X_Q the set $\{(C, R, N) \in A_P \mid C = Q\}$. It holds that (Δ^+, Δ^-) sceptically (resp. credulously) explains Q iff $F_{(\Delta^+,\Delta^-)}$ sceptically (resp. credulously) explains X_Q.*

Proof of theorem 5.4.2. Follows directly from theorem 5.4.1 and definitions 5.4.5 and 5.4.6. □

This theorem shows that our model of abduction in argumentation can indeed be seen as an abstraction of abductive logic programming.

Example 5.4.2. *Let $ALP = (P, U)$ be the ALP as defined in example 5.4.1. All arguments generated by ALP are:*

$$\begin{array}{lcllcl} a & = & (p, \{(p \leftarrow {\sim}s, r), r\}, \{s\}) & d & = & (r, \{r\}, \emptyset) \\ b & = & (q, \{(q \leftarrow {\sim}p)\}, \{p\}) & e & = & (s, \{s\}, \emptyset) \\ c & = & (p, \{(p \leftarrow {\sim}s, {\sim}q)\}, \{s, q\}) & & & \end{array}$$

Given these definitions, the abductive argumentation framework in example 5.2.1 is equivalent to M_{ALP}. In example 5.4.1 we saw that the query q is sceptically explained by the hypotheses $(\{s\}, \emptyset)$ and $(\{s\}, \{r\})$, while $(\emptyset, \{r\})$ only credulously explains it. Indeed, looking again at example 5.2.1, we see that $G_1 = F_{(\{s\},\emptyset)}$ and $G_3 = F_{(\{s\},\{r\})}$ explain sceptical support for the observation $\{b\} = X_q$, while $G_2 = F_{(\emptyset,\{r\})}$ only explains credulous support.

This method of instantiation shows that, on the abstract level, hypotheses cannot be represented by independently selectable abducible arguments. The running example shows e.g. that a and d cannot be added or removed independently.

5.5 Related Work

Some of the ideas we applied also appear in the model of Sakama [83]. In his model of abduction in argumentation, both additions and removals of arguments from an abstract argumentation framework act as explanations for the observation that an argument is accepted or rejected. The main difference between Sakama's model of abduction in abstract argumentation and the one presented here, is that he takes an explanation to be a set of independently selectable abducible arguments, while we take it to be a change to the argumentation framework that is applied as a whole. We have demonstrated, however,

that this is necessary when applying the abstract model in an instantiated setting. Furthermore, Sakama did address computation in his framework, but his method was based on translating abstract argumentation frameworks into logic programs. Sakama did not explore the instantiation of his model.

Some of the ideas we applied also appear in work by Wakaki et al. [90]. In their model, an ALP generates an instantiated argumentation framework and each hypothesis yields a different division into active/inactive arguments. Unlike our model, as well as Sakama's [83], Wakaki et al. do not consider removal of arguments as explanation.

Kontarinis et al. [61] use term rewriting logic to compute changes to an abstract argumentation framework with the goal of changing the status of an argument. There are two similarities between their approach and ours. Firstly, we use production rules to generate dialogues and these rules can be seen as a kind of term rewriting rules. Secondly, their approach amounts to rewriting goals into statements to the effect that certain attacks in the argumentation framework are enabled or disabled. These statements resemble the moves **PRO:** $x \rightsquigarrow^{+} y$ and **PRO:** $x \rightsquigarrow^{-} y$ in our system. However, they treat attacks as entities that can be enabled or disabled independently. As discussed, different arguments (or in this case attacks associated with arguments) cannot be regarded as independent entities, if the abstract model is instantiated.

Other work dealing with the change of an argumentation framework with the goal of changing the status of arguments include Baumann [10], Baumann and Brewka [12], Bisquert et al. [17] and Perotti et al. [20]. Furthermore, Booth et al. [23] and Coste-Marquis et al. [37] frame it as a problem of *belief revision*. None of these works, however, make a connection between change of abstract argumentation and abduction.

5.6 Conclusion and Future Work

We developed a model of abduction in abstract argumentation, in which changes to an argumentation framework act as explanations for sceptical/credulous support for observations. We presented sound and complete dialogical proof procedures for the main reasoning tasks, i.e., finding explanations for sceptical/credulous support. In addition, we showed that our model of abduction in abstract argumentation can be seen as an abstract form of abduction in logic programming.

As a possible direction for future work, we consider the incorporation of additional criteria for the selection of good explanations, such as minimality with respect to the added and removed arguments/attacks, as well as the use of preferences over different abducible argumentation frameworks. An interesting question is whether the proof theory can be adapted so as to yield only the preferred explanations.

Chapter 6

Change in Preference-Based Argumentation

6.1 Introduction

Many works have recognized the importance of *preferences* in argumentation. Preferences over arguments may be derived, e.g., from their relative specificity or from the relative strength of the beliefs with which they are built. On the abstract level preferences can be represented by *preference-based* argumentation frameworks, which instantiate argumentation frameworks with a preference relation over the set of arguments [2, 86]. An attack of an argument x on y then *succeeds* only if y is not strictly preferred over x. *Value-based* argumentation frameworks provide yet another account of how preferences are derived [14]. The idea here is that arguments promote certain *values* and that different *audiences* have different preferences over values, from which the preferences over arguments are derived.

An underexposed aspect in these models is change of preferences. Preferences are usually assumed to be fixed and no account is provided of how or why they may change. We address this aspect by applying Dietrich and List's recently introduced model of *property-based preference* [41, 40]. In this model, preferences over alternatives are derived from preferences over sets of properties satisfied by the alternatives. Furthermore, agents are assumed to have a motivational state, consisting of the properties on which the agent focuses in a given situation, when forming preferences over alternatives. The authors present an axiomatic characterization of their model, in terms of a number of reasonable constraints on the relationship between motivational states and preferences.

Our contribution is a new, dynamic model of preferences in argumentation, centering on what we call *property-based* argumentation frameworks. It is based on the model of Dietrich and List and provides an account of how and why preferences in argumentation may change. Our model generalizes preference-based argumentation frameworks as well as value-based argumentation frameworks, if properties are used to represent values. We look at two types of acceptance,

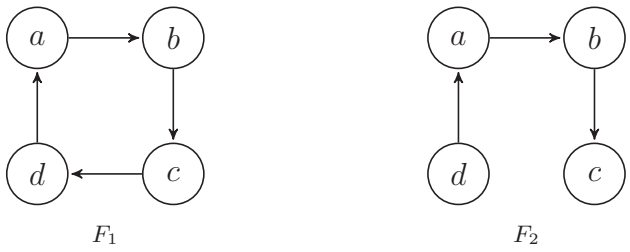


Figure 6.1: Two argumentation frameworks.

called *weak* and *strong* acceptance (i.e., acceptance in *some* or *all* motivational states). We also provide a dialogical proof theory that establishes whether an argument is weakly accepted. It is based on the grounded game [68] and extends it with dialogue moves consisting of properties.

The outline of this chapter is as follows. In section 6.2 we first give a brief outline of preference-based and value-based abstract argumentation. Then we give in section 6.3 an overview of the relevant parts of Dietrich and List's model of property-based preferences. We move on to our own work in section 6.4, where we present our model of property-based argumentation frameworks, followed by a dialogical proof procedure for weak acceptance in section 6.5. We discuss some related work in section 6.6 and we conclude in section 6.7.

The results presented in this chapter are based on joint work with Richard Booth and Souhila Kaci [22].

6.2 Preferences and Values in Argumentation

The idea of *Preference-based* argumentation frameworks [2] is to extend the notion of an argumentation framework with a preference relation over arguments, which is used to represent the relative strength of arguments. The idea is that an attack of an argument x on y succeeds only if y is not strictly preferred over (i.e., not stronger than) x. A preference-based argumentation framework *represents* a unique argumentation framework $(A, \rightsquigarrow)$, where the attack relation $\rightsquigarrow$ consists only of the attacks that succeed [60]. The extensions of a preference-based argumentation framework are those of the argumentation framework that it represents. Formally:

Definition 6.2.1. A preference-based argumentation framework (abbreviated as *PAF*) is a triple $PAF = (A, \rightarrow, \preceq)$ where A is a finite set of arguments, $\rightarrow$ an attack relation and $\preceq$ a partial pre-order (i.e., a reflexive and transitive relation) or a total pre-order (i.e., a reflexive, transitive and complete relation) over A. A *PAF* $(A, \rightarrow, \preceq)$ *represents* the argumentation framework $(A, \rightsquigarrow)$ where $\rightsquigarrow$ is defined by $\forall x, y \in A, x \rightsquigarrow y$ iff $x \rightarrow y$ and not $(x \prec y)$.

Example 6.2.1. *Consider the PAF* $(A, \rightarrow, \preceq)$ *where* A *and* $\rightarrow$ *are as in* F_1 *in figure 6.1 and* $\preceq$ *is a total pre-order defined by* $x \preceq y$ *iff* $x \in \{b, c\}$ *or* $y \in \{a, d\}$.

We have that $(A, \rightarrow, \preceq)$ *represents* F_2*, shown in figure 6.1. This argumentation framework has one complete, grounded, stable and preferred extension, namely* $\{d, b\}$*.*

Preference-based argumentation frameworks give—at least at the abstract level—no account of where preferences over arguments come from, or how they are formed. Bench-Capon's [14] model of *value-based* argumentation frameworks does. In a value-based argumentation framework, the idea is that arguments may promote certain *values* and that different *audiences* have different preferences over values, from which the preferences over arguments are derived. An *audience specific* value-based argumentation framework encodes a single audience's preferences over values.

Definition 6.2.2. A *value-based argumentation framework* (*VAF* for short) is a 5-tuple $(A, \rightarrow, V, val, U)$, where A is a set of arguments, $\rightarrow$ an attack relation, V a set of *values*, $val : A \rightarrow V$ a mapping from arguments to values and U a set of *audiences*. An *audience specific value-based argumentation framework* (*aVAF* for short) is a 5-tuple $(A, \rightarrow, V, val, <_a)$ where $a \in U$ is an audience and $<_a$ a partial order (i.e. an irreflexive and transitive relation) over V.

An *aVAF* represents a unique *PAF* [60]:

Definition 6.2.3. An *aVAF* $(A, \rightarrow, V, val, <_a)$ *represents* the *PAF* $(A, \rightarrow, \preceq)$, where $\preceq$ is defined by $\forall x, y \in A, x \preceq y$ iff $val(x) <_a val(y)$ or $val(x) = val(y)$.

Since a *PAF* represents a unique argumentation framework, an *aVAF* also represents a unique argumentation framework. The extensions of an *aVAF* are the extensions of this argumentation framework.

Example 6.2.2. *Consider the aVAF* $(A, \rightarrow, V, val, <_a)$ *where* A *and* $\rightarrow$ *are as shown in figure 6.1,* $V = \{blue, red\}$*,* $val(a) = val(d) = blue, val(b) = val(c) = red$ *and* $<_a$ *is defined by* $x <_a y$ *iff* $x = red$ *and* $y = blue$*. It can be checked that this aVAF represents the PAF from example 6.2.1 and thus the argumentation framework* F_2 *shown in figure 6.1.*

6.3 Dietrich and List's Property-Based Preference Model

Dietrich and List's model of *property-based preference* [41, 40] aims at giving an account of rational choice that explains how preferences are formed and how they may change. This is opposed to traditional models that assume an agent's preferences over alternatives to be given and fixed. In this model, every alternative $x \in X$ is associated with a set $P(x)$ of *properties* satisfied by x, each $P(x)$ being a subset of a set $\mathcal{P}$ of possible properties. Furthermore, a set $\mathcal{M} \subseteq 2^{\mathcal{P}}$ of *motivational states* encodes sets of properties on which an agent may focus in a given situation. That is, if $M \in \mathcal{M}$ is the agent's state then only the properties in M matter to the agent when forming preferences over X. Change of preferences can then be understood as being caused by moving from

one motivational state to another. Note that $\mathcal{M}$ may coincide with $2^{\mathcal{P}}$ but in general this need not be the case, as certain combinations of properties may be deemed inconsistent.

Every state $M \in \mathcal{M}$ gives rise to a preference order (i.e., a total pre-order) $\preceq_M$ over X representing the agent's preferences in the state M. There is thus a family $(\preceq_M)_{M\in\mathcal{M}}$ of preference orders over X. Strict and indifference relations $\prec_M$ and $\sim_M$ are defined as usual.

According to the model of property-based preference, preferences over X are formed using an underlying *weighing relation* $\leq$ over combinations of properties. This relation can be thought of as a 'betterness' relation, i.e., if $S \leq S'$ then the set of properties S' is at least as good as the set of properties S.

Definition 6.3.1. A family $(\preceq_M)_{M\in\mathcal{M}}$ of preference orders is called *property-based* if there is a *weighing relation* $\leq \subseteq 2^{\mathcal{P}} \times 2^{\mathcal{P}}$ such that, for every $M \in \mathcal{M}$ and $x, y \in X$, $x \preceq_M y$ iff $P(x) \cap M \leq P(y) \cap M$.

The authors present an axiomatic characterization of their model, in terms of two constraints on the relationship between motivational states and preferences.

Theorem 6.3.1. *[An axiomatic characterization [41]] Let $(\preceq_M)_{M\in\mathcal{M}}$ be a family of preference orders. Consider the following axioms:*

Axiom 1 *$\forall x, y \in X$, $\forall M \in \mathcal{M}$, if $P(x) \cap M = P(y) \cap M$, then $x \sim_M y$.*

Axiom 2 *$\forall x, y \in X$, $\forall M, M' \in \mathcal{M}$ s.t. $M \subseteq M'$, if $P(x) \cap (M' \setminus M) = P(y) \cap (M' \setminus M) = \emptyset$ then $x \preceq_M y \leftrightarrow x \preceq_{M'} y$.*

It holds that if $\mathcal{M}$ is intersection-closed (i.e. $M, M' \in \mathcal{M}$ implies $M \cap M' \in \mathcal{M}$) then a family of preference orders $(\preceq_M)_{M\in\mathcal{M}}$ satisfies axioms 1 and 2 iff it is property-based.

Axiom 1 says that the preference relation is indifferent on pairs of alternatives that have the same properties that are at the same time motivational, while axiom 2 says that preferences on pairs of alternatives change only if additional properties become motivational that are satisfied by at least one of the alternatives. A third axiom, strengthening the second and concerned with the class of *separable* weighing relations may be considered as well. The reader is referred to Dietrich and List [41] for details.

6.4 Property-Based Argumentation Frameworks

The value-based argumentation framework model gives an account of where an agent's (or audience's) preferences over arguments come from, namely the relative importance of the values they promote. However, it gives no account of how or why they may change. This motivates us to apply the model of property-based preference in argumentation, giving rise to what we call *property-based*

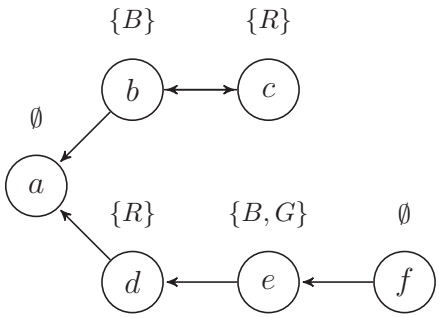


Figure 6.2: An argumentation framework with properties of arguments.

argumentation frameworks. In a property-based argumentation framework, each argument is associated with a set of properties that it satisfies. Among the types of properties we may consider are values promoted by the argument.

Furthermore, a property-based argumentation framework consists of a set of motivational states $\mathcal{M}$ and a weighing relation $\leq$ over sets of properties. The idea is as before: $\leq$ encodes the agent's preferences over sets of properties but only properties in the agent's state $M \in \mathcal{M}$ matter when forming preferences over arguments.

Definition 6.4.1. A *property-based argumentation framework* is a 6-tuple $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ where A is a set of arguments, $\rightarrow$ an attack relation, $\mathcal{P}$ is a set of properties, $P : A \rightarrow 2^{\mathcal{P}}$ a mapping of arguments to sets of properties, $\mathcal{M} \subseteq 2^{\mathcal{P}}$ is an intersection-closed set of motivational states and $\leq \subseteq 2^{\mathcal{P}} \times 2^{\mathcal{P}}$ a reflexive, transitive and complete weighing relation.

Note that there are cases where $\leq$ does not need to be transitive and complete over all sets of properties. For simplicity, however, we assume that it is. The reader is referred to Dietrich and List [41, Remark 1] for details.

If we focus on values as properties then the weighing relation can be understood as encoding the relative importance that an agent associates with different combinations of values, and the motivational state as consisting of the values of which an agent is aware in a given situation.

Given a property-based argumentation framework, each motivational state $M \in \mathcal{M}$ represents a unique *PAF* which we denote by PAF_M. Preferences in PAF_M are formed by comparing sets of properties satisfied by the arguments, that are at the same time motivational. The argumentation framework according to which the agent determines the extensions in the motivational state M, denoted by F_M, is the argumentation framework represented by PAF_M.

Definition 6.4.2. Given a property-based argumentation framework $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ and a motivational state $M \in \mathcal{M}$ we say that:

- M represents the $PAF_M = (A, \rightarrow, \preceq)$, where $\preceq$ is defined by $\forall x, y \in A, x \preceq y$ iff $P(x) \cap M \leq P(y) \cap M$.
- M represents the argumentation framework $F_M = (A, \rightsquigarrow_M)$, which is the argumentation framework represented by PAF_M.

Given an attack $x \to y$ and state $M \in \mathcal{M}$, we say that $x \to y$ is *enabled* (otherwise *disabled*) in M iff $x \rightsquigarrow_M y$.

Let us illustrate the definitions with an example.

Example 6.4.1. *Consider the property-based argumentation framework* $(A, \to, \mathcal{P}, P, \mathcal{M}, \leq)$ *where* A *and* $\to$ *and the properties assigned by* P *to the arguments are as shown in figure 6.2. Furthermore,* $\mathcal{P} = \{R, G, B\}$, $\mathcal{M} = 2^{\mathcal{P}}$ *and* $\leq$ *is defined via a weight function* $w : \mathcal{P} \to \mathbb{Z}$ *with* $w(R) = w(G) = 1$ *and* $w(B) = -2$ *as follows:* $X \leq X'$ *iff* $\sum_{x \in X} w(x) \leq \sum_{x \in X'} w(x)$. *This gives rise to the weighing relation* $\{B\} < \{R, B\} = \{G, B\} < \{R, G, B\} = \emptyset < \{R\} = \{G\} < \{R, G\}$, *where* $<$ *is the strict counterpart of* $\leq$.

Figure 6.3 shows the argumentation frameworks represented by all possible motivational states. We have, e.g., that in $PAF_{\{G\}}$ *the argument* e *is strictly preferred over* f, *so that the attack from* f *to* e *is disabled* $F_{\{G\}}$. *On the other hand, in* $PAF_{\emptyset}$ *and* $PAF_{\{B,G\}}$ *the argument* e *is not preferred over* f. *Here, the attack from* f *to* e *succeeds and is therefore enabled in* $F_{\emptyset}$ *and* $F_{\{B,G\}}$.

Arguments in the argumentation frameworks in figure 6.3 that are a member of the grounded extension of the respective argumentation frameworks are coloured white. We can see, e.g., that a *is accepted only in the motivational state* $\{R, G\}$.

We should remark that in many systems of argumentation, arguments have (in)formal 'logical content'. As a result, conflicts between arguments cannot generally be disregarded, on pain of inconsistency of the argumentation framework's outcome. This can be taken into account by requiring, for example, the relation $\to$ to be symmetric, representing a conflict relation over two arguments, i.e. both arguments cannot be accepted together. In this way one attack between a pair of arguments always remains enabled.

Apart from looking at acceptance of arguments in a given motivational state, we can look at acceptance of arguments in some or all possible states. We will say that an argument is *weakly* (resp. *strongly*) accepted iff it is a member of the grounded extension given some (resp. all) motivational states. Weak acceptance thus means that the agent may accept an argument, namely when she moves to the right motivational state, whereas strong acceptance means that an agent accepts an argument regardless of her motivational state.

Definition 6.4.3. Let $(A, \to, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $x \in A$ an argument. We say that x is *weakly accepted* (resp. *strongly accepted*) iff x is a member of the grounded extension of F_M for some (resp. all) $M \in \mathcal{M}$.

Example 6.4.2 (Continued from example 6.4.1)**.** *All arguments except* b *are weakly accepted. Only* f *is strongly accepted.*

The following properties follow directly from theorem 6.3.1.

Proposition 6.4.1. *Let* $(A, \to, \mathcal{P}, P, \mathcal{M}, \leq)$ *be a property-based argumentation framework. We have:*

Property 1 $\forall x, y \in A$ *s.t.* $x \to y$, $\forall M \in \mathcal{M}$ *s.t.* $P(x) \cap M = P(y) \cap M, x \rightsquigarrow_M y$.

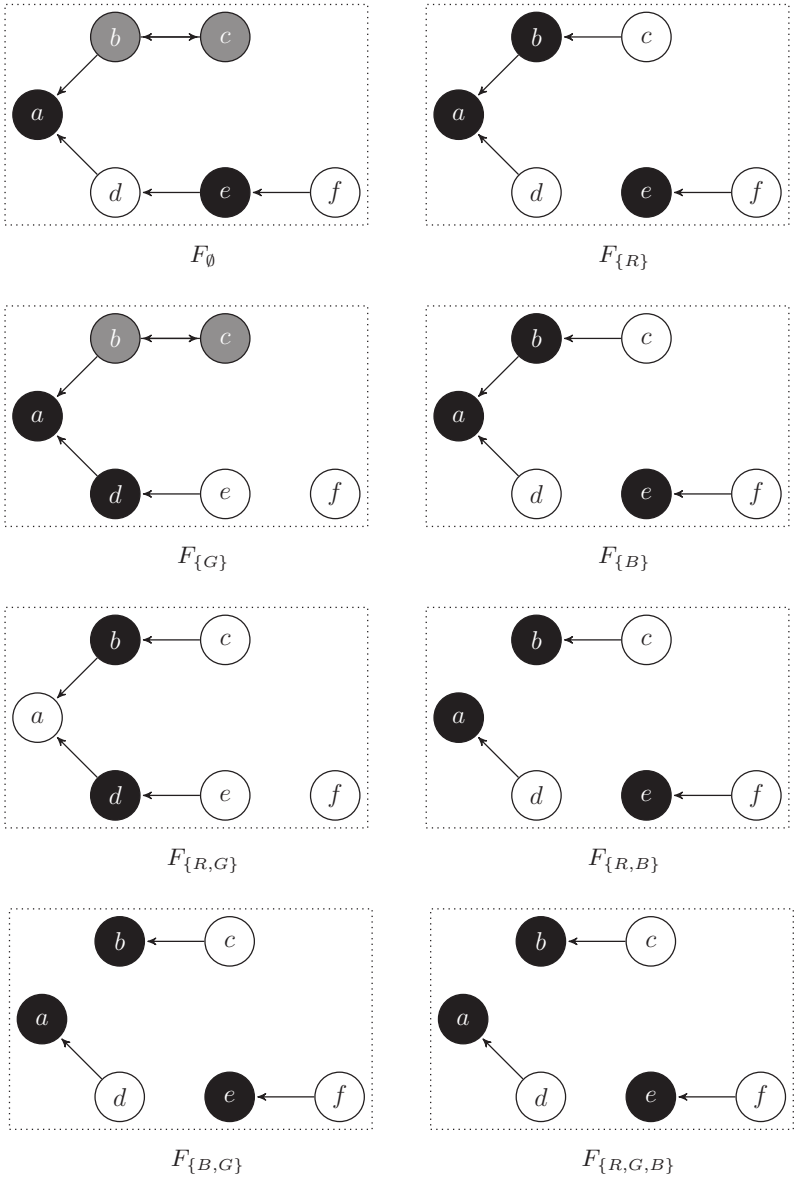


Figure 6.3: Different motivational states (example 6.4.1).

Property 2 *$\forall x, y \in A$, $\forall M, M' \in \mathcal{M}$ s.t. $M \subseteq M'$, if $P(x) \cap (M' \setminus M) = P(y) \cap (M' \setminus M) = \emptyset$ then $x \rightsquigarrow_M y$ iff $x \rightsquigarrow_{M'} y$.*

Property 1 states that an attack $x \to y$ is enabled in a motivational state M if x and y have the same set of properties that are also motivational in M, while property 2 states that an attack between x and y changes only if additional properties become motivational that are satisfied either by x or by y.

6.5 A Dialogical Proof Theory for Weak Acceptance

In this section we present a proof procedure to establish weak acceptance of an argument in a property-based argumentation framework. It is a dialogical proof procedure because it is based on generating dialogues where two players (PRO and OPP) take alternating turns in putting forward attacks according to a certain set of rules. This is similar in spirit to the *grounded game*, a dialogical proof procedure that establishes an argument's membership of the grounded extension [68]. In the grounded game, PRO repeatedly puts forward arguments (either as an initial claim or in defence against OPP's attacks) and OPP can initiate different disputes by putting forward possible attacks on the arguments put forward by PRO. PRO wins iff it can end every dispute in its favor according to a "last-word" principle.

By contrast, the proof procedure we present simply generates dialogues won by PRO. Such dialogues represent proofs that the initial argument is weakly accepted, and are structured as single sequences of moves where PRO and OPP put forward attacks and, in addition, PRO puts forward properties. If the procedure generates no dialogues then the argument is not weakly accepted. The procedure is based essentially on production rules, and is in this sense similar to the dialogical proof procedures that we presented in chapter 5.

Dialogical proof procedures make it possible to relate a semantics to a stereotypic pattern of dialogue. It has been shown, e.g., that the grounded and preferred credulous semantics can be related to persuasion and Socratic style dialogue [30, 32]. Dialogues generated by our procedure can also be thought of as persuasion dialogues, where PRO has the additional freedom to change the motivational state of the players by putting forward properties. Intuitively, this may benefit PRO in two ways: PRO can enable attacks necessary to put up a successful line of defence, and disable attacks put forward by the opponent from which PRO cannot defend its own arguments. PRO thus persuades OPP to accept an argument, where PRO decides which properties become motivational. Dialogues are structured as follows.

Definition 6.5.1. Let $(A, \to, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework. A *dialogue* is a sequence $S = (m_1, \ldots, m_n)$, where each m_i is either:

- an *attack move* "**OPP:** $x \rightsquigarrow y$", where $x, y \in A$ and $x \to y$,
- a *defence move* "**PRO:** $x \rightsquigarrow y$", where $x, y \in A$ and $x \to y$,

- an *enabling property move* "**PRO:** $P+$", where $P \subseteq \mathcal{P}$,
- a *disabling property move* "**PRO:** $P-$", where $P \subseteq \mathcal{P}$,
- a *conceding move* "**OPP: ok**",
- a *success claim move* "**PRO: win**".

We denote by $S \cdot S'$ the concatenation of S and S' and we say that S is a *subsequence* of S' iff $S' = S'' \cdot S \cdot S'''$ for some S'', S''', and that S is a *proper* subsequence of S' iff $S' = S'' \cdot S \cdot S'''$ for nonempty S'' or S'''.

Definition 6.5.2. Let $S = (m_1, \ldots, m_n)$ be a dialogue. We denote the *motivational state in S at index i* by M_i^S, defined recursively by:

$$M_i^S = \begin{cases} \emptyset & \text{if } i = 0, \\ M_{i-1}^S \cup P & \text{if } m_i = \textbf{PRO:}\ P+ \text{ or } m_i = \textbf{PRO:}\ P-, \\ M_{i-1}^S & \text{otherwise.} \end{cases}$$

We now define a set of production rules that generate *weak x-acceptance dialogues*. Note that argumentation frameworks containing cycles may generate infinite sequences of moves. We prevent this by requiring dialogues to be finite.

Definition 6.5.3 (Weak acceptance dialogue)**.** Let $(A, \to, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and let $x \in A$.

- A *weak x-acceptance dialogue* is a finite sequence

  $$S_1 \cdot (\textbf{PRO: win})$$

  where S_1 is an x-attack sequence.
- An *x-attack sequence* is a sequence

  $$(\textbf{OPP:}\ y_1 \rightsquigarrow x) \cdot S_1 \cdot \ldots \cdot (\textbf{OPP:}\ y_n \rightsquigarrow x) \cdot S_n \cdot (\textbf{OPP: ok})$$

  where $\{y_1, \ldots, y_n\} = \{y \mid y \to x\}$ and each S_i is a y_i-defence sequence.
- An *x-defence sequence* is either:
  - a *regular x-defence sequence*

    $$(\textbf{PRO:}\ y \rightsquigarrow x) \cdot S_1$$

    for some $y \in A$ s.t. $y \to x$, where S_1 is a y-attack sequence,
  - an *enabling property defence sequence*

    $$(\textbf{PRO:}\ P+) \cdot S_1$$

    for some $P \subseteq \mathcal{P}$, where S_1 is a regular x-defence sequence,
  - a *disabling property defence sequence*

    $$(\textbf{PRO:}\ P-)$$

    for some $P \subseteq \mathcal{P}$.

Intuitively, a disabling property move can be interpreted as saying "the preceding move is invalid considering the properties P." An enabling property move, on the other hand, says "the following move is valid considering the properties P." Not every weak x-acceptance dialogue, generated by the production rules in definition 6.5.3, will follow this interpretation. We need to impose a number of additional constraints to ensure that property moves make sense.

Definition 6.5.4 (Property-consistency)**.** Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $S = (m_1, \ldots, m_n)$ a sequence. We say that S is *property-consistent* iff for all $i \in [1, \ldots, n]$, we have:

1. $M_i^S \in \mathcal{M}$

2. If $m_i = \textbf{PRO:}\, x \rightsquigarrow y$ then for all $j \in [i, \ldots, n]$, $x \rightsquigarrow_{M_j^S} y$,

3. If $m_i = \textbf{PRO:}\, P-$ and $m_{i-1} = \textbf{OPP:}\, x \rightsquigarrow y$ then for all $j \in [i, \ldots, n]$, $x \not\rightsquigarrow_{M_j^S} y$.

Condition 1 ensures that property moves are valid in the sense that they actually lead to a new motivational state $M \in \mathcal{M}$. Conditions 2 and 3 ensure that a property move does not undermine preceding property moves. That is, condition 2 ensures that attacks put forward by PRO remain enabled in subsequent states and condition 3 ensures that disabled attacks remain disabled.

Example 6.5.1 (Continued from example 6.4.1)**.** *Consider the following two property-consistent weak acceptance dialogues for the argument a shown in table 6.1 and 6.2.*

Index	Move	State
1	**OPP:** $b \rightsquigarrow a$	$\emptyset$
2	**PRO:** $c \rightsquigarrow b$	$\emptyset$
3	**OPP:** $b \rightsquigarrow c$	$\emptyset$
4	**PRO:** $\{R\}-$	$\{R\}$
5	**OPP: ok**	$\{R\}$
6	**OPP:** $d \rightsquigarrow a$	$\{R\}$
7	**PRO:** $\{G\}+$	$\{R, G\}$
8	**PRO:** $e \rightsquigarrow d$	$\{R, G\}$
9	**OPP:** $f \rightsquigarrow e$	$\{R, G\}$
10	**PRO:** $\emptyset-$	$\{R, G\}$
11	**OPP: ok**	$\{R, G\}$
12	**OPP: ok**	$\{R, G\}$
13	**PRO: win**	$\{R, G\}$

Table 6.1: A weak acceptance dialogue for the argument a.

Index	Move	State
1	**OPP:** $b \rightsquigarrow a$	$\emptyset$
2	**PRO:** $c \rightsquigarrow b$	$\emptyset$
3	**OPP:** $b \rightsquigarrow c$	$\emptyset$
4	**PRO:** $\{R, G\}-$	$\{R, G\}$
5	**OPP: ok**	$\{R, G\}$
6	**OPP:** $d \rightsquigarrow a$	$\{R, G\}$
7	**PRO:** $e \rightsquigarrow d$	$\{R, G\}$
8	**OPP:** $f \rightsquigarrow e$	$\{R, G\}$
9	**PRO:** $\emptyset-$	$\{R, G\}$
10	**OPP: ok**	$\{R, G\}$
11	**OPP: ok**	$\{R, G\}$
12	**PRO: win**	$\{R, G\}$

Table 6.2: A weak acceptance dialogue for the argument a.

*Explanation: In the dialogue shown on figure 6.1, the initial exchange of attacks consists of $b \rightarrow a, c \rightarrow b$ and $b \rightarrow c$. PRO must end this line of argument by making a disabling property to disable the attack $b \rightarrow c$. PRO moves **PRO:**$\{R\}-$ and as a result, the motivational state of the dialogue becomes $\{R\}$. OPP's next attack is $d \rightarrow a$. PRO cannot move $e \rightarrow d$ because this attack is disabled in the current motivational state. PRO moves **PRO:**$\{G\}+$, changing the motivational state of the dialogue to $\{R, G\}$, so that $e \rightarrow d$ is enabled. To OPP's attack $f \rightarrow e$ PRO responds with an empty disabling move, as $f \rightarrow e$ is already disabled in the current motivational state. The dialogue shown in figure 6.2 is similar with the exception that PRO immediately moves both R and G when making a disabling property move on line 4. As a result, no enabling property move is needed on line 7 because the attack $d \rightarrow e$ is already enabled.*

The existence of a property-consistent weak x-acceptance dialogue implies weak acceptance of x, i.e., it is a sound proof procedure:

Lemma 6.5.1 (Soundness)**.** *Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $x \in A$. If there exists a property-consistent weak x-acceptance dialogue $S = (m_1, \ldots, m_n)$ then x is a member of the grounded extension of the argumentation framework represented by M_n^S. Hence x is weakly accepted.*

Proof. Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property based argumentation framework, $x \in A$ and S a property-consistent weak x-acceptance dialogue. A subsequence S' of S that is a y-attack sequence (for some $y \in A$) will be called a *y-attack subsequence*. We denote the *depth* of an attack subsequence S' by $D(S')$ and define it by $D(S') = 0$, if $S' = (\textbf{OPP: ok})$ and $1 + k$ otherwise, where $k = max(\{D(S'') \mid S'' \in T\})$, where T is the set of attack sequences that are proper subsequences of S'. Furthermore from hereon we denote the grounded extension of $(A, \rightsquigarrow_{M_n^S})$ by G. We show that for every y-attack subsequence S' it holds that $y \in G$. We prove this by strong induction on the depth of S'. Let the induction hypothesis $H(k)$ stand for "if S' is a y-attack subsequence with depth k then $y \in G$."

- Base case ($H(0)$): Here $S' = (\textbf{OPP: ok})$, thus y has no attackers in $(A, \rightarrow)$, hence no attackers in $(A, \rightsquigarrow_{M_n^S})$. It follows that $y \in G$.
- Induction step: Assume $H(0), \ldots, H(k-1)$ holds. We need to prove $H(k)$. It can be checked that for every z s.t. $z \rightarrow y$, either:
  - There is a z'-attack sequence S'' that is a proper subsequence of S'. Thus $D(S'') < k$ and $z' \rightarrow z$. From $H(D(S''))$ and the fact that S is property-consistent it follows that z is attacked by G.
  - S' contains a disabling property move. Hence $z \not\rightsquigarrow_{M_n^S} y$.

  This means that for every z such that $z \rightsquigarrow_{M_n^S} y$, G attacks z, hence $y \in G$.

By the principle of strong induction it follows that if there is a y-attack subsequence then $y \in G$. Thus we have $x \in G$, hence x is weakly accepted. □

Conversely, if x is weakly accepted then a property-consistent weak x-acceptance dialogue exists:

Lemma 6.5.2 (Completeness)**.** *Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $x \in A$ be weakly accepted. There exists a weak x-acceptance dialogue S that is property-consistent.*

Proof. Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $x \in A$ be weakly accepted. Let $F = (A, \rightsquigarrow_M)$. Then there is some $M \in \mathcal{M}$ s.t. x is a member of the grounded extension of F. From hereon we use M to refer to any such motivational state and G to refer to the grounded extension of F.

We now prove, by strong induction over the degree of an argument $y \in G$ that there exists a property consistent weak y-acceptance dialogue. Let $H(k)$ stand for "If $y \in G$ and $Deg_F(y) = k$ then there exists a property consistent weak y-acceptance dialogue."

- Base case ($H(0)$): If $y \in G$ and $Deg_F(y) = 0$ then there is no $z \in A$ s.t. $z \rightsquigarrow_M y$ and we can define S by $(\textbf{OPP:} z_1 \rightsquigarrow y) \cdot S' \cdot \ldots \cdot (\textbf{OPP:} z_n \rightsquigarrow y) \cdot S' \cdot (\textbf{OPP: ok}) \cdot (\textbf{PRO: win})$, where $\{z_1, \ldots, z_n\} = \{z' \mid z' \rightarrow y\}$ and $S' = (\textbf{PRO:} M-)$. It can be checked that S is a property consistent weak y-acceptance dialogue.
- Induction step: Assume $H(0), \ldots, H(k-1)$ holds. Thus if $y' \in G$ and $Deg_F(y') < k$ then there exists a property consistent weak y'-acceptance dialogue. We denote this dialogue by $S(y')$. We need to prove $H(k)$.

  Assume that $y \in G$ and $Deg_F(y) = k$. It follows that for every $z \in A$ s.t. $z \rightsquigarrow_M y$, there exists an argument which we denote by $def(z, y)$ such that $def(z, y) \in G$ and $def(z, y) \rightsquigarrow_M z$. Furthermore from the fixpoint construction it follows that $Deg_F(def(z, y)) < k$, so that $S(def(z, y))$ is well defined.

  Now, for every $z \in A$ s.t. $z \rightarrow y$ we define $T_y(z)$ by (1) $T_y(z) = (\textbf{OPP:} z \rightsquigarrow y) \cdot (\textbf{PRO:} M-)$, if $z \not\rightsquigarrow_M y$ and (2) $T_y(z) = (\textbf{OPP:} z \rightsquigarrow y) \cdot (\textbf{PRO:} M+) \cdot S'$, if $z \rightsquigarrow_M y$—where S' is defined by $S(def(z, y)) = S' \cdot (\textbf{PRO: win})$. It

can be checked that $T_y(z_1) \cdot \ldots \cdot T_y(z_i) \cdot (\textbf{OPP: ok}) \cdot (\textbf{PRO: win})$ (where $\{z_1, \ldots, z_i\} = \{z' \mid z' \rightarrow y\}$) is a property consistent weak y-acceptance dialogue.

By the principle of strong induction it follows that for every $y \in G$, there exists a property consistent weak y-acceptance dialogue. Hence, there exists a property consistent weak x-acceptance dialogue. □

Notice that in the fourth move of in the second dialogue in example 6.5.1, PRO puts forward both R and G in a disabling property move. However, it suffices to put forward just R, as in the first dialogue, because G is not relevant with respect to disabling the attack $b \rightarrow c$. We call a dialogue in which property moves are relevant a *property-relevant* dialogue. Property moves in a property-relevant dialogue consist only of properties satisfied by one of the arguments involved in the attack that is enabled or disabled.

Definition 6.5.5 (Property-relevance)**.** Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $S = (m_1, \ldots, m_n)$ a weak acceptance dialogue. We say that S is *property-relevant* iff for all $i, j \in [1, \ldots, n]$ s.t. $j = i + 1$, we have:

1. If $m_i = \textbf{OPP:}\, x \rightsquigarrow y$ and $m_j = \textbf{PRO:}\, P-$ then $P \subseteq P(x) \cup P(y)$.
2. If $m_i = \textbf{PRO:}\, P+$ and $m_j = \textbf{PRO:}\, x \rightsquigarrow y$ then $P \subseteq P(x) \cup P(y)$.

Note that in example 6.5.1 the first dialogue is property-relevant, whereas the second one is not. Focusing on property-relevant dialogues can be used to optimize the algorithm. Furthermore, it makes sense intuitively: when persuading an opponent to accept an argument, one does not refer to properties not relevant to this objective.

As a final result we show that weak acceptance of an argument implies the existence of a property-consistent weak x-acceptance dialogue that is, in addition, property relevant. However, this requires that $\mathcal{M}$ is sufficiently rich to ensure that PRO is not forced to put forward irrelevant properties. This can be achieved by assuming that $\mathcal{M} = 2^{\mathcal{P}}$, but note that there are cases where a weaker assumption is sufficient.

Lemma 6.5.3 (Property-relevant completeness)**.** *Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework where $\mathcal{M} = 2^{\mathcal{P}}$, and let $x \in A$ be weakly accepted. There exists a weak x-acceptance dialogue S that is property-consistent and property-relevant.*

Proof. Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework and $x \in A$ be weakly accepted. Let $S = (m_1, \ldots, m_n)$ be the property-consistent weak x-acceptance dialogue (for x a member of the grounded extension of $(A, \rightsquigarrow_M)$) as constructed in the proof of lemma 6.5.2. That is, every property move in S is either of the form $\textbf{PRO:}\, M+$ or $\textbf{PRO:}\, M-$. Using property 6.4.1 (2) it can be checked that the dialogue S' formed by

- replacing every move $m_i = \textbf{PRO:}\, M+$ in S by $\textbf{PRO:}\, M'+$, where $M' = M \cap P(x) \cup P(y)$ where x, y are defined by $m_{i+1} = \textbf{PRO:}\, x \rightsquigarrow y$, and

- replacing every move $m_i = \textbf{PRO:}\, M-$ in S by $\textbf{PRO:}\, M'-$, where $M' = M \cap P(x) \cup P(y)$ where x, y are defined by $m_{i-1} = \textbf{OPP:}\, x \rightsquigarrow y$,

is also a property-consistent weak x-acceptance dialogue, that is in addition property-relevant. □

Summarizing, we have the following result.

Theorem 6.5.4. *Let $(A, \rightarrow, \mathcal{P}, P, \mathcal{M}, \leq)$ be a property-based argumentation framework.*

- *An argument $x \in A$ is weakly accepted iff there exists a weak x-acceptance dialogue that is property-consistent.*
- *If $\mathcal{M} = 2^{\mathcal{P}}$ then an argument $x \in A$ is weakly accepted iff there exists a weak x-acceptance dialogue that is property-consistent and property-relevant.*

Proof. Follows from lemmas 6.5.1, 6.5.2 and 6.5.3. □

6.6 Related Work

We already mentioned the relation of our model with that of preference and value-based argumentation frameworks [2, 14]. Also related is a study of value-based argumentation frameworks where arguments promote multiple values [60], concerned mainly with the problem of deriving a unique preference order over arguments from a preference relation over individual values. Note that in our approach, a property-based argumentation framework together with a motivational state already defines a unique preference order over arguments.

Furthermore, Bench-Capon et al. have considered dialogues in which a proponent can make moves consisting of value preferences [15]. In this approach, the outcome of a winning dialogue corresponds to the specification of an audience (i.e., a preference order over values) such that some initial set of arguments is accepted in the corresponding *aVAF*.

Also related is Modgil's model of *extended argumentation frameworks*, in which arguments attack and disable attacks between other arguments [67]. Such arguments can be seen as meta-level arguments expressing preferences over object level arguments. Whereas we take the agent's state (which determines whether individual attacks are enabled) to be external to the argumentation framework, here it is part of argumentation framework itself. That is, whether an attack is enabled depends on the status of a metalevel argument.

Our work shares methodological similarities with work of Kontarinis et al. [61], who present a goal-oriented procedure to determine which attacks to disable or enable in order to make an argument accepted under a given semantics. While the procedure that they present is designed to be implemented as a term rewriting system, our procedure is defined simply by a set of production rules, amenable to implementation using e.g. PROLOG.

6.7 Conclusion and Future Work

We presented a dynamic model of preferences in argumentation, based on Dietrich and List's model of property-based preference. This model provides an account of how and why preferences in argumentation may change and generalizes both preference-based argumentation frameworks and value-based argumentation frameworks, if properties are taken to be values. We consider a number of directions for future work. First, we plan to complete the proof-theoretic picture by looking at the problem of deciding whether an argument is strongly accepted. In addition, we will consider other semantics in addition to grounded. Second, we plan to investigate the possibility of axiomatizing property-based argumentation frameworks, in the spirit of Dietrich and List's axiomatization as presented in section 6.3. Finally, we intend to look at connections between property-based argumentation frameworks and Modgil's model of extended argumentation frameworks.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The overall aim of this work was to model and study the notion of change in the context of abstract argumentation. Our motivation was that, while the abstract argumentation formalism is essentially static, argumentation is a dynamic activity. The dynamic perspective we have taken improved our understanding of the behaviour and applicability of the abstract argumentation formalism.

In chapter 1 we identified two types of change in abstract argumentation, which we call intervention and observation. We explained that they are conceptually similar to the similarly named types of change in the theory of causal Bayesian networks. Intervention represents action, which amounts to the manipulation of the argumentation framework, leading to a 'bottom-up' revised evaluation of the argumentation framework. Observations, on the other hand, are pieces of information from the environment, that require a 'top-down' revision of the evaluation of the argumentation framework. We regard the two types of change as two forms of entailment: intervention-based entailment and observation-based entailment.

In chapter 3 we proposed a formal model of intervention-based entailment. We focussed on two types of actions: defeat (addition of an attacker) and provisional defeat (addition of a self-attacking attacker). The resulting notion of entailment allowed us to study the behaviour of semantics for argumentation under change. We studied this behaviour by proposing a number of properties for well-behaved intervention-based entailment, and by systematically checking the conditions under which these properties are satisfied. The properties we proposed were direct translations of a number of properties that have been considered in the context of non-monotonic inference (the so called KLM properties). The results we obtained provide insight into the behaviour of semantics for argumentation under change. For example, the complete and grounded semantics satisfy both Cautious Monotony and Cut—two intuitive properties that an intervention-

based entailment relation can be expected to satisfy—but the preferred, semi-stable and stable semantics do not.

In chapter 4 we developed a formal model for observation-based entailment. We first addressed the question of how a rational agent should revise the evaluation of an argumentation framework to account for an observation. We proposed a model, based on an abductive principle, in which observations are accounted for by the most preferred interventions that make the observation true. Such a model determines an observation-based entailment relation that captures how an argumentation framework is evaluated given an observation. We proved a representation result which links the notion of observation-based entailment to preferential entailment. We also studied the role of the directionality principle in the behaviour of observation-based entailment.

In sum, we have shown in chapters 3 and 4 that there are two distinct ways in which change in abstract argumentation can be modelled. While different aspects of these two types of change have been investigated in the literature, we modelled them in a uniform way and proved a number of novel results concerning change in argumentation.

In chapter 5 we developed a model of abduction in abstract argumentation. In this model, changes to an argumentation framework act as explanations for sceptical/credulous support for observations. We presented sound and complete dialogical proof procedures for the main reasoning tasks, i.e., finding explanations for sceptical/credulous support. In addition, we showed that our model of abduction in abstract argumentation can be seen as an abstract form of abduction in logic programming.

In chapter 6 we developed a dynamic model of preference-based argumentation, based on what we call property-based argumentation frameworks. Here, the idea is that preferences over arguments are derived from preferences over properties of arguments and change as the result of moving to different motivational states. This model is based on Dietrich and List's model of property-based preference and it provides an account of how and why preferences in argumentation may change. Our model generalizes both preference-based argumentation frameworks and value-based argumentation frameworks. We presented sound and complete dialogical proof procedures, similar to the procedures presented in chapter 5, for the task of checking whether an argument is accepted given some or all motivational states of a given property-based argumentation framework.

7.2 Future Work

At the end of each chapter in this thesis we have discussed a number of open issues for future research. Here we discuss a number of further directions for future work.

7.2.1 Application to Extensions of Dung's Formalism

First of all, we have studied various aspects of change in argumentation using Dung's original formalism. However, many extensions of Dung's formalism have

been proposed that aim at going beyond the simple notion of an argumentation framework as consisting of a set of arguments and a binary attack relation. Examples are bipolar argumentation frameworks, which model attack relations as well as support relations between arguments [35]; extended argumentation frameworks, which allow higher-order attacks [49, 5]; and argumentation frameworks with attack relations over sets of arguments [70]. In all these extensions, the form of the evaluation process is similar to that of Dung's original formalism. That is, the input is an argumentation framework (bipolar, extended, set-based, etc.) and the output is a set of extensions or labellings, which are computed according to a semantics that is tailored to the respective extension of Dung's formalism. This means that many of the ideas that we discussed can be applied to these formalisms as well. For example, the notion of intervention and observation, being particular generalizations of how an argumentation framework is evaluated, can also be studied in the context of these extensions.

7.2.2 Extend Results to Other Semantics

In this thesis we focussed on the main admissibility-based semantics, namely the complete, grounded, preferred, stable and semi-stable semantics. There are a number of semantics that we have left out of consideration, such as the stage semantics [89], the ideal semantics [43], the prudent semantics [36], the eager semantics [28] and the CF2 semantics [8]. We plan to address the analysis of the behaviour of these semantics in terms of intervention and observation in future work.

Furthermore, several semantics have been proposed that generalize the strict distinction between acceptance, rejection and undecidedness. This includes quantitative approaches, where arguments are associated with numbers rather than discrete labels. These numbers may indicate probability [56, 64, 79, 88] as well as some type of strength [46, 50]. Applying the ideas discussed in this thesis to these approaches is another possibility for future research.

7.2.3 Synthesis of New Semantics

In chapter 3 we proposed a number of properties for intervention-based entailment. We checked the semantics under which these properties are satisfied and we also considered a number of conditions with respect to the topology of the argumentation framework. The results we obtained say something about the behaviour of the semantics that we considered. Alternatively, desirable properties, such as Cautious Monotony and Cut, can be used to *define* new semantics for argumentation. This is the approach taken, for example, by Baroni et al. [8], who defined a number of new semantics (the most interesting one being the CF2 semantics) by taking the property of SCC-recursiveness as a starting point. Similarly, one may ask: are there semantics (possibly admissibility-based) that satisfy the intuitive properties of Cautious Monotony and Cut, other than the complete and grounded semantics?

7.2.4 Iterated Revision

Dynamics in argumentation arises from the fact that argumentation goes hand in hand with dialogue. In terms of abstract argumentation, a dialogue can be seen as a *sequence* of changes to an argumentation framework. The notion of intervention-based entailment, however, captures change due to *one* intervention, rather than a sequence of changes. Bridging this gap gives rise to a number of questions. First of all, dealing with sequences of changes means that we deal with a form of iterated revision. In the area of belief revision, one of the challenges in modelling iterated revision was to come up with an operator that revises an initial belief state so as to obtain a revised belief state (rather than a revised belief set, as was done in the AGM approach). This problem does not arise in argumentation, because the state is captured by the argumentation framework, and the (iteratively) revised state is simply the (iteratively) revised argumentation framework. However, another question addressed in the area of belief revision is: which postulates should a well-behaved iterated revision operator satisfy? [38] Similarly, we can ask: which postulates for iterated change of an argumentation framework can a semantics be expected to satisfy? In this context, there is a seeming connection between the property of reinstatement in argumentation and recovery in iterated revision. This has been noted by Boella et al. [19]. Making this connection formal, as well as addressing the question of what other postulates are relevant in this context, are issues for future research.

Index

Bibliography

[1] Ernest W. Adams. *The Logic of Conditionals: An Application of Probability to Deductive Logic*. D. Reidel Pub. Co., 1975.

[2] Leila Amgoud and Claudette Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34(1-3):197–215, 2002.

[3] Pietro Baroni, Guido Boella, Federico Cerutti, Massimiliano Giacomin, Leendert van der Torre, and Serena Villata. On the input/output behavior of argumentation frameworks. *Artificial Intelligence*, 217:144–197, 2014.

[4] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *Knowledge Engineering Review*, 26(4):365–410, 2011.

[5] Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Giovanni Guida. ¡ i¿ afra¡/i¿: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37, 2011.

[6] Pietro Baroni and Massimiliano Giacomin. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10-15):675–700, 2007.

[7] Pietro Baroni and Massimiliano Giacomin. Semantics of abstract argument systems. In *Argumentation in Artificial Intelligence*, pages 25–44. Springer, 2009.

[8] Pietro Baroni, Massimiliano Giacomin, and Giovanni Guida. Scc-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168(1-2):162–210, 2005.

[9] Pietro Baroni, Massimiliano Giacomin, and Beishui Liao. On topology-related properties of abstract argumentation semantics. a correction and extension to dynamics of argumentation systems: A division-based method. *Artificial Intelligence*, 212:104–115, 2014.

[10] Ringo Baumann. Normal and strong expansion equivalence for argumentation frameworks. *Artificial Intelligence*, 193:18–44, 2012.

[11] Ringo Baumann. What does it take to enforce an argument? minimal change in abstract argumentation. In Luc De Raedt, Christian Bessière,

Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 127–132. IOS Press, 2012.

[12] Ringo Baumann and Gerhard Brewka. Expanding argumentation frameworks: Enforcing and monotonicity results. In Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Guillermo Ricardo Simari, editors, *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 75–86. IOS Press, 2010.

[13] Trevor J. M. Bench-Capon. Value-based argumentation frameworks. In Salem Benferhat and Enrico Giunchiglia, editors, *9th International Workshop on Non-Monotonic Reasoning (NMR 2002), April 19-21, Toulouse, France, Proceedings*, pages 443–454, 2002.

[14] Trevor J. M. Bench-Capon. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[15] Trevor J. M. Bench-Capon, Sylvie Doutre, and Paul E. Dunne. Audiences in argumentation frameworks. *Artificial Intelligence*, 171(1):42–71, 2007.

[16] Salem Benferhat, Sylvain Lagrue, and Odile Papini. Revision of partially ordered information: Axiomatization, semantics and iteration. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, pages 376–381. Professional Book Center, 2005.

[17] Pierre Bisquert, Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Enforcement in argumentation is a kind of update. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, volume 8078 of *Lecture Notes in Computer Science*, pages 30–43. Springer, 2013.

[18] Alexander Bochman. Credulous nonmonotonic inference. In Thomas Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. 2 Volumes, 1450 pages*, pages 30–35. Morgan Kaufmann, 1999.

[19] Guido Boella, Celia da Costa Pereira, Andrea Tettamanzi, and Leon van der Torre. Dung argumentation and agm belief revision. Fifth International Workshop on Argumentation in Multi-Agent Systems, ArgMAS 2008, 2008.

[20] Guido Boella, Dov M. Gabbay, Alan Perotti, Leendert van der Torre, and Serena Villata. Conditional labelling for abstract argumentation. In Sanjay Modgil, Nir Oren, and Francesca Toni, editors, *Theory and Applications*

of Formal Argumentation - First International Workshop, TAFA 2011. Barcelona, Spain, July 16-17, 2011, Revised Selected Papers, volume 7132 of *Lecture Notes in Computer Science*, pages 232–248. Springer, 2011.

[21] Richard Booth, Dov M. Gabbay, Souhila Kaci, Tjitze Rienstra, and Leendert W. N. van der Torre. Abduction and dialogical proof in argumentation and logic programming. In Torsten Schaub, Gerhard Friedrich, and Barry O'Sullivan, editors, *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, volume 263 of *Frontiers in Artificial Intelligence and Applications*, pages 117–122. IOS Press, 2014.

[22] Richard Booth, Souhila Kaci, and Tjitze Rienstra. Property-based preferences in abstract argumentation. In Patrice Perny, Marc Pirlot, and Alexis Tsoukiàs, editors, *Algorithmic Decision Theory - Third International Conference, ADT 2013, Bruxelles, Belgium, November 12-14, 2013, Proceedings*, volume 8176 of *Lecture Notes in Computer Science*, pages 86–100. Springer, 2013.

[23] Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert van der Torre. A logical theory about dynamics in abstract argumentation. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *Scalable Uncertainty Management - 7th International Conference, SUM 2013, Washington, DC, USA, September 16-18, 2013. Proceedings*, volume 8078 of *Lecture Notes in Computer Science*, pages 148–161. Springer, 2013.

[24] Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert van der Torre. Monotonic and nonmonotonic inference for abstract argumentation. In Chutima Boonthum-Denecke and G. Michael Youngblood, editors, *Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2013, St. Pete Beach, Florida. May 22-24, 2013*. AAAI Press, 2013.

[25] Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert W. N. van der Torre. Conditional acceptance functions. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 470–477. IOS Press, 2012.

[26] Martin Caminada. On the issue of reinstatement in argumentation. In Michael Fisher, Wiebe van der Hoek, Boris Konev, and Alexei Lisitsa, editors, *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006, Liverpool, UK, September 13-15, 2006, Proceedings*, volume 4160 of *Lecture Notes in Computer Science*, pages 111–123. Springer, 2006.

[27] Martin Caminada. Semi-stable semantics. In Paul E. Dunne and Trevor J. M. Bench-Capon, editors, *Computational Models of Argument: Proceedings of COMMA 2006, September 11-12, 2006, Liverpool, UK*, volume 144 of *Frontiers in Artificial Intelligence and Applications*, pages 121–130. IOS Press, 2006.

[28] Martin Caminada. Comparing two unique extension semantics for formal argumentation: ideal and eager. In *Proceedings of the 19th Belgian-Dutch conference on artificial intelligence (BNAIC 2007)*, pages 81–87, 2007.

[29] Martin Caminada and Gabriella Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1):64–102, 2011.

[30] Martin Caminada and Mikolaj Podlaszewski. Grounded semantics as persuasion dialogue. In Bart Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument - Proceedings of COMMA 2012, Vienna, Austria, September 10-12, 2012*, volume 245 of *Frontiers in Artificial Intelligence and Applications*, pages 478–485. IOS Press, 2012.

[31] Martin Caminada, Samy Sá, and João Alcântara. On the equivalence between logic programming semantics and argumentation semantics. In Linda C. van der Gaag, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 12th European Conference, ECSQARU 2013, Utrecht, The Netherlands, July 8-10, 2013. Proceedings*, volume 7958 of *Lecture Notes in Computer Science*, pages 97–108. Springer, 2013.

[32] Martin W. A. Caminada. Preferred semantics as socratic discussion. In Alfonso E. Gerevini and Alessandro Saetti, editors, *Proceedings of the eleventh AI*IA symposium on artificial intelligence*, pages 209–216, 2010.

[33] Martin W. A. Caminada and Dov M. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109–145, 2009.

[34] Claudette Cayrol, Florence Dupin de Saint-Cyr, and Marie-Christine Lagasquie-Schiex. Change in abstract argumentation frameworks: Adding an argument. *Journal of Artificial Intelligence Research*, 38:49–84, 2010.

[35] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In Lluis Godo, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 8th European Conference, ECSQARU 2005, Barcelona, Spain, July 6-8, 2005, Proceedings*, volume 3571 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2005.

[36] Sylvie Coste-Marquis, Caroline Devred, and Pierre Marquis. Prudent semantics for argumentation frameworks. In *17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2005), 14-16 November 2005, Hong Kong, China*, pages 568–572. IEEE Computer Society, 2005.

[37] Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014.

[38] Adnan Darwiche and Judea Pearl. On the logic of iterated belief revision. *Artificial intelligence*, 89(1-2):1–29, 1996.

[39] Marc Denecker and Antonis Kakas. Abduction in logic programming. In *Computational Logic: Logic Programming and Beyond*, pages 402–436. Springer, 2002.

[40] Franz Dietrich and Christian List. A reason-based theory of rational choice. *Noûs*, 47(1):104–134, 2013.

[41] Franz Dietrich and Christian List. Where do preferences come from? *International Journal of Game Theory*, 42(3):613–637, 2013.

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

[43] Phan Minh Dung, Paolo Mancarella, and Francesca Toni. A dialectic procedure for sceptical, assumption-based argumentation. In Paul E. Dunne and Trevor J. M. Bench-Capon, editors, *Computational Models of Argument: Proceedings of COMMA 2006, September 11-12, 2006, Liverpool, UK*, volume 144 of *Frontiers in Artificial Intelligence and Applications*, pages 145–156. IOS Press, 2006.

[44] Paul E. Dunne. Computational properties of argument systems satisfying graph-theoretic constraints. *Artificial intelligence*, 171(10-15):701–729, 2007.

[45] Paul E. Dunne, Wolfgang Dvorák, Thomas Linsbichler, and Stefan Woltran. Characteristics of multiple viewpoints in abstract argumentation. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference, KR 2014, Vienna, Austria, July 20-24, 2014*. AAAI Press, 2014.

[46] Paul E Dunne, Anthony Hunter, Peter McBurney, Simon Parsons, and Michael Wooldridge. Weighted argument systems: Basic definitions, algorithms, and complexity results. *Artificial Intelligence*, 175(2):457–486, 2011.

[47] Wolfgang Dvořák. *Computational aspects of abstract argumentation*. PhD thesis, Technischen Universität Wien, 2012.

[48] Dov M. Gabbay. Theoretical foundations for non-monotonic reasoning in expert systems. In KrzysztofR. Apt, editor, *Logics and Models of Concurrent Systems*, volume 13 of *NATO ASI Series*, pages 439–457. Springer Berlin Heidelberg, 1985.

[49] Dov M Gabbay. Semantics for higher level attacks in extended argumentation frames part 1: Overview. *Studia Logica*, 93(2-3):357–381, 2009.

[50] Dov M. Gabbay. Introducing equational semantics for argumentation networks. In Weiru Liu, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 11th European Conference, ECSQARU 2011, Belfast, UK, June 29-July 1, 2011. Proceedings*, volume 6717 of *Lecture Notes in Computer Science*, pages 19–35. Springer, 2011.

[51] Peter Gärdenfors and David Makinson. Nonmonotonic inference based on expectations. *Artificial Intelligence*, 65(2):197–245, 1994.

[52] Allen Van Gelder, Kenneth A. Ross, and John S. Schlipf. The well-founded semantics for general logic programs. *Journal of the ACM*, 38(3):620–650, 1991.

[53] Michael Gelfond and Vladimir Lifschitz. The stable model semantics for logic programming. In Robert A. Kowalski and Kenneth A. Bowen, editors, *Logic Programming, Proceedings of the Fifth International Conference and Symposium, Seattle, Washington, August 15-19, 1988 (2 Volumes)*, pages 1070–1080. MIT Press, 1988.

[54] Guido Governatori, Michael J. Maher, Grigoris Antoniou, and David Billington. Argumentation semantics for defeasible logic. *Journal of Logic and Computation*, 14(5):675–702, 2004.

[55] Peter Grdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press, 1988.

[56] Anthony Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.

[57] Katsumi Inoue and Chiaki Sakama. Abductive framework for nonmonotonic theory change. In *IJCAI*, pages 204–210. Morgan Kaufmann, 1995.

[58] Katsumi Inoue and Chiaki Sakama. Computing extended abduction through transaction programs. *Annals of Mathematics and Artificial Intelligence*, 25(3-4):339–367, 1999.

[59] Hadassa Jakobovits and Dirk Vermeir. Robust semantics for argumentation frameworks. *Journal of Logic and Computation*, 9(2):215–261, 1999.

[60] Souhila Kaci and Leendert van der Torre. Preference-based argumentation: Arguments supporting multiple values. *International Journal of Approximate Reasoning*, 48(3):730–751, 2008.

[61] Dionysios Kontarinis, Elise Bonzon, Nicolas Maudet, Alan Perotti, Leon van der Torre, and Serena Villata. Rewriting rules for the computation of goal-oriented changes in an argumentation system. In João Leite, Tran Cao Son, Paolo Torroni, Leon van der Torre, and Stefan Woltran, editors, *Computational Logic in Multi-Agent Systems - 14th International Workshop, CLIMA XIV, Corunna, Spain, September 16-18, 2013. Proceedings*, volume 8143 of *Lecture Notes in Computer Science*, pages 51–68. Springer, 2013.

[62] Sarit Kraus, Daniel J. Lehmann, and Menachem Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44(1-2):167–207, 1990.

[63] Daniel J. Lehmann and Menachem Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[64] Hengfei Li, Nir Oren, and Timothy J. Norman. Probabilistic argumentation frameworks. In Sanjay Modgil, Nir Oren, and Francesca Toni, editors, *Theory and Applications of Formal Argumentation - First International Workshop, TAFA 2011. Barcelona, Spain, July 16-17, 2011, Revised Selected Papers*, volume 7132 of *Lecture Notes in Computer Science*, pages 1–16. Springer, 2011.

[65] Bei Shui Liao, Li Jin, and Robert C. Koons. Dynamics of argumentation systems: A division-based method. *Artificial Intelligence*, 175(11):1790–1814, 2011.

[66] Jorge Lobo and Carlos Uzcátegui. Abductive consequence relations. *Artificial Intelligence*, 89(1-2):149–171, 1997.

[67] Sanjay Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.

[68] Sanjay Modgil and Martin W.A. Caminada. Proof theories and algorithms for abstract argumentation frameworks. In I. Rahwan and G. Simari, editors, *Argumentation in Artificial Intelligence*, pages 105–129. Springer, 2009.

[69] Martín O. Moguillansky, Nicolás D. Rotstein, Marcelo A. Falappa, Alejandro Javier García, and Guillermo Ricardo Simari. Argument theory change through defeater activation. In Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Guillermo Ricardo Simari, editors, *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 359–366. IOS Press, 2010.

[70] Søren Holbech Nielsen and Simon Parsons. A generalization of dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In Nicolas Maudet, Simon Parsons, and Iyad Rahwan, editors, *Argumentation in Multi-Agent Systems, Third International Workshop, ArgMAS 2006, Hakodate, Japan, May 8, 2006, Revised Selected and Invited Papers*, volume 4766 of *Lecture Notes in Computer Science*, pages 54–73. Springer, 2006.

[71] Emilia Oikarinen and Stefan Woltran. Characterizing strong equivalence for argumentation frameworks. *Artificial Intelligence*, 175(14-15):1985–2009, 2011.

[72] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge University Press, 2000.

[73] Ramón Pino Pérez and Carlos Uzcátegui. Jumping to explanations versus jumping to conclusions. *Artificial Intelligence*, 111(1-2):131–169, 1999.

[74] John L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

[75] John L. Pollock. *Cognitive Carpentry: A Blueprint for how to Build a Person*. MIT Press, Cambridge, MA, USA, 1995.

[76] Henry Prakken and Gerard A. W. Vreeswijk. Logics for defeasible argumentation. In *Handbook of Philosophical Logic, Second Edition*, 2001.

[77] Teodor C. Przymusinski. The well-founded semantics coincides with the three-valued stable semantics. *Fundamenta Informaticae*, 13(4):445–463, 1990.

[78] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1-2):81–132, 1980.

[79] Tjitze Rienstra. Towards a probabilistic dung-style argumentation system. In Sascha Ossowski, Francesca Toni, and George A. Vouros, editors, *Proceedings of the First International Conference on Agreement Technologies, AT 2012, Dubrovnik, Croatia, October 15-16, 2012*, volume 918 of *CEUR Workshop Proceedings*, pages 138–152. CEUR-WS.org, 2012.

[80] Nico Roos. Preferential model and argumentation semantics. In *Proceedings of the 13th International Workshop on Non-Monotonic Reasoning (NMR-2010)*, 2010.

[81] Nicolás D. Rotstein, Martín O. Moguillansky, Marcelo A. Falappa, Alejandro Javier García, and Guillermo Ricardo Simari. Argument theory change: Revision upon warrant. In Philippe Besnard, Sylvie Doutre, and Anthony Hunter, editors, *Computational Models of Argument: Proceedings of COMMA 2008, Toulouse, France, May 28-30, 2008*, volume 172 of *Frontiers in Artificial Intelligence and Applications*, pages 336–347. IOS Press, 2008.

[82] Nicolás D. Rotstein, Martín O. Moguillansky, Alejandro Javier García, and Guillermo Ricardo Simari. A dynamic argumentation framework. In Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Guillermo Ricardo Simari, editors, *Computational Models of Argument: Proceedings of COMMA 2010, Desenzano del Garda, Italy, September 8-10, 2010*, volume 216 of *Frontiers in Artificial Intelligence and Applications*, pages 427–438. IOS Press, 2010.

[83] Chiaki Sakama. Abduction in argumentation frameworks and its use in debate games. In *Proceedings of the 1st International Workshop on Argument for Agreement and Assurance (AAA2013), Kanagawa, Japan.*, 2013.

[84] Chiaki Sakama. Counterfactual reasoning in argumentation frameworks. In Simon Parsons, Nir Oren, Chris Reed, and Federico Cerutti, editors, *Computational Models of Argument - Proceedings of COMMA 2014, Atholl Palace Hotel, Scottish Highlands, UK, September 9-12, 2014*, volume 266 of *Frontiers in Artificial Intelligence and Applications*, pages 385–396. IOS Press, 2014.

[85] Yoav Shoham. A semantical approach to nonmonotic logics. In *Proceedings, Symposium on Logic in Computer Science, 22-25 June 1987, Ithaca, New York, USA*, pages 275–279. IEEE Computer Society, 1987.

[86] Guillermo R. Simari and Ronald P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53:125–157, 1992.

[87] Robert Stalnaker. What is a nonmonotonic consequence relation? *Fundamenta Informaticae*, 21(1-2):7–21, 1994.

[88] Matthias Thimm. A probabilistic semantics for abstract argumentation. In Luc De Raedt, Christian Bessière, Didier Dubois, Patrick Doherty, Paolo Frasconi, Fredrik Heintz, and Peter J. F. Lucas, editors, *ECAI 2012 - 20th European Conference on Artificial Intelligence. Including Prestigious Applications of Artificial Intelligence (PAIS-2012) System Demonstrations Track, Montpellier, France, August 27-31 , 2012*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pages 750–755. IOS Press, 2012.

[89] Bart Verheij. Two approaches to dialectical argumentation: admissible sets and argumentation stages. In *Proceedings of the Eighth Dutch Conference on Artificial Intelligence (NAIC96)*, pages 357–368, 1996.

[90] Toshiko Wakaki, Katsumi Nitta, and Hajime Sawamura. Computing abductive argumentation in answer set programming. In Peter McBurney, Iyad Rahwan, Simon Parsons, and Nicolas Maudet, editors, *Argumentation in Multi-Agent Systems, 6th International Workshop, ArgMAS 2009, Budapest, Hungary, May 12, 2009. Revised Selected and Invited Papers*, volume 6057 of *Lecture Notes in Computer Science*, pages 195–215. Springer, 2009.

[91] Yining Wu, Martin Caminada, and Dov M. Gabbay. Complete extensions in argumentation coincide with 3-valued stable models in logic programming. *Studia Logica*, 93(2-3):383–403, 2009.