For the Sake of the Argument

explorations into argument-based reasoning

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 612.060.005.

SIKS Dissertation Series No. 2004-09

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Printed by Ponsen & Looijen
ISBN 90-6464-520-5

VRIJE UNIVERSITEIT

For the Sake of the Argument
explorations into argument-based reasoning

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. T. Sminia,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der Exacte Wetenschappen
op dinsdag 22 juni 2004 om 13.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Martinus Wigbertus Antonius Caminada

geboren te Amsterdam

promotor: prof.dr. R.P. van de Riet
copromotor: dr.mr. H. Prakken

Preface

Part of the idea of this thesis originated during a lecture of Arno Lodder at the Vereniging Voor Logica (VVL). To me, the idea of a mathematical description of what rational dialogue or argumentation should look like sounded very appealing. I then studied a number of formalisms in this field, but to my surprise, the informal kinds of dialogues that I observed around me often turned out to have a totally different structure and nature than what was specified by the various dialectical formalisms. It is this deviation that inspired me to write this thesis.

There are a few people who directly or indirectly contributed to this thesis that I would like to thank. First of all, there is my promotor, Reind van de Riet, who despite the fact that I moved out of his field of expertise continued to support me. Furthermore, I would like to thank my copromotor, Henry Prakken, for it is his advice that helped to give the thesis its current form. Much informal input has also been gathered by carefully examining the various of discussions that took place during lunch and coffee time, at which Jan Broersen, Radu Serban, Mehdi Dastani and Joris Hulstijn participated. Special thanks go to Leon van der Torre, who also helped to develop a formal account of the original idea. Also very helpful was a discussion with Gerard Vreeswijk. His warning about "hacking", or tuning a particular formalism so that a certain desirable outcome is obtained resulted in an extensive treatment of criteria for the construction of logical formalisms. Last, but not least, I would like to thank Jan-Willem Klop for his support during the difficult last months of writing this thesis. Although the process of producing this thesis has been a difficult one, I hope the result will be worth reading.

Contents

Chapter 1

An introduction into the problem

1.1 On the topic of argumentation

This thesis is about formal reasoning and argumentation. People who try to convince each other often use (informal) arguments to support their respective point of views. The field of formal reasoning and argumentation attempts to provide a mathematical model of this reasoning and argumentation, so that its relevant characteristics become explicit. The modeling is done in such a way that it enables one to rigorously determine (to calculate, or prove) what, given a number of facts or premises, should be the valid conclusions. The system of formal reasoning or argumentation thus acts as an abstraction, or idealization, of the reasoning as performed by humans (more on this in section 2.2). Applications of formal reasoning can be found in the area of expert systems and in AI and law.

Many formalisms of argumentation (such as [Vree93], [PrSa97] and [Poll95]) regard an argument as a structured chain of rules (sometimes also called *reasons*). An argument begins with one or more premises — statements that are simply regarded as true by all involved parties, such as directly observable facts.[1] After this follows the repeated application of various rules, which generate new conclusions and therefore enable the application of additional rules. An example of such an argument is as follows:

> "Sjaak probably went to the soccer game, since people claim his car was parked nearby the stadium, and Sjaak is known to be a soccer fan."

$$claimed(car\_at\_stadium),\ soccer\_fan,$$
$$claimed(car\_at\_stadium) \Rightarrow car\_at\_stadium,$$
$$car\_at\_stadium \wedge soccer\_fan \Rightarrow at\_game$$

Arguments are often *defeasible*, meaning that the argument by itself is not a conclusive reason for the conclusions it brings about. Whether or not an argument should be accepted depends on its possible counterarguments. For the above argument, a possible counterargument would be:

> "Sjaak did not go to the soccer game, since his friends claim he was watching the game with them in a bar."

[1] The formalization of Pollock [Poll95] also proposes *suppositional arguments* (see section 5.3) but for simplicity we restrict the above discussion to the restricted form of *linear arguments*.

$friends\_claim(at\_bar)$,
$friends\_claim(at\_bar) \Rightarrow at\_bar$,
$at\_bar \rightarrow \neg at\_game$

The issue of determining the arguments and conclusions that are considered to be *justified* then becomes a matter of weighting and evaluating the given arguments.

Most systems for formal argumentation take arguments to be grounded in the premises; that is, each rule of the argument is ultimately (directly or indirectly) based on premises only. In human argumentation, however, one can often observe arguments which are not based on premises only, but which are instead at least partly based on the conclusions of the other person's argument. As an illustration, consider the following example of a discussion between the opponent and proponent of a certain thesis:

> P: "Guus did not go to the game because his mobile phone record shows he was in his mother's house at the time of the game."
>
> $phone\_record$,
> $phone\_record \Rightarrow at\_mothers\_house(phone)$,
> $at\_mothers\_house(phone) \Rightarrow at\_mothers\_house(Guus)$,
> $at\_mothers\_house(Guus) \rightarrow \neg at\_game(Guus)$
>
> O: "Then he would not have watched the game at all, since his mother's TV has been broken for quite a while. Don't you think that's a little odd? Guus is known to be a soccer fan and would have loved to watch the game."
>
> $soccer\_fan(Guus)$,
> $at\_mothers\_house(Guus) \Rightarrow \neg watch\_game(Guus)$,
> $soccer\_fan(Guus) \Rightarrow watch\_game(Guus)$

Here, the opponent takes the propositions as uttered by the proponent as a starting point and then uses these to (defeasibly) derive a contradiction, thus illustrating the (implicit) absurdity of the proponent's original argument.

Socrates and the elenchus

The idea of taking the other party's opinion and then deriving a contradiction (or something else that is undesirable to the other party) is not new. One of the first well known examples of this style of reasoning can be found in the philosophy of Socrates, as written down by Plato. Socrates's form of reasoning — also called the elenchus — consists of letting the opponent make a statement, and then taking this statement as a starting point to derive more statements, each of which is committed by the opponent. The ultimate aim is to let the opponent commit himself to a contradiction, which shows that the beliefs of the opponent obviously cannot hold and that therefore he should reject his beliefs.

As an example of how Socrates's form of dialectical reasoning worked, consider the following dialogue, in which Socrates questions Menexenus about the nature of friendship [Plato1, pp. 212-213]

> (...) Answer me this. As soon as one man loves another, which of the two becomes the friend? the lover of the loved, or the loved of the lover? Or does it make no difference?
>
> None in the world, that I can see, he replied.
>
> How? said I; are both friends, if only one loves?
>
> I think so, he answered.
>
> Indeed! is it not possible for one who loves, not to be loved in return by the object of his love?
>
> It is.
>
> Nay, is it not possible for him even to be hated? treatment, if I mistake not, which lovers frequently fancy they receive at the hands of their favorites. Though they love their darlings as dearly as possible, they often imagine that they are not loved in return, often that they are even hated. Don't you believe this to be true?
>
> Quite true, he replied.
>
> Well, in such a case as this, the one loves, the other is loved.
>
> Just so.
>
> Which of the two, then, is the friend of the other? the lover of the loved, whether or not he be loved in return, and even if he be hated, or the loved of the lover? or is neither the friend of he other, unless both love each other?
>
> The latter certainly seems to be the case, Socrates.
>
> If so, I continued, we think differently now from what we did before. Then it appeared that if one loved, both were friends; but now, that unless both love, neither are friends.
>
> Yes, I'm afraid we have contradicted ourselves.

Socrates's method is that of asking questions. The questions, however, are often meant to direct the dialogue partner into a certain direction. It is the questions that force the dialogue partner to make certain inferences, as these seem to logically follow from the dialogue partner's own position. The inferences are not completely deductive, as they are usually based on common sense and what is reasonable. The inference is therefore more of a defeasible than of a strict nature.

After Menexenus admits he has run into trouble, Socrates continues the discussion from a revised standpoint [Plato1, pp. 213-215].

> This being the case then, the lover is not a friend to anything that does not love him in return.
>
> Apparently not.
>
> People, then, are not friends to horses, unless their horses love them in return, nor friends to quails or to dogs, nor again, to wine or gymnastics, unless their love be returned; nor friends to wisdom, unless wisdom loves them in return. But in each of these cases, the individual loves the object, but is not a friend

to it, and the poet is wrong who says: "Happy the man who, to whom he's a friend, has children, and horses. Mettlesome, dogs of the chase, guest in a far away land."

I don't think he is wrong, Socrates.

But do you think he's right?

Yes, I do.

The lover, then, it appears, Menexenus, is a friend to the object of his love, whether the object loves, or even hates him. Just as to quite young children, who are either not yet old enough to love, or who are old enough to feel hatred when punished by father or mother, their parents, all the time even that they are being hated, are friends in the very highest degree.

Yes, such appears to be the case.

By this reasoning, then, it is not the object of love that is the friend, but the lover.

Apparently.

And so, not the object of hatred that is the enemy, but the hater.

Clearly.

It frequently happens, then, that people are enemies to those who love them; that is, are enemies to their friends, and friends to their enemies; if it be true that the lover is the friend, but not the loved. But surely, my dear friend, it were grossly unreasonable, nay, I think altogether impossible, for a man to be a friend to his enemy and an enemy to his friend.

Yes, Socrates, it does seem impossible.

Well, then, if this is impossible, it must be the object of the love that is the friend of the lover.

Clearly.

And so again, the object of the hatred that is the enemy to the hater.

Necessarily.

But if this is true, we cannot help arriving at the same conclusion as we did in the former case; namely, that it often happens that a man is not a friend, but even an enemy to a friend; as often, that is, as he is not loved, but even hated by the man whom he loves; and often again, that he is not an enemy, but even a friend to an enemy, as often, in fact, as he is not hated, but even loved by the man whom he hates.

No, I'm afraid we can't.

What are we to do then, said I, if neither those who love are to be friends, nor those who are loved, nor, again, those who both love and are loved? Are there any other people beside these that we can say become friends to each other?

To tell you the truth, Socrates, said he, I don't see my way at all.

This time, Menexenus has no escape. He has to admit that, apparently, some of his reasoning has been wrong. As a consequence, none of the statements that led to the untenable position can be relied upon.

Socrates's elenchus is not meant for the derivation of new facts. On the contrary, its purpose is primarily destructive, meant to destroy someone's pretension of knowledge [Nels94]. In "The Sophist", Plato provides the following definition of the elenchus [Plato2]:

> They [those that apply the elenchus] cross-examine a man's words, when he thinks that he is saying something and is really saying nothing, and easily convict him of inconsistencies in his opinions; these they then collect by the dialectical process, and placing them side by side, show that they contradict one another about the same things, in relation to the same things, and in the same respect. He, seeing this, is angry with himself, and grows gentle towards others, and thus is entirely delivered from great prejudices and harsh notions, in a way that is most amusing to the hearer, and produces the most lasting effect to the person who is the subject of the operation.

The destruction of knowledge is best pursued by showing it to be incompatible with other knowledge, as argued by the French scholar Chaïm Perelman [Pere82, p. 24]:

> How do we disqualify a fact or truth? The most effective way is to show its incompatibility with other facts and truths which are more certainly established, preferably with a *bundle* of facts and truths which we are not willing to abandon.

Of course, an obvious way to show incompatibility is by means of a classical counterargument, but there are also forms of incompatibility that require argumentation beyond classical arguments (see section 3.2.2).

Some modern examples

It should be mentioned that the kind of reasoning in which one confronts the other party with the (defeasible) consequences of its statements is also widely used in modern times. Consider the following dialogue between politician EB and interviewing journalist IJ:

> EB: In two years time, the waiting lists in health care will be as good as resolved.
> IJ: Then you are actually saying that the insurance fees will be increased, because the government has already decided not to put more money into the health care system, and you have promised not to lower the coverage of the standard insurance.

In general, one may say that many of today's interviews in which the interviewer takes a critical stance, the interviewer tries to force the interviewee to draw conclusions or make statements that the interviewee may wish to avoid.

On a more philosophical level, James Skidmore discusses the issue of *transcendental arguments*, which are meant to combat various forms of (philosophical) scepticism. The aim of a transcendental argument is "to locate something that the sceptic must presuppose in order for her challenge to be meaningful, then to show that from this presupposition it follows that the skeptic's challenge can be dismissed." [Skid02, p. 121] Skidmore gives various (rather long) examples of these kind of arguments — we will not repeat them here.

To summarize, the technique of using statements from the other party's argument against him is still common in modern times, both in popular as well as in philosophical argumentation. It is the author's opinion that therefore the question of how these arguments can be formally modeled is a relevant one.

1.2 Research questions

A hang yourself (HY) argument[2], as we preliminarily define it, is an argument that shows the problematic nature of another argument by taking the opinion of the other party as a starting point, and then deriving something the other party cannot accept — such as a contradiction or a self-defeating argument.

Our main research questions are the following:

1. To which extent is a HY-style of argumentation already supported by existing formalisms for reasoning and argumentation?

2. How should HY be seen in the context of (formal) dialogue and related concepts (commitments) ?

3. How should one in general pursue the conversion from informal reasoning to formal reasoning, and according to which principles should this be done for HY?

4. How can HY-arguments be included in particular systems for formal argumentation and reasoning, and what are the formal properties of the resulting systems?

5. Are HY-arguments suitable for every domain of reasoning, or are there domains of reasoning where HY is not appropriate?

The topic of formalized argumentation is relevant not only from the perspective of philosophy, but also from the perspective of computer science. One possible application can be found in the field of AI and Law, where various forms of legal reasoning and argumentation are studied, often with the purpose of developing computerized tools that can assist lawyers and other legal workers with their activities. Another application can be found in the field of multi-agent systems, where agents use argumentation or formal dialogue in order to persuade each other to carry out certain activities (more on this in the epilogue).

1.3 Outline

This thesis is structured as follows. First, in chapter 2, the necessary preliminaries are provided. This is done in two steps. In section 2.1, an overview is provided of several formalisms for nonmonotonic reasoning, argumentation and dialogues. The discussion will be relatively brief, and emphasis is laid on those aspects that are used in other parts of this thesis. In section 2.2, the issue is studied according to what principles systems for formal reasoning or argumentation are or should be constructed.

[2]We use the term *hang yourself* (or HY) because from the perspective of, say, a Socratic dialogue, the idea is that a party is persuaded to make an inference that discredits its own reasoning; it is like the other party "hangs himself".

In chapter 3, the concept of HY-arguments is formalized in a simplified version of the logic of Prakken and Sartor. This is done as follows. First, the simplified logical formalism of Prakken and Sartor is stated. It is then shown that this formalism by itself does not support the kind of reasoning as implemented by HY-arguments. This is illustrated by various examples. In section 3.2, an analysis of the HY-style of argumentation is given. The aim is to provide guidelines according to which HY-arguments can be modeled in systems for formal argumentation. The next step, in section 3.3.1 is then to actually implement this particular type of arguments, at first in the earlier stated logic of Prakken and Sartor. Some properties of the resulting logic are also studied.

In chapter 4, two different issues are studied. First, in section 4.1, some of the differences and similarities between HY and the concept of contraposition are discussed. Using this discussion, the question is then studied under which conditions and interpretations contraposition and HY can be considered suitable forms of reasoning. It is argued that for epistemical reasoning, HY is suitable (section 4.2.1), whereas for constitutive reasoning, it is not (section 4.2.2).

In chapter 5, it is shown that the concept of HY-arguments can also be included in other systems than that of Prakken and Sartor. First, it is shown that the concept of HY-arguments is in fact compatible with the argument-based semantics as stated by Dung (section 5.1). It is then shown how HY-arguments can be implemented in respectively Reiter's default logic (section 5.2) and Pollock's system (section 5.3), as well as what the effects are of doing so.

Chapter 6 then provides a comprehensive overview of the main results of this thesis, and states a few candidate topics that could be further investigated.

A treatment of some applications of formal dialogue and argumentation is provided in the epilogue, with the aim of putting the results of this thesis into a broader perspective.

Chapter 2

Logic, arguments and dialogues

This chapter provides a brief introduction into nonmonotonic logic, argument systems and formal dialogue, as well as an overview of the principles according to which systems for formal reasoning can be constructed. The treatment will necessarily be concise, with an emphasis on the aspects that are actually used in other parts of this thesis.[1]

Another aim of this chapter is to deal with questions regarding the nature of nonmonotonic reasoning and research methodology. The opinions of various researchers on this topic are taken into account partly by providing quotes from their respective work. The advantage of using quotes, apart from its directness, is that it allows a clear distinction between what opinions are ours, and what opinions are of the respective researcher.

2.1 About defeasible logic, argumentation and dialogues

The aim of this section is to give a brief overview of the fields of nonmonotonic logic, formal argumentation and dialogue systems, as well as to specify some relationships between them.

2.1.1 Nonmonotonic logic

Logic can be described as the field of research that is concerned with reasoning. One of the characteristics of logic is that the kind of truths studied by it are of a very general nature; they are not inherently connected to any particular field of application.[2] Susan Haack summarizes the general nature of logic as follows [Haac78, p. 5]:

> The traditional idea is that logic is concerned with the validity of arguments as such, irrespective, that is, of their subject-matter — that logic is, as Ryle neatly puts it, 'topic-neutral' — could be thought to offer a principle on which to delimit the scope of logic.

Nonmonotonic logic — sometimes also called *defeasible logic* — is a field of logic whose inferences are *defeasible*, that is, the inferences can be defeated when additional information

[1] For a more elaborate discussion of the various systems for nonmonotonic formal reasoning, we refer to [GHRN94] and [GaGu02]

[2] Tarski, for instance, tries to illustrate the generality of logic by claiming that logical notions are those that are invariant under any transformation [Tars86].

is available. As an example, take the following statements[3]:

Birds usually fly: $Bird(x) \Rightarrow Fly(x)$
Penguins do not fly: $Penguin(x) \supset \neg Fly(x)$

Based on the fact that Tweety is a bird, one may infer that it flies. This inference becomes defeated, however, under the extra information that Tweety is a penguin.

Nonmonotonic logic owes its name to the fact that it is not monotonic, that is, the following property (monotonity) does not hold:[4]

If $\Gamma \mid\!\sim q$ then $\Gamma \cup \{p\} \mid\!\sim q$

Thus, under nonmonotonic logic, the conclusions are not deductively valid; it is possible that the premises are true while the conclusion is not. The idea is not to entail what is necessary true, but what is *normally* true. Perelman puts it as follows [Pere82, p. 25]:

> Although with facts and truths we often depend on presumptions, which, although they are not as certain as our facts and truths, nevertheless furnish a sufficient basis upon which to rest a reasonable conviction. We habitually associate presumptions with what *normally* [my emphasis, MC] happens and with what can reasonably counted upon. Although these presumptions, tied to common experience and common sense, permit one to function reasonably well, they can be contradicted by the facts, because the unexpected can never be excluded.

A first impression may be that nonmonotonic logic is a somewhat strange form of formal reasoning in which correctness is compromised for the sake of practical purposes. One should take into account, however, that many forms of everyday human reasoning have an inherently defeasible nature. From the timetables at the railway station, for instance, I may infer that tomorrow at 10:14h there is a train from Amsterdam to The Hague, an inference that is defeated if I find out that tomorrow will be the start of a railway strike. Our view of defeasible reasoning is shared with Pollock [Poll95, pp. 41-42]:

> A common impression in AI is that defeasible reasoning consists of jumping to conclusions or making "tentative guesses". It is supposed that defeasible reasoning is less secure than normal reasoning, and should be countenanced only for the sake of computational efficiency. What is overlooked is that defeasible reasoning *is* normal reasoning. Its use is not just a matter of computational efficiency. It is logically impossible to reason successfully about the world around us using only deductive reasoning. All interesting reasoning outside mathematics involves defeasible steps. For instance, our basis information about the world comes from sense perception. Things appear certain to us, and we take that to be a reason for believing that they are that way. Clearly this reasoning is defeasible, but our reasoning in this way is no mere matter of convenience. (...) If a regularity has been observed to hold in every case, that gives us a reason for thinking that it holds in general. If it has only been observed to hold in most cases, that gives us a reason for thinking it will continue to hold in most

[3]Here, "$\Rightarrow$" stands for defeasible implication and "$\supset$" stands for material implication.
[4]The symbol $\mid\!\sim$ stands for nonmonotonic derivability.

> cases. Such reasoning is defeasible, and it cannot be replaced by deductive reasoning. There is no way to deduce general conclusions from finitely many observations. (...) The upshot is that defeasible reasoning is not just common; it is thoroughly pervasive and absolutely essential. Almost everything we believe is believed at least indirectly on the basis of defeasible reasoning, and things could not have been any other way.

Some researchers have argued that the term *nonmonotonic logic* does not suit the subject, as this term has an inherently negative nature; it emphasizes on what nonmonotonic logic is not (monotonic). It has been proposed that a better name can be found in the terms "defeasible logic" and "commonsense reasoning". In this thesis, we use the terms nonmonotonic logic, defeasible logic and commonsense reasoning as synonyms.

Some examples of nonmonotonic logics

One of the oldest formalisms for nonmonotonic reasoning is circumscription [McCa86, Lifs94]. In circumscription, the Tweety example can be formalized as follows:

$$\begin{array}{ll} \mathit{Bird}(x) \wedge \neg \mathtt{ab}(x) \supset \mathit{Flies}(x) & (\text{"Birds fly, as long as they are not abnormal."}) \\ \mathit{Penguin}(x) \supset \neg \mathit{Flies}(x) & (\text{"Penguins do not fly; no exceptions."}) \end{array}$$

Circumscription by itself is not a totally new logic; it is merely a different way of applying standard first-order logic. The idea is that instead of taking into account all models in which the given formulas are true, one states as additional condition that a certain predicate (in this case, the ab-predicate) is *minimized*. To illustrate how this works, suppose that Tweety is a bird (Bird(Tweety)). Minimizing the ab-predicate results in models in which the above formulas are true and no object has the ab-property. Now suppose that in addition Tweety is also a penguin. Then it follows from the first rule above that ab(Tweety). Minimizing the ab-predicate then results in models in which the above formulas are true and the only object that has the ab-property is Tweety.

Circumscription's idea of not taking into account all possible models, but instead taking a *subset* of them, is generalized by the formalism of *preferential entailment* [Shoh87, Shoh88]. Preferential entailment is centered around a triple $\mathcal{M} = (M, \models, <)$ where M is a set of models, $\models$ is a satisfaction relation between models and formulas (that is: $\models \subseteq M \times L$), and $<$ is the preference relation (that is: $< \subseteq M \times M$). A model m *preferentially entails* a set of formulas A (written $m \models_< A$) iff m is a minimal element (under $<$) of $|A|$ (where $|A|$ stand for $\{n \mid n \models A\}$). The notion of preferential entailment is then defined as follows [Maki94, p. 72]:

$$A \mathrel{|\!\sim_<} x \text{ iff for all } m \in M, \text{ if } m \models_< A \text{ then } m \models_< x$$

That is, a formula x is preferentially entailed from a set of formulas A iff x is satisfied in all preferred models in which A is satisfied. Preferential entailment is abstract in the sense that it requires very few restrictons regarding the satisfaction relation ($\models$) and the preference relation ($<$). An overview of some properties of preferential entailment can be found in [Maki94].

Hector Geffner and Judea Pearl distinguish two approaches for the implementation of default reasoning [GePe92]. In the *extensional* approach (as is implemented by, for instance, Reiter's default logic), defaults are regarded as prescriptions for extending one's

set of beliefs. The extensional approach has as advantage that irrelevant information does not influence the entailment. If birds fly and Tweety is a red bird, then Tweety also flies unless there is reason to believe that red birds do not fly. A disadvantage is that additional mechanisms are needed in order for, say, more specific defaults to be preferred in case of conflicts. In the *conditional* aproach (as implemented by ε-semantics; see section 4.2.1), defaults are regarded as beliefs whose validity is bounded to a particular context. A property of the conditional approach is that more specific information is preferred. The default that birds fly is not applicable anymore when the bird in question is a penguin. In that case, the more specific default that penguins do not fly is applied. A disadvantage of the conditional approach is that the addition of irrelevent information can also cause defaults not to be applicable anymore. Geffner and Pearl then define a formalism (conditional entailment) that aims to combine the best of these two approaches [GePe92]

Another well-known example of a nonmonotonic logic is Reiter's default logic [Reit80]. A summary of Reiter's logic is provided in section 5.2.

Donald Nute has specified a formalism for nonmonotonic reasoning that has a visual implementation. In Nute's system, entailment is based on more specific antecedents, and a closed intertwining of the definitions of derivability and non-derivability [NuEr98, NuHH98].

2.1.2 Argumentation systems

Human argumentation, like nonmonotonic logic, has an inherently defeasible nature; this is because the support of a certain thesis not only depends on the existence of a supporting argument, but also negatively depends on the existence of possible counterarguments. The defeasible nature of argumentation can be illustrated by a quote of Perelman [Pere82, p. 2]:

> We can immediately see that dialectical reasoning begins from theseses that are generally accepted, with the purpose of gaining acceptance of other theseses which could be or are controversial. Thus it aims either to persuade or convince. But instances of dialectical reasoning are not made up of series of valid and compelling inferences; rather, they advance *arguments* which are more or less strong, more or less convincing and which are never purely formal.

Perelman argues that concrete argumentation is usually audience-dependent in that it aims to act upon a specific audience. Our interest, however, is from a more general perspective, abstracting away as much as possible from any particular audience. That is, our view of argumentation is that of the philosopher [Pere82, p. 17]:

> While the specialist who addresses a learned society and the priest who preaches in his church know the theseses upon which they can base their expositions, the philosopher is in an infinitely more difficult situation. In principle his discourse is addressed to everyone, to a universal audience composed of those who are disposed to hear him and are capable of following his argumentation.

Rebutting and undercutting

In order to understand the full complexity of argumentation, it is desirable to distinguish between different ways in which arguments can defeat each other. Pollock argues

that there are in fact two main forms of defeat: *rebutting* defeat and *undercutting* defeat [Poll87, Poll92, Poll95]. As an example of the difference between rebutting and undercutting, consider the following argument:

A: The object is red, because Alice says it looks red.
$Says(Alice, Looks(Object, Red)) \Rightarrow Looks(Object, Red)$,
$Looks(Object, Red) \Rightarrow Is(Object, Red)$

Now consider two (separate) possible counterarguments against A:

B_1: The object does not look red, because Bob says it doesn't.
$Says(Bob, \neg Looks(object, red)) \Rightarrow \neg Looks(object, red)$
B_2: Charley says the object is illuminated by a red light.
$Says(Charly, Illuminated(object, red\_light)) \Rightarrow Illuminated(object, red\_light)$

Argument B_1 attacks A on one of its (intermediate) conclusions. In fact, it is an argument for the negation of one of A's conclusions. Argument B_2, however, is more subtle. It does not claim the opposite of some conclusion of A, but merely states additional information under which the reason for believing "Is(object, red)" is no longer a correct one.

If an argument defeats another argument by deriving the negation of one of the other argument's conclusions, then the argument is called a *rebutting* defeater. If, on the other hand, an argument defeats another argument merely by deriving additional information showing that one of the other argument's rules is not a valid reason for its conclusions, then the argument is called an *undercutting* defeater.

Pollock claims that the issue of undercutting defeaters has traditionally been largely ignored in philosophy and AI [Poll87, p. 485]. One exception is Stephen Toulmin, who has explicitly included an "except" clause in his analytical model of reasoning [Toul58].

Reinstatement

To determine whether or not an argument can be considered justified, it is not sufficient to merely look at its defeaters; also relevant is whether the defeaters are defeated themselves. Consider the following example (taken from [Kono88]):

> Suppose Ralph normally goes fishing on Sundays, but on the Sunday which is Mother's day, he typically visits his parents. Furthermore, in the spring of each leap year his parents take a vacation, so that they cannot be visited.

Suppose it is Sunday, Mother's day and a leap year. Then, one can formulate three arguments related to whether Ralph goes fishing or not:

Argument A:
Ralph goes fishing because it is Sunday.

Argument B:
Ralph does not go fishing because it is Mother's day, so he visits his parents.

Argument C:
Ralph cannot visit his parents, because it is a leap year, so they are on vacation.

We say that an argument B *defeats* argument A iff B is a reason against A.

If one abstracts from the internal structure of an argument, as well as from the reasons *why* they defeat each other, what is left is called an *argumentation framework*. An argumentation framework simply consists of a set of (abstract) arguments and a binary defeat relation between these arguments.

Definition 2.1. *An* argumentation framework *is a pair* $AF = \langle Args, defeats \rangle$ *where* $Args$ *is a set of arguments and* $defeats \subseteq Args \times Args$*. We say that* A defeats B *iff* $(A, B) \in defeats$*.*

The argumentation framework of the "Ralph goes fishing" example is shown in figure 2.1.

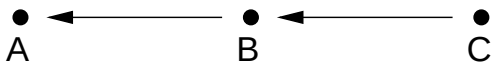


Figure 2.1: Arguments and reinstatement (AF_1).

In figure 2.1, A is defeated by B and B is defeated by C. An interesting question is which of the arguments should be *justified*, that is, which of the arguments should be accepted as the overall result of the defeat relation. Because A is defeated by B, it would at first seem that A should not be justified, since it has a counterargument. If one looks further, however, it turns out that this counterargument (B) is itself defeated by an argument (C) that is not defeated by anything. So, at least C should be justified. But if C is justified, then B is ultimately defeated and does not form a reason against A anymore. Therefore, A should also be justified.

In figure 2.1, we say that argument C *reinstates* argument A.[5] Because of the issue of reinstatement it is necessary to state some formal criterion that takes an argumentation framework and determines which of the arguments are justified and which are not.

Grounded semantics

In his 1995 paper, Dung discusses various principles that can be used to determine whether an argument is justified or not [Dung95]. These principles are referred to as the *semantics* of the argumentation system (the reason for calling it semantics is explained in section 2.2.3). In this thesis, three of these principles will be discussed, starting with what Dung calls *grounded semantics*.

In order to define grounded semantics, it is first necessary to define the notion of acceptability.

Definition 2.2. *Let* $AF = \langle Args, defeats \rangle$ *be an argumentation framework. An argument* $A \in Args$ *is* acceptable *with respect to a set* S *of arguments iff for each argument* $B \in Args$*: if* B *defeats* A *then* B *is defeated by some argument in* S*.*

[5] Although reinstatement is directly or indirectly implemented by many systems for defeasible reasoning, it has been mentioned that in some cases reinstatement can also cause problems; see for instance [Hort01].

Thus, an argument A is acceptable with respect to S iff it is sort of "defended" by S, that is, iff each that defeats A is defeated by some argument in S. As an example, in AF_1 (figure 2.1) A is acceptable with respect to any set containing C.

The next notion to be defined is that of a characteristic function.

Definition 2.3. *The* characteristic function*, denoted by F_{AF}, of an argumentation framework $AF = \langle Args, defeats \rangle$ is defined as follows:*

$$F_{AF} : 2^{Args} \rightarrow 2^{Args}$$

$$F_{AF}(S) = \{A \mid A \textit{ is acceptable with respect to } S\}$$

Definition 2.4. *The* grounded extension *of an argumentation framework AF, denoted by GE_{AF} is the least fixed point of F_{AF}.*

As an example, AF_1 has the grounded extension $\{A, C\}$. It is also possible to define grounded semantics in an inductive way, instead of making use of a fixed point definition.

Theorem 2.1. *Let $AF = \langle Args, defeats \rangle$ be an argumentation framework. Consider the following sequence of sets of arguments:*

- $F^0 = \emptyset$
- $F^{i+1} = \{A \in Args \mid A \textit{ is acceptable with respect to } F^i\}$

Then it holds that:

1. *all arguments in $\cup_{i=0}^{\infty}(F^i)$ are in GE_{AF}*
2. *iff each argument is defeated by at most a finite number of arguments, then an argument is in GE_{AF} iff it is in $\cup_{i=0}^{\infty}(F^i)$*

Proof. See [Dung95]. □

As an illustration of how theorem 2.1 works, take AF_1:

$$\begin{aligned}
F^0 &= \emptyset \\
F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\
&= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\
&= \{C\} \\
F^2 &= \{A \in Args \mid A \text{ is acceptable with respect to } \{C\}\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } C\} \\
&= \{A, C\} \\
F^3 &= \{A \in Args \mid A \text{ is acceptable with respect to } \{A, C\}\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } A \text{ or } C\} \\
&= \{A, C\} \\
F^{i+1} &= F^i \text{ (for } i \geq 3)
\end{aligned}$$

So again, we find that the grounded extension of AF_1 is $\{A, C\}$. As another example to illustrate the working of grounded semantics, take argumentation framework AF_2 (figure 2.2). An intuitive interpretation is as follows [PrVr02]:

A: Dixon is no pacifist because he is a republican.
B: Dixon is a pacifist because he is a quaker,
and he has no gun because he is a pacifist.
C: Dixon has a gun because he lives in Chicago.

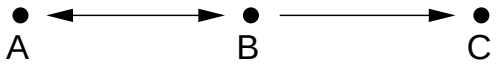


Figure 2.2: Argumentation framework AF_2.

In AF_2 the grounded extension is empty. This can be seen as follows:

$$\begin{aligned} F^0 &= \emptyset \\ F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\ &= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\ &= \emptyset \\ F^{i+1} &= F^i \text{ (for } i \geq 1) \end{aligned}$$

Stable semantics

The second principle to be discussed is that of *stable semantics*. In order to define stable semantics, it is first necessary to define the notion of a conflict-free set of arguments.

Definition 2.5. *Let $AF = \langle Args, defeats \rangle$ be an argumentation framework. A set $S \subseteq Args$ is said to be* conflict-free *iff there are no arguments $A, B \in S$ such that B defeats A.*

A *stable extension* is essentially a conflict-free set of arguments that defeats every argument that does not belong to it.

Definition 2.6. *Let $AF = \langle Args, defeats \rangle$ be an argumentation framework. A conflict-free set of arguments $S \subseteq Args$ is called a* stable extension *iff S defeats each argument in $Args \backslash S$.*

To illustrate how stable semantics works, take the example AF_1. Here, there exists exactly one stable extension: $\{A, C\}$. In AF_2, there exist two stable extensions: $\{A, C\}$ and $\{B\}$.

The possibility of more than one extension poses problems if one is interested in the "overall" conclusions of an argumentation theory. A possible solution is the *sceptical* approach: an argument is justified iff it is part of every extension.

Stable semantics has the problem that it is possible that no stable extension exists. Take for instance the following situation:

A: Alice says that Bob is unreliable.
B: Bob says that Charley is unreliable.
C: Charley says that Alice is unreliable.

This situation is depicted as argumentation framework AF_3 in figure 2.3.

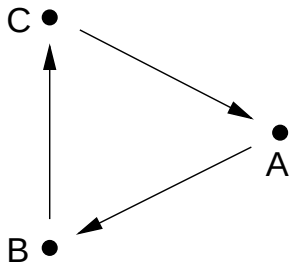


Figure 2.3: Argumentation framework AF_3.

In AF_3, no stable extension exists. This can be seen as follows. Suppose there would be a stable extension S. Then $\{A, B\} \not\subseteq S$, $\{B, C\} \not\subseteq S$ and $\{C, A\} \not\subseteq S$ (this is because otherwise S would not be conflict-free). Therefore, S contains at most one element of $\{A, B, C\}$. But then S does not attack each argument not belonging to S, so S is again no stable extension. Contradiction.

The possible absence of stable extensions was originally seen as indicating that there was something wrong with the argumentation framework that led to this absence. Dung, however, argues that there are perfectly legitimate situations and knowledge representations in which no stable extensions exist [Dung95].
As an aside, under grounded semantics, there is always exactly one extension (which in case of AF_3 happens to be the empty set).

Preferred semantics

The last principle to be discussed is that of *preferred semantics*. Preferred semantics is based on the concept of an *admissible* set of arguments.

Definition 2.7. *Let $AF = \langle Args, defeats \rangle$ be an argumentation framework. A conflict-free set of arguments $S \subseteq Args$ is* admissible *iff each argument in S is acceptable with respect to S.*

Thus, an admissible set S "defends" itself against all possible counterarguments. That is, for each argument A that defeats some argument in S, S contains a (possibly different) argument that defeats A.

Definition 2.8. *A* preferred extension *of an argumentation framework AF is a maximal (with respect to set inclusion) admissible set of AF*

To illustrate how preferred semantics works, take the example of AF_1; here there is exactly one preferred extension: $\{A, C\}$. In example AF_2, there are two preferred extensions: $\{A, C\}$ and $\{B\}$. In example AF_3, there is exactly one preferred extension: $\emptyset$.

An advantage of preferred semantics compared to stable semantics is that there always exists at least one preferred extension.

Theorem 2.2. *Every argumentation framework possesses at least one preferred extension.*

Proof. See [Dung95]. □

Furthermore, the set of all stable extensions is included in the set of all preferred extensions.

Theorem 2.3. *Each stable extension is a preferred extension, but not vice versa.*

Proof. See [Dung95] □

Some examples of argumentation systems

The issue of defeasible argumentation has been extensively studied by the American researcher John Pollock, who throughout the years produced many different versions of his formalism [Poll87, Poll91b, Poll95]. A brief treatment of some of Pollock's work is provided in section 5.3.

Another early formalism for defeasible argumentation is that of Simari and Loui [SiLo92]. In this formalism, priority among arguments is based on specificity, and Pollock's definition of grounded semantics is used to deal with the issue of reinstatement.

An alternative well-known formalism for argument-based reasoning is that of Gerard Vreeswijk [Vree97]. The idea of Vreeswijk's system is to keep the process of argumentation as abstract as possible. For instance, Vreeswijk does not commit himself to one specific principle for dealing with the comparative strength of competing arguments (such as specificity, weakest link or last link). Vreeswijk also provides an implementation that makes it possible that, given a specific knowledge base, is able to construct and evaluate arguments [Vree94]. Baroni, Giacomin and Guida further enhance the framework of Vreeswijk by also allowing the validity of the premises to be discussed [BaGG00].

The last formalism for argument-based reasoning to be mentioned here is that of Prakken and Sartor [PrSa97]. One of the main characteristics of this formalism is that the relative strength of arguments is not fixed in advance and is itself open to argumentation. The reason for this is that in law, one often encounters conflicting principles, and one then has to argue why a certain principle should take preference above another. Prakken and Sartor have also extended their framework to allow for argumentation using precedents [PrSa98].

2.1.3 Dialogue systems

The third concept to be discussed, after nonmonotonic logics and argumentation systems, is that of a *dialogue system*. In their 1995 book, Walton and Krabbe distinguish six main types of dialogues [WaKr95]:

1. persuasion, which is centered around conflicting points of view
2. negotiation, in which participants aim to achieve a settlement that is particularly advantageous for individual parties
3. inquiry, in which the aim is to collectively discover more information, as well as to destroy incorrect information

4. deliberation, which is driven by the need to take a collective decision
5. information seeking, in which one party asks for information known by another party
6. eristics, in which two parties combat each other in a quarrel

Walton and Krabbe claim that the Socratic dialogue is primarily of the information seeking type. This, however, is not completely true. The main purpose of a Socratic dialogue is to convince the other party that its beliefs are incorrect; it is therefore probably better classified as a dialogue of the persuasion type.

Formal models of dialogues are useful not only as an analytical model of the workings of informal dialogues, but can also serve as a proof theory of formal logics. Lorenz and Lorenzen, for instance, have constructed a dialogue system that implements first order logic [LoLo78]. A dialogue system implementing intuitionistic logic is also available [Fels86]. For defeasible reasoning, one of the first researchers to make the connection between nonmonotonic logic and dialectics is Ron Loui [Loui98]. Loui's work has an abstract nature; it does not specify the particular form of the arguments in dialogue. For Loui, an important reason behind the dialogue approach is its ability to deal with resource-bounded reasoning. The idea is to allocate the resources to the party that risks to lose the debate, thus making the reasoning process fair and effective [Loui98, p. 7]. Another researcher who studied defeasible dialectics is Gerard Vreeswijk [Vree95b]. Vreeswijk's system for defeasible dialectics is based on his abstract argumentation formalism; the dialectical approach is proved to be equivalent with the declarative formalism. Like Loui, Vreeswijk acknowledges the computational advantages of the dialectical approach [Vree93, p. 121]. The idea of dialectics as a proof theory for a declarative argumentation formalism is also endorsed by Prakken and Sartor [PrSa97]; they have shown that an argment is justified in their grounded semantics based formalism iff there exists a winning strategy for it in the accompanied dialectical system.

In general, one can distinguish between static dialectics and dynamic dialectics [Hage00]. In static dialectics, all premises are given beforehand, while in dynamic dialectics premises (and rules) can be introduced in the course of the dialogue. For static dialectics, dialectical systems can serve as a proof theory of an underlying logic [Prak97]. The idea is that a thesis is justified in the underlying logic iff there exists a winning strategy in the dialectical proof theory. Thus, the emphasis is in principle on *all* possible dialogues. In dynamic dialectics, the emphasis is on the particular dialogue in question.

One particular difficulty with dynamic dialectics is that it is possible that information introduced by one party may also have been useful to the other party during an earlier stage of the dialogue. For a dialogue to be sound and fair, this requires the possibility that newly introduced information can be used wherever it is relevant, regardless of the state of the dialogue when this new information was introduced [Prak01].

Regarding defeasible argumentation, one can distinguish two different forms of formal dialogues: argument dialectical systems and dialogue systems. One of the main differences is that in an argument dialectical system, moves consist of entire arguments, while dialogue systems for defeasible reasoning also allow for arguments to be rolled out gradually, much akin to a Hamblin-MacKenzie style of arguing. As a result of this, the steps in an argument dialectical system are essentially all of the same type; at each step a complete argument is given. For a dialogue system, at the other hand, steps can involve different kinds of speech acts, like "claim", "why", "concede" and "retract". Figures 2.4 and 2.5 illustrate this

difference. Examples of argument dialectical systems are [PrSa97], [PrVr00] and [Prak01]. Examples of a dialogue systems for defeasible reasoning are [Lodd98] and [Prak00].

P: $a \Rightarrow b \Rightarrow c$
O: $d \Rightarrow e \Rightarrow \neg b$

Figure 2.4: The workings of an argument dialectical system.

P: claim c
O: why c
P: because $b \Rightarrow c$
O: claim $\neg b$
P: why $\neg b$
O: because $e \Rightarrow \neg b$
P: why e
O: because $d \Rightarrow e$
P: retract b, concede $d, e, \neg b$

Figure 2.5: The workings of a dialogue system for defeasible reasoning.

An additional dialogue system for defeasible reasoning that is worth mentioning is Thomas Gordon's Pleadings Game [Gord95]. In the pleadings game, the emphasis is not so much on the weighting and evaluation of arguments in order to resolve the main dispositional issue, but to carefully determine the exact differences of opinion between parties. These differences of opinion are then presented to the judge, whose task it is to resolve them.

A dialectical proof theory for grounded semantics

As was stated earlier, the idea of applying dialectical proof theory for a declarative argumentation formalism has peen persued by Prakken and Sartor for their specific system [PrSa97]. The underlying principles are, however, not bound to the specific features of their system. In fact, it is possible to give *any* grounded semantics based formalism a dialectical proof theory. The now following definitions and proofs are a generalization of [PrSa97] and are relevant for chapter 3.

The first notion to be defined is that of a dialogue tree, which essentially views a dialogue in a gametheoretic way. Recall that the *level* of a tree-node is its distance to the root of the tree (which is 0 in case the node is the root itself).

Definition 2.9. *A* dialogue tree *is a tree of arguments such that each argument (except the root) defeats its parent. Nodes of even level are called* P-moves*; nodes of an odd level are called* O-moves*.*

A *winning strategy* is a restricted form of a dialogue tree, in which all possible countermoves of the opponent are included and the proponent uses exactly one counterargument to address each of them. Furthermore, it is required that each possible dialogue is won by the proponent, that is, the proponent ultimately makes a move to which the opponent has no answer.

Definition 2.10. *A* winning strategy *for argument A is a dialogue tree that has A as root, such that:*

1. *each O-move has exactly one P-move as child,*
2. *the children of each P-move consist of all possible defeaters of the P-move and*
3. *every root-originated path is finite*

We now state and prove that there exists a winning strategy for argument A iff argument A is in the grounded extension.

Theorem 2.4. *Let $AF = \langle Args, defeats \rangle$ be an argumentation framework. There exists a winning strategy for argument $A \in Args$ iff A is an element of the grounded extension of AF.*

Proof.
"$\Longrightarrow$":
Let $A \in Args$ be an argument for which there exists a winning strategy. We now prove that A is in the grounded extension. This is done by induction on the depth (that is: on the highest level) of the winning strategy. Notice that the depth of a winning strategy is always an even number.

base depth = 0. In this case, the winning strategy consists only of the root (the main argument A). So apparently, there are no counterarguments against A. But then, A is in F^1 (using the inductive definition of the grounded extension), and therefore also in the grounded extension.

step Suppose that for each winning strategy with a depth less or equal to i it holds that its root (A) is in the grounded extension. We now prove that each winning strategy of depth $i + 2$ also has its root in the grounded extension. Let WS be a winning strategy of depth $i + 2$. Now consider alle subtrees starting at level 2. These are again winning strategies, with a depth less or equal to i. The induction hypothesis states that the main argument of each of these winning strategies is in the grounded extension. Let F^j be the first set of $F^0, F^1, F^2, \ldots$ that contains all main arguments of these winning strategies. Because A is acceptable with respect to F^j, A is an element of F^{j+1} and is therefore an element of the grounded extension.

"$\Longleftarrow$":
Let $A \in Args$ be an element of the grounded extension. We now prove that A has a winning strategy. This is done by induction on i in F^i.

base $i = 1$. In this case A does not have any counterarguments, so it has a winning strategy consisting of a single node (A itself).

step Suppose that all elements in F^i have a winning strategy. We now prove that alle elements of F^{i+1} also have a winning strategy. Let A be an arbitrary element of F^{i+1} and let $\{B_1, B_2, \ldots, B_n\}$ be the set of defeaters of A. From the definition of F^{i+1} it follows that every B_i $(1 \leq i \leq n)$ is defeated by an element (say C_i) of F^i. According to the induction hypothesis, every C_i $(1 \leq i \leq n)$ has a winning strategy for it. Now take the tree with A as root, $B_1, B_2, \ldots, B_n$ as children and for each i $(1 \leq i \leq n)$ the winning strategy of C_i inserted as a subtree under the associated B_i. It now holds that this tree is a winning strategy for A.

□

Prakken and Sartor state two additional restrictions on an argument game (and therefore also on a winning strategy): the proponent may not repeat earlier moves and the proponent's moves must *strictly* defeat the opponent's preceeding move.

As for the first restriction, disallowing the proponent repeating earlier steps does not affect the existance of a winning strategy. This is because upon the proponent rests the task to ultimately make a move to which the opponent has no answer, and repeating earlier steps does not serve this purpose because the opponent could simply respond with the same answer as before. As for the second restriction, requiring that the proponent's arguments *strictly* defeat the opponent's arguments does not affect the existance of a winning strategy either. This is because when the proponent reacts to an argument (A) with an argument (B) that non-strictly defeats A, then the opponent could react by stating the same argument (A) again, which does not bring the proponent any closer to its aim of winning the dialogue.

Although the above two restrictions do not affect the existance of a winning strategy, they do prevent winning strategies from becoming unnecesary long and complex. In this thesis we therefore apply the above restrictions to all relevant dialogues and winning strategies.

2.2 Research methodology and criteria for evaluation

In this part we ask ourselves the questions what makes a certain system for formal reasoning "appropriate" or "suitable", and according to which methods such a system should be constructed. Apart from practical considerations such as computability and computational complexity, the main reason for constructing a formal logic is its ability to provide an exact and mathematically precise specification of what, given a set of premises, the valid conclusions are. Theoretically, many of such specifications are possible, depending on the specific axioms and derivation rules of the logic. Mathematics provides one with the instruments to specify different forms of (formal) reasoning; it does not by itself specify which forms of reasoning can be considered as "correct".

Often, the aim of a logic is to provide a means of analyzing a certain concept, such as permissions and obligations (deontic concepts), knowledge and beliefs (epistemic and doxastic concepts) and commonsense. Donald Nute states this function as follows [Nute80, p. 2]:

> When we use formalization as an analytic tool, we typically focus upon some part of ordinary language which we wish to clarify, or upon some concept, or set of concepts, which we wish to understand better. (...) To connect the non-formal linguistic structure or concept with the formalization, we devise some uniform method for pairing off the informal claims which might be made using the structure or concept being analyzed with expressions in the formal language. This also has the effect of pairing off informal arguments with sequences of formal expressions and of pairing off the meanings of our informal claims with structures or parts of structures in our formal semantics.

Interpreted in this way, the aim of the logical system then becomes validating exactly those formal arguments whose accompanying informal argument is considered to be valid

[Haac78, p. 15]:

> Formal logical systems aim to formalize informal arguments, to represent them in precise, rigorous and generalizable terms; and an acceptable formal logical system ought to be such that, if a given informal argument is represented in it by a certain formal argument, then that formal argument should be valid in the system just in case the informal argument is valid in the extra-systematic sense.

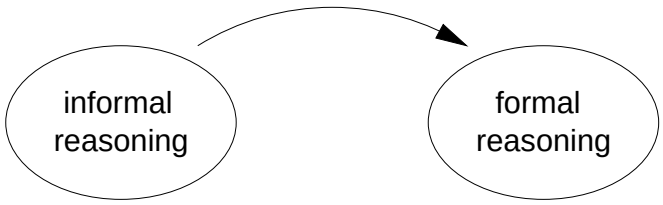


Figure 2.6: The mapping from informal reasoning to formal reasoning.

The problem of constructing a formal logic can to a great extent be seen as a modeling problem. One aims to use formal means in order to represent informal reasoning (see figure 2.6). The problem is that our human knowledge of informal reasoning is usually implicit; it is *procedural* as Pollock argues [Poll95, p. 2]:

> An important fact about procedural knowledge is that although we are rarely in a position to articulate it precisely, we do have the ability to judge, whether, in a particular case, we are conforming to the rules describing that knowledge. Thus, in language processing, although I cannot articulate the rules of my grammar, I can tell whether a particular utterance is grammatical. (...) It just "feels right" or "feels wrong". And similarly, in reasoning, although we may be unable to articulate the rules for reasoning with any precision, we can still recognize cases of good or bad reasoning. These "intuitive" or "introspective" judgments provide the data for constructing a theory about the content of the rules governing our reasoning. In this respect, the construction of philosophical [or formal, MC] theories of reasoning is precisely parallel to the construction of linguistic theories of grammar. In each case, our data are "intuitive", but the process of theory construction and confirmation is inductive, differing in no way from theory construction and confirmation in any other science.

In order to map informal reasoning to formal reasoning, several techniques are present, as well as several pitfalls. The remaining part of this section provides a brief overview of these.

2.2.1 Examples

One of the most simple criteria available is that of the example. The idea is to provide a certain informal example, usually in natural language, that has a certain intuitive conclusion. The idea is then that this informal example can be represented using a system for

formal reasoning, and this system of formal reasoning should then derive conclusions that are in line with the intuitive conclusions.

Examples can play an important role in the construction of formal logic; research in the area of nonmonotonic reasoning, for instance, seems to be driven largely by examples created to show up a perceived strength (or weakness) of a particular formalism [Ethe89]. Haack gives the following description of the role of examples in the construction of formal systems [Haac78, p. 32]

> One could think of a formal logical system as being devised in something like the following way. Some informal arguments are intuitively judged as valid, others invalid. One then constructs a formal language in which the relevant structural features of those arguments can be systematically represented, and axioms/rules which allow the intuitively approved and the intuitively disapproved, arguments. (...)

Vreeswijk describes the method of using examples (which he calls benchmark problems) as follows [Vree93, p. 99]:

> A benchmark problem (in defeasible reasoning) is a problem with several outcomes. Some of these outcomes are in accordance with common sense, while others are counterintuitive, or at least undesired. The outcomes that are in line with commonsense are considered to be the 'right' outcomes of the benchmark problem.
>
> Now, if one wants to test the adequacy of a specific nonmonotonic formalism, the procedure is to let the formalism solve some of the benchmark problems, to see if it delivers the right conclusions. If it does, our conjecture is strengthened that the present formalism is ruled by a sound mechanism of defeat; if it does not, then we have found at least one counterexample that contradicts the adequacy of the formalism in question. This is the way in which benchmark problems play their role in the nonmonotonic research program.

Examples at work: deontic logic

One field in which the use of examples has become truly dominant as a research method is that of deontic logic.[6]

The field of modern deontic logic was more or less started with the introduction of what has become known as Von Wright's Old System. Soon, it was realized that Von Wright's Old System is very close to a KD modal logic, which is nowadays known as Standard Deontic Logic (SDL).

Over the years, many so called "paradoxes" have been stated to criticize SDL as well as other types of deontic logics. A paradox is a small counterexample illustrating that a certain logic needs to be adjusted in order not to derive a counterintuitive result.[7]

Some well-known paradoxes in the field are:

[6]Deontic logic is the field of (formal) reasoning that is concerned with concepts like permission and obligation.

[7]Another interpretation of paradoxes is that the paradox is inherently connected to the style of reasoning. The aim is then not to resolve the paradox, but to use logic as a tool to analyze it. An example of this is the liar's paradox. We will, however, not further go into this interpretation of paradoxes, and instead focus on paradoxes as a method to "tune" logics in order to derive the desired results.

- Ross paradox [Ross41]: $O(p) \vdash O(p \vee q)$
  "If I ought to mail a letter, I also ought to mail or burn it."

- Chisholm paradox [Chis63]: $O(p)$, $O(p \supset q)$, $\neg p \supset O(\neg q)$, $\neg p \vdash \bot$
  "It ought to be that a certain man goes to the assistance of his neighbors."
  "It ought to be that if he does go he tells them he is coming."
  "If he does not go then he ought not to tell them he is coming."
  "He does not go."

One of the researchers who, inspired by the various paradoxes, tried to provide a logic that properly deals with them is John-Jules Meyer. Meyer shows that his dynamic deontic logic properly handles a significant number of paradoxes (except the Ross-paradox) [Meye88]. Unfortunately, after publication of Meyer's work, another paradox was stated, this time in Meyer's system [Meyd90]:

- $[\alpha]P(\beta) \vdash P(\alpha;\beta)$
  "If after shooting the president it is permitted to remain silent, then it is permitted to shoot the president and remain silent".

John-Jules Meyer and Roel Wieringa then propose an alteration of the original dynamic deontic logic to properly deal with this example [MeWi93].

In general, one can say that the desire to correctly deal with a number of standard paradoxes has motivated a great part of the deontic logic related research. Examples hereof are [Cupp94, DiMW94, TaTo96, ToTa97, PrSe97, CaJo97, Torr97], as well as many others.

Criticism of the use of examples

One advantage of using paradoxes for the construction of logics is that paradoxes can play a similar role as experiments in empirical sciences. That is, they allow a logical formalism to be falsified (Popper) and one may say that one logical formalism is preferred to another if it correctly solves a proper superset of the paradoxes solved by the other formalism.

Still, the method of using paradoxes or small intuitive examples is not without disadvantages. As observed by Prakken [Prak02], some apparent counterexamples are in fact based on "hidden", implicit information. Consider the following example (taken from [Pear92]):

> If you marry Ann, you will be happy; if you marry Nancy, you will be happy as well. Does this mean that you will be happy if you marry both of them?

At first, this may suggest that from a knowledge base $\{A \Rightarrow H,\ N \Rightarrow H\}$ one should not be allowed to infer $A \wedge N \Rightarrow H$, because the above example appears to argue against it. Thus, the property of *left conjunction* (see page 91) should be given up. But in fact, the reason that one intuitively rejects $A \wedge N \Rightarrow H$ in the above example is that one has the implict background knowledge that marrying both of them does not make one happy at all; it will rather make one end up in jail instead. If *this* particular piece of information is explicitly added, then the resulting knowledge base becomes $\{A \Rightarrow H,\ N \Rightarrow H,\ A \wedge N \Rightarrow H\}$ and most logics for default reasoning would not entail $A \wedge N \Rightarrow H$ in the first place. Notice that the use of examples with implicit, "hidden" information is more problematic in defeasible logic than it is for classical, monotonic logic. If certain information has not formally been

modeled, then a monotonic logic may entail *no* answer, whereas nonmonotonic logic may entail the *wrong* answer.

Another criticism against the use of examples is that it can lead research to become focussed on a relatively small set of examples well-known in the research community[8], which may not be representative for the full complexity of the problem of (formal) reasoning. Gerard Vreeswijk puts it as follows [Vree93, p. 97]:

> Nowadays, the research program seems to be driven largely by examples that are created to demonstrate a perceived strength of a particular formalism. Ultimately, the enterprise seems to come down to obtaining the 'right' answers to a small set of simple problems. Such a pragmatic approach is not entirely satisfactory, even more if one comes to think of the possibility that unsuitable examples may be ignored, so to speak, 'for the sake of convenience'.

Ginsberg notices essentially the same problem [Ethe89, p. 502]:

> Instead of focusing on the result of applying our intuitions to a broad range of problems, we have focussed on a few trivially simple examples. (...) I'm especially troubled by what seems to me to be a trend not to formalize our intuitions, but to get the 'right' answers to a small set of simple problems.

A more fundamental objection against the use of examples for "tuning" a logical formalism is stated by Vreeswijk in his Ph.D. thesis. In order to understand Vreeswijk's objection, one has to be aware that a logical formalism, at the most abstract level, can be seen as a "derivation function" that takes a set of input formulas (the premises) and outputs a set of conclusions. A paradox or example then essentially states that a certain input should be associated with a certain output; say, an input consisting of formulas A, B and C should result in an output containing (or not containing) D. The more paradoxes a logical formalism solves, the more restrictions have been implemented on its "derivation function".

To some extent, one can compare the task of implementing a number of paradoxes into a system for formal reasoning with the task of defining a geometric function that goes through a set of predefined points (see figure 2.7). This observation leads to what Vreeswijk calls the *interpolation theorem*.

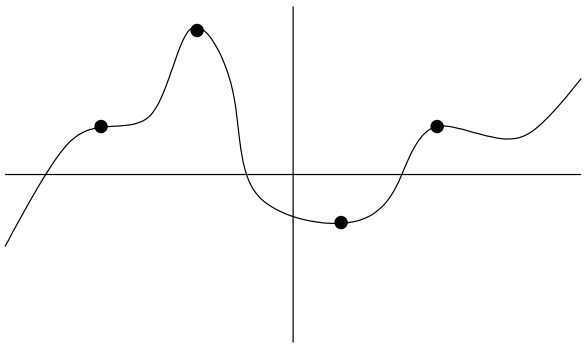


Figure 2.7: Defining a formalism that complies with a pre-given set of examples.

[8] An overview of these examples is provided in [Vree95c].

Theorem 2.5 (Interpolation theorem for defeasible reasoning[Vree95a]).
For every finite number of benchmark problems with predefined but non-interfering conclusions, there exists a defeasible logic that complies with every benchmark problem and its conclusion.

The significance of the interpolation theorem is explained as follows [Vree93, p. 103]:

> A comparison with the theory of program correctness might be helpful here. As is known, a computer program may proven to be correct in different ways, varying from a mathematical correctness proof (the only right way to do it, but nobody does it) to a nearly exhaustive execution of all possible inputs (a not-so-good way to do it, but everybody does it). Elaborating on this comparison, we might say that tuning a particular mechanism of defeat against a collection of benchmark problems is as questionable as performing sample-data checks to show up the correctness of a particular computer program. One might pursue the comparison even further by stating that the use of benchmark problems is only appropriate in situations where the *in*correctness of a particular mechanism or feature of defeat is at issue. The [interpolation] theorem was used to show that such an inductive proof method , i.e. a method that moves from particular instances to a general conclusion, actually proves very little.

To summarize our discussion, we do think that examples can be an interesting starting point of discourse about when a system for formal reasoning has properly implemented a particular form of informal reasoning, although the method (of using examples) by itself may not be sufficient to provide a definite answer to this question.

2.2.2 Postulates

In the previous section, it was argued that the method of examples by itself is insufficient to validate a certain system for formal reasoning. This leads to the standpoint that perhaps, what is needed are criteria that are more general. A possible approach would be to formulate a number of plausibly sounding principles that the entailment-relation has to fulfill. These kind of principles are also known as *postulates*.

Some examples of possible postulates for nonmonotonic logic are given by Makinson [Maki94]. Each postulate is given both in the form using the defeasible entailment relation ($\mid\!\sim$), as in the form using the defeasible consequence set ($\mathcal{C}(\ldots)$):

- reflexivity / inclusion
  $A \mid\!\sim x$ whenever $x \in A$
  $A \subseteq \mathcal{C}(A)$

- cut
  $A \mid\!\sim x$ whenever $A \mid\!\sim y_i$ for all $i \in I$ and $A \cup \{y_i \mid i \in I\} \mid\!\sim x$
  $A \subseteq B \subseteq \mathcal{C}(A)$ implies $\mathcal{C}(B) \subseteq \mathcal{C}(A)$

- cautious monotonity
  $A \cup \{y_i \mid i \in I\} \mid\!\sim x$ whenever $A \mid\!\sim y_i$ for all $i \in I$ and $A \mid\!\sim x$
  $A \subseteq B \subseteq \mathcal{C}(A)$ implies $\mathcal{C}(A) \subseteq \mathcal{C}(B)$

- cumulativity (= cut + cautious monotonity)
  If $A \mathrel{|\!\sim} y_i$ for all $i \in I$ then $A \mathrel{|\!\sim} x$ iff $A \cup \{y_i \mid i \in I\} \mathrel{|\!\sim} x$
  $A \subseteq B \subseteq \mathcal{C}(A)$ implies $\mathcal{C}(A) = \mathcal{C}(B)$

Makinson gives the following treatment on the relevance of these postulates [Maki94, p. 43]:

> Why should these conditions be seen as important? Because they correspond to certain very natural and useful ways of organizing our reasoning. They tell us that when we are reasoning, we may accumulate our conclusions into our premises without loss of inferential power (cautious monotonity) or amplification of it (cut). In this sense, the reasoning process is taken to be stable.

Although postulates often have an intuitive ring to them, they may not always be satisfied by existing formalisms for defeasible reasoning (see the discussion on cautious monotony on page 70). Apart from that, postulates share the same disadvantage as examples in that the set of postulates one uses to evaluate a formal logic is not necessary a complete one.

A field where the use of postulates plays a prominent role is that of belief revision, where the AGM postulates have become a touchstone [AlMa82, Gärd82, AlGM85].

2.2.3 Semantics

The aim of formal semantics is to determine what it actually *means* for a certain formal statement to be valid. The idea is that derivability in a logical formalism should coincide with validity in underlying mathematical structures (the formal semantics).

Ideally, formal semantics can act as a bridge between the informal, intuitive reasoning of humans and the formal, mathematical reasoning of the logical formalism (see figure 2.8).

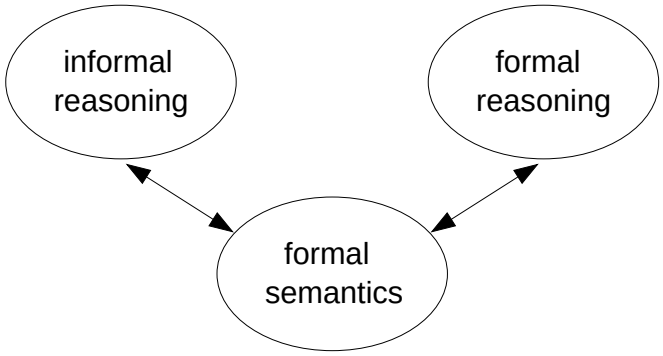


Figure 2.8: Formal semantics as bridge between informal and formal reasoning.

In this section, two types of formal semantics are to be discussed: model-based semantics, and dialogue- or argument-based semantics.

Model-based semantics

The notion of model-based semantics as is applied in modern logics was originally developed by Tarski [Tars56]. The idea is that a formula F should be derivable from a set of formulas S iff in every model in which all formulas of S are true, F is also true. Semantics in the form of model theory has been successfully applied in fields including propositional logic,

predicate logic and modal logic. One can say that the development of model theory has to a great extent solved the foundational crisis in the development of formal logics in the first part of the 20th century.

This is not to say that semantics in the form of model theory can be applied equally well in every particular field of logic. For defeasible reasoning, for instance, the application of model theory results in a semantics of preferred models, which is not always easy to be understood or worked with. The relevance of any kind of formal semantics, however, largely depends on the "intuitive appeal" of the semantics. If one cannot make clear that the formal semantics is closely related to intuitive notions of what is true and what is not, then the whole issue of semantics becomes nothing but a technical exercise.

One of the critics of formal semantics in the field of nonmonotonic reasoning is Pollock [Poll91b, p. 40]:

> My own opinion is that the importance of model theoretic semantics is vastly overrated. Experience in formal logic has indicated that it is possible to construct model theoretic semantics for even the most outlandish logical theories. The mere existence of a model theoretic semantics shows nothing at all about the correctness of the theory. If a theory is already known to be correct, then the discovery of a formal semantics for it can be a useful technical tool in its investigation. But the formal semantics is not itself an argument for the correctness of the theory unless there is some independent reason for thinking that the semantics is correct. The systems of formal semantics that define various species of nonmonotonic logic do indeed have initial plausibility, but I would argue that their authors have sometimes been driven more by considerations of formal elegance than by an appreciation of the subtleties required of defeasible reasoning for a fullblown epistemology. Before getting carried away with a semantical investigation, we should be sure that the theory described by the semantics is epistemologically realistic, and that requires attending to the nuts and bolts of how defeasible reasoning actually works.

A general remark on this issue is made by Vreeswijk [Vree93, p. 81] [Vree92]:

> Within the school of defeasible argumentation, the problem of finding an adequate model theoretic semantics is traditionally considered to be a problem of only minor significance. In fact, several authors have strongly questioned the relevance of model theoretic semantics [Loui90, Poll91b]. Their plea against such 'exercises in mathematics' (Ron Loui) is clear and convincing.

Dialogue and argument-based semantics

An alternative way of defining formal semantics is by regarding how people try to convince each other of their respective point of view. A possible instance of such a semantics comes in the form of a *dialogue game*. The earlier mentioned dialogue game of Lorenz and Lorenzen [LoLo78] is an example of such a dialogue-based semantics.

For nonmonotonic logic, a possible semantics comes in the form of an argument framework and accompanying justification-principle (see section 2.1.2). An important result came when Dung *et. al.* showed that many formalisms for nonmonotonic reasoning — like Reiter's Default Logic, Circumscription and Pollock's system — can be described in terms

of argumentation [Dung95, BDKT97]. Some argumentation systems, like that of Prakken and Sartor [PrSa97], make the link between argumentation and dialogues. One particular effect of this is that the thus newly derived knowledge becomes what rational agents can agree upon in a fair debate. This, however, requires that the rules of the debate are in fact fair and that agents are free to state all possible arguments and counterarguments that can be regarded as relevant.[9]

2.2.4 Intuitions

Until now, much has been said about the possible ways for informal reasoning to be modeled by formal reasoning. One question that has barely been touched, however, is what we actually mean by informal reasoning. Or, to put it in other words, what is it that we are actually trying to model?

Informal reasoning, for the lack of any workable alternative, can be seen as the kind of reasoning that people intuitively regard valid. Therefore, like Pollock [Poll95] and Haack [Haac78], we base our formalization on human intuitions. The next question then becomes: *whose* human intuitions? Are we basing our formalism on the intuitions of "ordinary" people (the *logica utens*) or on the intuitions of those who have systematically studied the field of logic (the *logica docens*[10]). Both of these choices have their downsides.

logica utens

The logica utens can be described as one's unreflective judgment of the validity of informal arguments [Haac78, p. 15]. Thus, one (rather naive) way to construct a logic would be to investigate what people do when they reason, and try to formulate a formal model of this.

The problem with this is that people do not always reason correctly [Poll95, p. 2]. As an example, people are sometimes persuaded by fallacious arguments, such as an *argumentum ad hominem* (personal attacks) or an *argumentum ad verecundiam* (illegitimate claim to authority). An overview of various kinds of falleous arguments is provided in [EeGK87].

Other examples of erroneous human reasoning are provided by Ross and Anderson [RoAn82, pp. 147-149]. They mention various psychological experiments in which people were provided with feedback that had no correspondance to how they actually performed their task during the experiment, After the experiment was finished, the subjects were informed of this. It turned out, however, that even in cases where the subjects explicitly acknowledged understanding the the feedback they had received beared no actual correspondance to how they performed their task, part of their thus discredited knowledge persevered. For example, those who received positive feedback (that is, they were originally told that they performed well on their assigned task) turned out to have signifiantly more confidence in their performance and ability than those who received negative feedback. This phenomenon has been observed in various different experiments.

There is evidence that people's mental theories can persist, even when the grounds they are based on are refuted [RoAn82]:

> [Other] studies first manipulated and then attempted to undermine subjects' theories about the functional relationship between two measured variables: the

[9] ...and that includes the concept of HY-arguments.

[10] The terms *logica utens* and *logica docens* are borrowed from medieval logics.

> adequacy of firefighters' professional performances and their prior scores on a paper and pencil test of risk performance. (...) [S]uch theories survived the revelations that the cases in question had been totally fictitious and the different subjects had, in fact, received opposite pairings of riskiness scores and job outcomes. (...) [O]ver 50% of the initial effect of the "case history" information remained after debriefing.
>
> In summary, it is clear that beliefs can survive (...) the total destruction of their original evidential bases.

Based on these results, Harman treats the *foundational theory* and the *coherence theory* of belief revision. The key difference is whether one keeps track of one's original justifications for beliefs; foundational theory does whereas coherence theory does not [Harm86, p. 30]:

> (...) It turns out that the theories are most easily distinguished by the conflicting advice they occasionally give concerning whether one should *give up* a belief P from which many other of one's beliefs have been inferred, when P's original justification has to be abandoned. Here a surprising contrast seems to emerge — "is" and "ought" seem to come apart. The foundations theory seems, at least at first, to be more or less in line with our intuitions about how people *ought* to revise their beliefs; the coherence theory is more in line with what people *actually do* in such situations. Intuition seems strongly to support the foundations theory over the coherence theory as an account of what one is *justified* in doing in such cases; but *in fact* one will tend to act as the coherence theory advises.

Harman then studies how this discrepancy can be resolved; we will, however, not go into this.

Another example of incorrectness of human reasoning is Wason's well-known card experiment [Waso66, p. 145]. In this psychological experiment the situation was as follows. The subjects (students) were presented with an array of cards. They were told that every card had a letter at one side and a number at the other side; the situation was such that only one side was visible. They were then asked which cards they would *need* to turn over to determine whether the experimenter was lying in uttering the following statement: "if a card has a vowel on one side, then it has an even number on the other side." (see also figure 2.9)

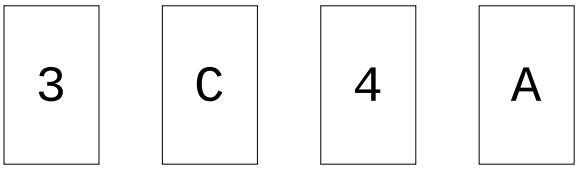


Figure 2.9: Which cards to turn?

The correct response, of course, is to turn around the cards displaying vowels (as there may be an odd number on the other side), as well as the cards displaying odd numbers (as there may be a vowel at the other side). The rationale is that only the combination "vowel and odd number" can falsify the given statement, so one should only regard the cards in

which this combination is possible at the first place. Wason, however, reports that this answer was given by only a minority of the subjects; the most frequent response was to select cards displaying vowels and cards displaying even numbers [Waso66, p. 146].

Other researchers have done similar experiments in which people turned out to apply incorrect forms of reasoning; an overview is given in [ChHo85] and will not be repeated here. Cheng and Holyoak conclude that "the view that people typically reason in accord with formal logic has been overwhelmingly refuted by the evidence based on experiments in conditional reasoning" [ChHo85, p. 394].

logica docens

Given the fact that ordinary people often commit fallacies when reasoning,[11] it may be a better idea to base a system for formal reasoning on the intuitions of those who have systematically studied the field of logics. The more rigorous and precise judgments on the validity of informal arguments, as, through reflections on these judgments, formal systems are devised, are called the *logica docens* [Haac78, p. 15].

The idea is that people who have allowed for their intuitions to be criticized and examined are more likely to come up with correct forms of reasoning than those who have not. There is, however, one important pitfall. Most people that have nowadays studies the field of logics are accustomed with (possibly several) forms of formal reasoning. The conversion of intuitions into a system for formal reasoning, however, is not always a one-way process, as pointed out by Susan Haack [Haac78, p. 15]:

> (...) One may begin to develop a formal system on the basis of intuitive judgments of the extra-systematic validity of informal arguments, representing those arguments in a symbolic notation, and devising rules of inference in such a way that the formal representation of informal arguments judged (in)valid would be (in)valid in the system. Given these rules, though, other formal arguments will turn out to be valid in the system, perhaps formal arguments which represent informal arguments intuitively judged invalid; and then one may revise the rules of the system, or one may, instead, especially if the rule is agreeably simple and plausible and the intuition of informal invalidity not strong, revise one's opinion of the validity of the informal argument or else one's opinion of the appropriateness of representing that informal argument in this particular way. And once a formal logical system becomes well-established, of course, it is likely that it will in turn tutor one's intuitions about the validity or invalidity of informal arguments.

Another researcher who observes this tutoring effect is Donald Nute [Nute80, p. 3]

> To the extent that these constructions are guided by ordinary usage and intuitions, our formalization is descriptive. But, as I have already hinted, our formalization may also turn out to be normative. We may find that certain sequences of sentences in our formal language turn out to be (or not to be) deviations in our logic at the same time that informal arguments corresponding to

[11] The point is that these fallacies are often not incidental. As a comparison, someone involved in a complex mathematical calculation might make an incidental mistake, but still be aware what arithmatical principles do and do not hold. For informal reasoning, the evidence treated earlier indicates that falleous reasoning results not from incidental mistakes, but from applying fundamentally flawed principles.

> these sequences of artificial sentences are not (or are) intuitively valid. We may find that some artificial sentence is (or is not) satisfied by some structure in our formal semantics even though the corresponding English sentence is not (or is) intuitively true. If and when this happens, we will discover that our initial preformal or unreflective intuitions and our formalization do not exactly 'fit' each other. We might try to improve the fit by putting them together a bit differently. We do this by changing our method for symbolizing English sentences into the artificial language. We also may try to get a better fit by altering the logic and/or the formal semantics. But there is also a third alternative. If we are strongly committed to the symbolization and also strongly committed to the axioms and rules of the logic, and to the basic structures of the semantics, we may alter our intuitions to better fit the formalization. (Less drastically, we may form intuitions to fit some feature of the formalization where we had no preformal intuitions at all) Where we actually change our intuitions, we could see what we are doing as bringing our personal understanding of the conventions governing the use of conditionals more closely in line with the implicit, communal conventions, or we could see what we are doing as actually *reforming* our conventions in order to bring *them* into line with some general principles about linguistic conventions to which we have become committed, and with which our original conventions do not agree. (...) To the extent that our exercise in formalization results in an alternation of our original intuitions about conditionals, our formalization results in an alteration of our original intuitions about conditionals, our formalization will be normative.

As a personal note, the author of this thesis has observed the process of formal systems tutoring human intuitions through the years that he assisted teaching an introductory logic-course to first-year students of computer science. After finishing the treatment of the regular exercises I often provided the students with an additional problem for which I asked their opinions. One of these problems was: "all unicorns are purple".[12] The first impression, often, was that this was a weird example for which the students had no intuitions. I then reminded them that $\forall x \in U : P(x) \equiv \neg\exists x \in U : \neg P(x)$ (a rule that has by that time been illustrated by a few examples from non-empty domains to show its intuitiveness), and that because there are no unicorns that are not purple, all unicorns are in fact purple. The students usually accepted this argument, as would most people trained in formal logic.

Haack argues that in general, one should seek to find a balance between the objective of getting all original intuitions represented and the objective of producing a formalism that is not overly complex [Haac78, p. 33]:

> However, if formal logic faithfully followed informal arguments in all their complexity and vagueness there would be little point in formalization; one aims, in formalizing, to generalize, to simplify, and to increase precision and rigor. This means, I think, that one should neither expect nor desire a direct formal representation of all the informal arguments judged, extra-systematically, to be valid. Rather, pre-systematic judgments of validity will supply data for the

[12] Well, actually the problem was "Every girl in this classroom has red hair", but as there were usually no girls attending the class (computer science, like any other technical study, attracts an almost exclusively male audience) the problem is essentially of the same type.

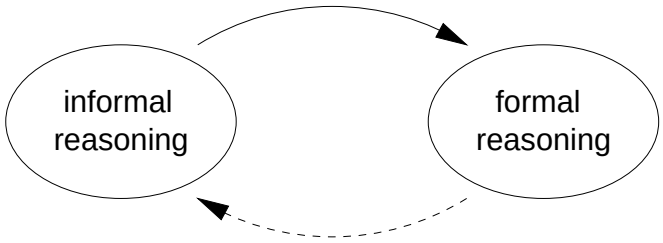


Figure 2.10: Formal systems having an effect upon human intuitions.

> construction of a formal logic, but considerations of simplicity, precision and rigor may be expected to lead to discrepancies between informal arguments and their formal representations, and even in some cases perhaps to a reassessment of intuitive judgments. One uses intuitive judgments of some arguments to construct a formal theory which gives verdicts, perhaps quite unexpected verdicts, on other arguments; and one might eventually sacrifice some of the original judgments to considerations of simplicity and generality. (...)
>
> One should recognize, then, that a failure on the part of a formal system to represent *all* the knobs and bumps of the informal argument is not necessarily objectionable. On the other hand, one must be wary of assuming that *all* adjustments are acceptable; one needs to ask whether the gains in simplicity and generality compensate the discrepancy.

To round off our discussion, the issue of which intuitions to model is a particularly tricky one. Obviously, the unreflective intuitions of ordinary people need not to be correct. The intuitions of the "experts", however, are likely to have been influenced by the formalisms they have been working with [Gins94, p. 16], and one has to be aware that the particular quirks of the formal systems may have found a way into the intuitions of the researcher (we come back on this on page 97).

2.2.5 Mirror examples

The last technique to be discussed is what we call a *mirror example*. The purpose of a mirror-example, which we propose as an addition to already existing techniques, is to gain more insight into the specific field of reasoning, as well as in the pitfalls of logical modeling.

To illustrate this concept, we refer to a publication of Horty [Hort01], in which he — coming from the field of inheritance networks — criticizes the commonly used logics for defeasible reasoning.[13] One of the paradoxes Horty specifies is the following; we call it the "Microsoft Millionaires":

$$
\begin{array}{rcrcll}
\mathcal{S} & = & \{ & \rightarrow & NME, & \\
 & & NME & \rightarrow & ME, & \\
 & & LHT & \rightarrow & \neg MLN\} & \\
\mathcal{D} & = & \{ & ME & \Rightarrow MLN, & (r_1) \\
 & & & NME & \Rightarrow LHT, & (r_2) \\
 & & & & \Rightarrow \neg LHT & (r_3)\}
\end{array}
$$

$r_1 < r_2 < r_3$

[13] Much of Horty's criticism is answered in [Prak02].

P: Ann is a millionaire (MLN), because she is an employee of Microsoft (ME).
$\rightarrow NME,\ NME \rightarrow ME,\ ME \Rightarrow MLN$ (argument A_1)

O: Not at all, she has only been working there since a couple of days (NME), so she probably doesn't even have \$100.000 (LHT).
$\rightarrow NME,\ NME \Rightarrow LHT,\ LHT \rightarrow \neg MLN$ (argument A_2)

P: I think she does own \$100,000 because a few months ago she inherited \$200,000 from her parents.
$\Rightarrow \neg LHT$ (argument A_3)

The opponent now does not have any counterarguments anymore, so the main thesis — that Ann is a millionaire — is considered justified, even though this is contrary to our intuitions; argument A_3 does not seem to be a proper reason to reinstate argument A_1.

If our criterion for the evaluation of logics would be solely based on the application of examples or paradoxes, then the above example would be a reason to start adjusting the entailment properties of the logic in order to suit the paradox. Before we proceed, however, it can be interesting also to look at another example, which we will call "drunk Anton":

$$
\begin{array}{lcl}
\mathcal{S} & = & \{ \quad \rightarrow\ BK, \\
 & & \quad BK \rightarrow\ NL, \\
 & & \quad AB \rightarrow\ \neg DG\} \\
\mathcal{D} & = & \{ \quad NL \Rightarrow\ DG, \quad (r_1) \\
 & & \quad BK \Rightarrow\ AB, \quad (r_2) \\
 & & \quad \Rightarrow\ \neg AB \quad (r_3)\}
\end{array}
$$

$r_1 < r_2 < r_3$

P: Anton drinks gin (DG) because he is Dutch (NL).
$\rightarrow BK,\ BK \rightarrow NL,\ NL \Rightarrow DJ$

O: No, he doesn't, because he is a member of the Dutch association against alcohol abuse ("Blauwe Knoop", BK) so he probably abstains (AB) from drinking any alcohol at all.
$\rightarrow BK,\ BK \Rightarrow AB,\ AB \rightarrow \neg DG$

P: Anton definitely drinks, people even saw him drunk on the street.
$\Rightarrow \neg AB$

Here, the existing logical formalism (in this case, the logic of P&S) *does* provide the intuitive answer: Anton probably drinks gin. This is because the only reason why he would not do so is that he doesn't drink at all, and this reason is defeated. And if we cannot hold that he doesn't drink at all, then he probably (also) drinks gin, since the Dutch tend to like it.

A similar phenomenon is observed by Vreeswijk [Vree91], who provides two examples which we refer to as "young adults" and "bankrupt conservatives" [14]:

Adults are usually employed: $a > e$
University students are usually not employed: $u > \neg e$

[14] The examples "young adults" and "bankrupt conservatives" were originally introduced by, respectively, Pearl [Pear88] and Prakken [Prak93].

Young adults are usually university students: $(y \wedge a) > u$
Sam is a young adult: $y \wedge a$

Conservatives are usually selfish: $c > s$
Poor people are usually not selfish: $p > \neg s$
Bankrupt conservatives are usually poor: $(b \wedge c) > p$
John is a bankrupt conservative: $b \wedge c$

Vreeswijk argues that, intuitively, in the "young adults" example, the outcome should be $\neg e$, while the intuitive outcome of the "bankrupt conservatives" example should be s. The problem, however, is that the examples are essentially one and the same, apart from the syntactical details (see figure 2.11).

$a > e$	$u > \neg e$	$(y \wedge a) > u$	$y \wedge a$	($\neg e$ is an intuitive conclusion)
$c > s$	$p > \neg s$	$(b \wedge c) > p$	$b \wedge c$	($\neg s$ is *not* an intuitive conclusion)

Figure 2.11: "young adults" and "bankrupt conservatives" as mirror example.

A pair of examples, like "Microsoft Millionaires" and "drunken Anton", or like "young adults" and "bankrupt conservatives", we call *mirror examples*. The key property of a mirror example is that there are two intuitive examples IE_1 and IE_2 which (apart from syntactical sugar) share the same formalization FE ("formalized example"), while in IE_1 C is an "intuitive" conclusion, whereas in IE_2 C is *not* an intuitive conclusion.[15]

Mirror examples can provide great difficulties when using the method of examples or paradoxes for evaluating the appropriateness of a certain logic. Recall that the idea of this method is to find a derivation function in such a way that, provided with the premises of the example, it outputs the "desired" conclusions. With mirror-examples, however, we have two similar (apart from syntactic sugar) "inputs" that require a different "output". It is like we would have to find a geometrical function that given one particular x-value would have to cross two points with different y-values, which is by definition not possible.

The situation of mirror-examples, like the one illustrated above, allows for several different forms of diagnosis (note that we are not interested in the particular mirror-example described above, but in the concept of mirror-examples in general).

1. One of the formalizations — perhaps even both of them — is "wrong" in the sense that it does not adequately represent the associated intuitive example.[16] The example should be modeled in a different way — still using the syntactical constructs of the existing logical system — so that the existing logical system derives the "right" answer. While this approach is in itself perfectly legitimate, it does, however, require an explanation of *why* aforementioned representation is wrong, and requires

[15] We use the term "mirror example" because both examples appear similar, yet at the same time they can be seen as each other's opposite.

[16] Take for instance the "young adults" / "bankrupt conservatives" mirror example. One distinction that has not been modelled here is that whereas university students are usually adults ($u > a$), it is not the case that poor people are usually conservative. If $u > a$ is added to "young adults" and specificity is applied as criterion in case of conflicts, then $u > \neg e$ is preferred to $a > e$, so $\neg e$ is entailed. In the "bankrupt conservatives" example, on the other hand, the lack of a default $p > c$ makes that $p > \neg s$ is not preferred to $c > s$, so $\neg s$ is not entailed. In general, one of the requirements for a formalization to be adequate is that all relevant information has explicitly been modeled.

guidelines about how intuitive examples should be modeled in the formal system, for if those guidelines are lacking, one has to adjust the representation of the problem based on the outcome of the logic, which is clearly not a desirable thing to do.

2. The difference between the two intuitive examples is related to a concept that is not present in the logical system that is being used. If, for instance, IE_1 and IE_2 at first seem to share the same formalization (in say, propositional logic), but on closer inspection one of the implications of, say IE_2 turns out to be a counterfactual, then trying to change one of the example's formalizations using the same logical system will not provide the desired results. Instead, what is needed is a construct or form of reasoning that is beyond the scope of the original logical system. For this, there are two options:

   (a) the construction of a new logical system that can deal with both types of reasoning represented by the intuitive examples IE_1 and IE_2

   (b) the construction of two separate logical systems, one that can deal with the type of reasoning represented by IE_1, and one that can deal with the type of reasoning represented by IE_2. While this is less ambitious than the construction of a single unifying logical system, it still provides a conceptual understanding of two different styles of reasoning, as well as a proposal how these can be formalized.

The notion of mirror-examples is relevant because it *forces* those confronted with them to carefully think about the concepts that are implemented by the logical system(s) in question, as well as about the various concepts that play a role in intuitive reasoning.[17] In our view, the approach of analyzing what lies behind the apparent conflict of he mirror example is to be preferred to Vreeswijk's original approach that in such situations the formalism should not draw a conclusion at all.

[17]The method of stating and analyzing mirror examples can be compared with the dialectical method as described by Hegel [Webe08]. In the Hegelian dialectic, one concept (thesis) together with its opposite (antithesis) generates a new concept (synthesis) that unifies both thesis and antithesis. This is akin to the resolution of mirror examples by means of method 2a; one devises a logical formalism that can deal with both parts of the mirror example.

Chapter 3

HY-arguments and their formalization

In this chapter, the concept of HY-arguments is illustrated using an existing formalism for defeasible argumentation. First, it is shown how the lack of HY-arguments can yield unintuitive results (section 3.1). The next step is then a conceptual analysis of the workings of this kind of arguments (section 3.2). Using the results of this, a proposal is made for including HY-arguments into the sample formalism for defeasible argumentation (section 3.3).

3.1 The problem

The first step is to illustrate the difficulties that can result if HY-arguments are not supported in a certain formalism for defeasible argumentation. The problem will be explained using a simplified version of the argumentation system of Prakken and Sartor (P&S). The reasons for choosing the argumentation framework of P&S are as follows.

1. It supports rebutting as well as undercutting; both concepts are needed in order to illustrate the workings of HY-arguments.

2. It supports priorities among rules; this means that it can also be shown how priorities work with respect to HY-arguments.[1]

3. It is relatively simple; arguments are constructed of rules and nothing else. Non-defeasible entailment is obtained by applying strict rules, defeasible entailment by applying defeasible rules. This results in a fair level of uniformity between strict and defeasible reasoning.

In order for the discussion not to become overly complex, some particular features of the original argumentation system of P&S have been altered or left out. In particular, we assume that priorities among rules are fixed and not open for discussion. Furthermore, we take *weakest link* as the principle for lifting the priority ordering among rules to a priority

[1] Despite the fact that HY can work with priorities, we often chose to have equal priority of all defeasible rules for the sake of simplicity.

ordering among arguments.[2] A third alteration is that rebutting and undercutting are considered to be equal, that is, one does not take priority above the other in the overall definition of defeat. The resulting system will for simplicity be called "the formalism of P&S", although strictly, what is meant is a somewhat simplified variant.

Formalization

Given the above considerations, the reference system can be described using the definitions 3.1 until 3.11. The first notion to be defined is that of a literal.

Definition 3.1. *Let props be a set of atomic propositions. Now define a set literals = props* $\cup \{\neg p \mid p \in props\}$*. A negation function (*$- : literals \longrightarrow literals$*) is defined by* $-P = \neg P$ *and* $-\neg P = P$*.*

Notice that the negation function has a purely syntactic nature; its input as well as output is a syntactic literal. The advantage is that in this way negation can be applied without the need to define a rule of inference stating equivalence between $\neg\neg L$ and L; this is because the above definition makes that $--L = L$.

As was noticed earlier, rules come in two forms: strict and defeasible. Syntactically, the difference between strict and defeasible rules is indicated by the particular type of arrow. A short, single lined arrow ("$\rightarrow$") indicates a strict rule, while a short double lined arrow ("$\Rightarrow$") indicates a defeasible rule. Another difference is that defeasible rules can have weakly negated literals in the antecedent whereas strict rules cannot. In cases where the difference between strict and defeasible rules is not relevant, or where both kinds of rules are meant, a long single lined arrow is used ("$\longrightarrow$").

Definition 3.2. *A* rule *is an expression of the form:* $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L_n$ *where r is the name of the rule and each* L_i $(0 \leq i \leq n)$ *is a literal. The conjunction at the left of the arrow is called the* antecedent *and the literal at the right side of the arrow is called the* consequent*. In the antecedent,* $\sim$ *is used to denote* weak negation*. The following kinds of rules are distinguished:*

1. *strict rules:* $r : L_0 \wedge \ldots \wedge L_{n-1} \rightarrow L_n$
2. *defeasible rules:* $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \Rightarrow L_n$

A strict rule with an empty antecedent is called a premise*.*

An argument consists of a sequence of rules such that each rule has the non weakly negated part of its antecedent fulfilled by some previous rule(s).

Definition 3.3. *An* argument *is a finite sequence* $A = [r_0, ..., r_n]$ *of rules such that:*

1. *for every i* $(0 \leq i \leq n)$*, for every literal in the antecedent of* r_i *that is not preceded by a weak negation sign there is a* $h < i$ *such that L is the consequent of* r_h*;*

[2] One reason for choosing *weakest link* instead of P&S's original principle of *last link* has to do with the difference between epistemical reasoning and constitutive reasoning. Weakest link is suitable for epistemical reasoning (for which HY-arguments are applicable) while last link is more suitable for constitutive reasoning (for which HY-arguments are not applicable). More on the difference between epistemical and constitutive reasoning in section 4.2.2.

2. no two distinct rules in the sequence have the same consequent.

The idea of a consistent set S of strict rules is that one cannot use the rules of S to construct an inconsistent argument.

Definition 3.4. *A set S of strict rules is* consistent *iff there is no argument A whose rules are a subset of S so that A has a rule with consequent L and a rule with consequent $-L$.*

A defeasible theory consists of a set of strict rules, a set of defeasible rules and a priority ordering among the defeasible rules.

Definition 3.5. *A* defeasible theory *is a triple $(\mathcal{S}, \mathcal{D}, <)$ where $\mathcal{S}$ is a consistent finite set of strict rules, $\mathcal{D}$ is a finite set of defeasible rules and $<$ a partial strict order among the defeasible rules. An argument is* based on *the defeasible theory $(\mathcal{S}, \mathcal{D}, <)$ iff all rules in A are in $\mathcal{S} \cup \mathcal{D}$.*

Notice that since $\mathcal{S}$ and $\mathcal{D}$ are finite, every argument based on $(\mathcal{S}, \mathcal{D}, <)$ has a finite length. Furthermore, there are only finitely many arguments based on $(\mathcal{S}, \mathcal{D}, <)$.

Definition 3.6. *Let A be an argument and L a literal:*

1. A is strict *iff it does not contain any defeasible rule; it is defeasible otherwise*

2. L is a conclusion *of A iff L is the consequent of some rule in A.*

3. L is an assumption *of A iff $\sim -L$ occurs in some rule in A.*

Notice that a conclusion is not necessarily the consequent of the *last* rule in the argument. It can be the consequent of *any* rule in the argument.

The aim of the following definition is to "typecast" a list of elements to a set of elements.

Definition 3.7. *Let S be a list of elements $[e_1, \ldots, e_n]$. We define set(S) to be the set $\{e_1, \ldots, e_n\}$.*

At first sight, the most obvious way to define, say, rebutting would be to determine whether two arguments have opposite conclusions. The idea of the following definition is to do just that, with one small modification: arguments rebut-attack each other iff their closures (under strict rules) have opposite conclusions, that is, iff one could add sequences of strict rules to each argument such that the extended arguments have opposite conclusions.

Definition 3.8. *Let A_1 and A_2 be two arguments based on a defeasible theory $(\mathcal{S}, \mathcal{D}, <)$.*

1. A_2 rebut-attacks *A_1 on L iff there are two lists S_1 and S_2 with set$(S_1) \subseteq \mathcal{S}$ and set$(S_2) \subseteq \mathcal{S}$ such that $A_1; S_1$ is an argument with conclusion L and $A_2; S_2$ is an argument with conclusion $-L$.*

2. A_2 undercut-attacks *A_1 on L iff there is a list S_2 with set$(S_2) \subseteq \mathcal{S}$ such that A_1 has an assumption L and $A_2; S_2$ is an argument with conclusion $-L$.*

A_2 attacks *A_1 iff A_2 rebut-attacks or undercut-attacks A_1. An argument is* coherent *iff it does not attack itself.*

The previous definition (def. 3.8) is concerned with the pre-defeat notion of *attack*. In order to provide the full definition of defeat, it is necessary to take into account the priorities among the arguments. This is done in the following three definitions.

First, the set of rules relevant to a conclusion L is defined. The idea is to recursively include all relevant rules (topdown). This starts with the rule having L as its direct consequent (say r_0), followed by the rules that have the (non-weakly negated part of) the antecedent of r_0 as their consequents, etc.

Definition 3.9. *Let A be an argument with a conclusion L. The set of rules* relevant *to L — written as $R_L(A)$ — is the smallest set such that:*

1. $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L \in R_L(A)$ *where r is a rule in A*
2. *if* $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L_n \in R_L(A)$ *then also* $R_{L_0}(A) \cup \ldots \cup R_{L_j}(A) \subseteq R_L(A)$

The following definition lifts the strict ordering between rules into a strict ordering between arguments according to the weakest link principle.

Definition 3.10. *For any two sets R and R' of rules, $R < R'$ iff for some defeasible rule $r \in R$ and all defeasible rules $r' \in R'$ it holds that $r < r'$.*

The definition of defeat is relatively straightforward. One particular feature, however, is that incoherent arguments are always defeated by an argument that is itself undefeated (the empty argument). The idea is that self-defeating arguments should be prevented from keeping other arguments from becoming justified; it would appear very odd for someone to be able to refute others, while at the same time he is refuting himself (more on this in section 5.3.2).

Definition 3.11. *Let A_1 and A_2 be two arguments based on $(\mathcal{S}, \mathcal{D}, <)$. Then A_2* defeats *A_1 iff A_2 is empty and A_1 is incoherent, or else if there exists an L such that:*

1. *A_2 undercut-attacks A_1 on L; or*
2. *A_2 rebut-attacks A_1 on L, and not $R_{-L}(A_2; S_2) < R_L(A_1; S_1)$.*

We say that A_1 strictly defeats *A_2 iff A_1 defeats A_2 and A_2 does not defeat A_1.*

Given a defeasible theory, the above definitions make sure that a set of arguments is defined, as well as a defeat relation between these arguments. When one abstracts from the internal structure of an argument and from the specifics of the defeat relation, what remains is a Dung-style argumentation framework $(Args, defeats)$. Based on this argumentation framework, Prakken and Sartor use grounded semantics to determine which arguments and conclusions are *justified*.

Examples

As was shown in section 2.1.3, grounded semantics can be given a dialectical proof theory. The idea is that a proponent and an opponent are involved in a discussion about the validity of a main argument. They take turns, and in every turn they provide a counterargument against the other party's argument. The rules of the dialogue are such that each

argument of the opponent should defeat the previous argument of the proponent, while each argument of the proponent should *strictly* defeat the previous argument of the opponent.[3] Furthermore, the proponent is not allowed to repeat any earlier moves, to prevent the dialogue from non-termination.

Given these rules, the following are examples of dialogues using the formalism of P&S:

1. $\mathcal{S} = \{\rightarrow A,\ \rightarrow D\}$
   $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C, D \Rightarrow E,\ E \Rightarrow \neg C\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C \quad (A_1)$
   O: $\rightarrow D,\ D \Rightarrow E,\ E \Rightarrow \neg C \quad (A_2)$

   Here, A_2 rebut-attacks and defeats A_1 (definition 3.8 i and 3.11 ii)

2. $\mathcal{S} = \{\rightarrow A,\ \rightarrow E\}$
   $\mathcal{D} = \{A \wedge \sim B \Rightarrow C,\ C \Rightarrow D,\ E \Rightarrow B\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \wedge \sim B \Rightarrow C,\ C \Rightarrow D \quad (A_1)$
   O: $\rightarrow E,\ E \Rightarrow B \quad (A_2)$

   Here, A_2 undercut-attacks and defeats A_1 (definition 3.8 ii and 3.11 i)

3. $\mathcal{S} = \{\rightarrow A,\ \rightarrow D,\ B \rightarrow C,\ E \rightarrow \neg C\}$
   $\mathcal{D} = \{A \Rightarrow B,\ D \Rightarrow E\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B \quad (A_1)$
   O: $\rightarrow D,\ D \Rightarrow E \quad (A_2)$

   Here, A_2 rebut-attacks and defeats A_1 (using $S_1 = [B \rightarrow C]$ and $S_2 = [E \rightarrow \neg C]$ to produce concatenations $A_1' = A_1; S_1$ and $A_2' = A_2; S_2$, so definition 3.8 i can be applied)

4. A special example, involving an incoherent argument, is as follows:
   $\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
   $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C,\ C \rightarrow D,\ \rightarrow \neg D \quad (A_1)$
   O: $\emptyset \quad (A_2)$

   Here, the proponent puts forward an incoherent (or self-defeating) argument. According to definition 3.11, an incoherent argument can be strictly defeated by an empty argument, which is exactly what is put forward by the opponent.

5. An incoherent argument can also take a more concealed form:

[3]The idea is that there is a certain asymmetry in the dialogue. The opponent has the relatively easy task of casting doubt on a certain thesis (for which non-strict defeat is already enough). The proponent, however, should make sure that the thesis is casted away from all doubt (for which it needs to *strictly* defeat the possible counterarguments).

$\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
$\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
$< = \emptyset$

P:	$\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$	(A_1)
O:	$\emptyset$	(A_2)

Here, the proponent puts forward an incoherent (or self-defeating) argument. Argument A_1 is incoherent because it can be concatenated with $S_1 = [C \rightarrow D,\ \rightarrow \neg D]$, which results in an attack in the sense of definition 3.11 i.

6. A disadvantage of the formalization of example 5 is that the conflict is somewhat hidden. A third party observer has to reconstruct the set S_1 of strict rules in order to see that A_1 is incoherent.

   It may be desirable to provide an explicit explanation of why the opponent thinks A_1 is incoherent. Unfortunately, in the formalism of P&S it is not straightforward to provide this kind of explicit information in the course of a dialogue, as is illustrated by the following example:

   $\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
   $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
   $< = \emptyset$

P:	$\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$	(A_1)
O:	$\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C,\ C \rightarrow D,\ \rightarrow \neg D$	(A_2)
P:	$\emptyset$	(A_3)

   In this example, the opponent explicitly wants to show that A_1 is incoherent, so it extends A_1 with the strict ruleset $S_1 = [C \rightarrow D,\ \rightarrow \neg D]$. The problem, however, is that by showing that A_1 is incoherent, the opponent itself makes an incoherent argument (A_2), which according to definition 3.11 can be defeated by the empty argument (A_3). So, instead of the opponent winning the dialogue by successfully showing that A_1 is incoherent, it *loses* the dialogue because by showing this, it makes an incoherent argument itself.

7. In example 6, the opponent tries to show the problematic nature of argument A_1 by extending A_1 with a set of strict rules. In the examples 7, 8, 9 and 10 we illustrate that one can also show the problematic nature of an argument A_1 by extending it with a set of defeasible rules. The main problem is that the party that illustrates the problematic nature of the other party's argument ends up *losing*, not winning, the dialogue.

   P: "There is a threat that tonight's soccer game will lead to riots (t), because Ajax plays against Feijenoord (af), there are no indications that there will be any extra police on the streets $(\sim p)$."

   O: "Don't worry, if there is really a threat, then the government will of course send extra police."

   $\mathcal{S} = \{\rightarrow af\}$
   $\mathcal{D} = \{af \wedge \sim p \Rightarrow t,\ t \Rightarrow p\}$
   $< = \emptyset$

P: $\rightarrow af,\ af \wedge \sim p \Rightarrow t$ (A_1)
O: $\rightarrow af,\ af \wedge \sim p \Rightarrow t,\ t \Rightarrow p$ (A_2)
P: $\emptyset$ (A_3)

8. P: "The shipment of goods must have arrived in the Netherlands by now (a), because we placed an order three months ago (tma)"
   O: "I don't think so. If the goods would really have arrived in the Netherlands, then there would be a customs declaration (cd), and I can't see any such declaration in our information system ($\neg is$)."

$\mathcal{S} = \{\rightarrow tma,\ \rightarrow \neg is\}$
$\mathcal{D} = \{tma \Rightarrow a,\ \neg is \Rightarrow \neg cd,\ a \Rightarrow cd\}$
$< = \emptyset$

P: $\rightarrow tma,\ tma \Rightarrow a$ (A_1)
O: $\rightarrow tma,\ tma \Rightarrow a,\ a \Rightarrow cd,\ \rightarrow \neg is,\ \neg is \Rightarrow \neg cd$ (A_2)
P: $\emptyset$ (A_3)

9. P: "Jack must be member of the tuff-tuff-club (ttc). He drives a Ferrari (f) and is known to be involved in criminal affairs (c), so he can be assumed to be an illegally fast driver (s), probably member of the ttc (a semi-illegal league of people who are speeding in the Netherlands)."
   O: "I don't think so. For if he were member, then the police would keep an eye on him (p) to catch him on speeding (s) and, as criminals do not like to be caught he would not dare to speed in the first place ($\neg s$)."

$\mathcal{S} = \{\rightarrow f,\ \rightarrow c\}$
$\mathcal{D} = \{f \wedge c \Rightarrow s,\ s \Rightarrow ttc,\ ttc \Rightarrow p,\ p \wedge c \Rightarrow \neg s\}$
$< = \emptyset$

P: $\rightarrow f,\ \rightarrow c,\ f \wedge c \Rightarrow s,\ s \Rightarrow ttc$ (A_1)
O: $\rightarrow f,\ \rightarrow c,\ f \wedge c \Rightarrow s,\ s \Rightarrow ttc,\ ttc \Rightarrow p,\ p \wedge c \Rightarrow \neg s$ (A_2)
P: $\emptyset$ (A_3)

10. P: "Next year, we are going to get a tax-relief (tr), because our politicians promised so (pmp)."
    O: "But in the current situation, you can only implement a tax-relief by accepting a significant budget deficit (bd), which means we will also get a huge fine from Brussels (fb). There goes our tax-relief."

$\mathcal{S} = \{\rightarrow pmp\}$
$\mathcal{D} = \{pmp \Rightarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr\}$
$< = \emptyset$

P: $\rightarrow pmp,\ pmp \Rightarrow tr$ (A_1)
O: $\rightarrow pmp,\ pmp \Rightarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr$ (A_2)
P: $\emptyset$ (A_3)

What the examples 6, 7, 8, 9 and 10 have in common is that the opponent wants to indicate that the proponent drew its conclusions too fast. If the proponent would have thought more about it, then after some additional reasoning steps, it would have discovered that the argument would lead to self-defeat. In the formalism of P&S, however, it is the opponent who is blamed for making a self defeating argument (which, according to definition 3.11 can be defeated by an empty argument), while in fact the opponent is only confronting the proponent with the consequences of proponent's own reasoning.[4] From a conceptual point of view, the whole issue is that the statements in a party's argument are not necessarily endorsed by the party himself. It is perfectly possible that some of the statements, borrowed from the other party's argument, are just temporarily assumed, for the sake of the argument.

3.2 Analysis

In this section, an informal analysis is made of how HY-arguments are applied. As we originally introduced the concept of HY using Socratic dialogues, our discussion starts with a treatment of some existing theory in the field of dialogues, which in particular includes the notion of a *commitment* (section 3.2.1). The next step, then, is to use some of the results of this discussion for specifying how HY-arguments interact with classical (non-HY) arguments, as well as with each other (section 3.2.2). The overall aim of this section is to provide the principles on which a formalization of HY-argumentation can take place.

3.2.1 Commitments

In their book "Commitment in Dialogues", Walton and Krabbe specify how dialogues can be seen from the perspective of commitments [WaKr95]. Among the types of commitment that can be distinguished are *action commitment* and *propositional commitment*. Action commitment means that a party is committed to take a certain action. According to Hamblin [Hamb87], the fulfillment of an action commitment can be *extensional*, meaning that the committed action has simply taken place, or *wholehearted*, which has the additional property that the action that led to fulfillment was deliberately meant to do so. That is, the fulfillment was not merely "by coincidence", or as an accidental side-effect.

Propositional commitment means that a party is committed to a certain proposition (a descriptive statement that can be assigned a truth value). Propositional commitments, according to Walton and Krabbe, are in fact special forms of action commitments. The action here is that the content of a propositional commitment should be defended if anyone challenges it. The challenge, as well as the defense, can lead to a dialogue, and the notion of a propositional commitment is therefore strongly connected to that of a dialogue. As the main interest of this thesis is in logic, arguments and dialogues, we focus on the notion of *propositional* commitment — all occurrences of "commitment" should therefore be read as "propositional commitment" unless explicitly stated otherwise.

A commitment in dialogue can be seen as a party's "official" point of view, a public statement of what the party holds to be true. Commitment is not the same as belief. A

[4]The problems as sketched by examples 6 until 10 are not restricted to the formalism of P&S. They also play a role in other formalisms, as will be shown in section 5.2 and 5.3.

party can commit itself to a certain statement without necessarily believing it (such as telling a lie), or believing something without being committed to it (like a lawyer who believes his client to be guilty without openly saying so in court).

In dialogue systems, such as Hamblin's H[5] [Hamb70, Hamb87], MacKenzie's DC [MacK79, MacK90], Rescher's Dialectics [Rsch77], Gordon's Pleadings Game [Gord95], or Lodder's Dialaw [Lodd98, Lodd99a, Lodd99b], commitments are made explicit using the concept of a *commitment store*. Every time a party makes a new claim or concedes a claim of the other party, a new commitment is added to the relevant commitment store.

As an example of how commitments come into existence and are disposed of, we take the following dialogue from [Lodd98]:

P: It was not allowed to search Tyrell ($\neg sa$)
O: Why do you think so?
P: Only if someone is a suspect, he may be searched, and Tyrell was not a suspect ($\neg s$)
O: I agree, but Tyrell was on probation (pc) and had to allow a search at any time.
P: You are right, the search was allowed.

As an aside, in the formalized dialogue below, $reason(p, q)$ stands for "p is a reason for q", and $outweights(S_1, S_2, p)$ means that the propositions of set S_1 are reasons for p, the propositions of set S_2 are reasons against p, and that S_1 outweights S_2. The terms *reason* and *outweights* are borrowed from Hage and Verheij's reason-based logic [HaVe94].

(1) P: claim, $\neg sa(ty)$
"It was not allowed to search Tyrell."
(2) O: question, $\neg sa(ty)$
"Why not?"
(3) P: claim, $reason(\neg s(ty), \neg sa(ty))$
"A reason why it was not allowed to search Tyrell is that he was not a suspect."
(4) O: accept, $reason(\neg s(ty), \neg sa(ty))$
"I can agree with that."
(5) P: claim, $outweights(\{\neg s(ty)\}, \emptyset, \neg sa(ty))$
"...and since there is one reason in favour of $\neg sa(ty)$ ($\{\neg s(ty)\}$) and no reason against $\neg sa(ty)$ ($\emptyset$), the reasons for $\neg sa(ty)$ outweight the reasons against $\neg sa(ty)$."
(6) O: claim, $reason(pc(ty), sa(ty))$
"There is a reason that the search was allowed: Tyrell was on probation."
(7) P: withdraw, $\neg sa(ty)$
"I agree that I can no longer hold that the search was not allowed."

Regarding commitments, the formalized dialogue can be seen as follows:

[5] We use H as an abbreviation of "Why-Because system with questions"

(1) P: $C_P(\neg sa(ty))$
O: (no commitments)
(2) P: $C_P(\neg sa(ty))$
O: (no commitments)
(3) P: $C_P(\neg sa(ty), reason(\neg s(ty), \neg sa(ty)))$
O: (no commitments)
(4) P: $C_P(\neg sa(ty), reason(\neg s(ty), \neg sa(ty)))$
O: $C_O(reason(\neg s(ty), \neg sa(ty)))$
(5) P: $C_P(\neg sa(ty), reason(\neg s(ty), \neg sa(ty)), outweights(\{\neg s(ty)\}, \emptyset, \neg sa(ty)))$
O: $C_O(reason(\neg s(ty), \neg sa(ty)))$
(6) P: $C_P(\neg sa(ty), reason(\neg s(ty), \neg sa(ty)), outweights(\{\neg s(ty)\}, \emptyset, \neg sa(ty)))$
O: $C_O(reason(\neg s(ty), \neg sa(ty)), reason(pc(ty), sa(ty)))$
(7) P: $C_P(reason(\neg s(ty), \neg sa(ty)))$
O: $C_O(reason(\neg s(ty), \neg sa(ty)))$

In the above example, every time a party makes a claim ((1), (3), (4), (6)), a commitment is created. A commitment is also created when a party accepts (sometimes also called *concedes*) a commitment of the counterparty (4). Commitment is lost when a party withdraws it (7).

Different dialogue systems have different principles of dealing with commitments. Hamblin's H, for instance, does not have an explicit concede (or accept). If a party makes a claim and the counterparty remains silent (that is, the counterparty does not utter any doubts or objections), then the counterparty automatically commits himself to what the other party has claimed (silence implies consent). In Lodder's Dialaw, a withdrawal of proposition p does not only remove p from the party's commitment store, but also removes all disagreements (differences in commitments between parties) that have been created since p was claimed. The reason for this is that in Dialaw, these disagreements are considered not to be relevant anymore, since they are related to an issue (whether p is true or not) that has now been settled. It is also possible for a dialogue system to force commitments to be closed under logical consequence.

Despite the differences in various dialogue systems, it can be said that in general, a commitment is created by either an assertion or a concession (explicit or implicit), and a commitment is removed by retraction.

Nested commitments

The question is whether it is always the case that whenever a party claims p, it becomes committed to p. The following example has been taken from Walton and Krabbe [WaKr95, p. 58].

> Thomson and Thompson doubt whether Captain Haddock did accurately calculate the ship's position. They show the captain some improved calculations of their own. Haddock: "You are right... I have made a mistake. Gentlemen, please take off your hats..." [takes off cap, and stands in prayer for some time]. (...) Thomson: "But Captain, tell us what you mean..." Haddock: "I mean, gentlemen, that according to your calculations we are now standing inside Westminster Abbey!"

Walton and Krabbe give the following treatment of this example.

> Haddock's provocative thesis, confronting Thomson and Thompson in their capacity as calculators, determines the ship's position as inside Westminster Abbey. This is what the captain should defend, if challenged, starting from the other party's concessions, in this case the improved calculations made by Thomson and Thompson. Obviously, Haddock himself does not hold the ship to be in that position.

A somewhat similar example was given earlier.

> EB: In two years time, the waiting lists in health care will be as good as resolved (wlr).
>
> IJ: Then you are actually saying that the insurance fees will be increased (*ifi*), because the government has already decided not to put more money into the health care system ($\neg mmhs$), and you have promised not to lower the coverage of the standard insurance ($\neg lcsi$).

Here, the interviewed minister makes a certain claim (assert(wlr)). The interviewing journalist then confronts the minister with what he thinks are the consequences of the minister's own words, namely that the insurance fees will be increased. Does this, however, also mean that the *interviewer* is committed to the fact that the insurance fees will be increased? The answer depends on whether the interviewer is committed to the minister's original statement. Let us suppose that the interviewer is *not* committed to the prediction that in two years time, the waiting lists will be as good as resolved — perhaps the interviewer believes that the cabinet will fall within several months, and that most of the government's plans will not be realized. In that case, the interviewer is not necessarily committed to the fees increase either ($\neg C_{IJ}(ifi)$).

So, what is it that the interviewer is committed to, if any? The first thing to notice is that the interviewer uses a rule $wlr \wedge \neg mmhs \wedge \neg lcsi \Rightarrow ifi$ (wlr, together with $\neg mmhs$ and $\neg lcsi$ is a reason for *ifi*), so it seems that the interviewer is at least committed to the fact that wlr, together with $\neg mmhs$ and $\neg lcsi$ is a reason for *ifi*. What the interviewer claims, however, is not just the validity of the piece of information $wlr \wedge \neg mmhs \wedge \neg lcsi \Rightarrow ifi$, but also that this information is *applicable* to the specific statement as uttered by the minister. Instead of the interviewer saying that the insurance fees will be increased, he claims that the *minister* implicitly says this, since it follows from the minister's own words. Thus, instead of $C_{IJ}(ifi)$ we have $C_{IJ}(C_{EB}(ifi))$.

Regarding the Thomson and Thompson example, the same analysis can be made. Captain Haddock is not really committed to the ship's position being inside Westminster Abbey. Instead, he claims that this follows from Thomson and Thompson's explicit commitments ($C_H(C_{T\&T}(WA))$), and that Thomson and Thompson are therefore committed to something that is obviously not true.

If we adapt Walton and Krabbe's view that propositional commitments are to be interpreted as statements to be defended when being challenged, then what Haddock should defend is $C_{T\&T}(WA)$. That is, Haddock should maintain that WA follows from Thomson and Thompson's calculations. The only way Thomson and Thompson could defend themselves is by claiming that they are not committed to WA at all; perhaps they can claim that the captain has misinterpreted their calculations.

Some properties of nested commitments

The fact that commitments, like beliefs, can be nested makes it interesting to examine which properties hold for them.[6] Beliefs, for instance, satisfy a KD45 axiomatization, but as commitments are not beliefs, this axiomatization does not necessarily apply to commitments.

The K-axiom seems reasonable. One should be committed to the direct consequences of one's own commitments. This means that for parties P and O we have:

$$\begin{aligned} C_P(K) \wedge C_P(K \supset L) &\supset C_P(L) \\ C_O(K) \wedge C_O(K \supset L) &\supset C_O(L) \end{aligned}$$

The K-axioms are closely related to Walton and Krabbe's notion of *subcommitment* [WaKr95, p. 44]. In MacKenzie's DC it is possible for commitments not to be closed under logical consequence, but closure can be enforced by means of the "resolve" statement.

At first sight, the D-axiom also seems reasonable. Ideally, reasoners should not be committed to an inconsistency.

$$\begin{aligned} &\neg C_P(\bot) \\ &\neg C_O(\bot) \end{aligned}$$

It must be noticed, however, that although inconsistent commitments can lead to incredibility of the party that holds them, situations in which a party turns out to have inconsistent commitments *can* occur; these situations can for instance be the result of a Socratic dialogue. Or, as Walton and Krabbe formulate it: "A party thus committed may suffer a loss immediately after, or may be forced to retract some statement, but the position itself should not be declared illegal" [WaKr95, p. 58]. In MacKenzie's DC, it is possible to use the "resolve" statement to confront a party with an inconsistency in its belief. The party must then retract one or more of the commitments leading to the inconsistency.

Thus, the K- and D-axioms are not necessarily hard and fast rules. Temporary violations can occur, and it is up to the violating party to restore these properties, for if it does not, it risks losing the dialogue.

Another point is what we call the awareness axioms.

$$\begin{aligned} C_P(L) &\supset C_O(C_P(L)) \\ C_O(L) &\supset C_P(C_O(L)) \end{aligned}$$

The rationale of the awareness-axioms is as follows. In a dialogue, whenever a new commitment comes into existence, there is an accompanying move in the dialogue (such as *assert* or *concede*) that created it. In natural language dialogues, parties often provide feedback to notice whether or not they have understood each other. This feedback can take the form of a simple "uhum", or of a more elaborate "yes, I see what you mean". This feedback is closely related to the principle of *grounding* [ClSc89, Clar96]. Grounding is the process in which a common ground is established, knowledge that is shared among parties (common knowledge). Let us assume that every commitment comes into existence by an explicit speech act (assert or concede), and that for every speech act the appropriate feedback takes place. Then, after each claim, there is an acknowledgment that this claim

[6] Although the discussion in this section is not vital for the remaining part of this thesis, we believe that the treatment of properties of nested commitments is interesting in itself.

has been understood (which, by the way, does not necessarily imply agreement with the content of the claim). Suppose, party P makes a claim (p) and party O acknowledges that it understood this claim. Then, P is committed to p ($C_P(p)$), and O has openly uttered that it understands that P is committed to p ($C_O(C_P(p))$). Thus, if every commitment is the result of one or more explicit speech acts and if sufficient feedback takes place (which we, by the way, often abstract from in our formalized dialogues) then the awareness-axioms are valid.

The last point to be discussed is what we call the confidence axioms. The confidence axioms roughly state that if a party is committed to a certain proposition, it is also committed to the fact that the other party is committed to that statement. The rationale of the confidence axioms can perhaps best be illustrated by a quote from Jaap Hage about social reasons [Hage97, p. 50]:[7]

> (...) we must distinguish between personal reasons, and what might be called *social reasons*. 'Social reason' is a term of art; I will use it for facts that are considered as reasons by most members of a social group. Social reasons are reasons based on *social rules*.
>
> If the wind suddenly becomes stronger, and the sky is quickly darkening with clouds, this is for most Dutch people a reason to believe that it will soon rain. These facts are *for a typical Dutchman* a reason to believe that it will soon rain. In other words, they are a personal reason, but not only for one person, but for most persons that partake in the group.
>
> Moreover, most group members not only consider the increasing wind and darkening sky as personal reasons to believe that it will rain, but also as reasons why others should believe that it will rain, Social reasons are perhaps the most important category of reasons in daily life. Humans that live together must harmonize their behavior, and this brings the demand with it that behavior is predictable. The best basis for harmonization of behavior is if people consider the same things as reasons for behavior or for belief, that is, if they use the same rules. That is why members of a group not only often consider the same facts as reasons, but also exercise some social pressure for conformity. This pressure for conformity can take many different forms, from avoidance and ridiculing to punishment and banishment from the group. It is this pressure that distinguishes social rules and reasons from mere personal principles and personal reasons that count for many group members.

Although Hage treats the effects that social reasons have on beliefs, the notion of social reason is relevant for commitments as well. Social reasons are imposed upon the individual members of a group, and it would be difficult for an individual member to publicly argue against their appliance or consequences. The confidence axioms can be stated as follows:

$$\begin{aligned} C_P(L) &\supset C_P(C_O(L)) \\ C_O(L) &\supset C_O(C_P(L)) \end{aligned}$$

It must be mentioned, however, that the confidence axioms are only valid if L is the conclusion of an argument constructed of social reasons related to a group of which both

[7] According to Hage, his view of social reasons is in turn based on that of Hart [Hart61].

parties are member. Or, as Hage formulates it: "An argument is valid if it is in accordance with the social rules or principles of inference which are valid in the group to which the reasoner belongs" [Hage97, p. 251].

To illustrate the workings of the confidence axioms, consider a field of empirical science, in which a certain scientist has performed an experiment of which statement L can be concluded. If both the way in which the experiment was carried out and the way in which from the results the conclusions were drawn confirms with the standards of the scientist's research community, the scientist can then assume that upon publication fellow scientists also become committed to the obtained conclusion.

A dialogue perspective

It is interesting to see how nested commitments can play a role in (semi-)formal dialogue. Take for instance the "shipment of goods" example provided earlier. In a dialogue system that does not support a HY-style reasoning, the dialogue could look as follows:

P:	claim a	$C_P(a)$	"I think that a.
O:	why a		"Why do you think so?"
P:	because $tma \Rightarrow a$	$C_P(a, tma)$	"Because of tma."
O:	concede tma, concede a	$C_O(tma, a)$	"OK, you are right."

Here, the opponent concedes the main claim, so the proponent wins the dialogue. If, during the cause of a dialogue, players can confront each other with the (defeasible) consequences of their opinions, then a different dialogue may result:

P:	claim a	$C_P(a)$	"I think that a."
O:	but-then $a \Rightarrow cd$	$C_O(C_P(cd))$	"Then you implicitly also hold that cd."
P:	concede cd	$C_P(a, cd)$	"Apparently..."
O:	claim $\neg cd$	$C_O(\neg cd)$[8]	"But cd is not the case."
P:	why $\neg cd$	[unchanged]	"Why not?"
O:	because $\neg is \Rightarrow \neg cd$	$C_O(\neg cd, \neg is)$	"Because of $\neg is$."
P:	concede $\neg is$, concede $\neg cd$	$C_P(a, cd, \neg cd, \neg is)$	"Oops, you're right; I caught myself in..."

Here, much akin to a Socratic dialogue, the opponent wins the dialogue because it forces the proponent to commit himself to an inconsistency.

A key feature in the above dialogue is the *but-then* statement, with which the opponent confronts the proponent with the defeasible consequences of the proponent's commitments. A but-then statement is a special form of claim, in which the speaker does not become committed himself in the consequent of the rule being claimed applicable. In general, in order to use a "but-then $\varphi_1 \wedge \ldots \wedge \varphi_n \Rightarrow \phi$", the other player has to be committed to $\varphi_1 \wedge \ldots \wedge \varphi_n$. The immediate aim of a but-then statement is to commit him to ϕ as well. The final aim is then to get the other player to the point where it is obvious that his commitments are inconsistent.

Another example of how the but-then statement can be used is as follows (ttc):

[8] we no longer explicitly mention $C_O(C_P(cd))$ since $C_P(cd)$

P:	claim ttc	$C_P(ttc)$
O:	why ttc	
P:	because $s \Rightarrow ttc$	$C_P(s, ttc)$
O:	why s	
P:	because $f \wedge c \Rightarrow s$	$C_P(f, c, s, ttc)$
O:	but-then $ttc \Rightarrow p$	$C_O(C_P(p))$
P:	concede p	$C_P(f, c, s, ttc, p)$
O:	but-then $p \wedge c \Rightarrow \neg s$	$C_O(C_P(\neg s))$
P:	concede $\neg s$	$C_P(f, c, s, ttc, p, \neg s)$ (inconsistent)

An interesting question is how the style of reasoning of the "because" statement can be compared with that of the "but-then" statement (see also figure 3.1):

1. With the because statement, reasoning goes *backwards* (abduction); the party being questioned tries to find reasons to support its thesis. With the but-then statement, on the other hand, reasoning goes *forward* (deduction); the party being questioned can be forced to make additional reasoning steps.

2. With the because statement, the *proponent* of a thesis (like ϕ in figure 3.1) tries to find a path (or tree) from the premises to ϕ (the opponent's task is then to try to defeat this path). With the but-then statement, on the other hand, it is the *opponent* of the thesis that tries to find a path (or tree).

3. The path (or tree) constructed using because statements should ultimately originate from statements that are accepted to be *true* (such as premises), whereas the path constructed using but-then statements should ultimately lead to statements that are considered *false* (contradictions)

4. With a successfully constructed because path (or tree), but the proponent and opponent become committed to the propositions on the path, whereas with a successfully constructed but-then path (or tree), it is possible that only the proponent becomes committed to the propositions on the path.

Figure 3.1: because and but-then

In the above analysis, it may seem that an opponent of ϕ has two options: either trying to construct a but-then path from ϕ, or trying to prevent the proponent from successfully constructing an undefeated because path. In many occasions, however, a mixed strategy is also possible. In the ttc-example, for instance, the opponent starts with asking for a because path. When the proponent starts doing so, the proponent has to make several commitments, which the opponent then uses against him for the construction of a but-then path.

The use of a but-then statement does not automatically lead to a new commitment on the side of the other party. Sometimes, it can be successfully argued why the counterparty

does not have to become committed. To illustrate why, consider again the ttc-example, but now with $ttc \Rightarrow p$ replaced by $ttc \wedge \sim sv \Rightarrow p$ (say, in small villages (sv) the police has not heard of the ttc, so membership is not a reason for extra police attention), and a rule $\Rightarrow sv$ is available.

P:	claim ttc	$C_P(ttc)$
O:	why ttc	
P:	because $s \Rightarrow ttc$	$C_P(s, ttc)$
O:	why s	
P:	because $f \wedge c \Rightarrow s$	$C_P(f, c, s, ttc)$
O:	but-then $ttc \wedge \sim sv \Rightarrow p$	$C_O(C_P(p))$
P:	claim sv	$C_P(f, c, s, ttc, sv)$
O:	concede sv, retract $C_P(p)$	$C_O(sv)$

Here, the opponent again tries to construct a successful but-then path. This path, however, is undercut by the proponent. What happens next depends on the nature of the dialogue. When backtracking is allowed, the opponent may pursue another strategy. When backtracking is not allowed, the opponent has lost the game.

In the above examples, the use of nested commitments has been illustrated. The following general remarks can be made:

1. A but-then statement is in essence a special form of a claim statement (see the captain Haddock example). A claim statement has as effect that a new commitment comes into existence, and such should also be the case for a but-then statement.

2. But-then statements do not in general create "flat" commitments (at least, not immediately). Suppose party O utters "but-then $\varphi_1 \wedge \ldots \wedge \varphi_n \wedge \sim \psi_1 \wedge \ldots \wedge \sim \psi_m \Rightarrow \phi$". This does of course not mean that O becomes committed to ϕ (so we don't have $C_O(\phi)$. It also does not mean that P is actually committed to ϕ (that is, we don't automatically have $C_P(\phi)$), because P may avoid commitment by successfully defending ψ_i $(1 \leq i \leq m)$ The only thing thing that can be said is that O feels that P is implicitly committed to ϕ (so $C_O(C_P(\phi))$), but whether P is actually committed to ϕ is still open for discussion.

3. In general, the party that makes a claim bears the responsibility of defending the contents of the claim. For instance, if P utters "claim ϕ" then upon P rests the task of defending ϕ. Similarly, if O utters "but-then $\varphi_i \wedge \ldots \wedge \varphi_n \wedge \sim \psi_1 \wedge \ldots \wedge \sim \psi_m \Rightarrow \phi$" then upon O rests the task of defending $C_P(\phi)$, and if O is unable to do so, it can loose the dialogue game.

To summarize: a nested commitments is a quite natural concept when in a dialogue it becomes possible to confront parties with the consequences of their own commitments.

As an aside, it may be interesting to compare the but-then statement with the resolve statement of MacKenzie's DC [MacK79]. In DC, if party A claims proposition q ("claim q"), which is then questioned ("why q") by party B, and party B is committed to p, from which q directly follows, then party A may utter "resolve $p \supset q$", which forces party B to become committed to q as well (or alternatively, B may retract its commitment to p).[9] See

[9] Another use of MacKenzie's resolve statement is to notice the other party that its commitments are inconsistent, thus forcing the other party to retract one or more of them.

the following example.

P:	claim p	$C_P(p)$
O:	concede p	$C_O(p)$
P:	claim $p \supset q$	$C_P(p,\ p \supset q)$
O:	concede $p \supset q$	$C_O(p,\ p \supset q)$
P:	claim q	$C_P(p,\ p \supset q,\ q)$
O:	why q	[unchanged]
P:	resolve "If $p \wedge (p \supset q)$ then q"	[unchanged]
O:	concede q	$C_O(p,\ p \supset q,\ q)$

One obvious difference between the resolve and the but-then statement is that after a successful resolve statement *both* parties are committed to the proposition in question, whereas with a successful but-then statement it is possible that only one party becomes committed. Furthermore, the but-then statement also has an inherently defeasible nature; it is possible that the other party has a reason (exception) against applying the rule in question. This reason can then be discussed in the remainder of the dialogue. In short, although the resolve statement can be sufficient for classical, monotonic reasoning, for defeasible reasoning it is desirable to have a statement that has special, nonmonotonic properties. We think that the but-then statement fulfills this requirement.

3.2.2 HY-arguments

An HY-argument can preliminarily be defined as an argument that illustrates the problematic nature of another argument by assuming one or more of the other argument's conclusions. In this section we study the structure and behavior of HY-arguments. Our aim is to formulate some general properties which these arguments should adhere to, in order for them to be properly formalized.

Rebutting HY-arguments

Rebutting, in a classical sense, means that argument A_2 attacks argument A_1 by deriving a conclusion (L) that conflicts with a conclusion ($-L$) of A_1. Rebutting essentially involves showing a contradiction; a contradiction that is at least partly caused by the other party's argument.

In the case of classical rebutting, at least one *direct conclusion* of the other party's argument is involved in the conflict. In the system of P&S, this is already broadened such that at least one *strict consequence* of the conclusions of the other party's argument is involved in the conflict. With HY-arguments, we will broaden it even more, and require that at least one *strict or defeasible consequence* of the other party's argument is involved in the conflict. This does not mean that *every* conclusion of a HY-argument A_2 needs to be based on one or more conclusions of the argument it attacks (A_1). As an illustration, consider the "shipment of goods" example.

$$\begin{array}{l|l} tma \Rightarrow a & a \Rightarrow cd \\ & \neg is \Rightarrow \neg cd \end{array}$$

First, some notation must be explained. The argument left of the bar, we call A_1. The two-line argument right of the bar, we call A_2. A_2 contains two lines; the line that is at

the same level of A_1 is based on one or more conclusions of A_1 (in this case: the conclusion a); the second line of A_2 is completely independent of A_1; it is based on premises only. In A_2, proposition a is called a *foreign commitment* because it is based on an actual commitment in A_1. A conclusion or part of an argument that is based on one or more foreign commitments is called *fc-based*.

What the above example makes clear is that, in general, a HY-argument has two parts, a part that is fc-based and a part that is not. What matters is that at least part of the conflict should take place in the fc-based part of A_2, otherwise A_1 cannot be blamed for deriving a contradiction. As the system of P&S is centered around literals, a conflict will have the shape of L versus $-L$. Let us systematically examine all possibilities how this conflict can occur (see table 3.1).

$L \quad -L$	A_1	$A_2^{not\ fc-based}$	$A_2^{fc-based}$
A_1	self-defeating	I	III
$A_2^{not\ fc-based}$	*I*	self-defeating	II
$A_2^{fc-based}$	*III*	*II*	IV

Table 3.1: Rebutting HY-arguments and possible conflicts

First, the situations in which A_1 contains L and $-L$, as well as in which A_2 contains L and $-L$ in the part that is not fc-based, essentially boil down to classically incoherent arguments. In these cases, the inconsistency cannot be attributed to any other argument, so there should be no reason to defeat the other argument.

There are four remaining situations[10], which are depicted in example I to IV. Recall that the notational convention is that if a line of A_2 is at the *same* level as A_1, then it makes use of one or more of A_1's conclusions, whereas if a line of A_2 is *not* at the same level as A_1, then it does not make use of any of A_1's conclusions.

Example I: $\Rightarrow \Rightarrow \Rightarrow L \quad | \quad \Rightarrow \Rightarrow \Rightarrow -L$

Example II: $\Rightarrow \Rightarrow \Rightarrow \quad | \quad \Rightarrow \Rightarrow \Rightarrow L \quad \Rightarrow \Rightarrow \Rightarrow -L$

Example III: $\Rightarrow \Rightarrow \Rightarrow L \quad | \quad \Rightarrow \Rightarrow \Rightarrow -L$

Example IV: $\Rightarrow \Rightarrow \Rightarrow \quad | \quad \Rightarrow \Rightarrow \Rightarrow L \Rightarrow \Rightarrow \Rightarrow -L$

In example I, A_1 contains L, and the part of A_2 that is not fc-based contains $-L$; this situation boils down to classical rebut. In example II, A_2 contains L in the fc-based part and $-L$ in the non fc-based part; a more concrete situation of this would be the "shipment of goods" example given earlier. In example III, A_1 contains L, and A_2 contains $-L$ in the fc-based part. A more concrete situation hereof would be the "tax relief" example given earlier. In example IV, A_2 contains L as well as $-L$ in the fc-based part.

[10]Well, actually there are 7 remaining situations, but 6 of them are pairwise equivalent (by letting L and $-L$ change positions).

All of the examples I until IV can be regarded as correct instances of (HY-)rebutting. Therefore, we believe that a formalization of an extended argumentation system that allows for HY-arguments should allow all of the above four instances as correct forms of rebutting.

Undercutting HY-arguments

When it comes to undercutting arguments, we can roughly distinguish the same four examples as with the rebutting arguments.

Example V:	$\Rightarrow K\ (K \wedge \sim L \Rightarrow M)\ M \Rightarrow$	|	$\Rightarrow \Rightarrow \Rightarrow L$
Example VI:	$\Rightarrow \Rightarrow \Rightarrow$	|	$\Rightarrow K\ (K \wedge \sim L \Rightarrow M)$ $\Rightarrow \Rightarrow \Rightarrow L$
Example VII:	$\Rightarrow K\ (K \wedge \sim L \Rightarrow M)\ M \Rightarrow$	|	$\Rightarrow \Rightarrow \Rightarrow L$
Example VIII:	$\Rightarrow \Rightarrow \Rightarrow$	|	$\Rightarrow K\ (K \wedge \sim L \Rightarrow M)\ M \Rightarrow \Rightarrow \Rightarrow L$

Example V illustrates the "traditional" undercut, as is implemented by many systems for defeasible reasoning. The opponent has successfully attacked the proponent's argument by showing that one of the assumptions of proponent's argument is wrong.

The opponent's arguments at example VI and VIII, however, do not attack the proponent's argument at all. The opponent does not show that there is anything wrong with proponent's conclusions. The only thing that the opponent shows is that further reasoning based on proponent's conclusions does not always lead to additional conclusions. The opponent has indicated no contradiction in the proponent's argument at all. Therefore, example VI and VIII are not valid attacks.

Example VII is a real HY-undercut. It builds on the conclusions of the proponent to show that these conclusions in essence undercut the very same grounds they are based on.

As only examples V and VII can be regarded as correct instances of undercutting, an HY-enriched formal argumentation system should make sure that if A_2 is an undercutting argument for A_1, then A_2 undercuts a rule in A_1.

Interaction between arguments

So far, the question that has been asked was in which possible ways can an HY-argument attack classical arguments.[11] Another interesting question, however, is whether HY-arguments are capable of attacking other HY-arguments. It will turn out that under certain restrictions, it is indeed possible for HY-arguments to attack and defeat each other.[12]

To illustrate the difficulties that play a role, it us useful to show an example in which two HY-arguments should not defeat each other.

[11] With *classical arguments* we mean non-HY-arguments, that is, arguments that are based on premises only and not on the conclusions of other arguments.

[12] We assume that the rule priorities are such that the third argument (which is called A_3) strictly defeats the second argument (which is called A_2); otherwise no reinstatement of the first argument (which is called A_1) can happen in the first place.

Example IX:

Can an argument A_3 rebut a rebutting HY-argument A_2 (that rebuts an argument A_1)?

$$\Rightarrow\Rightarrow\Rightarrow L \Rightarrow\Rightarrow\Rightarrow \;\mid\; \Rightarrow\Rightarrow\Rightarrow K \Rightarrow\Rightarrow\Rightarrow -L \;\mid\; \Rightarrow\Rightarrow\Rightarrow -K$$

Argument A_3, although it seems to rebut A_2, should not attack it. For if it would, this would lead to reinstatement of A_1, whereas A_3 is in fact just an additional reason to reject A_1 (A_3 shows that from A_1 an additional contradiction K and $-K$ can be derived). In essence, it often makes little sense to rebut a rebutting HY-argument, as the sole reason for the rebutting HY-argument is to show a contradiction in A_1. Additional contradictions (A_3) will only reemphasize what A_2 has already found, namely that the conclusions of A_1 can lead to contradictions.

The observation in example IX, that a rebutting argument A_3 need not to attack a rebutting HY-argument A_2 is not a hard and fast rule, as is illustrated by the following example.

Example IX bis:

We again ask ourselves the question: can an argument A_3 rebut a rebutting HY-argument A_2 (that rebuts an argument A_1)?

$$\Rightarrow\Rightarrow\Rightarrow L \;\mid\; \begin{array}{l}\Rightarrow\Rightarrow\Rightarrow M\ (M \wedge N \Rightarrow K)\ K \Rightarrow\Rightarrow\Rightarrow -L \;\mid\; \\ \Rightarrow\Rightarrow\Rightarrow N\end{array}$$
$$\Rightarrow\Rightarrow\Rightarrow -N$$

Here, argument A_2 can be divided into two parts. The upper part is based on one or more conclusions of A_1 while the lower part (leading to conclusion N) is based on the premises only. Upper and lower part of A_2 are linked by the rule "$M \wedge N \Rightarrow K$". From K, finally, $-L$ can be derived.

A_2 is now attacked by A_3 ($\Rightarrow\Rightarrow -N$), which is based on premises only (that is, it is not built "on top" of A_2). The reason that A_3 can attack A_2 is that A_3 rebuts a conclusion of A_2 that is based on premises only. A_3 makes that the opponent can no longer hold A_2 as true, since one of its intermediate results is rebutted. At the same time, A_3 is *not* an extra reason to reject A_1 (as was the case in our original example IX). The reason for this is that A_3 rebuts an argument (the argument for N) that does not build upon the conclusions of A_1.

The difference between the examples IX and IX bis is that in example IX, A_3 is based on a part of A_2 that is itself based on A_1. In example IX bis, on the other hand, A_3 is not based on a part of A_2 that is itself based on A_1 (in fact, A_3 is not based on A_2 at all).

Also, from the perspective of commitments, the situation is different. Normally, a HY-argument goes as follows: "If you think that p, then you must also think that q, since from p it follows that q". This requires that the other agent is actually committed to p itself, and not to the fact that our own agent should be committed to p. Therefore, a HY-argument should be based on conclusions of another argument that are not fc-based themselves.

Principles

From the above discussion, the following general principles regarding HY-arguments can be stated:

1. An HY-argument contains at least one foreign commitment, that is, a conclusion imported from another argument. All foreign commitments should have their origin in conclusions (of the other argument) that are themselves not fc-based.
2. An HY-argument A_2 HY-rebuts an argument A_1 if
   (a) all foreign commitments of A_2 have their origin in A_1
   (b) A_2 contains conflicting conclusions, of which at least one conclusion is fc-based
3. An HY-argument A_2 HY-undercuts an argument A_1 if
   (a) all foreign commitments of A_2 have their origin in A_1
   (b) A_2 contains an fc-based conclusion that undercuts some rule in A_1

3.3 Formalization and properties

3.3.1 Formalization

In this section we include the concept of HY-arguments in an example system for defeasible reasoning, for which we have chosen Prakken and Sartor's work. In particular, we modify Prakken and Sartor's definitions in order to be able to specify HY-arguments, as well as to extend the defeat relation for this new class of arguments.

The original argumentation system, as axiomatized in section 3.1, will be referred to as $DS_{classic}$. The HY-enriched argumentation system to be defined in the current section will be referred to as DS_{HY}. The idea of DS_{HY} is, roughly, to take $DS_{classic}$, include a new type of arguments (HY-arguments) and extend the defeat-relation so that this new type of arguments is taken into account.

There exists a certain overlap between $DS_{classic}$ and DS_{HY}. That is, while most of the definitions of DS_{HY} are different from those of $DS_{classic}$, some are also the same. To allow for easy reference, the general approach is to include every relevant definition of DS_{HY}, even in cases where there is an overlap with $DS_{classic}$.

An example of a definition that remains unchanged is that of a literal.

Definition 3.12. *Let props be a set of atomic propositions. Now define a set literals* $=$ *props* $\cup \{\neg p \mid p \in props\}$. *A negation function* $(- : literals \longrightarrow literals)$ *is defined by* $-P = \neg P$ *and* $-\neg P = P$.

With HY-arguments, three kinds of rules are distinguished: strict rules, defeasible rules and foreign commitments. Foreign commitments are used as "imported" conclusions from other arguments and serve a similar purpose as assumptions in classical logic.

Definition 3.13. *A rule is an expression of the form:*
$r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L_n$
where r *is the name of the rule and each* L_i $(0 \leq i \leq n)$ *is a literal. The conjunction at*

the left of the arrow is the antecedent and the literal at the right side of the arrow is the consequent of the rule. In the antecedent, $\sim$ stands for weak negation. The following kinds of rules are distinguished:

1. *strict rules:* $r : L_0 \wedge \ldots \wedge L_{n-1} \rightarrow L_n$
2. *defeasible rules:* $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \Rightarrow L_n$
3. *foreign commitments:* $r : \rightsquigarrow L_n$

To illustrate the purpose of foreign commitments, consider the following example.

$\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
$\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
$< = \emptyset$
P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C \quad (A_1)$
O: $\rightsquigarrow C,\ C \rightarrow D,\ \rightarrow \neg D \quad (A_2)$

In A_2, $\rightsquigarrow C$ is a foreign commitment, based on the conclusion C in A_1. The requirement that for an argument to make sense, every foreign commitment should be based on an actual conclusion of another argument, will be formalized in the definition of attack.

An argument, then, is constructed using the three kinds of rules of definition 3.13.

Definition 3.14. *Let $(\mathcal{S}, \mathcal{D}, <)$ be a defeasible theory. An* argument *based on this defeasible theory is a finite sequence $A = [r_0, \ldots, r_n]$] of rules such that:*

1. *for every i ($0 \leq i \leq n$), for every literal L in the antecedent of r_i not preceded by a weak negation sign, there is a $h < i$ such that L is the consequent of r_h*
2. *no two distinct rules in the sequence have the same consequent*
3. *All strict rules of A are in $\mathcal{S}$, all defeasible rules of A are in $\mathcal{D}$, and for all foreign commitments $\rightsquigarrow C$ in A it holds that C is a conclusion of a rule in $\mathcal{S} \cup \mathcal{D}$.*

An argument is classical *iff it does not contain any foreign commitments.*

The notion of a conclusion and an assumption remains unchanged.

Definition 3.15. *Let A be an argument and L be a literal.*

1. *L is a conclusion of A iff L is the consequent of some rule in A*
2. *L is an assumption of A iff $\sim -L$ occurs in some rule in A*

The notion of relevance itself remains unchanged, although we now also apply it to determine which rules are based on foreign commitments and which rules are not.

Definition 3.16. *Let A be an argument with a conclusion L. The set of rules relevant to L — written as $R_L(A)$ — is the smallest set such that:*

1. *$r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L \in R_L(A)$ where r is a rule in A*
2. *if $r : L_0 \wedge \ldots \wedge L_j \wedge \sim L_{j+1} \wedge \ldots \wedge \sim L_{n-1} \longrightarrow L_n \in R_L(A)$ then also $R_{L_0}(A) \cup \ldots \cup R_{L_j}(A) \subseteq R_L(A)$*

We say that conclusion L is based on *a rule r in argument A iff $r \in R_L(A)$.*
We say that L is fc-based *iff $R_L(A)$ contains at least one foreign commitment.*

The next step is then to define when one argument attacks the other. Several types of attacks are distinguished. The first types are what we call *classical* attacks.

Definition 3.17. *Let A_1 and A_2 be two arguments of which at least A_2 is classical.*

1. A_2 classically rebut-attacks A_1 *on L iff A_1 has a conclusion L that is not fc-based and A_2 has a conclusion $-L$.*

2. A_2 classically undercut-attacks A_1 *on L iff A_1 has an assumption L and A_2 has a conclusion $-L$.*

When all arguments are classical, then the above definition boils down to "traditional" rebutting and "traditional" undercutting. A difference with definition 3.8 of $DS_{classic}$ is that the above definition does not implicitly make arguments closed under strict consequence. That is, it does not make use of lists of strict rules S_1 and S_2 to form "extended" arguments $A_1; S_1$ and $A_2; S_2$. The reason for this decision is that using S_1 and S_2 actually implies a limited form of HY-reasoning; one confronts the other party with the *consequences* of its own reasoning. Classical attacks, however, are only concerned with the conclusions of the argument itself.

The notions of HY-attacks are formalized in accordance with the principles as stated earlier at the end of section 3.2.2.

Definition 3.18. *Let A_1 and A_2 be two arguments.*

1. A_2 HY-rebut-attacks A_1 *on L iff:*

   - *A_2 has a conclusion L and a conclusion $-L$, where at least L is fc-based, and*
   - *for every foreign commitment $\rightsquigarrow C$ in A_2: C is a conclusion of A_1 that is not fc-based.*

2. A_2 HY-undercut-attacks A_1 *on L iff*

   - *A_1 has an assumption L and A_2 has a conclusion $-L$ that is fc-based.*
   - *for every foreign commitment $\rightsquigarrow C$ in A_2: C is a conclusion of A_1 that is not fc-based*

At first, it may appear that the above definitions of attack (def. 3.17 and 3.18) are all relevant forms of attack. There is, however, one additional complication. Consider the following example.

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A,\ \rightarrow D,\ \rightarrow \neg F,\ \rightarrow G\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C,\ D \Rightarrow E,\ E \Rightarrow \neg C,\ \neg C \Rightarrow F,\ G \Rightarrow \neg E\} \\ < & = & \emptyset \end{array}$$

Here, the following arguments can be constructed:

$$\begin{array}{ll} A_1: & \rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C \\ A_2: & \rightarrow D,\ D \Rightarrow E,\ E \Rightarrow \neg C \\ A_3: & \rightsquigarrow \neg C,\ \neg C \Rightarrow F,\ \rightarrow \neg F \end{array}$$

A_1 is defeated by A_2, while A_2 is (strictly) defeated by A_3. Therefore, under the above notions of attack, A_1 is reinstated and becomes justified. This may not be what one wants, since the HY-argument A_3 is not stronger than A_2. Yet, according to definition 3.18, A_3 attacks A_2 but A_2 does not attack A_3, meaning that A_3 would *strictly defeat* A_2.

To illustrate why A_3 should not strictly defeat A_2, consider the following alternative counterargument against A_2:

$$A_3' : \quad \rightarrow G,\ G \Rightarrow \neg E$$

This argument is not a reason for reinstating A_1 either, since it is not any stronger than A_2. The point is that, as far as attacks are concerned, classical rebutting is symmetrical. If A classically rebut-attacks B then B also classically rebut-attacks A.

The rationale of the following definition, therefore, is to restore symmetry for HY-rebutting, in order to avoid undesirable instances of reinstatement like sketched above.

Definition 3.19. *Let A_1 and A_2 be two arguments. We say that A_2* reverse HY-rebut-attacks *A_1 on L iff A_1 HY-rebut-attacks A_2 on L.*

The principle of weakest link for comparing the relative strength of arguments remains unchanged.

Definition 3.20. *For any two sets R and R' of rules, $R < R'$ iff for some defeasible rule $r \in R$ and all defeasible rules $r' \in R'$ it holds that $r < r'$.*

The definition of defeat distinguishes five types of attacks: classical rebutting, classical undercutting, HY-rebutting, reverse HY-rebutting and HY-undercutting. The way in which priorities are dealt with might seem quite complicated at first, but can be made clear using the following example.

$$\begin{array}{lll} \mathcal{S} & = & \{\rightarrow a\} \\ \mathcal{D} & = & \{a \Rightarrow b,\ a \Rightarrow c,\ b \wedge c \Rightarrow d,\ b \Rightarrow e,\ c \Rightarrow \neg e\} \\ \text{P:} & \rightarrow a,\ a \Rightarrow b,\ a \Rightarrow c,\ b \wedge c \Rightarrow d & (A_1) \\ \text{O:} & \rightsquigarrow b,\ b \Rightarrow e,\ \rightsquigarrow c,\ c \Rightarrow \neg e & (A_2) \end{array}$$

Of which rules should the priorities be taken into account? For A_2, the relevant rules are $b \Rightarrow e$ and $c \Rightarrow \neg e$, since these lead up to the contradiction. For A_1, the relevant rules are $a \Rightarrow b$ and $a \Rightarrow c$ (but *not* $b \wedge c \Rightarrow d$) since these produce the conclusions imported by A_2 as foreign commitments. In general, when evaluating the priorities in a HY-setting, one should ask oneself two questions: (1) how strong is derivation of the contradiction in A_2 and (2) how strong is the derivation of the conclusions in A_1 that are used as foreign commitments in A_2 to derive the contradiction. That is, we ask ourself the question what do we trust more: the reason for rejecting the foreign commitments, or the reason for accepting the foreign commitments. This is formalized in the following definition.

Definition 3.21. *Let A_1 and A_2 be two arguments based on $(\mathcal{S}, \mathcal{D}, <)$.* A_2 defeats A_1 *iff there exists an L such that:*

1. *A_2 classically rebut-attacks A_1 on L and*
   not $R_{-L}(A_2) < R_L(A_1)$, or

2. A_2 *classically undercut-attacks* A_1 *on* L*, or*

3. A_2 *HY-rebut-attacks* A_1 *on* L *and*
not $R_L(A_2) \cup R_{-L}(A_2) < \cup_{c_i \in \{c_i | \rightsquigarrow c_i \in R_L(A_2) \cup R_{-L}(A_2)\}} R_{c_i}(A_1)$*, or*

4. A_2 *reverse HY-rebut-attacks* A_1 *on* L *and*
not $\cup_{c_i \in \{c_i | \rightsquigarrow c_i \in R_L(A_1) \cup R_{-L}(A_1)\}} R_{c_i}(A_2) < R_L(A_1) \cup R_{-L}(A_1)$*, or*

5. A_2 *HY-undercut-attacks* A_1 *on* L*.*

We say that A_2 strictly *defeats* A_1 *iff* A_2 *defeats* A_1 *and* A_1 *does not defeat* A_2*.*

Definition 3.22. *If* A *is a list of rules or a set of rules, then* $strict(A)$ *is the set of strict rules in* A *and* $defeasible(A)$ *is the set of defeasible rules in* A*.*

Definition 3.23. *An argument is* coherent *iff it does not have a counterargument without defeasible rules.*

For now, justified arguments are defined in terms of dialogue games. A semantical treatment will be provided in section 5.1.

Definition 3.24. *A formula is* justified *iff it is the conclusion of a classical argument for which exists a winning strategy in dialogue.*

Examples

To illustrate the workings of DS_{HY}, consider again the earlier mentioned examples of section 3.1.

1. $\mathcal{S} = \{\rightarrow A, \rightarrow D\}$
$\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C, D \Rightarrow E,\ E \Rightarrow \neg C\}$
$< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$ (A_1)
   O: $\rightarrow D,\ D \Rightarrow E,\ E \Rightarrow \neg C$ (A_2)

   Here, A_2 classically rebut-attacks and defeats A_1.

2. $\mathcal{S} = \{\rightarrow A, \rightarrow E\}$
$\mathcal{D} = \{A \wedge \sim B \Rightarrow C,\ C \Rightarrow D,\ E \Rightarrow B\}$
$< = \emptyset$

   P: $\rightarrow A,\ A \wedge \sim B \Rightarrow C,\ C \Rightarrow D$ (A_1)
   O: $\rightarrow E,\ E \Rightarrow B$ (A_2)

   Here, A_2 classically undercut-attacks and defeats A_1.

3. $\mathcal{S} = \{\rightarrow A, \rightarrow D,\ B \rightarrow C,\ E \rightarrow \neg C\}$
$\mathcal{D} = \{A \Rightarrow B,\ D \Rightarrow E\}$
$< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B$ (A_1)
   O: $\rightarrow D,\ D \Rightarrow E,\ \rightsquigarrow B,\ B \rightarrow C$ (A_2)

   Here, A_2 HY-rebut-attacks and defeats A_1. Notice that because both arguments

have the same priority, A_1 also reverse HY-rebut-attacks A_2. Thus, the defeat is not strict. The defeat becomes strict if $(A \Rightarrow B) < (D \Rightarrow E)$. In that case, A_1 does not reverse HY-rebut-attacks A_2.

4. $\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
   $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C,\ C \rightarrow D,\ \rightarrow \neg D$ (A_1)
   O: $\rightsquigarrow D,\ \rightsquigarrow \neg D$ (A_2)

   Here, the incoherent argument A_1 is strictly defeated by HY-argument A_2. Because A_2 does not have any defeasible rules, it cannot be defeated. This is akin to the empty argument in $DS_{classic}$.

5. $\mathcal{S} = \{\rightarrow A,\ C \rightarrow D,\ \rightarrow \neg D\}$
   $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
   $< = \emptyset$

   P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$ (A_1)
   O: $\rightsquigarrow C,\ C \rightarrow D,\ \rightarrow \neg D$ (A_2)

   Here, like in the previous example, A_1 is strictly defeated by HY-argument A_2. The advantage of using HY-arguments instead of just the empty argument (the approach of $DS_{classic}$) is that it makes clear *why* A_1 is incoherent.

6. Same as example 5.

7. "Ajax-Feijenoord"
   $\mathcal{S} = \{\rightarrow af\}$
   $\mathcal{D} = \{af \wedge \sim p \Rightarrow t,\ t \Rightarrow p\}$
   $< = \emptyset$

   P: $\rightarrow af,\ af \wedge \sim p \Rightarrow t$ (A_1)
   O: $\rightsquigarrow t,\ t \Rightarrow p$ (A_2)
   Here, A_2 HY-undercut-attacks and defeats A_1.

8. "shipment of goods"
   $\mathcal{S} = \{\rightarrow tma,\ \rightarrow \neg is\}$
   $\mathcal{D} = \{tma \Rightarrow a,\ \neg is \Rightarrow \neg cd,\ a \Rightarrow cd\}$
   $< = \emptyset$

   P: $\rightarrow tma,\ tma \Rightarrow a$ (A_1)
   O: $\rightsquigarrow a,\ a \Rightarrow cd,\ \rightarrow \neg is,\ \neg is \Rightarrow \neg cd$ (A_2)
   Here, A_2 HY-rebut attacks and defeats A_1.

9. "tuff-tuff-club"
   $\mathcal{S} = \{\rightarrow f,\ \rightarrow c\}$
   $\mathcal{D} = \{f \wedge c \Rightarrow s,\ s \Rightarrow ttc,\ ttc \Rightarrow p,\ p \wedge c \Rightarrow \neg s\}$
   $< = \emptyset$

   P: $\rightarrow f,\ \rightarrow c,\ f \wedge c \Rightarrow s,\ s \Rightarrow ttc$ (A_1)
   O: $\rightsquigarrow ttc,\ ttc \Rightarrow p,\ \rightarrow c,\ p \wedge c \Rightarrow \neg s,\ \rightsquigarrow s$ (A_2)
   Here, A_2 HY-rebut-attacks and defeats A_1.

10. "tax relief"

$\mathcal{S} = \{\rightarrow pmp\}$
$\mathcal{D} = \{pmp \Rightarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr\}$
$< = \emptyset$

P: $\rightarrow pmp,\ pmp \Rightarrow tr$ (A_1)
O: $\rightsquigarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr$ (A_2)

3.3.2 Conclusion maximality versus rule maximality

In this section the concepts of conclusion maximality and rule maximality are introduced. It is explained what these concepts are and how they relate to $DS_{classic}$ and DS_{HY}.

The difference between conclusion maximality and rule maximality is perhaps best explained using an example.

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A,\ \rightarrow C\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow Z,\ C \Rightarrow D,\ D \Rightarrow \neg Z\} \\ < & = & \emptyset \end{array}$$

In this defeasible theory, an argument for Z can be constructed, as well as an argument for $\neg Z$.

$$\begin{array}{lllllll} \rightarrow & A, & A & \Rightarrow & B, & B & \Rightarrow & Z \\ \rightarrow & C, & C & \Rightarrow & D, & D & \Rightarrow & \neg Z \end{array}$$

It is not difficult to see that in $DS_{classic}$ A, B, C and D are justified. In DS_{HY}, on the other hand, only A and C are justified. This is because B for instance now has a counterargument ($\rightsquigarrow B,\ B \Rightarrow Z,\ \rightarrow C,\ C \Rightarrow D,\ D \Rightarrow \neg Z$).

Regarding the preservation of consistency, there are essentially two approaches for analyzing this example. If one views the example from the perspective of Reiter's default logic, and represent the defeasible rules in $\mathcal{D}$ by normal defaults, then two extensions would exist: $\{A, B, Z, C, D\}$ and $\{A, B, C, D, \neg Z\}$. Both extensions represent a (maximal) defensible point of view. The sets are maximal in the sense that the addition of any derivable conclusions not in the set would lead to a contradiction.

What Reiter, as well as many other researchers do is that they aim to derive as many conclusions as possible, without running into inconsistency. If one would regard the problem from the perspective of rules instead of conclusions, then default logic tries to apply as many rules as possible, except those rules that would *immediately* lead to a conflict.

There is also a different way to look at the conflict in the above example. There are four defeasible rules that are involved in the conflict. These four rules cannot all be applied at the same time, since this would lead to inconsistency, so at least one rule has to be omitted. As no rule has priority above another rule, all rules can equally be blamed for the conflict. Therefore, there are four maximal conflict-free sets of rules:

$$\begin{array}{ll} \{A \Rightarrow B,\ B \Rightarrow Z,\ C \Rightarrow D\} & \{A \Rightarrow B,\ B \Rightarrow Z,\ D \Rightarrow \neg Z\} \\ \{A \Rightarrow B,\ C \Rightarrow D,\ D \Rightarrow \neg Z\} & \{B \Rightarrow Z,\ C \Rightarrow D,\ D \Rightarrow \neg Z\} \end{array}$$

These sets have the following associated sets of conclusions:

$$\begin{array}{ll} \{A, B, Z, C, D\} & \{A, B, C, Z\} \\ \{A, B, C, D, \neg Z\} & \{A, C, D, \neg Z\} \end{array}$$

Under sceptical semantics (a proposition is justified iff it is contained in every extension) only A and C are justified. Compare this with the result of default logic, where A, B, C and D are justified.

We call the principle of taking maximal sets of rules — regardless whether they can be applied or not — such that no conflicts can occur *rule maximality.*[13] The idea of taking maximal sets of conclusions such that there exists a coherent argument for all of them (the approach that is for instance taken by default logic) we call *conclusion maximality.*

In the remainder of this section, a formal account is given of these two concepts. In order to simplify matters, we restrict ourselves to the case where no priorities are defined and the only form of defeat is rebutting.

Definition 3.25. *A defeasible theory* $(\mathcal{S}, \mathcal{D}, <)$ *is* simple *iff* $< = \emptyset$ *and* $\forall r \in \mathcal{D} : r$ *is a rule without any weak negation signs.*

We often abbreviate a simple defeasible theory $(\mathcal{S}, \mathcal{D}, <)$ to $(\mathcal{S}, \mathcal{D})$.

The advantage of simple defeasible theories is that reinstatement becomes an almost trivial issue. Consider the following (abstract) dialogue:

P: A_1
O: A_2
P: A_3

A_3 can reinstate A_1 by strictly defeating A_2. In $DS_{classic}$, for A_3 to strictly defeat A_2, at least one of the following conditions has to be fulfilled (see definition 3.11):

1. A_3 undercuts A_2
2. A_3 rebuts A_2 and is stronger
3. A_2 is incoherent and $A_3 = \emptyset$

In a simple defeasible theory, no undercutting can take place, so option 1 is ruled out. As for option 2, the only way how A_2 can be stronger when no priorities have been defined is when A_3 consists of strict rules only. This would mean that A_2 is inconsistent; the same that is the case at option 3. Overall, an argument is justified in $DS_{classic}$ under a simple defeasible theory, iff it does not have a coherent counterargument. A similar observation can be made about DS_{HY}.

Before proceeding with the main treatment, some preliminaries are to be defined. The first notion to be defined is that of a maximal set. A maximal set can roughly be described as a set that satisfies a certain property and that cannot be extended with additional elements without losing this property.

Definition 3.26. *Let E be a finite universe, $S \subseteq E$ and P a property. S is a* maximal set *such that property P holds iff S satisfies property P and there is no set S' with $S \subsetneqq S' \subseteqq E$ such that S' satisfies property P.*

Lemma 3.1. *Let E be a finite universe. If there is a set S that satisfies property P then there is also a maximal set S_{max} that satisfies property P.*

[13] An example of an existing formalism that implements rule maximality is input/output logic (IOL) under elimination of excess output [MaTo01]. IOL's concept of a *maxfamily* can be compared to a rule-maximal set of rules (definition 3.32).

Proof. Trivial. □

Given two lists A and B. We say that A is a *weak sublist* of B iff removing some elements of B makes it equal to A.

Definition 3.27 (weak sublist). *Let $A = [a_1, a_2, \ldots, a_m]$ and $B = [b_1, b_2, \ldots, b_n]$ be lists of elements. A is called a* weak sublist *of B if there exist zero or more $b_{i_1}, b_{i_2}, \ldots, b_{i_k} (1 \leq i_1 < i_2 < \ldots < i_k \leq n)$ such that $[b_1, b_2, \ldots, b_{i_1-1}, b_{i_1+1}, \ldots, b_{i_2-1}, b_{i_2+1}, \ldots, b_{i_k-1}, b_{i_k+1}, \ldots, b_n]$ is equal to A. A is a* strict weak sublist *of B iff A is a weak sublist of B and $A \neq B$.*

Conclusion maximality

In order to define conclusion maximality, we first define the sets of all coherent classical arguments that can be constructed using a (simple) defeasible theory.

Definition 3.28. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory.*

- $CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) :=$
  $\{A \mid A$ *is a coherent classical argument with* $defeasible(A) \subseteq \mathcal{D}$ *and* $strict(A) \subseteq \mathcal{S}\}$
- $IncoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) :=$
  $\{A \mid A$ *is an incoherent classical argument with* $defeasible(A) \subseteq \mathcal{D}$ *and* $strict(A) \subseteq \mathcal{S}\}$
- $Args_{classic}(\mathcal{S}, \mathcal{D}) :=$
  $CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) \cup IncoherentArgs_{classic}(\mathcal{S}, \mathcal{D})$

Conclusion maximality can be defined in two ways. The first way is centered around a maximal set (in the sense of definition 3.26) of *conclusions* that has a coherent argument.

Definition 3.29 (C_{cmax}). *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. A* conclusion-maximal set of conclusions *C_{cmax} is a maximal set of literals such that*
$\exists A \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : conclusions(A) = C_{cmax}$.

Notice that every C_{cmax} is consistent because otherwise A would be incoherent.

Lemma 3.2. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. There exists at least one conclusion-maximal set of conclusions C_{cmax}.*

Proof. Let C be the empty set; then it holds that C is a set of literals such that there is a coherent argument A (the empty argument) with $conclusions(A) = C$. According to lemma 3.1, there also exists a *maximal* set of literals C_{cmax} such that there is a coherent argument A with $conclusions(A) = C_{cmax}$. This satisfies the definition of a conclusion-maximal set of conclusions (definition 3.29). □

Notice that since $\mathcal{S}$ and $\mathcal{D}$ are finite, every argument based on $(\mathcal{S}, \mathcal{D}, <)$ contains finitely many conclusions. Therefore, C_{cmax} is also finite.

Another way to define conclusion maximality is by taking maximal sets of rules that can be applied in a coherent argument.

Definition 3.30 (R_{cmax}). *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. A* conclusion-maximal set of rules *R_{cmax} is a maximal set of defeasible rules such that*
$\exists A \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : defeasible(A) = R_{cmax}$.

Lemma 3.3. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. There exists at least one conclusion-maximal set of rules R_{cmax}.*

Proof. Similar to the proof of lemma 3.2. □

A maximal coherent argument is a coherent argument that cannot be extended with additional rules without making it incoherent. This is stated in the following definition.

Definition 3.31 (maximal argument). *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory and $A \in CoherentArgs_{classic}$.*
A is a maximal coherent classical argument *iff*
$\neg\exists r \in (\mathcal{S} \cup \mathcal{D}) \backslash (defeasible(A) \cup strict(A)) \exists B \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D})$
$(r \in defeasible(B) \cup strict(B) \wedge A$ *is a weak sublist of* $B)$.

The following theorem states that every C_{cmax} has an associated R_{cmax}.

Theorem 3.1 (from C_{cmax} to R_{cmax}). *For every conclusion-maximal set of conclusions C_{cmax}, there is a coherent classical argument A with $conclusions(A) = C_{cmax}$ and $defeasible(A) = R_{cmax}$, where R_{cmax} is a conclusion-maximal set of rules.*

Proof. Let C_{cmax} be a conclusion-maximal set of conclusions. Then, according to he definition of a conclusion-maximal set of conclusions (definition 3.29), C_{cmax} is a *maximal* set of literals such that $\exists A \in CoherentArgs_{classic} : conclusions(A) = C_{cmax}$. This means that:

1. $\exists A \in CoherentArgs_{classic} : conclusions(A) = C_{cmax}$ and
2. $\neg\exists C'_{cmax} \supsetneq C_{cmax} \exists A' \in CoherentArgs_{classic} : conclusions(A') = C_{cmax}$

A is a maximal argument, that is, a coherent classical argument satisfying (definition 3.31): $\neg\exists r \in (\mathcal{S}, \mathcal{D}) \backslash (defeasible(A) \cup strict(A)) \exists B \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D})(r \in defeasible(B)$ $\cup\, strict(B) \wedge A$ is a weak sublist of $B)$.

Proof. Suppose $\exists r \in (\mathcal{S}, \mathcal{D})$ $(defeasible(A) \cup strict(A)) \exists B \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D})(r \in$ $defeasible(B) \cup strict(B) \wedge A$ is a weak sublist of $B)$. Let r be such a rule. As an argument does not have double conclusions, the conclusion of r should not occur as a conclusion of any other rule in B. As A is a weak sublist of B, the conclusion of r should also not occur as a conclusion of any rule in A. This means that c is a conclusion that is not in A (and therefore also not in C_{cmax}). Hence: $\exists C'_{cmax} \supsetneq C_{cmax} \exists A' \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D})$: $conclusions(A') = C'_{cmax}$ (take a C'_{cmax} containing c and a A' equal to B). Contradiction. □

Now, let A'' be an argument with $conclusions(A'') = conclusions(A)$ and $defeasible(A'')$ be a maximal superset of $defeasible(A)$ (such an A'' always exists). Let $R_{cmax} = defeasible(A'')$. It now holds that R_{cmax} is a maximal set of defeasible rules such that $\exists A \in Coherent$-$Args_{classic}(\mathcal{S}, \mathcal{D})$: $defeasible(A) = R_{cmax}$ (take A'' for A). This means that A'' has a conclusion-maximal set of rules. As $conclusions(A'') = conclusions(A) = C_{cmax}$, it also means that A'' has a conclusion-maximal set of conclusions. □

The following theorem states that every R_{cmax} has an associated C_{cmax}.

Theorem 3.2 (from R_{cmax} to C_{cmax}). *For every conclusion maximal set of rules R_{cmax}, there is a classical coherent argument A with $defeasible(A) = R_{cmax}$ and $conclusions(A) = C_{cmax}$, where C_{cmax} is a conclusion-maximal set of conclusions.*

Proof. Let R_{cmax} be a conclusion-maximal set of rules. According to the definition of a conclusion-maximal set of rules (definition 3.30), this means that R_{cmax} is a maximal set of defeasible rules such that $\exists A \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : defeasible(A) = R_{cmax}$. Let A be such that $defeasible(A) = R_{cmax}$. Let B be a coherent argument such that A is a weak sublist of B and $strict(B)$ is maximal. B is then a maximal argument (we cannot "add" any strict or defeasible rules anymore): $\neg\exists r \in (\mathcal{S}, \mathcal{D})$ $(defeasible(B) \cup strict(B))\exists C \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D})(r \in defeasible(C) \cup strict(C) \wedge B$ is a weak sublist of $C)$. We now have to prove that the conclusions of B form a conclusion-maximal set of conclusions. For this, according to the definition of a conclusion-maximal set of conclusions (definition 3.29), we have to prove two things:

1. $\exists E \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : conclusions(E) = conclusions(B)$
   This is trivial, just take $E = B$.

2. $defeasible(B)$ is a *maximal* set such that
   $\exists E \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : conclusions(E) = conclusions(B)$

For the second point, we have to prove that there is no $E \in CoherentArgs_{classic}(\mathcal{S}, \mathcal{D}) : defeasible(E) = defeasible(B)$. Suppose, such an E would exist. Then E would have a rule with a conclusion that is not in B. Let r be the "first" of such a rule, that is, a rule with all premises in B and the conclusion not in B (such a "first" rule always exists if there is any rule in E that is not in B). This, however, would mean that there exists an argument B' (containing rule r) such that B is a weak sublist of B'. Contradiction. □

Summarizing, one can say that there are two possible ways to define conclusion maximality (one from the perspective of conclusions and one from the perspective of rules) and that these ways are related to each other.

One of the properties of conclusion maximality is that it blames the *last* rule in case of possible conflicts. The following theorem basically says that if there is a minimal incoherent set of rules (and there are no other defeasible rules except those in this minimal incoherent set of rules), then only by leaving out a *last* rule of A, a conclusion-maximal set of rules can be constructed.

Theorem 3.3. *Let $(\mathcal{S}, \mathcal{D})$ be a defeasible theory where $\mathcal{D}$ is a minimal set of defeasible rules such that an incoherent argument A can be constructed with $defeasible(A) = \mathcal{D}$. Let r be a defeasible rule in D such that there is a conclusion-maximal set of rules R_{cmax} with $D - \{r\} \subseteq R_{cmax}$ and $r \notin R_{cmax}$. Then there does not exist a r' such that $r' \neq r$ and $r \in R_{r'}(A)$.*

Proof. Let $(\mathcal{S}, \mathcal{D})$ be an defeasible theory where $\mathcal{D}$ is a minimal set of defeasible rules such that an incoherent classical argument A can be constructed with $defeasible(A) = \mathcal{D}$. Let r be a defeasible rule in $\mathcal{D}$ such that there is a conclusion-maximal set of rules $R_{cmax} = \mathcal{D} - \{r\}$.

Suppose r is *not* the last defeasible rule in A, that is, suppose there exists an r' with $r' \neq r$ and $r \in R_{r'}(A)$. $\mathcal{D} - \{r\}$ is a conclusion-maximal set of rules, that is, a maximal set of defeasible rules such that there is a coherent classical argument A' with $defeasible(A') = R_{cmax}$. A' does not contain r, nor does A' contain any other rule with the conclusion of r.

Proof. Suppose A' would contain a rule with the conclusion of r. Then the conclusion of r would be derivable in A'. Hence, the conclusion of r also follows from $\mathcal{D} - \{r\}$ (since

$defeasible(A') = R_{cmax} = \mathcal{D} - \{r\}$). This, however, means that everything that follows from $\mathcal{D}$ also follows from $\mathcal{D} - \{r\}$, and since $\mathcal{D}$ allows for the construction of an incoherent argument, $\mathcal{D} - \{r\}$ also allows for it. This contradicts with the fact that $\mathcal{D} - \{r\}$ is a conclusion-maximal set of rules. □

In A', however, r' is based on r. It is not possible to base r' on a set of rules not including r, because then rule r would not be necessary and $\mathcal{D}$ would not be minimal. Contradiction. This means that r must be the *last* rule in A. □

Rule maximality

The next concept to be defined is that of rule maximality.

Definition 3.32 (R_{rmax}). *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory.*
A rule-maximal set of rules R_{rmax} *is a maximal set of rules such that* $R_{rmax} \subseteq \mathcal{D}$ *and* $\neg\exists A \in IncoherentArgs(\mathcal{S}, \mathcal{D}) : defeasible(A) \subseteq R_{rmax}$.

Lemma 3.4. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. There exists at least one rule-maximal set of rules R_{rmax}.*

Proof. Let A be an arbitrary argument without defeasible rules. From the fact that $\mathcal{S}$ is consistent, it follows that A is a coherent argument. Thus, it follows that there exists an $R(= \emptyset) \subseteq \mathcal{D}$ such that there is no incoherent argument A' with $defeasible(A') \subseteq R$. According to lemma 3.1, there also exists a *maximal* R with this property. Thus, there exists a maximal set $R_{rmax} \subseteq \mathcal{D}$ such that there does not exist an incoherent argument A' with $defeasible(A') \subseteq R_{rmax}$. This satisfies the definition of rule maximality (definition 3.32). □

Contrary to conclusion maximality, rule maximality blames an *arbitrary* rule in case of possible conflicts.

Theorem 3.4. *Let I be a minimal set of defeasible rules such that there is an incoherent argument A with $defeasible(A) \subseteq I$. For every $r \in I$ there is a rule-maximal set of rules R_{rmax} with $I - \{r\} \subseteq R_{rmax}$ and $r \notin R_{rmax}$.*

Proof. Let I be a minimal set of defeasible rules such that there is an incoherent argument A with $defeasible(A) \subseteq I$. Now, let r be an arbitrary element of I. Because I is minimal, it holds that no incoherent argument A can be constructed from $I - \{r\}$. This (according to lemma 3.1 about maximal sets) means that there is also a *maximal* set of rules R_{rmax} with $I - \{r\} \subseteq R_{rmax}$ such that no incoherent argument can be constructed with the rules of R_{rmax}. It also holds that $r \notin R_{rmax}$ (as otherwise one could construct an incoherent argument from R_{rmax}). □

Cautious monotonity

One way to analyze the differences between conclusion maximality and rule maximality is from the perspective of postulates. Take for instance the postulate of cautious monotonity. Using the syntax of P&S, this postulate can be formulated as follows. Recall that $\mid\sim$ stands for nonmonotonic derivability.

$$(\mathcal{S},\mathcal{D}) \mid\!\sim L \;\&\; (\mathcal{S},\mathcal{D}) \mid\!\sim M \implies (\mathcal{S} \cup \{\rightarrow L\},\mathcal{D}) \mid\!\sim M$$

As for conclusion maximality, the postulate of cautious monotonity does not hold. This is stated in the following theorem:

Theorem 3.5. *Let "$(\mathcal{S},\mathcal{D}) \mid\!\sim_{cmax} L$" stand for "every conclusion-maximal set of conclusions of $(\mathcal{S},\mathcal{D})$ contains L". It does* not *hold that if $(\mathcal{S},\mathcal{D}) \mid\!\sim_{cmax} L$ and $(\mathcal{S},\mathcal{D}) \mid\!\sim_{cmax} M$ then $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D}) \mid\!\sim_{cmax} M$.*

Proof. Take the following counterexample:
$\mathcal{S} = \{\rightarrow p\}$
$\mathcal{D} = \{p \Rightarrow q,\ q \Rightarrow r,\ r \Rightarrow \neg q,\ \neg q \Rightarrow s,\ \Rightarrow \neg s\}$
Notice that the only (minimal) argument for s is $\rightarrow p,\ p \Rightarrow q,\ q \Rightarrow r,\ r \Rightarrow \neg q,\ \neg q \Rightarrow s$, which is inconsistent. Thus, there is only one conclusion-maximal setof conclusions: $\{p, q, r, \neg s\}$. This corresponds with the maximal argument $\rightarrow p,\ p \Rightarrow q,\ q \Rightarrow r,\ \Rightarrow \neg s$. This means that $(\mathcal{S},\mathcal{D}) \mid\!\sim_{cmax} r$ and $(\mathcal{S},\mathcal{D}) \mid\!\sim_{cmax} \neg s$. Now consider what happens when $\rightarrow r$ is added to the set of strict rules. Let $\mathcal{S}' = \mathcal{S} \cup \{\rightarrow r\}$. Under the defeasible theory $(\mathcal{S}',\mathcal{D})$, there *does* exist a coherent argument for s: $\rightarrow r,\ r \Rightarrow \neg q,\ \neg q \Rightarrow s$. There are now two conclusion-maximal sets of conclusions: $\{p, r, q, \neg s\}$ and $\{p, r, \neg q, s\}$. These correspond with the maximal arguments $\rightarrow p,\ \rightarrow r,\ p \Rightarrow q,\ \Rightarrow \neg s$ and $\rightarrow p,\ \rightarrow r,\ r \Rightarrow \neg q,\ \neg q \Rightarrow s$. This means that it does not hold that $(\mathcal{S} \cup \{\rightarrow r\},\mathcal{D}) \mid\!\sim_{cmax} \neg s$. Therefore, cautious monotonity under conclusion-maximization does not hold. □

Under rule-maximization, however, cautious monotonity *does* hold.

Theorem 3.6. *Let "$(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} L$" stand for "every rule-maximal set of rules R_{rmax} has an argument with conclusion L". It then holds that if $(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} L$ and $(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} M$ then $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D}) \mid\!\sim_{rmax} M$.*

Proof. Suppose that $(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} L$ and $(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} M$. We now have to prove that $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D}) \mid\!\sim_{rmax} M$. For this, it suffices to show that every rule-maximal set of rules under $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D})$ is also a rule-maximal set of rules under $(\mathcal{S},\mathcal{D})$ (if this holds, then from $(\mathcal{S},\mathcal{D}) \mid\!\sim_{rmax} M$ it would follow that $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D}) \mid\!\sim_{rmax} M$). We prove this *reductio ad absurdum*. Let R'_{rmax} be a rule-maximal set of rules under $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D})$. Suppose R'_{rmax} is *not* a rule-maximal set of rules under $(\mathcal{S},\mathcal{D})$. There are two possibilities:

1. R'_{rmax} is not a rule-maximal set of rules under $(\mathcal{S},\mathcal{D})$ because an incoherent argument A can be constructed under $(\mathcal{S},\mathcal{D})$. That is, there is an incoherent argument A with $\mathit{defeasible}(A) \subseteq R'_{rmax}$. But then A would also be incoherent under $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D})$. This means that R'_{rmax} is not a rule-maximal set of rules under $(\mathcal{S} \cup \{\rightarrow L\},\mathcal{D})$. Contradiction.

2. R'_{rmax} is not a rule-maximal set of rules under $(\mathcal{S},\mathcal{D})$ because it is not maximal. That is, there exists a rule-maximized set of rules R''_{rmax} under $(\mathcal{S},\mathcal{D})$ that is a strict superset of R'_{rmax}. For this R''_{rmax}, there exists an argument A with $\mathit{defeasible}(A) \subseteq R''_{rmax}$ and $-L$ as a conclusion (that is, $-L$ follows from R''_{rmax}).

   Proof. This can be seen as follows. Suppose R''_{rmax} does not have an argument A with $\mathit{defeasible}(A) \subseteq R''_{rmax}$ and $-L$ as conclusion under $(\mathcal{S},\mathcal{D})$. Because R''_{rmax} is

a rule-maximal set of rules, it is not possible to construct an incoherent argument from R''_{rmax}. This, together with the fact that $-L$ does not follow from R''_{rmax} means that R''_{rmax} is also a rule-maximal set of rules under $(\mathcal{S} \cup \{\rightarrow L\}, \mathcal{D})$. But then R'_{rmax} would not have been maximal (because no incoherent arguments can be constructed with it, and it is also maximal, because if one can add something, then it could also be added in $(\mathcal{S}, \mathcal{D})$)! Contradiction. □

However, from $(\mathcal{S}, \mathcal{D}) \mid\sim_{rmax} L$ follows that every rule-maximal set of rules under $(\mathcal{S}, \mathcal{D})$ should have an argument with conclusion L. However, R''_{rmax} has an argument with conclusion $-L$. This means that there is also an argument with conclusion L and $-L$ (just put the two arguments together, leaving out double conclusions), so R''_{rmax} is incoherent. Contradiction.

□

From $DS_{classic}$ to conclusion maximality

The next thing to do is to show that conclusions that are justified in $DS_{classic}$ are also justified under conclusion maximality.

Theorem 3.7 (from $DS_{classic}$ to conclusion-maximization).
If conclusion c is justified in $DS_{classic}$, then c is an element of every conclusion-maximal set of conclusions.

Proof. Suppose that c is not an element of every conclusion-maximal set of conclusions. Then there is a conclusion-maximal set of conclusions without c. There are two possibilities:

1. There does not exist a coherent argument for c. Then c is indeed not justified.

2. There is a coherent argument for c. Let C be an arbitrary coherent argument for c. The fact that there is a C_{cmax} without c means that there is a maximal argument $A_{C_{cmax}}$ where c is not a conclusion. Let r be a rule in C with the antecedent in $A_{C_{cmax}}$ but the conclusion not in $A_{C_{cmax}}$ Such a rule *always* exists.

   Proof. C contains at least one conclusion (c) that $A_{C_{cmax}}$ does not contain. Therefore, C contains at least one rule with a conclusion that is not in $A_{C_{cmax}}$. This also means that C contains a "first" rule (r) with a conclusion that is not in $A_{C_{cmax}}$. The fact that r is the *first* rule in C with a conclusion that is not in $A_{C_{cmax}}$ means that all rules in C before r have conclusions that are in $A_{C_{cmax}}$. As the antecedent of r is contained in the conclusions before r, it also holds that the antecedent of r is contained in the conclusions of $A_{C_{cmax}}$. Thus, we have a rule (r) whose antecedent is in $A_{C_{cmax}}$, but whose conclusion is not in $A_{C_{cmax}}$. □

   Because $A_{C_{cmax}}$ is a maximal argument, the inclusion of this rule would make it incoherent. Thus (as we only have rebutting), the conclusion of r is in conflict with a conclusion in $A_{C_{cmax}}$. Thus, C is a counterargument against $A_{C_{cmax}}$, so (because defeat is symmetrical by lack of priorities) $A_{C_{cmax}}$ is a counterargument against C so c is not justified.

   □

From DS_{HY} to rule maximality

The next step is to show that if an argument is justified in DS_{HY}, it is also justified under rule maximality.

First, it is defined when a literal *follows from* a defeasible set of rules.

Definition 3.33. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. A literal c* follows from *a set of defeasible rules $R \subseteq \mathcal{D}$ iff there is a classical argument A with conclusion c and $defeasible(A) \subseteq R$ and $strict(A) \subseteq \mathcal{S}$.*

Lemma 3.5. *Let $(\mathcal{S}, \mathcal{D})$ be a simple defeasible theory. Let R_{rmax} be a rule-maximal set of rules such that c does not follow from R_{rmax}. Let A be a (classical) argument for c. Then, there exists a (HY-)counterargument that defeats A.*

Proof. A contains at least one defeasible rule (r) that is not in R_{rmax}.

Proof. Suppose R_{rmax} contains every defeasible rule of A (that is, suppose $defeasible(A) \subseteq R_{rmax}$). Then, one can use R_{rmax} to construct argument A. Hence, c would follow from R_{rmax}. Contradiction. □

If we would "add" this rule to R_{rmax}, this results in the set $R = R_{rmax} \cup \{r\}$. Because R_{rmax} is a *maximal* set such that no incoherent arguments can be constructed, it holds that from R an incoherent argument can be constructed. That is: $\exists A' \in IncoherentArgs(\mathcal{S}, \mathcal{D})$: $defeasible(A') \subseteq R$. Because A' is inconsistent, it holds that there is an S, such that $A'; S$ has conclusion L and conclusion $-L$ (this is because we only have rebutting). Furthermore, A' contains r (since otherwise A' could not have been incoherent).

Now define A'' as A' with r replaced by $\rightsquigarrow conclusion(r)$. It now holds that:

1. L or $-L$ is based on $\rightsquigarrow conclusion(r)$. Thus, either L or $-L$ is fc-based.
2. every foreign commitment from A'' is based on a conclusion of A that is not fc-based (this is because r is a rule in A and A is a classical argument).

These two properties mean that A'' is a HY-counterargument against A (in the sense of definition 3.18 (1)). Furthermore, A'' is also coherent; the part of A'' that is not fc-based is a subset of R_{rmax}, so no strict HY-counterargument can be made against A''. □

Theorem 3.8 (from DS_{HY} to R_{rmax}).
If c is justified in DS_{HY}, then c follows from every rule-maximal set of rules.

Proof. Suppose c does not follow from every rule-maximal set of rules. Then there are two possibilities:

1. There is no coherent classical argument for c. Then c is indeed not justified.
2. There is a coherent classical argument for c (let A be an arbitrary coherent classical argument for c). Let R_{rmax} be a rule-maximal set of rules such that c does not follow from R_{rmax}. According to lemma 3.5, a HY-counterargument can be constructed that defeats A, and since this counterargument cannot be strictly defeated, we have that A is not justified. As this holds for any arbitrary A that is a classical coherent argument for c, it follows that c is not justified.

□

From rule maximality to conclusion maximality

In this subsection, it is shown that with rule-maximality additional extensions are created when compared with conclusion maximality. More formally: the extensions of conclusion maximization can be seen as a subset of the extensions of rule maximization. The consequence of this is that all conclusions that follow from rule-maximization, also follow from conclusion maximization.

Lemma 3.6. *For each conclusion-maximal set of conclusions C_{cmax}, there exists a rule-maximal set of rules R_{rmax} with $C_{cmax} = \{c \mid c \text{ follows from } R_{rmax}\}$.*

Proof. Let C_{cmax} be an arbitrary conclusion-maximal set of conclusions. Then, according to theorem 3.1 (from C_{cmax} to R_{cmax}) there exists a R_{cmax} with $conclusions(R_{cmax}) = C_{cmax}$ This R_{cmax} is coherent; no incoherent arguments can be constructed with it. This also means that R_{cmax} can be extended with extra rules in order to become an rule-maximal set of rules R_{rmax} (more formally: there is an R_{rmax} with $R_{cmax} \subseteq R_{rmax}$). Now take an arbitrary R_{rmax} such that $R_{cmax} \subseteq R_{rmax}$. It now holds that if c follows from R_{rmax}, then c follows from R_{cmax}.

Proof. (reductio ad abdurdum) Suppose c follows from R_{rmax}, but c does not follow from R_{cmax}. Then there is an argument B with conclusion c that can be constructed from R_{rmax} but not from R_{cmax} (that is: $defeasible(B) \subseteq R_{rmax}$ but $defeasible(B) \not\subseteq R_{cmax}$). The fact that c does not follow from R_{cmax} means that the rules necessary to derive c are not included in R_{cmax}. But because there is an argument for c it would be possible to "add" these rules to the maximal argument (say A) of R_{cmax}. This would result in a new argument (say C). But because R_{cmax} was already a conclusion-maximal set of rules, it follows that C is incoherent! But because all rules in C are also in R_{rmax} it would also hold that R_{rmax} is incoherent. Contradiction. □

□

Theorem 3.9 (From rule maximality to conclusion maximality).
If $c \in \cap_{i=0}^{n}$ conclusions of R_{rmax_i} then $c \in \cap_{i=0}^{n} C_{cmax_i}$.

Proof. (by modus tollens) Suppose $c \notin \cap_{i=0}^{n} C_{cmax_i}$. Then there is a C_{cmax} without conclusion c. From the lemma above it follows that there is a rule-maximal set of rules R_{rmax} without conclusion c. Thus $c \notin \cap_{i=0}^{n}$conclusions of R_{rmax_i}. □

From DS_{HY} to $DS_{classic}$

Theorem 3.10 (From DS_{HY} to $DS_{classic}$). *If c is a justified conclusion under DS_{HY}, then c is a justified conclusion under $DS_{classic}$.*

Proof. (by modus tollens) Suppose c is not a justified conclusion under $DS_{classic}$. This can mean two things:

1. There is no argument A with conclusion c at all. Then c is also not a justified conclusion under DS_{HY}.

2. There is an argument A with conclusion c, but every argument $(A_1, A_2, \ldots A_n)$ with conclusion c has a coherent classical counterargument $(B_1, B_2, \ldots B_n)$. But then there also exist coherent HY-arguments $(C_1, C_2, \ldots C_n)$ against $(A_1, A_2, \ldots A_n)$ where each C_i consists of B_i with a foreign commitment added to the conclusion it attacks. Thus, c is not justified under DS_{HY}.

□

Overview

The relationship between conclusion maximality, rule maximality, $DS_{classic}$ and DS_{HY} can be summarized as follows:

- if a conclusion is justified under rule maximality, then it is justified under conclusion maximality (theorem 3.9);
- if a conclusion is justified in $DS_{classic}$, then it is justified under conclusion maximality (theorem 3.7);
- if a conclusion is justified in DS_{HY}, then it is justified under rule maximality (theorem 3.8);
- if a conclusion is justified in DS_{HY}, then it is justified in $DS_{classic}$ (theorem 3.10).

Of each of the above four properties, the converse does not hold. This can be made clear using the following counterexample:

$$\begin{array}{rcl} \mathcal{S} & = & \{\rightarrow a\} \\ \mathcal{D} & = & \{a \Rightarrow b,\ a \Rightarrow \neg b,\ a \Rightarrow d,\ b \Rightarrow c, \neg b \Rightarrow c,\ d \Rightarrow \neg a\} \end{array}$$

graphical representation:

$$\begin{array}{rrrrr} \rightarrow & a & \Rightarrow & d & \Rightarrow\ \neg a \\ \rightarrow & a & \Rightarrow & b & \Rightarrow\ c \\ \rightarrow & a & \Rightarrow & \neg b & \Rightarrow\ c \end{array}$$

conclusion-maximal sets of conclusions:

$$\begin{array}{llllll} \{ & a, & b, & c, & d & \} \\ \{ & a, & \neg b, & c, & d & \} \end{array}$$

justified conclusions under conclusion maximality: $\{a, c, d\}$

rule-maximal sets of conclusions:

$$\begin{array}{llllll} \{ & a, & \neg b, & c & & \} \\ \{ & a, & b, & c & & \} \\ \{ & a, & d, & \neg b, & c & \} \\ \{ & a, & d, & b, & c & \} \end{array}$$

justified conclusions under rule maximality: $\{a, c\}$

justified conclusions in $DS_{classic}$: $\{a, d\}$
justified conclusions in DS_{HY}: $\{a\}$

The overall relationships between the four aforementioned systems are summarized in the Hasse-diagram of figure 3.2.

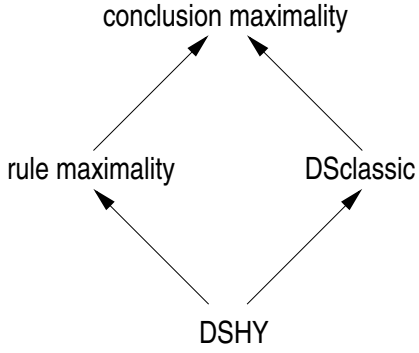


Figure 3.2: The relationship between conclusion maximality, rule maximality, $DS_{classic}$ and DS_{HY}.

3.3.3 On the replacement of strict rules

In this subsection, we explain that in our HY-enriched dialogue system, there is no longer a need for two different types of rules: strict rules and defeasible rules. Instead, strict rules can be modeled as defeasible rules (with a priority higher than any original defeasible rule) without any change of the conclusions that are considered justified. To put it in other words: if we replace the strict rules by high prioritized defeasible rules, we still derive the same outcome. This property holds for DS_{HY} in general, and not just for simple defeasible theories. The property also depends on the existence of HY-arguments; it does not hold in $DS_{classic}$. [14]

First, we define a function f that replaces strict rules by defeasible rules.

Definition 3.34. *Let f be a function ($f : rules \longrightarrow rules$) such that:*

- $f(r) = r$ *(where r is a defeasible rule)*
- $f(\rightsquigarrow L) = \rightsquigarrow L$
- $f(L_1 \wedge L_2 \wedge \ldots \wedge L_{n-1} \rightarrow L_n) = L_1 \wedge L_2 \wedge \ldots \wedge L_{n-1} \Rightarrow L_n$

We will sometimes abuse notation and write $f(A)$, where A is an argument $[r_1, r_2, \ldots, r_n]$, meaning $[f(r_1), f(r_2), \ldots, f(r_n)]$. We also write $f(R)$, where R is a set of rules $\{r_1, r_2, \ldots, r_n\}$ meaning $\{f(r_1), f(r_2), \ldots, f(r_n)\}$.

The next step is then to define the mapping of strict rules to defeasible rules for an entire defeasible theory. Before doing so, it is important to make sure that none of the strict rules is mapped to a rule that is already in $\mathcal{D}$.[15] This requires the defeasible theory in question to be rule-disjunct.

Definition 3.35. *A defeasible theory $(\mathcal{S}, \mathcal{D}, <)$ is called* rule-disjunct *iff there are no two rules $R_s \in \mathcal{S}$ and $r_d \in \mathcal{D}$ that have syntactically the same antecedent and consequent.*

[14] To see why, consider $\mathcal{S} = \{\rightarrow a,\ b \rightarrow c,\ \rightarrow \neg c\}$, $\mathcal{D} = \{a \Rightarrow b\}$ and $<= \emptyset$. Here, $DS_{classic}$ entails $\{a\}$. If the strict rules in $\mathcal{S}$ are replaced by defeasible rules, however, b is also entailed, no matter how high the priorities of the former strict rules are taken.

[15] There are various reasons why this should be avoided. The lack of rule-disjunction makes, for instance, that the function F would not be 1-1.

The requirement that a defeasible theory needs to be rule-disjunct can easily be met by removing every defeasible rule from $\mathcal{D}$ that violates rule-disjunction. The resulting defeasible theory has the same entailment (that is, the same set of justified conclusions) as the original one.

The following definition introduces a function that, given a defeasible theory $(\mathcal{S}, \mathcal{D}, <)$, produces a defeasible theory $(\mathcal{S}', \mathcal{D}', <')$, where $\mathcal{S}'$ is empty, $\mathcal{D}'$ consists of $\mathcal{D}$ plus some additional defeasible rules (the "mapped" strict rules in $\mathcal{S}$) and $<'$ is an enhanced version of $<$ that also works on the additional defeasible rules in $\mathcal{D}'$ (it assigns the "mapped" strict rules in $\mathcal{S}$ a maximal priority, higher than any rule originally in $\mathcal{D}$).

Definition 3.36. *Let $(\mathcal{S}, \mathcal{D} <)$ be a rule-disjunct defeasible theory. We define a function F from defeasible theory to defeasible theory as follows: $F(\mathcal{S}, \mathcal{D}, <) := (\emptyset,\ \mathcal{D} \cup \{f(s) \mid s \in \mathcal{S}\}, < \cup \{(d, f(s)) \mid d \in \mathcal{D}, s \in \mathcal{S}\})$*

All notions of attack are preserved by the function F. This can be proved using the following lemma.

Lemma 3.7. *Let A be an argument. It holds that:*

- *A has a conclusion c iff $f(A)$ has a conclusion c*
- *A has an assumption L iff $f(A)$ has an assumption L*
- *A's conclusion c is fc-based iff $f(A)$'s conclusion c is fc-based*

Proof. The mapping function f preserves the conclusions and assumptions of each argument. Furthermore, because the structure of an argument is preserved by f as well, the notion of fc-based is also preserved. □

Lemma 3.8. *Let A_1 and A_2 be two arguments in the rule-disjunct defeasible theory $(\mathcal{S}, \mathcal{D}, <)$ under DS_{HY}.*

1. *A_2 classically rebut-attacks A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ classically rebut-attacks $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*
2. *A_2 classically undercut-attacks A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ classically undercut-attacks $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*
3. *A_2 HY-rebut-attacks A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ HY-rebut-attacks $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*
4. *A_2 HY-undercut-attacks A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ HY-undercut-attacks $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*
5. *A_2 reverse HY-rebut-attacks A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ reverse HY-rebut-attacks $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*

Proof.

1. This follows from definition 3.17 (1) and lemma 3.7.
2. This follows from definition 3.17 (2) and lemma 3.7.
3. This follows from definition 3.18 (1) and lemma 3.7.

4. This follows from definition 3.18 (2) and lemma 3.7.

5. This follows directly from point 3 and definition 3.19.

□

Another feature of the mapping function is that it preserves the priority ordering among sets of rules. This is stated in the following lemma.

Lemma 3.9. *Let $(\mathcal{S}, \mathcal{D}, <)$ be a defeasible theory, $(\mathcal{S}', \mathcal{D}', <') = F(\mathcal{S}, \mathcal{D}, <)$ and $R_1 \subseteq \mathcal{S} \cup \mathcal{D}$, $R_2 \subseteq \mathcal{S} \cup \mathcal{D}$ sets of rules such that R_1 or R_2 contains at least one defeasible rule. It holds that $R_1 < R_2$ (using the priority ordering of $(\mathcal{S}, \mathcal{D}, <)$) iff $f(R_1) <' f(R_2)$ (using the priority ordering of $F(\mathcal{S}, \mathcal{D}, <)$).*

Proof.
"$\Longrightarrow$":
Suppose $R_1 < R_2$ (using the priority ordering of $(\mathcal{S}, \mathcal{D}, <)$). Then (definition 3.20) it holds that: $\exists r_1 \in \mathit{defeasible}(R_1) \forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$.

Let r_1 be a defeasible rule of R_1 such that $\forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$. Then it also holds that $r_1 \in \mathit{defeasible}(f(R_1))$ (because function f leaves defeasible rules unchanged).

Furthermore, let r_2' be an arbitrary defeasible rule of $f(R_2)$. Then there are two possibilities for this r_2':

1. $r_2' = r_2$ for some defeasible rule $r_2 \in R_2$. Then, it holds that $r_1 < r_2$ (this follows from (i)), and because $<'$ is an extension of $<$, it also holds that $r_1 <' r_2'$ in $F(\mathcal{S}, \mathcal{D}, <)$

2. r_2' is the result of a strict rule $r_2 \in R_2$. Then, according to the definition of F, we also have $r_1 <' f(r_2)$ (because applying F gives the former strict rules a priority that is higher than that of the former defeasible rules). Thus, we have $r_1 <' r_2'$

So, we have $\forall r_2' \in \mathit{defeasible}(f(R_2)) : r_1 <' r_2'$. And because $r_1 \in \mathit{defeasible}(R_1)$ it also holds that $r_1 \in \mathit{defeasible}(f(R_1))$. Thus, we have $\exists r_1' \in \mathit{defeasible}(f(R_1)) \forall r_2' \in \mathit{defeasible}(f(R_2)) : r_1' < r_2'$. Thus, we have $R_1' < R_2'$.

"$\Longleftarrow$":
Suppose $f(R_1) <' f(R_2)$ (using the priority ordering of $F(\mathcal{S}, \mathcal{D}, <)$. Then (definition 3.20) it holds that: $\exists r_1' \in \mathit{defeasible}(f(R_1)) \forall r_2' \in \mathit{defeasible}(f(R_2)) : r_1' <' r_2'$. Let r_1' be a defeasible rule in $f(R_1)$ such that $\forall r_2' \in \mathit{defeasible}(f(R_2)) : r_1' <' r_2'$. Because $\mathit{defeasible}(R_2) \subseteq \mathit{defeasible}(f(R_2))$ it also holds that $\forall r_2 \in \mathit{defeasible}(R_2) : r_1' <' r_2$. Because r_2 is not a "degenerated strict rule", it also holds that $\forall r_2 \in \mathit{defeasible}(R_2) : r_1' < r_2$. Therefore, we have:

$$\text{(i)} \quad \exists r_1' \in \mathit{defeasible}(f(R_1)) \forall r_2 \in \mathit{defeasible}(R_2) : r_1' < r_2$$

Now, our next thing to prove is that (regarding the above formula) there is an r_1' that is not only an element of $\mathit{defeasible}(f(R_1))$, but also an element of $\mathit{defeasible}(R_1)$ an element of R_1. We prove this *reductio ad absurdum*. Suppose: $\neg \exists r_1 \in \mathit{defeasible}(R_1) \forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$ This, together with (i) means that $r_1' \in \mathit{defeasible}(f(R_1))$ has to be the result of a strict rule in R_1. This, however, is a contradiction if there is at least one defeasible rule (r_2) in R_2 (then $r_1' < r_2$ because of (i) but also $r_2 <' r_1'$ because of the fact

that "degenerated strict rules" get a priority higher than any other rules). This means that there are no defeasible rules in R_2. Then, according to the premises of our lemma (either R_1 or R_2 contains at least one defeasible rule), there has to be at least one defeasible rule in R_1 (say: r_1). But, as there are no defeasible rules in R_2, it automatically holds that $\forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$. Thus $\exists r_1 \in \mathit{defeasible}(R_1) \forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$. Contradiction.

Therefore, we have $\exists r_1 \in \mathit{defeasible}(R_1) \forall r_2 \in \mathit{defeasible}(R_2) : r_1 < r_2$ and therefore $R_1 < R_2$. □

Because the mapping function preserves the notion of attack, as well as the priority among sets of rules (and therefore also the priority among arguments), it also preserves the notion of defeat.

Lemma 3.10. *Let $(\mathcal{S}, \mathcal{D}, <)$ be a rule-disjunct defeasible theory under DS_{HY}. A_2 defeats A_1 in $(\mathcal{S}, \mathcal{D}, <)$ iff $f(A_2)$ defeats $f(A_1)$ in $F(\mathcal{S}, \mathcal{D}, <)$.*

Proof. This follows from the lemma 3.8, lemma 3.9 and definition 3.21. □

Theorem 3.11. *Let $(\mathcal{S}, \mathcal{D}, <)$ be a rule-disjunct defeasible theory. Argument A is justified in $(\mathcal{S}, \mathcal{D}, <)$ iff argument $f(A)$ is justified in $F(\mathcal{S}, \mathcal{D}, <)$.*

Proof. The fact that $(\mathcal{S}, \mathcal{D}, <)$ is a rule-disjunct defeasible theory means that f is a bijective function from arguments in $(\mathcal{S}, \mathcal{D}, <)$ to arguments in $F(\mathcal{S}, \mathcal{D}, <)$. From lemma 3.10 it follows that the defeat relationship is preserved among the mapped arguments. So actually, what the function F does is that it produces a defeasible theory of which the Dung-style argumentation framework is isomorf with the argumentation framework of the original $(\mathcal{S}, \mathcal{D}, <)$. This means that, regardless of the particular semantics being applied (in this case: grounded semantics), argument A is justified in $(\mathcal{S}, \mathcal{D}, <)$ iff argument $f(A)$ is justified in $F(\mathcal{S}, \mathcal{D}, <)$. □

In essence, the point is that with HY-arguments there is no need for the distinction between strict and defeasible rules, as strict rules can be replaced by defeasible rules. This is quite an exceptional feature; usually, when one enriches a certain logic with new concepts, one ends up with more, and not less, syntactical constructs. With HY-arguments, the situation is the other way around; in the HY-enriched formalism all that is needed is just one type of rule.

Chapter 4

On the application of HY-arguments

In this chapter, two main issues are dealt with. First, in section 4.1, some differences and similarities between HY and the principle of contraposition are given. Then, in section 4.2, the question is studied whether or not contraposition and HY are desirable principles to be implemented in formalisms for defeasible reasoning.

4.1 HY and contraposition

An interesting question is to which extent the effect of adding HY-arguments can be compared with the effect of adding contraposition. In this section, some technical differences between contraposition and HY are given and it is discussed what the added value of HY-arguments is compared to arguments based on contraposition.

Technical differences

In this section, we show several examples to illustrate the similarities and differences between these principles. Notice that examples 1 to 4 concern simple argumentation structures, so an argument is justified iff it does not have a coherent counterargument.

example 1

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A,\ \rightarrow \neg C\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C\} \\ < & = & \emptyset \end{array}$$

In P&S's original system, there are justified arguments for A ($\rightarrow A$) and B ($\rightarrow A,\ A \Rightarrow B$). The (only) argument for C ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$) is defeated by the strict argument $\rightarrow \neg C$ so C is not justified.

If we look at a system with HY-arguments, only A and $\neg C$ are justified, B is not. The reason is that the argument for B ($\rightarrow A,\ A \rightarrow B$) now has a HY-counterargument $\rightsquigarrow B,\ B \Rightarrow C,\ \rightarrow \neg C$.

Suppose we have a system without HY-arguments, but with contraposition. Contraposition, essentially means that whenever we have a rule $A \Rightarrow B$ (or $A \rightarrow B$) we may also use it in the contraposed way of $\neg B \Rightarrow \neg A$ (or $\neg B \rightarrow \neg A$). This means that the number of usable rules can (at most) be doubled. If we allow contraposition in example 1, we thus

obtain the following effective rule-bases:

$$\begin{array}{lcl} \mathcal{S}' & = & \{\mathit{true} \rightarrow A,\ \mathit{true} \rightarrow \neg C, \neg A \rightarrow \mathit{false},\ \neg C \rightarrow \mathit{false}\} \\ \mathcal{D}' & = & \{A \Rightarrow B,\ B \Rightarrow C,\ \neg B \Rightarrow \neg A,\ \neg C \Rightarrow \neg B\} \\ <' & = & \emptyset \end{array}$$

If we apply P&S's original system to this extended rule-bases, we obtain justified conclusions A and $\neg C$, and nothing else. B is not justified because there is now a coherent (non-HY) argument ($\rightarrow \neg C,\ \neg C \Rightarrow \neg B$) against B.

The results, as far as justified conclusions are concerned, of example 1 can therefore be summarized as follows:

- P&S: $\{A, B, \neg C\}$
- HY: $\{A, \neg C\}$
- contrapos: $\{A, \neg C\}$

In example 1, we see that the effect of adding HY-arguments is the same as the effect of adding contraposition. The question is whether this is always the case.

example 2

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow \neg C\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C\} \\ < & = & \emptyset \end{array}$$

Here, P&S's original system allows us to derive nothing but justified conclusion $\neg C$ (using argument $\rightarrow \neg C$), as there are simply no arguments for any other conclusion. If we allow HY-arguments, then still no other conclusions than $\neg C$ can be derived; as there are still no arguments for anything else. If, on the other hand, we allow contraposition, then we can also derive $\neg B$ ($\rightarrow \neg C,\ \neg C \Rightarrow \neg B$) and $\neg A$ ($\rightarrow \neg C,\ \neg C \Rightarrow \neg B,\ \neg B \Rightarrow \neg A$), and since these arguments do not have any counterarguments, both of them are justified.

The results of example 2 can therefore be summarized as follows:

- P&S: $\{\neg C\}$
- HY: $\{\neg C\}$
- contrapos: $\{\neg C, \neg B, \neg A\}$

Example 2 makes clear that the outcome of a system with HY arguments can be different from the outcome of a system with contraposition. This is not surprising, since HY-arguments are not able to generate new conclusions. They can merely cast doubt on other conclusions. HY-arguments share the property of Socrates's elenchus in that they are destructive instead of constructive.

example 3

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A,\ \rightarrow \neg D\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ \neg C \Rightarrow \neg B,\ C \Rightarrow D\} \\ < & = & \emptyset \end{array}$$

Here, the original system of P&S entails only justified conclusions A ($\rightarrow A$), B ($\rightarrow A,\ A \Rightarrow B$) and $\neg D$ ($\rightarrow \neg D$). A system with HY-arguments entails the same justified conclusions, since none of these conclusions have a HY-counterargument. A system with contraposition, on the other hand, only entails justified conclusions A and $\neg D$, since B now has a coherent counterargument $\rightarrow \neg D,\ \neg D \Rightarrow \neg C,\ \neg C \Rightarrow \neg B$.

The results of example 3 can therefore be summarized as follows:

- P&S: $\{A, B, \neg D\}$
- HY: $\{A, B, \neg D\}$
- contrapos: $\{A, \neg D\}$

Example 3 makes clear that the justified conclusions of a system with contraposition can also be *less* than the justified conclusions of a system with HY-arguments. The point is that in this case no (non-trivial) HY-argument can be constructed. The outcome $\{A, B, \neg D\}$ of the system with HY-arguments can be seen from the perspective of rule-maximality. Since $\mathcal{D}$ by itself is consistent (no contradictions can be derived) the rule-maximal set of defeasible rules is equal to $\mathcal{D}$. Therefore, everything that follows from $\mathcal{D}$ is justified.

example 4

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C,\ C \Rightarrow D,\ D \Rightarrow \neg B\} \\ < & = & \emptyset \end{array}$$

Here, the original system of P&S entails A, B, C and D. If we allow HY-arguments, however, then only A remains justified. In order to see why this is the case, take for instance the argument for D: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C,\ C \Rightarrow D$. It now has a HY counterargument $\rightsquigarrow D,\ D \Rightarrow \neg B,\ \rightsquigarrow B$. HY-counterarguments against B and C are also available, so only A remains justified.

Contraposition allows for the justified conclusions A and B (but not C or D). This can be seen as follows. Although there is an argument for C ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$). There is also a coherent counterargument ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow \neg D,\ \neg D \Rightarrow \neg C$) so C is not justified. For a similar reason, D is also not justified. B on the other hand *is* justified; it has an argument $\rightarrow A,\ A \Rightarrow B$ that has no coherent counterargument since $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow \neg D,\ \neg D \Rightarrow \neg C,\ \neg C \Rightarrow \neg B$ is incoherent!

The results of example 4 can therefore be summarized as follows:

- P&S: $\{A, B, C, D\}$
- HY: $\{A\}$
- contrapos: $\{A, B\}$

At first sight, one may wonder whether the kind of extreme scepticism that is shown by the HY-system can be warranted based on intuitive grounds. To answer this question, it may be interesting to look at a somewhat different example (let's call it example 4'):

$$\begin{array}{lcl} \mathcal{S}' & = & \{\rightarrow A,\ B \rightarrow C,\ C \rightarrow D,\ D \rightarrow \neg B\} \\ \mathcal{D}' & = & \{A \Rightarrow B\} \end{array}$$

Here, the defeasible rules $B \Rightarrow C$, $C \Rightarrow D$ and $D \Rightarrow \neg B$ have been replaced by strict rules. So now, any argument that entails B is incoherent (as from B it follows that $\neg B$). Hence, B should not be justified.

Now suppose the rules $B \Rightarrow C$, $C \Rightarrow D$ and $D \Rightarrow \neg B$ are not replaced by strict rules, but by rules that have an (extremely) high priority, at least (much) higher than $A \Rightarrow B$. In that case, the reason for believing that from B a contradiction follows is much stronger than the reason for believing B in the first place. Hence, it seems appropriate not to have B as a justified conclusion. This is in line with the principle that strict rules can be seen as defeasible rules with a very high priority (section 3.3.3).

Now suppose the rules $B \Rightarrow C$, $C \Rightarrow D$ and $D \Rightarrow \neg B$ have the same priority as (or are incomparable to) $A \Rightarrow B$. Then, one cannot determine what is stronger: the reason that led to B or the absurdity that results from it. It is this situation that takes place in example 4, and since priorities cannot resolve it, the conclusion B is not justified. It is, however, interesting to notice that if $B \Rightarrow C$, $C \Rightarrow D$ and $D \Rightarrow \neg B$ have a priority that is *lower* than $A \Rightarrow B$ then B *does* become justified.

As an aside, in section 3.3.3 it was shown that in a system with HY-arguments, strict rules can be replaced by defeasible rules with a very high priority. It is interesting to ask ourselves whether this property also holds in systems that support contraposition instead of HY. In other words, can we — given a system for defeasible reasoning that sanctions contraposition but not HY — replace the strict rules by defeasible rules with a very high priority while still having the same set of conclusions?

The answer to this question is "no". Our counterexample is example 4'. Here conclusion B is not justified. If we would replace the strict rules $B \rightarrow C$, $C \rightarrow D$ and $D \rightarrow \neg B$ by defeasible rules $B \Rightarrow C$, $C \Rightarrow D$ and $D \Rightarrow \neg B$ with a very high priority, then a system with only contraposition would allow us to derive justified conclusion B. Hence, the set of conclusions changes after the transformation takes place. We can therefore conclude that in a system with contraposition but without HY-arguments, the strict rules cannot simply be replaced in the way they could be in a system with HY-arguments.

example 5

$$
\begin{array}{lcll}
\mathcal{S} & = & \{\rightarrow A\} & \\
\mathcal{D} & = & \{A \Rightarrow B, & (1) \\
 & & B \Rightarrow C, & (2) \\
 & & A \Rightarrow \neg C & (3)\} \\
< & = & r_1 < r_2 \text{ iff } r_2 \text{ has a higher number than } r_1 &
\end{array}
$$

Example 5 is interesting because it involves the use of priorities. In P&S original system, there are three justified conclusions: A ($\rightarrow A$), B ($\rightarrow A,\ A \Rightarrow B$) and $\neg C$ ($\rightarrow A,\ A \Rightarrow \neg C$). There is also an argument for C ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$) but this is defeated by the (stronger) argument for $\neg C$ ($\rightarrow A,\ A \Rightarrow \neg C$).

If we allow contraposition, and extend the priority ordering among the contraposed rules, then this results in the following:

$$
\begin{array}{lcl}
\mathcal{S}' & = & \{true \rightarrow A, \\
 & & \neg A \rightarrow false\}
\end{array}
$$

$$
\begin{array}{lcll}
\mathcal{D}' & = & \{A \Rightarrow B, & (1)\\
 & & \neg B \Rightarrow \neg A, & (1)\\
 & & B \Rightarrow C, & (2)\\
 & & \neg C \Rightarrow \neg B, & (2)\\
 & & A \Rightarrow \neg C, & (3)\\
 & & C \Rightarrow \neg A & (3)\}\\
< & = & r_1 < r_2 \text{ iff } r_2 \text{ has a higher number than } r_1 &
\end{array}
$$

With contraposition, there are three justified conclusions: A ($\rightarrow A$), $\neg C$ ($\rightarrow A,\ A \Rightarrow \neg C$) and $\neg B$ ($\rightarrow A,\ A \Rightarrow \neg C,\ \neg C \Rightarrow \neg B$). Counterarguments against $\neg C$ and $\neg B$ exist ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$ and $\rightarrow A,\ A \Rightarrow B$) but these are weaker than the arguments for $\neg C$ and $\neg B$, so $\neg C$ and $\neg B$ remain justified.

In a system with HY-arguments, only A and $\neg C$ are justified. In particular, B is not derived, since the argument for B ($\rightarrow A,\ A \Rightarrow B$) is defeated by the (stronger) HY-argument ($\rightarrow A,\ A \Rightarrow \neg C,\ \rightsquigarrow B,\ B \Rightarrow C$) that has itself no strict defeaters.

The results of example 5 can be summarized as follows:

- P&S: $\{A, B, \neg C\}$
- HY: $\{A, \neg C\}$
- contrapos: $\{A, \neg C, \neg B\}$

We see that the P&S system derives B, whereas the system with contraposition derives $\neg B$. The system with HY-arguments, on the other hand, takes a more cautious approach and derives neither B nor $\neg B$.

Altogether, we can say that contraposition and HY are two different principles, that in general produce different results. Some of the differences can be explained from the perspective of rule-maximality and the transformation of strict rules to defeasible rules. This transformation is a property of systems with HY-arguments and does not generally hold in systems with contraposition.

On the value of HY-arguments

So far, the discussion about HY and contraposition has had a merely technical nature, focussing on the differences and similarities on an abstract level. Another relevant issue, however, is the intuitive and conceptual added value of HY-arguments, compared to arguments based on contraposition. Overall, one can distinguish three main reasons justifying the independent existence of HY-arguments:

1. Contraposition alone does not offer a solution for situations that require HY-undercutting. Take for instance the Ajax-Feijenoord example:
   $\mathcal{S} = \{\rightarrow af\}$
   $\mathcal{D} = \{af \wedge \sim p \Rightarrow t,\ t \Rightarrow p\}$
   With contraposition alone (without HY-arguments) there exists no counterargument against $\rightarrow af,\ af \wedge \sim p \Rightarrow t$, so t becomes justified.[1]

[1] A similar observation can be made about Pollock's pink elephant example, which is to be discussed in section 5.3.2.

2. Even in situations where undercutting does not play a role, contraposition by itself does not necessarily provide the desired outcome. Take for instance the tax relief example:[2]
   $\mathcal{S} = \{\rightarrow pmp\}$
   $\mathcal{D} = \{pmp \Rightarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr\}$
   With contraposition (but without HY) there exist exactly two arguments against tr:
   $\rightarrow pmp,\ pmp \Rightarrow tr,\ tr \Rightarrow bd,\ bd \Rightarrow fb,\ fb \Rightarrow \neg tr$
   $\rightarrow pmp,\ pmp \Rightarrow tr,\ tr \Rightarrow \neg fb,\ \neg fb \Rightarrow \neg bd,\ \neg bd \Rightarrow \neg tr$
   Both of these arguments are self-defeating, and are therefore ultimately defeated by the empty argument. Of course, it would be possible to disable the special working of the empty argument so that the above two arguments are still in force against tr, but doing so involves various other problems (see section 5.3.2).

3. One can say that the concept of HY is in a certain sense closer to how people intuitively experience reasoning, than the concept of contraposition. The author of this thesis has assisted teaching an introductory logic course for several years. One of my observations is that there is a difference between how the students regarded modus monens, and how they regarded modus tollens and contraposition. Although modus tollens and contraposition raraly gave the students any real trouble, they did have to think about *why* these principles are valid. Modus ponens, at the other hand, seemed so natural that one often had to remind the students to explicitly state when they were applying it (they were sometimes not even aware of it). HY-style reasoning, in essence, only requires the application of (defeasible) modus ponens, as well as the notion of a logical conflict. The fact that informal HY-style reasoning has been known since antiquity and is still in use today means that many people are familiar with it. Even in situations where contraposition and HY yield the same outcome, it can have certain advantages to use HY to explain the outcome.

4.2 Application domains

In this section, the applicability of HY and contraposition is studied from the perspective of two application domains: statistical (epistemical) reasoning and constitutive reasoning.

4.2.1 Contraposition and HY-arguments under statistical interpretation

In section 4.1, it was explained that contraposition and HY-arguments do not generally have the same effects, although similarities occur. An interesting question is whether the principle of contraposition should be regarded correct for defeasible reasoning, and if so, under what circumstances? This is relevant, as the question of validity of contraposition is related to validity of HY-arguments.

Although the issue of applicability of contraposition seems to be a fundamental one, the issue has until now received relatively little attention. Many authors treat the validity or invalidity of contraposition as an aside when describing their respective formalisms for

[2]Apart from the syntactical sugar, the tax relief example is essentially the same as example 4 given earlier.

non-monotonic reasoning. Our analysis will therefore begin with an overview of some of the comments by various authors in the field.

One of the proponents of the validity of contraposition for defeasible reasoning is Pollock [Poll95]. Pollock states the principle of contraposition (or *modus tollens*) as follows [Poll95, p. 67]:

> If G is projectible with respect to F and $r > 0.5$, then $\lceil \neg Gc \ \& \ prob(G/F) \geq r \rceil$ is a prima facie reason for $\lceil \neg Fc \rceil$, the strength of the reason depending upon the value of r.

To illustrate the validity of this rule, Pollock provides the example of carbon 14 analysis to determine the age of a certain item [Poll95, p. 67]:

> The exponential law of radioactive decay enables us to compute the probability, given that something is of a specified age, of the ratio of carbon 14 to carbon 13 lying in a certain envelope. When an anthropologist finds that the ratio of carbon 14 to carbon 13 in some artifact *does not* fall in the envelope, she concludes that it is not of the appropriate age.

Although the above example seems to be convincing, other authors do not agree with the validity of general contraposition or modus tollens. To illustrate invalidity, Brewka provides the following counterexample:

> Men usually do not have beards, but this does not mean that if someone does have a beard, it's usually not a man.

Other counterexamples are also available:

> If I buy a lottery ticket then I will normally not win any price, but this does not mean that if I *do* win a price, I did not buy a ticket.

Given the last two examples, it seems that there are perfectly legitimate situations in which contraposition does not hold, and that contraposition (or modus tollens) should therefore be rejected as a general principle for defeasible reasoning. This, however, is not all there is to say. An interesting analysis comes from Ginsberg [Gins94, page 16].

> Given a material implication $p \rightarrow q$ we can conclude $\neg q \rightarrow \neg p$. Given a rule allowing us to conclude q by default from p, should we be able to conclude $\neg p$ by default from $\neg q$? Should nonfliers, by default, not be birds?
>
> An analysis (...) gives us the answer yes, since it is legitimate to conclude (...) that:
>
> $$\neg f(x) \wedge \neg \mathtt{ab}(x) \rightarrow \neg b(x)$$
>
> Surprisingly, however, many workers in nonmonotonic reasoning would disagree.
>
> [A] reason is that the examples involving contraposition are somewhat more subtle (...). Consider the sentence, 'Humans tend not to be diabetics':
>
> $$h(x) \wedge \neg \mathtt{ab}(x) \rightarrow \neg d(x)$$

> Is it reasonable to conclude from this that diabetics tend not to be human?
>
> It appears not, although this may be somewhat shortsighted. Humans also tend not to have four legs, and it certainly *is* reasonable to conclude that most four-legged things are not people. What is happening is that we are using additional information about diabetes that is not present in [the above formula], namely that diabetes is, *in and on itself*, an abnormal condition:
>
> $$\neg \mathtt{ab}(x) \rightarrow \neg d(x)$$
>
> Given this, the contraposed default rule
>
> $$d(x) \wedge \neg \mathtt{ab}(x) \rightarrow \neg h(x)$$
>
> [has the status of being] valid but useless. So perhaps we should not be uncomfortable with contraposition after all.

Ginsberg's analysis shows that contraposition can be blocked by additional information being made explicit. Thus, even in a defeasible logic that sanctions contraposition, it is possible that this property is defeated in certain specific cases. Thus, to deal with contraposition, two approaches seem reasonable:

1. do not sanction contraposition as a general property of defeasible reasoning; instead let the user manually input the needed contraposed rules (an approach that is taken by for instance RDL)

2. sanction contraposition as a general property of defeasible reasoning; let the user manually input exceptions to the rules where contraposition is not applicable (this is the approach that is suggested by Ginsberg above)

The question then is which of the above approaches should be taken. To answer this question is not merely an issue of practical considerations (like take the approach that involves the least additional rules or exceptions) but also touches the essence of defeasible reasoning. It is therefore interesting to regard this question from a more fundamental point of view.

The question of whether or not contraposition should be allowed is also related to the applicability of HY-arguments. Consider the following example:

> "Sailors are usually men; most men do not have beards, but Captain Nemo does have a beard. Is Captain Nemo a man?"
> $\mathcal{S} = \{\rightarrow sailor(nemo),\ \rightarrow beard(nemo)\}$
> $\mathcal{D} = \{sailor(x) \Rightarrow man(x),\ man(x) \Rightarrow \neg beard(x)\}$
> $< = \emptyset$
> The argument-based dialogue would then go as follows:
> P: $\rightarrow sailor(nemo),\ sailor(nemo) \Rightarrow man(nemo)$
> O: $\rightsquigarrow man(nemo),\ man(nemo) \Rightarrow \neg beard(nemo)$

In general, one can turn several counterexamples against contraposition into counterexamples against HY. Therefore, the question whether or not contraposition can be regarded as a valid principle for defeasible reasoning is also relevant for our discussion of HY-arguments.

Unrestricted statistical interpretation

In order to determine whether contraposition and/or HY should be valid when reasoning with defaults, the first question that needs to be answered is "what is a default" or "how should one interpret a default?". These questions are important ones; in the current section, three different possible answers are provided, with an increasing level of sophistication.

The first possibility is the unrestricted statistical interpretation, under which a default "$A \Rightarrow B$" is interpreted as "most A's are B's". Notice that this interpretation allows for a traditional, model-based semantics. A default is true iff it holds in every model (probability distrbution) in which the premises are true, thus allowing for a classical notion of validity. The interpretation is monotonic because the addition of an extra premise can only result in *fewer* (or possibly the same) models satisfying the premises.

The example in which Brewka aims to illustrate that contraposition does not hold for default reasoning has the form $M \Rightarrow \neg B$ ("most men do not have beards"). For this rule, two probability distributions are provided (every dot is a separate world) in figure 4.1:

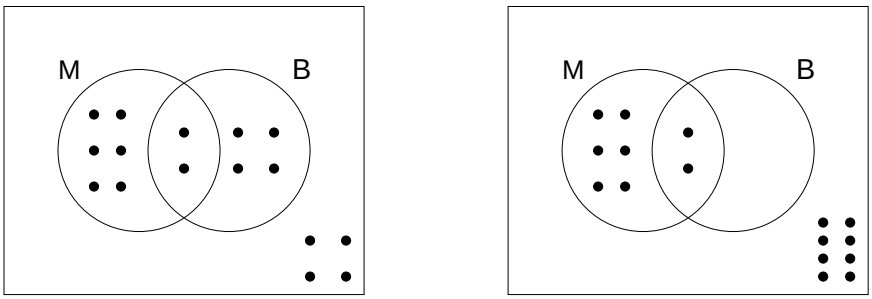


Figure 4.1: Men usually don't have beards

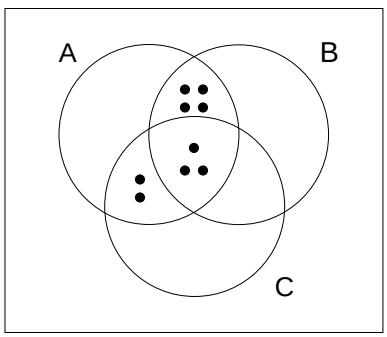
In the left-hand-side model of figure 4.1 the contraposed rule $B \Rightarrow \neg M$ is true, but in the right-hand-side model it is not. It should be noted that the right-hand-side model can be seen as a more appealing model for the thesis "most men have beards". This is because our background knowledge tells us that the ratio of men and women is about 50-50 and that no women have beards.

At first sight, this interpretation appears to confirm that contraposition should not be regarded as valid a valid principle. The problem, however, is that under the unrestricted statistical interpretation, many other reasonably sounding principles do not hold.

- cautious monotonity

  If $\mathcal{D} \vdash A \Rightarrow B$
  and $\mathcal{D} \vdash A \Rightarrow C$
  then $\mathcal{D} \vdash A \wedge B \Rightarrow C$.

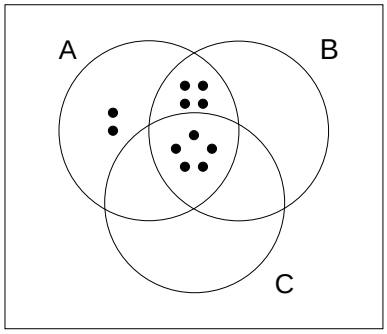
  Here it holds that most A's are B's,
  most A's are C's,
  but not most $A \wedge B$'s are C's.



- contraction

  If $\mathcal{D} \vdash A \Rightarrow B$
  and $\mathcal{D} \vdash A \wedge B \Rightarrow C$
  then $\mathcal{D} \vdash A \Rightarrow C$.

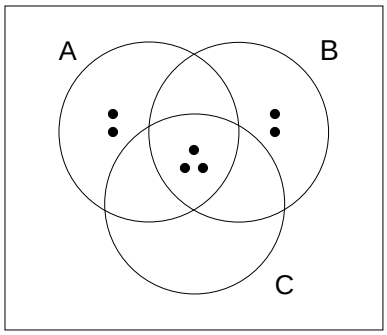
  Here it holds that most A's are B's,
  most $A \wedge B$'s are C's,
  but not most A's are C's.



- disjunction

  If $\mathcal{D} \vdash A \Rightarrow C$
  and $\mathcal{D} \vdash B \Rightarrow C$
  then $\mathcal{D} \vdash A \vee B \Rightarrow C$.

  Here it holds that most A's are C's,
  most B's are C's,
  but not most $A \vee B$'s are C's.



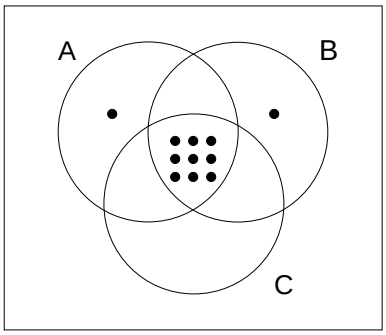
The bottom line is that under an unrestricted statistical interpretation of "$A \Rightarrow B$" as "most A's are B's", many inferences cannot be made.

ε-semantics

The fact that under an unrestricted statistical interpretation of defaults, principles like cautious monotonity, contraction and disjunction are not valid raises the question whether this interpretation is perhaps not strong enough. A possible way in which the statistical interpretation can be strengthened is proposed by ε-semantics [Adam75, Pear88]; here a default "$A \Rightarrow B$" is meant to be read as "nearly all A's are B's". The idea is that a conclusion is valid if it can be given a probability of $1-\varepsilon$ (that is: the conclusion is almost certain, apart from an arbitrary small uncertainty $\varepsilon > 0$) by interpreting the premises as almost certain. This is formalized in the following definition:

Definition 4.1 ([Pear92]). *Let $\mathcal{P}_{\mathcal{D},\varepsilon}$ stand for the set of distributions licensed by $\mathcal{D}$ for any given ε; that is:*
$\mathcal{P}_{\mathcal{D},\varepsilon} = \{P : P(v \mid u) \geq 1 - \varepsilon$ *and* $P(u) > 0$ *whenever* $u \Rightarrow v \in \mathcal{D}\}$
A conditional statement $p \Rightarrow q$ is said to be ε-entailed *by $\mathcal{D}$ if every distribution $P \in \mathcal{P}_{\mathcal{D},\varepsilon}$ satisfies $P(p \mid q) = 1 - O(\varepsilon)$ [ie, for every $\delta > 0$ there exists an $\varepsilon > 0$ such that every $P \in \mathcal{P}_{\mathcal{D},\varepsilon}$ would satisfy $P(q \mid p) \geq 1 - \delta$].*

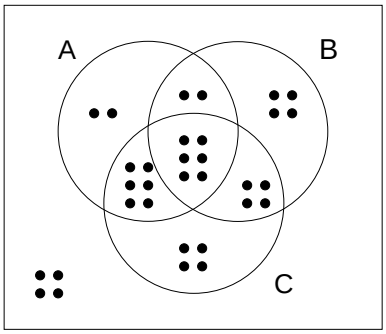
To see how ε-semantics works for the above three examples, consider the example of disjunction. In the counterexample against disjunction it holds that $P(C \mid A) = 0.6$, $P(C \mid B) = 0.6$ and $P(C \mid A \vee B) = 3/7 \approx 0.43$. It is interesting to see what happens if the probabilities $P(C \mid A)$ and $P(C \mid B)$ are increased, say to 0.9 (that is, we take ε to be 0.1), while still trying to minimize the probability $P(C \mid A \vee B)$. In the figure at the right hand side, it holds that $P(C \mid A) = 0.9$, $P(C \mid B) = 0.9$ and $P(C \mid A \vee B) = 9/11 \approx 0.72$

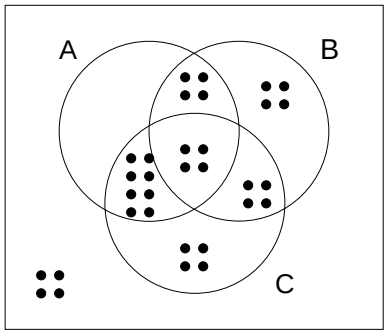


In order to minimize $P(C \mid A \vee B)$ it should be that case that $P(A \wedge \neg B \wedge C)$ and $P(\neg A \wedge B \wedge C)$ are as small as possible, while $P(A \wedge \neg B \wedge \neg C)$ and $P(\neg A \wedge B \wedge \neg C)$ are as big as possible. Under these constraints, the earlier given counterexample against disjunction is a probability distribution where $P(C \mid A) = P(C \mid B) = 0.6$ and $P(C \mid A \vee B)$ is as small as possible, and the example directly above is a possibility distribution where $P(C \mid A) = P(C \mid B) = 0.9$ and $P(C \mid A \vee B)$ is as small as possible. So we see that as $P(C \mid A)$ and $P(C \mid B)$ are increased, the smallest possible $P(C \mid A \vee B)$ is also increased. In a similar way it can also be illustrated that the properties cumulativity and contraction hold for ε-semantics. A formal proof of these properties is available [Adam75, Pear88].

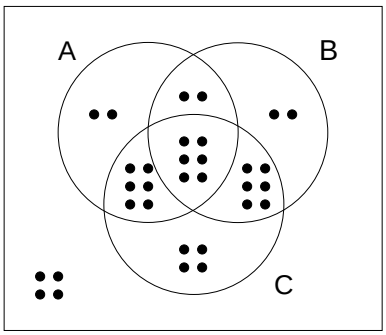
Unfortunately, not all plausibly sounding principles for defeasible reasoning are sanctioned by ε-semantics; that is, even under the interpretation of $A \Rightarrow B$ as "nearly all A's are B's" some properties do not hold. In the following discussion, four principles are given that are invalid under ε-semantics. For each principle, two probability distributions are provided, one in which the principle holds and one in which it does not.

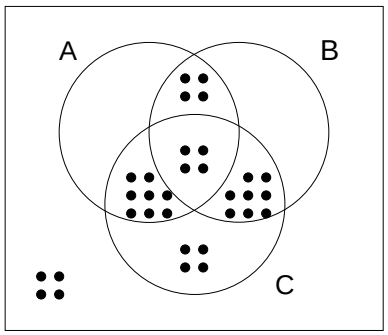
- irrelevance
  $\mathcal{D} = \{A \Rightarrow C\}$
  If $\mathcal{D} \vdash A \Rightarrow C$ then $\mathcal{D} \vdash A \wedge B \Rightarrow C$



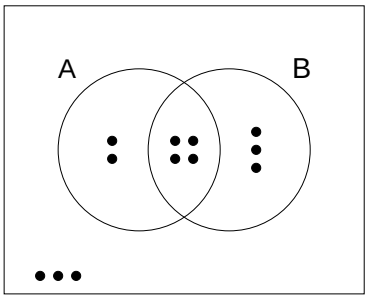


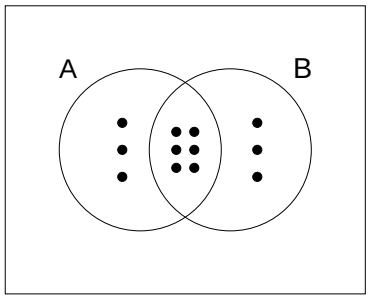
- left conjunction
  $\mathcal{D} = \{A \Rightarrow C,\ B \Rightarrow C\}$
  If $\mathcal{D} \vdash A \Rightarrow C$ and $\mathcal{D} \vdash B \Rightarrow C$ then $\mathcal{D} \vdash A \wedge B \Rightarrow C$



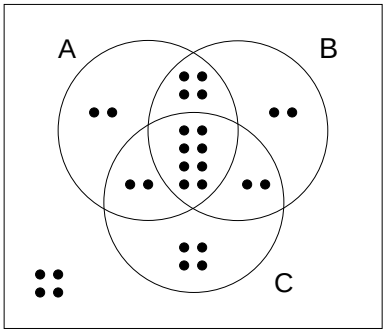


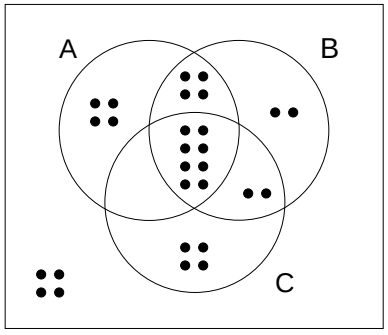
- contraposition
  $\mathcal{D} = \{A \Rightarrow B\}$
  If $\mathcal{D} \vdash A \Rightarrow B$ then $\mathcal{D} \vdash \neg B \Rightarrow \neg A$





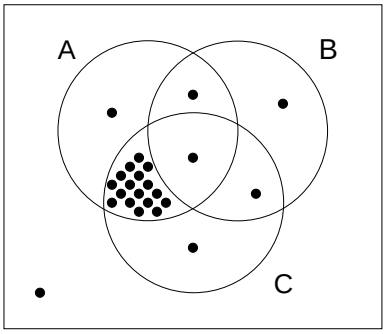
- transitivity
  $\mathcal{D} = \{A \Rightarrow B,\ B \Rightarrow C\}$
  If $\mathcal{D} \vdash A \Rightarrow B$ and $\mathcal{D} \vdash B \Rightarrow C$ then $\mathcal{D} \vdash A \Rightarrow C$





Whereas a property like disjunction can be sanctioned by interpreting the defaults $A \Rightarrow B$ as "nearly all A's are B's" (the approach taken by ε-semantics) this trick does not work for the properties irrelevance, left conjunction, contraposition or transitivity.

Take the example of irrelevance. Even when we take $P(C \mid A) = 0.9$, then there still exists a counterexample, as shown in the figure on the right-hand side.
It may be clear that no matter how high we take $P(C \mid A)$, as long as it is not exactly equal to 1, the property $A \wedge B \vdash C$ is not guaranteed to be true. As for left conjunction, contraposition and transitivity; these properties do not hold under ε-semantics for similar reasons.



For each of the properties irrelevance, left conjunction, contraposition and transitivity, intuitive counterexamples can be provided:

- irrelevance
  "Tux the bird":
  Birds fly and Tuxes are birds.[3] Do Tuxes fly? Perhaps not, because Tuxes may belong to a special subclass of birds that do not fly.

- left conjunction
  "jogging in the rain" [PrSa96]:
  If it is hot, I tend not to go out jogging. If it is raining I also tend not to go out jogging. Does this mean that if it is hot *and* it is raining, I tend not to go out jogging?
  "Marry both of them" [Pear92]
  If you marry Ann you will be happy, if you marry Nancy you will be happy as well. Does this mean you will be happy if you marry both of them?

- contraposition
  "men and beards": [Brew89]
  Men usually don't have beards. Does this mean that someone who has a beard is usually not a man?

- transitivity
  "unemployed students" [Pear88]
  Students are usually adults and adults are usually employed. Does this mean that students are usually employed?

One particular property of ε-semantics is that it is monotonic; conclusions are never withdrawn when new premises are introduced.

The maximum entropy approach

The above counterexamples against irrelevance, left conjunction, contraposition and transitivity look appealing at first sight. The point of each counterexample, however, is that it involves implicit background information. Tux does not fly because it is a penguin; marrying two persons generally does not make one happy (it makes one end up in jail instead); women have no beards at all; and students are a special class of adults that tend to be unemployed.

[3] Tux is the well-known penguin logo of the Linux-community.

Technically, all these counterexamples are valid. If birds generally fly and Tuxes are birds, than based on this information only, the fact that Tuxes fly may not be a valid inference, since it is unknown if Tuxes are a special subclass of birds or not. Yet, the general approach of default reasoning is to go ahead and make this inference anyway, under the assumption that everything is normal unless explicitly stated otherwise[4] So instead of considering every *possible* situation, one only considers the situations that are in some way considered as *normal*. It is this approach that is taken in for instance the semantics of circumscription; instead of all models, only the most normal models are taken into account. The idea of taking the most normal models is that situations like the above counterexamples are ruled out.

In circumscription, normality is usually implemented by minimizing the `ab`-predicates. An interesting question is how normality can be regarded from a statistical perspective. Can we regard the above counterexamples against irrelevance, left conjunction, contraposition and transitivity as "abnormal" models for their respective rule-bases? The first thing that strikes us is that while for every property the two probabilistic distributions assign the same chances to the defaults in $\mathcal{D}$, the right-hand side models (the counterexamples) appear to have a more unequal probability distribution than the left-hand side models (the ones that *do* comply with the properties in question). It is like in the left-hand side models the probabilities are more "spread out" while in the right-hand side models the probabilities are concentrated in a way that is not necessary to comply with the given defaults. A formalization of this observation is provided by the *maximum entropy approach* [Jayn79, PaVe90, PaVe97, Pari98], where entropy is defined as $-\Sigma_{w\in\mathcal{W}}P(w)\log P(w)$. Under this notion, the right-hand side models have an entropy lower than the left-hand side models, and should therefore not be taken into account as the most normal models.

The idea of extending ε-semantics with the notion of maximal entropy has been studied by Goldszmidt, Morris and Pearl [Gold93] and by Bourne [Bour99, BoPa99, BoPa00]. It is interesting to view the issue of probabilistic reasoning from their point of view. The first thing to notice is that although all defaults entailed by ε-semantics can be seen as acceptable, not every acceptable default is necessarily entailed, since some appealing forms of default reasoning (like transitivity) are missing. Bourne puts it as follows [Bour99, p. 38]:

> (...) despite its firm foundation in probability theory, the basic ε-semantics is clearly not sufficient to fully capture the kind of reasoning required of default systems. Being probabilistically sound, all ε-consequences are acceptable as default conclusions, but there are other defaults which, though not probabilistically sound, nevertheless ought, intuitively, to be entailed. These correspond to commonsense guidelines such as ignoring irrelevant information and assuming that the only exceptions to defaults which exist are those explicitly represented. (...) The ε-semantics therefore needs extending if it is to fully capture all the

[4] Vreeswijk puts it as follows: "It is worth pausing to wonder about [the] idea that 'what we usually expect from an inference relation is that it makes knowledge which is implicit in the premises explicit' [Brew91] because it is precisely *this* what makes inference non-nonmonotonic, and it is precisely *this* that frustrated logicians — at least logicians which were not interested in reasoning about mathematics only — for such a long time. Namely, that monotonic reasoning is nothing but tampering with the little knowledge we have, without making any addition to it. The original motivation for doing nonmonotonic reasoning, on the other hand, is just to surpass this level of 'rewriting' to *amplify* (my italics, MC) on what is in the premises." We agree with Vreeswijk on this point, and thus are interested in a suitable *amplification* principle for defaults (under a statistical interpretation).

default reasoning requirements.

Given the often expressed opinion that ε-semantics is not strong enough to serve as a complete basis for default reasoning, Bourne defends the ME-approach as a suitable addition [Bour99, p. 53]:

> (...) Because the ε-semantics sanctions conclusions which hold in *all* admissible probability distributions (PDs), the idea is to select that distribution which possesses the highest value of entropy as the most appropriate from which to make inferences. Given a problem in which a probability distribution is constrained to some extent but not uniquely determined, it makes sense to select that PD with the highest entropy since this is guaranteed to contain the most uncertainty, or to be the least biased or committed. In fact, to select any other PD means that additional assumptions have been made which are not justified by the data [Jayn79]. The problem, then, becomes one of optimizing the entropy function subject to the known constraints, leading to the maximum entropy (ME) distribution. This principle has been widely used across many fields and has been described as "a much needed extension to the established principles of rational inference in the sciences" [Buck91].

The system of Goldszmidt [Gold93], as well as the system of Bourne [Bour99] (which can be seen as an extension of the work of Goldszmidt that can deal with things like variable strength defaults) allow for the entailment of all defaults also entailed by ε-semantics[5], as well as additional defaults not entailed by ε-semantics. Examples of the latter are the following:

- $\{A \Rightarrow C\} \mid\sim_{\varepsilon,ME} A \wedge B \Rightarrow C$ (irrelevance)
- $\{A \Rightarrow C,\ B \Rightarrow C\} \mid\sim_{\varepsilon,ME} A \wedge B \Rightarrow C$ (left conjunction)
- $\{A \Rightarrow B\} \mid\sim_{\varepsilon,ME} \neg B \Rightarrow \neg A$ (contraposition)
- $\{A \Rightarrow B,\ B \Rightarrow C\} \mid\sim_{\varepsilon,ME} A \Rightarrow C$ (transitivity)

Yet, the above derivations are nonmonotonic; they can be defeated when additional information is available:

- $\{A \Rightarrow C,\ B \Rightarrow \neg C\} \not\mid\sim_{\varepsilon,ME} A \wedge B \Rightarrow C$
- $\{A \Rightarrow C,\ B \Rightarrow C,\ A \wedge B \Rightarrow \neg C\} \not\mid\sim_{\varepsilon,ME} A \wedge B \Rightarrow C$
- $\{A \Rightarrow B,\ \neg A \Rightarrow B\} \not\mid\sim_{\varepsilon,ME} \neg B \Rightarrow \neg A$
- $\{A \Rightarrow B,\ B \Rightarrow C,\ A \Rightarrow \neg C\} \not\mid\sim_{\varepsilon,ME} A \Rightarrow C$

[5]It should be mentioned that Goldszmidt restricts the applicability of his formalism to so called *minimal core sets* of defaults, a restriction that is not made by plain ε-semantics. For every minimal core set of defaults, all the defaults entailed by ε-semantics are also entailed by Goldszmidt's system.

Thus, properties like irrelevance, left conjunction, contraposition and transitivity should not be seen as hard and fast rules of inference, or as constraints. Instead, it is better to view them as properties one would expect to find, unless an exceptional circumstance exists, i.e. and observable phenomenon whose absence indicates an exception has occurred [Bour99, p. 80]. This is in contrast with ε-semantics, where the entailment has a monotonic nature.

The nonmonotonic nature of ME enriched ε-semantics requires that all relevant information has been explicitly modeled, for otherwise the results may be counterintuitive. In case of Brewka's "men and beards" example, one explicitly has to state that women do not have beards, and in the lottery example, one has to state that not buying a ticket results in not winning a price. With this extra information, the contraposed defaults $beard \Rightarrow \neg man$ and $win \Rightarrow \neg ticket$ can no longer be entailed.

Although the maximum entropy formalism can be a suitable approach for allowing nonmonotonic entailment in a statistical setting, the technical complexity of the ME-formalism can pose difficulties.[6] ME-entailment is calculated using a rather complex algorithm and in general, explaining *why* a certain proposition does or does not follow from a set of premises is not an easy task. Compare this with the argumentation approach, in which people who disagree with a certain outcome can be convinced by means of an argument-based dialogue.[7] In short, argument-based dialectics is as a procedure much closer to intuitive, human style reasoning than the somewhat complex declarative approach as implemented by ME-enriched ε-semantics. Our treatment of ME is therefore limited to an analysis of which general principles should or should not be held valid; principles that can then be "ported" to the field of argumentation systems.

Causal defaults

As an aside, it may be interesting to view Brewka's "men and beards" example, as well as the lottery example, from the perspective of *causal defaults*. Under causal interpretation a default "$A \Rightarrow B$" is read as "A causes B".

In the lottery example ("$ticket \Rightarrow \neg price$"), can we say that buying a ticket *causes* the fact that no price is won? It appears not. One of the reasons for this is that the notion of causality can be interpreted in terms of counterfactuals [Pear99]. If A has caused B, then if A was not the case, B would also not have been the case. In the lottery example, if one had not bought a ticket, then one would definitely not have won a price. Hence, the counterfactual interpretation fails. The point is that $ticket \Rightarrow \neg price$ is a default where the truth of $ticket$ does not have a positive influence on the truth of $\neg price$, the influence, if any, can better be seen as negative instead, for the chance of $\neg price$ is higher when $\neg ticket$ is the case. Thus, $ticket$ does not contribute to $\neg price$ at all. Similar remarks can be made with respect to $men \Rightarrow \neg beard$ and $human \Rightarrow \neg diabetics$. In the case of the lottery example, for instance, it would be appropriate to replace $ticket \Rightarrow \neg price$ by the more simpler default $true \Rightarrow \neg price$. In general, one should be aware that defaults where the antecedent does not have a positive influence on the consequent should not be candidates for contraposition, nor for HY.

[6] Because of this complexity, it was decided not to include Bourne's ME-entailment algorithm in this thesis.

[7] At least, if the semantical principle of the argumentation system allows for a dialectic proof theory, such as is the case for grounded semantics.

Contraposition and HY reviewed

We have now come to the main point of our discussion. Regarding contraposition, an often stated opinion is that it should not be valid because counterexamples exist. What we hope to have made clear is that if one allows counterexamples against contraposition, one also has to allow counterexamples against principles like left conjunction, transitivity and even irrelevance, since counterexamples against these are essentially of the same type. Yet, it is striking to see that formalisms for defeasible reasoning tend not to be based on any consistent choice on these issues. A similar observation can be made regarding the to contraposition related principle of *moduls tollens*. Both Reiter's default logic [Reit80] and the formalism of Prakken and Sartor [PrSa97] sanction a defeasible form of modus ponens, but do not sanction any form of modus tollens. A systematic analysis of the actual meaning of a default is often not provided. Yet, it is this analysis that should serve as a basis for determining which principles should or should not be sanctioned. The current trend seems to be to sanction various principles, but not those of (defeasible) modus tollens or contraposition. It is an anomaly that is rarely questioned, and one may wonder whether this is because many researchers have become acquainted with it. Or, as Ginsberg states when discussing the reasons behind the opposition against contraposition [Gins94, page 16]:

> (...) although almost all of the symbolic approaches to nonmonotonic reasoning do allow for the strengthening of the antecedents of default rules, many of them do *not* sanction contraposition of these rules. The intuitions of individual researchers tend to match the properties of the formal methods with which they are affiliated.

4.2.2 Epistemical reasoning versus constitutive reasoning

In the previous section, the question whether or not contraposition and/or HY-arguments should be allowed was discussed from the perspective of probabilistic empirical reasoning. It was pointed out that the validity of contraposition and HY requires an additional principle (like entropy) that is also needed for properties like transitivity and irrelevance to be sanctioned.

In this section, we again ask the question whether contraposition and HY should be sanctioned, this time not from the perspective of probabilistic empirical reasoning, but from the perspective of *constitutive* reasoning.

The difference between these two forms of reasoning can perhaps best be illustrated using a mirror example. A mirror example, as was explained in section 2.2.5, consists of two intuitive examples that, apart from syntax, share the same formalization, even though they have different desired outcomes.

Intuitive Example 1 (IE_1):

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow TMA,\ \rightarrow LIS\} \\ \mathcal{D} & = & \{TMA \Rightarrow A,\ A \Rightarrow CD,\ LIS \Rightarrow \neg CD\} \\ < & = & \emptyset \end{array}$$

P: The goods must have arrived (A) in the Netherlands now, since we placed our order three months ago (TMA).
$\rightarrow TMA,\ TMA \Rightarrow A$

O: I don't think so, for if the goods would have arrived, there would be a customs declaration (CD), and this declaration seems to be lacking in the customs's information system (LIS).
$\rightsquigarrow A,\ A \Rightarrow CD,\ \rightarrow LIS,\ LIS \Rightarrow \neg CD$

Intuitive Example 2 (IE_2):

$\mathcal{S} = \{\rightarrow SN,\ \rightarrow P\}$
$\mathcal{D} = \{SN \Rightarrow M,\ M \Rightarrow R,\ P \Rightarrow \neg R\}$
$< = \emptyset$

P: The person in question is misbehaving (M) since he is loudly snoring (SN) in the university library.
$\rightarrow SN,\ SN \Rightarrow M$

O: I don't think so, for if he was really misbehaving, he could be removed and he cannot since the person in question is a professor.
$\rightsquigarrow M,\ M \Rightarrow R,\ \rightarrow P,\ P \Rightarrow \neg R$

The two examples share the property that from a certain conclusion (A or M) a contradiction can be entailed. The reason to reject A in the first example, however, seems significantly more intuitive than the reason to reject M in the second example. In fact, example IE_2 has been taken from [Prak97, p. 185] where it is claimed that the intuitive outcome should be M, but not R or $\neg R$. Hence, the above pair (IE_1, IE_2) can be considered as a mirror example in the sense of section 2.2.5.

The next question then is how this situation should be dealt with. That is, do we (1) reject at least one of the formalizations as "incorrect", (2) enrich our logic and dialogue system with a new concept that enables it to distinguish between the formalizations of IE_1 and IE_2, or (3) accept the application of two different logical systems, one for IE_1 and one for IE_2?

In this essay, we choose for option 3 (and aim to make clear why this solution is suitable). That is, we claim that example IE_1 should be modeled in a system with HY-arguments, and IE_2 in a system without. This, however, leaves open the question of *why* the examples should be modeled using a different logical system. That is, what principles or intuitive concepts allow us *in advance* to determine that IE_1 should be modeled in a system with HY-arguments, while IE_2 should be modeled in a system without?

Direction of fit

In order to understand the nature of constitutive reasoning, it is illustrative to distinguish between statements that have a *word to world* direction of fit, and statements that have a *world to word* direction of fit [Sear79, Sear83]. This distinction can be illustrated with the following example, which is borrowed from Anscombe [Ansc57, p. 56]:

> Suppose I went to the supermarket with a shopping list. While I am shopping a detective is observing me and carefully writes down every item I put in my

shopping chart. At the end the detective's list will contain exactly the same items as my original shopping list. The purpose and nature of these lists, however, is different. If I use my list correctly, the shopping chart will contain the same items as on the list, whereas for the detective, the aim is that the list contains the same items as in the shopping chart. That is, my shopping list has a "world to word" direction of fit (the world should conform to the statements on the list) whereas the detective's list has a "word to world" direction of fit (the statements on the list should conform to the world).

It is interesting to notice that one of the differences between IE_1 and IE_2 concerns the direction of fit. The defeasible rules of IE_1 are meant to describe when a certain fact holds in the object-world. This object-world has an existence that is independent of the rules that express our knowledge about it. These rules, therefore, have a *word to world* direction of fit. Their correctness depends on a validity that has an independent existence.

In IE_2, on the other hand, the very nature of the rules is different. The rules do not merely describe the reality, but to some extent also construct it, especially if we assume these rules to be taken from, say, the library regulations. The rule $SN \Rightarrow M$, for instance, contributes to the definition of misbehavior in the context of the library regulations. The rule essentially *makes* it the case that snoring is considered to be misbehavior, as far as the library is concerned. The defeasible rules of IE_2, therefore, have a *world to word* direction of fit. Their application results in the creation of new (legal) facts.

Epistemic versus constitutive reasoning

Based on the direction of fit, one can distinguish two kinds of reasoning: epistemic and constitutive[8]. The nature of this distinction can be described as follows [Hage97, pp. 60]: "Epistemic reasons are reasons for believing in facts that obtain independent of the reasons that plead for or against believing them. Constitutive reasons, on the contrary, influence the very existence of their conclusions".

In order to understand the differences between epistemic and constitutive reasoning, we provide the following abstract example[9] (AE):

$$\begin{array}{lcl} \mathcal{S} & = & \{\rightarrow A,\ \rightarrow D\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C,\ D \Rightarrow \neg C\} \\ \text{conflict:} & & \rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C \\ & & \rightarrow D,\ D \Rightarrow \neg C \end{array}$$

Now, take the following two constitutive interpretations of this example.

1. deontic
   The following example is somewhat similar to that of the Christian Soldier. An artillery soldier is given the order to destroy an enemy military installation, and orders should generally be obeyed ($order \Rightarrow O(shoot)$). When the soldier looks through his binoculars, he observes some movements that probably mean that some people are really close to the target ($movements \Rightarrow people$), thus making it from an

[8]The term "constitutive rules" was originally introduced by Searle [Sear69]. In this essay, however, we use the term in the sense of [Hage96, Hage97].

[9]The reader will notice that the structure of this example is similar to IE_1 and IE_2.

ethical point of view imperative not to shoot ($people \Rightarrow O(\neg shoot)$). Thus, we have:

$\mathcal{S} = \{\rightarrow order,\ \rightarrow movements\}$

$\mathcal{D} = \{movements \Rightarrow people,\ people \Rightarrow O(\neg shoot),\ order \Rightarrow O(shoot)\}$

conflict: $\rightarrow movements,\ movements \Rightarrow people,\ people \Rightarrow O(\neg shoot)$
$\rightarrow order,\ order \Rightarrow O(shoot)$

Here, the conflict is between the obligation to shoot and the obligation not to do so. In some logics, like SDL[10], such a conflict would lead to an inconsistency. If we would allow for contraposition or HY, the effect would be that *people* is no longer justified. This is, of course, absurd; the belief in empirical statements should not depend on the presence or absence of deontic conflicts.

2. legal
An example of a legal interpretation is IE_2. Here, the reasoning concerns whether or not certain legal facts obtain. Even though the conflict could be described in deontic terms (is the library personnel permitted to remove the person in question or not), the conflict ($Permitted(remove)$ v.s. $\neg Permitted(remove)$) is essentially not of a deontic nature, like in the previous example ($Obliged(shoot)$ v.s. $Obliged(\neg shoot)$). The question is whether it is legally permitted to remove the person or not, and this question does not rely on the specifics of deontic reasoning. The fact that this conflict exists, however, is no reason to reject the intermediate conclusion of M. To make this point more clear, suppose that the library regulations contain an additional rule saying that those who misbehave have to pay a fine of ten euro ($M \Rightarrow F$) and that no rule is available that provides professors with exemption for this fine. Then, the fact that the intermediate conclusion M can lead to R (which conflicts with $\neg R$) is no reason to disallow the entailment of F.

The point is that, regarding contraposition and HY, constitutive reasoning obeys different principles than epistemic reasoning. Under epistemic reasoning it is perfectly reasonable to sanction contraposition and HY, as is illustrated in section 4.2.1 and by example IE_1. Under constitutive reasoning, on the other hand, contraposition and HY are *not* valid by default, as illustrated by example IE_2. In legal reasoning, for instance, the leading paradigm is that the law should be interpreted as consistently as possible. Hence, in the snoring professor example the potential conflict between R and $\neg R$ is not a reason to reject M or F. The idea is to keep the effects of possible conflicts as local as possible, or as Hage puts it [Hage97, pp. 109]: "Legal ideology will have it that rules of law do not conflict. If two rules seem to conflict, at least one of them is not applicable. The scopes of the conflicting rules are assumed to be disjoint".[11] A great deal of research has been dedicated at stating and formalizing meta-principles (such as *lex posterior*, *lex specialis* or *lex superior*) for determining which of the conflicting rules should be applied, and which should not. But even in the case that no determining meta-principle is available, the application of *both* rules is blocked and the conflict does not have consequences for

[10] *Standard Deontic Logic*, which is essentially a KD modal logic.

[11] A comparable observation is made by Perelman [Pere82, p. 44]: "Similarly, insofar as article 4 of the Napoleontic Code requires a judge to give a decision ('the judge who refuses to judge, under the pretext of the silence, the obscurity, or the inadequacy of the law, can be prosecuted as guilty of denying justice') the judge *having* to state the law, even in a case not foreseen by the legislator, will have to interpret the texts in such a way that his interpretation will allow him to settle the litigation, even if customary interpretation offers no solution."

conclusions that do not depend on it. Our snoring professor, even though he may not be removed, still has to pay his €10 fine.

(Im)perfect procedures versus pure procedures

The difference between epistemic and constitutive reasoning is comparable to the difference between (im)perfect procedures and pure procedures, as distinguished by Rawls.

To illustrate the concept of a *perfect procedure*, Rawls provides the example of cake-cutting [Rawl00, pp. 74]:

> A number of men are to divide a cake: assuming that the fair division is an equal one, which procedure, if any, will give this outcome? Technicalities aside, the obvious solution is to have one man divide the cake and get the last piece, the others being allowed their pick before him. He will divide the cake equally, since in this way he assures for him the largest share possible. This example illustrates the two characteristic features of perfect procedural justice. First, there is an independent criterion for what is a fair division, a criterion defined separately from and prior to the procedure which is to be followed. And second, it is possible to devise a procedure that is sure to give the desired outcome.

One of the assumptions of the above cake-cutting example is that the person cutting the cake can do so with great accuracy. As long as deviations in cutting are ignored, the result will be an equal distribution. If we assume that the deviations in cutting cannot be ignored, cake-cutting becomes an *imperfect procedure*. The characteristic mark of an imperfect procedure is that while there is an independent criterion for the correct outcome, there is no feasible procedure which is sure to lead to it [Rawl00, pp. 75].

A *pure procedure*, on the contrary, is the case when there is no independent criterion for the right result: instead there is a correct or fair procedure such that the outcome is likewise correct or fair, whatever it is, provided that the procedure has been properly followed. An example of a pure procedure is that of gambling [Rawl00, p. 75]:

> If a number of persons engage in a series of fair bets, the distribution of cash after the last bet is fair, or at least not unfair, whatever this distribution is. I assume here that fair bets are those having a zero expectation of gain, that the best are made voluntarily, that no one cheats, and so on. (...) Now any distribution of cash summing to the initial stock held by all individuals could result from a series of fair bets. In this sense all of these particular distributions are equally fair. A distinctive feature of pure procedural justice is that the procedure for determining the just result must actually be carried out; for in those cases there is no independent criterion by reference to which a definite outcome can be known to be just.

Other examples of pure procedures are free elections. The outcome of elections cannot be evaluated as "right" or "wrong" according to an outside objective standard. The idea is that any resulting outcome should be accepted, as long as the election process itself was carried out in a correct way. In general, one can only fight the outcome of a pure procedure by arguing that the procedure was not applied properly [Lodd98].

	Independent criterion	Procedure that is guaranteed to lead to desired result
Perfect procedure	Yes	Yes
Imperfect procedure	Yes	No
Pure procedure	No	Yes

Table 4.1: Three different kinds of procedures [Lodd98, p. 23]

An overview of the three different kinds of procedures is provided in table 4.1[12].

One type of processes are those concerned with reasoning, and it is interesting to evaluate how the kind of reasoning as performed in IE_1 and IE_2 can be regarded in terms of (im)perfect and pure procedures.

IE_1 is basically an instance of empirical (epistemic) reasoning. One uses potentially incomplete information and rules of thumb, with the idea that the reasoning process is likely to generate a correct result. Even though an outside criterion exists to evaluate correctness (the goods have either arrived or not), there is no guarantee that the reasoning process indeed obtains this result. Hence, the reasoning process as performed in IE_1 is essentially an imperfect procedure.

IE_2 is an instance of constitutive reasoning. The idea of the library regulations is that applying them defines which (legal) consequences hold in a particular situation. There is no outside criterion, other than the library regulations themselves, that allows us to evaluate the legal implications as far as the library is concerned. Hence, the reasoning process can be seen as a pure procedure[13].

The difference between epistemical and constitutive reasoning has implications for what principles do or do not hold in the reasoning process. Let us ask the question of whether a certain principle X (like contraposition) holds in constitutive reasoning. The answer, of course, is that it depends on how the particular form of constitutive reasoning is defined. This definition needs not to be explicit. It may very well be that a certain type of informal reasoning has become common in a certain community, and that it is the researcher's task to provide a formal model of this reasoning; this is essentially what happens in, for instance, AI & Law.

In empirical reasoning, an outside criterion is available for determining whether the results are considered correct or not. The task of the reasoner is to perform its reasoning in such a way that the outcome approximates the objective criterion as closely as possible. In a certain sense, the presence of an objective criterion *forces* the reasoning process to become of a certain shape, in which certain properties (like contraposition) hold and other properties do not hold.

In constitutive reasoning, such an objective criterion is absent. For the community of reasoners, there is nothing that forces their reasoning process to become of a certain shape. In essence, the reasoners rely only on their own opinions and and intuitions regarding what such a reasoning process should look like and which properties it should adhere to. Wason's experiment, however, makes clear that a large group of people has difficulties with the

[12] In [Lodd98] it is claimed that there also exists a fourth kind of procedure, the *legal procedure*, which is characterized by the absence of an independent criterion and the absence of a guaranteed correct result.

[13] "Rules should be considered as a kind of tools applied by humans to structure the (legal) world [Hage96, p. 220].

principle of contraposition; it should therefore not come as a surprise that, when no outside constraint or criterion is present that *forces* its validity, the type of unreflective reasoning that a group of people comes up with does not necessarily sanction contraposition. A similar remark can be made ragarding HY, as it has already been observed in Socrates's days that people are often not aware of the consequences of their own reasoning.

Epistemic reasoning as a form of constitutive reasoning

Legal reasoning involves reasoning with mostly constitutive rules. These rules are meant to define the applicability of the various legal concepts (force majeure, unsound mind, etc.) to a specific case. An interesting question is the exact role of epistemic reasoning. In order to come to a legal verdict, what is needed is not just a legal interpretation of the facts, but also a procedure to determine what the facts are, based on more basic facts. Is reasoning about evidence a particular form of epistemic reasoning?

To a certain extent it is, but the situation is more complex than it would seem at first sight. The reasoning process about evidence cannot simply be seen as an imperfect procedure, which aims to entail the same results as are true in the objective world. First, there is the issue of illegally obtained evidence. Even if it is generally believed that person X committed a certain (unlawful) deed, this does not necessarily lead to a legally defensible standpoint that this is the case, if the evidence of such was obtained using illegal means. Second, even in the absence of illegally obtained evidence, some legal rules of evidence are institutional (and thus constitutive). The Dutch law, for example, has the rule that an authentic deed made by a notary definitively proves the declaration described in it. A negative institutional rule of criminal evidence is that the testimony of only one witness does not count as sufficient evidence [Hage96].

The key point is that the notion of truth that results from the reasoning about legal and convincing evidence[14] is in a certain sense itself an abstraction, a construct of the human mind. This construct does not have a completely independent existence, but is instead at least partly defined by law. Therefore, the kind of reasoning that is connected with legal evidence should at least to a significant extent be considered as constitutive.

Overview

In [Hage96], reasons are divided into epistemic and constitutive reasons. Constitutive reasons can be further subdivided into classificatory reasons at one hand, and anakastic, deontic and epistemic reasons at the other hand. Classificatory reasons are reasons why a certain concept is applicable to a particular situation; the fact that John took Mary's purse is a reason why John is a thief. Deontic reasons are reasons for the existence of duties, obligations, prohibitions, etc. For example, the fact that taking Mary's purse would be theft is a reason why John is legally prohibited from doing so, and the fact that Eric promised to visit Derek is the reason why Erik has the duty to visit Derek. Anankastic reasons are facts that make other facts necessary or (im)possible. For instance, the Dutch constitution makes it possible for the parliament to make laws, and the fact that James is a minor makes it impossible for him to engage in a contract. Epistemic reasons are a special class, that sometimes also allows overlap with constitutive reasons. In figure 4.2 the situation is graphically depicted.

[14] "wettig en overtuigend bewijs"

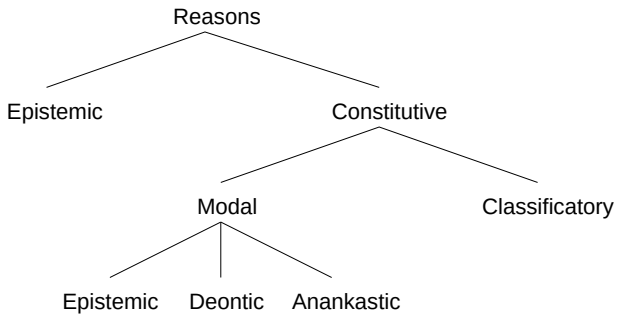


Figure 4.2: Kinds of reasons [Hage97, pp. 60]

Regarding the issue of contraposition and HY-arguments, we believe the most important distinction is between (purely) epistemic reasons, and constitutive reasons. We hope to have made clear that while contraposition and HY should be sanctioned in (purely) epistemic reasoning (recall that, as pointed out in section 4.1, there exist various situations in which contraposition is not enough and HY is also needed), they should *not* necessarily be sanctioned in constitutive reasoning. Hage — although not explicitly mentioning this issue — acknowledges the similarities between the various forms of constitutive reasoning [Hage97, pp. 61]: "The distinction between kinds of reasons and their conclusions is in particular important because I want to argue that seemingly different phenomena such as classifications and deontic judgments are based on the same underlying mechanism and that, as a consequence, one and the same logic can be used to deal with these seemingly heterogeneous phenomena."

Chapter 5

HY and other systems

So far, we have studied the concept of HY-arguments from the perspective of the defeasible logic of Prakken and Sartor, which has served as our reference system. An interesting question is whether problems illustrated in examples on page 44 also occur in other logics, and to which extent the concept of HY-arguments can be applied in order to solve these problems. That is, we ask ourselves the question whether the concept of HY-arguments is related to P&S's logic only, or whether it is equally relevant to other logics for defeasible reasoning.

In this chapter, we study two additional systems for defeasible reasoning: Reiter's default logic and Pollock's system. In addition, we also view the notion of HY-arguments from the perspective of Dung's argument-based semantics. Our aim is to make clear that the issue of HY-arguments is a general one, that is applicable to various kinds of logics for defeasible reasoning.

5.1 Semantical issues

In section 2.2.3 various approaches for argument based semantics were discussed. All these semantics are based on the notion of an *argumentation framework* (definition 2.1), consisting of a set of (abstract) arguments and a binary defeat relation between these arguments. Based on an argumentation framework, one can then apply a principle (like grounded semantics, stable semantics or preferred semantics) to determine which of the arguments are to be considered as *justified*.

HY and argumentation frameworks

To illustrate the working of an argumentation framework, consider the following example.

$$\begin{array}{rcl} \mathcal{S} & = & \{\rightarrow A\} \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow \neg A,\ A \Rightarrow C\} \\ < & = & \emptyset \end{array}$$

Now, if we assume a formalism whose arguments and defeat relation are purely classical,[1]

[1] $DS_{classic}$, for instance, is not completely classical, since S_1 and S_2 (definition 3.8) enable a limited form of (implicit) HY-style argumentation. Furthermore, the fact that $\emptyset$ defeats any incoherent argument is actually not a form of classical defeat either. Thus, with the above reference to "a formalism whose arguments and defeat relation are purely classical", we mean $DS_{classic}$ without S_1 and S_2 and without the empty argument.

this results in the following argumentation framework (we leave out arguments consisting of the same rules applied in different order).

Arguments:
$(A_1) \quad \rightarrow A$
$(A_2) \quad \rightarrow A,\ A \Rightarrow B$
$(A_3) \quad \rightarrow A,\ A \Rightarrow C$
$(A_4) \quad \rightarrow A,\ A \Rightarrow B,\ B \Rightarrow \neg A$
$(A_5) \quad \rightarrow A,\ A \Rightarrow B,\ A \Rightarrow C$
$(A_6) \quad \rightarrow A,\ A \Rightarrow B,\ B \Rightarrow \neg A,\ A \Rightarrow C$

Defeat relation:
$(A_4) < (A_1),\ (A_4) < (A_2),\ (A_4) < (A_3),\ (A_4) < (A_4),\ (A_4) < (A_5),\ (A_4) < (A_6)$
$(A_6) < (A_1),\ (A_6) < (A_2),\ (A_6) < (A_3),\ (A_6) < (A_4),\ (A_6) < (A_5),\ (A_6) < (A_6)$

This argumentation framework is graphically depicted in figure 5.1.

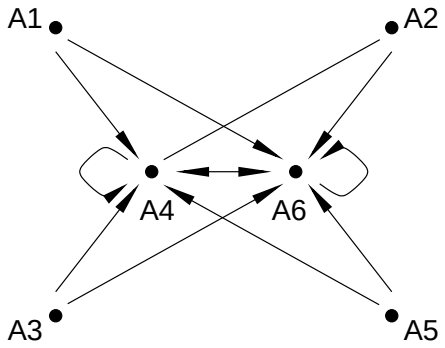


Figure 5.1: Argumentation framework without HY-arguments.

Now, if the argumentation formalism is changed to include HY-arguments, then this introduces a new class of arguments. In the above example, some of these new arguments are:

$(A_7) \quad \rightsquigarrow B,\ B \Rightarrow \neg A,\ \rightarrow A$
$(A_8) \quad \rightsquigarrow B,\ B \Rightarrow \neg A,\ \rightsquigarrow A$
$(A_9) \quad \rightsquigarrow \neg A,\ \rightarrow A$
$(A_{10}) \quad \rightsquigarrow \neg A,\ \rightsquigarrow A$

With the new arguments, the defeat relation is also extended:

$(A_2) < (A_7),\ (A_4) < (A_7),\ (A_6) < (A_7),\ (A_5) < (A_7)$
$(A_7) < (A_2),\ (A_7) < (A_4),\ (A_7) < (A_6),\ (A_7) < (A_5)$

$(A_2) < (A_8),\ (A_4) < (A_8),\ (A_6) < (A_8),\ (A_5) < (A_8)$
$(A_8) < (A_2),\ (A_8) < (A_4),\ (A_8) < (A_6),\ (A_8) < (A_5)$

$(A_4) < (A_9),\ (A_6) < (A_9)$
$(A_4) < (A_{10}),\ (A_6) < (A_{10})$

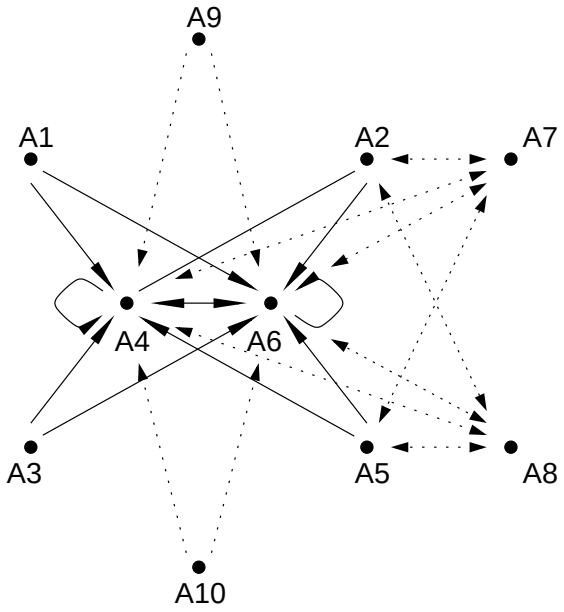


Figure 5.2: Argumentation framework with HY-arguments.

The new HY-enriched argumentation framework is depicted in figure 5.2.

Thus, one can see that the concept of HY-arguments is compatible with the notion of a Dung-style argumentation framework. Essentially, what happens is that both the set of arguments and the defeat relation among these arguments are extended. The extension is done in such a way that if one would leave out the HY-arguments, as well as each instance of the defeat relation involving at least one HY-argument, the result would be an argumentation framework of an argumentation formalism that allows only classical arguments and classical defeat.

HY and the notion of justified arguments

Given an argumentation framework, the next step is to determine which arguments are considered *justified*. For the formalism of P&S, the choice was made to implement grounded semantics, and for now, this choice is what we adhere to. To recall how grounded semantics works, consider figure 5.1. In this argumentation framework, grounded semantics can be applied as follows (using theorem 2.1 on page 15):

$$
\begin{aligned}
F^0 &= \emptyset \\
F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\
&= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\
&= \{A_1, A_2, A_3, A_5\} \\
F^2 &= \{A \in Args \mid A \text{ is acceptable with respect to } F^1\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } \{A_1, A_2, A_3, A_5\}\} \\
&= \{A_1, A_2, A_3, A_5, A_6\} \\
F^{i+1} &= F^i \text{ (with } i \geq 2)
\end{aligned}
$$

This means that $\cup_{i=0}^{\infty} = \{A_1, A_2, A_3, A_5\}$, so A_1, A_2, A_3 and A_5 are considered justified

under grounded semantics.

If we look at the HY-enriched formalism, on the other hand, then a different outcome results. Intuitively, with HY-arguments, A and C should not become justified, since the argument for $\neg A$ ($\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow \neg A$) is incoherent and the argument for B ($\rightarrow A,\ A \Rightarrow B$) has a HY-counterargument ($\rightsquigarrow B,\ B \Rightarrow \neg A,\ \rightarrow A$).

It is interesting to examine what happens when grounded semantics is applied straightforwardly to the HY-enriched argumentation framework of figure 5.2.

$$
\begin{aligned}
F^0 &= \emptyset \\
F^1 &= \{A \in Args \mid A \text{ is acceptable with respect to } \emptyset\} \\
&= \{A \in Args \mid A \text{ has no arguments defeating it }\} \\
&= \{A_1, A_3, A_9, A_{10}\} \\
F^2 &= \{A \in Args \mid A \text{ is acceptable with respect to } F^1\} \\
&= \{A \in Args \mid \text{ every argument defeating } A \text{ is defeated by } \{A_1, A_3, A_9, A_{10}\}\} \\
&= \{A_1, A_3, A_9, A_{10}\} \\
F^{i+1} &= F^i \text{ (with } i \geq 2)
\end{aligned}
$$

At least, this result is partly in line with what one expects. A_2 and A_5 are not justified anymore because they now have HY-counterarguments against them. A_1 and A_3, on the other hand, remain justified, as is desired. This result, however, comes with a price, since not only A_1 and A_3, but also the HY-arguments A_9 and A_{10} become justified. Worse yet, if one would just take the conclusions of justified arguments, then from $\rightsquigarrow \neg A,\ \rightarrow A$ being a justified argument, it would follow that both A and $\neg A$ would become justified conclusions!

Applying (grounded) semantics to DS_{HY} therefore involves two problems: (1) HY-arguments that can become justified and (2) the conclusions of HY-arguments that can become justified.

Let us first consider the point that HY-arguments can become justified. In the examples in section 3.3.1 and elsewhere, we saw that HY-arguments are meant as *counter*arguments. That is, they are not meant to yield conclusions on their own, but instead to prevent other conclusions from becoming justified. An HY-argument is context-dependent; it *needs* an argument that it defeats. One can say that an HY-argument does not have a meaning when standing on its own. A dialogue, for example, provides a proper environment for HY-arguments to make sense. The semantics as stated by Dung, on the other hand, treat arguments as having an existence that is independent from any other argument. Dung's semantics tries to answer the question "Is HY-argument A justified?", but this question does not make any sense if one considers that HY-arguments cannot stand on their own. It is simply not applicable to call HY-arguments "justified", "defeated" or "defensible", as these terms are applicable to classical arguments only. A possible solution would therefore be to reserve the terms "justified", "defeated" or "defensible" only to classical arguments, while leaving the rest of the specification of grounded semantics unchanged.

The second point to consider is that of justified conclusions. If our ultimate interest is in the justified conclusions, and we regard arguments only as a technical intermediate step to yield these conclusions, than in a certain sense it doesn't matter which arguments are justified and which are not, a long as we have the "right" justified conclusions. If we apply the notions of justified arguments as in theorem 2.1 without any changes, then there are two alternatives for defining when a conclusion is justified or not.

The first approach is to regard justified HY-arguments as making no sense at all. This means that HY-arguments should be left out completely when determining which conclusions are justified. Thus, for determining the justified conclusions, one should only look at classical arguments. It is this approach that we have taken in definition 3.24.

The second approach for determining which conclusions are justified is to try to interpret a justified HY-argument as "meaningful" as possible. HY-arguments can have a part that is not fc-based, and this part is not depending on a context with other arguments. In this vision, the fact that $\rightsquigarrow \neg A, \rightarrow A$ is justified would be a reason to make its conclusion A justified, since A is not fc-based. In general, this approach boils down to the following principle. Let A be a HY-argument. Take A' to be the part of A that is not fc-based (then A' is the largest weak sublist of A that is a classical argument). If A is justified, then all conclusions of A' are also defined as justified.

We now prove that these two approaches turn out to have the same effects. Notice that we use the notation *defeaters*(A) as an abbreviation of $\{A' \mid A' \text{ defeats } A\}$.

Lemma 5.1. *Let A and A' be two arguments in DS_{HY} such that defeaters$(A') \subseteq$ defeaters(A) and defeaters(A) is finite. Then A' is justified under grounded semantics if A is justified under grounded semantics.*

Proof. Suppose A is justified. Then (theorem 2.1) it holds that $A \in \cup_{i=0}^{\infty} F^i$. This means that there exists an F^i such that $A \in F^i$ (i is at least 1, since $F^0 = \emptyset$). Then $F^i = \{A \in Args \mid A \text{ is acceptable with respect to } F^{i-1}\}$. This means that A is acceptable with respect to F^{i-1}. This means that every argument defeating A is defeated by an argument in F^{i-1}. Since *defeaters*$(A') \subseteq$ *defeaters*(A), this means that every argument defeating A' is also defeated by F^{i-1}. Thus, it holds that $A' \in \{A \in Args \mid A \text{ is acceptable with respect to } F^{i-1}\}$. That is, it holds that $A' \in F^i$. This also means that $A' \in \cup_{i=0}^{\infty} F^i$, so according to theorem 2.1 A' is justified. □

Lemma 5.2. *Let A be an argument in DS_{HY} with a conclusion c such that c is not fc-based. Let A' be a minimal weak sublist of A such that A' is an argument with conclusion c. Then defeaters$(A') \subseteq$ defeaters(A).*

Proof. First, it should be noticed that $R_c(A') = R_c(A)$. Furthermore, A' is a classical argument; it does not contain any foreign commitments (otherwise c would have been fc-based). A' can be thought of as a "subargument" of A.

Now, let B be an arbitrary argument that defeats A'. According to the HY-enriched notion of defeat (definition 3.21 there are five possibilities:

1. B classically rebut-attacks A' and not $R_{-L}(B) < R_L(A')$. The fact that B classically rebut-attacks A' (definition 3.17 (1)) means that B also classically rebut-attacks A (this is because A contains all conclusions of A'). Furthermore, because $R_L(A') = R_L(A)$, it also holds that not $R_{-L}(B) < R_L(A)$. Thus, B defeats A.

2. B classically undercut-attacks A'. This means that (definition 3.17 (2)) B also classically undercut-attacks A (this is because A contains all assumptions of A'). Thus, B defeats A.

3. B HY-rebut-attacks A' and not $R_L(B) \cup R_{-L}(B) < \cup_{c_i \in \{c_i \mid \rightsquigarrow c_i \in R_L(B) \cup R_{-L}(B)\}} R_{c_i}(A')$. The fact that B HY-rebut-attacks A' (definition 3.18 (1)) means that B also HY-rebut-attacks A (this is because A contains all conclusions of A'). Furthermore, because A' is a subargument of A, it holds that $R_{c_i}(A') = R_{c_i}(A)$ for every conclusion c_i

in A'. Therefore, it also holds that $R_L(B) \cup R_{-L}(B) < \cup_{c_i \in \{c_i | \rightsquigarrow c_i \in R_L(B) \cup R_{-L}(B)\}} R_{c_i}(A)$. Thus, B defeats A

4. B reverse HY-rebut-attacks A'. The fact that B reverse HY-rebut-attacks A' means that A' HY-rebut-attacks B, but this is impossible because A' does not contain any foreign commitments. Therefore, B reverse HY-rebut-attacking A' is essentially a non-option.

5. B HY-undercut-attacks A'. This means that (definition 3.18 (2)) means that B also HY-undercut-attacks A (this is because A contains all assumptions and conclusions of A'). Thus, B defeats A.

The overall analysis of the above five forms of defeat is that if B defeats A' then B also defeats A. Thus $defeaters(A') \subseteq defeaters(A)$. □

Definition 5.1. *A conclusion is* justified1 *iff it is a conclusion of a classical argument (= an argument without any foreign commitments) that is justified in DS_{HY}.*

Definition 5.2. *A conclusion is* justified2 *iff it is a conclusion of an argument (classical or HY) that is justified in DS_{HY} and this conclusion is not fc-based.*

Theorem 5.1. *A conclusion is justified1 iff it is justified2.*

Proof.
"$\Longrightarrow$":
Let c be justified1. Then (definition 5.1) there is a justified classical argument A with conclusion c. Because A is a classical argument, c is not fc-based. Thus, c is also justified2.

"$\Longleftarrow$":
Let c be a conclusion that is justified2. Then there exists a justified argument A with a conclusion c that is not fc-based. Now take A' as a minimal weak sublist of A such that A' has conclusion c. A' is then a classical argument. Lemma 5.2 tells us that the defeaters of A' are a subset of the defeaters of A. Then, from lemma 5.1 it follows that A' is also justified. Thus, we have a classical argument (A') that has conclusion c and is justified. Thus, c is justified1. □

5.2 Default logic

As one of the older systems for defeasible reasoning, Reiter's default logic [Reit80] is nowadays regarded as a landmark system for nonmonotonic reasoning. Before continuing with our main point, we first recall some basics about Reiter's system.

Definition 5.3. *Given a language L that contains well formed formulas (wffs).*
A default *is an expression of the form:*

$$(\alpha : \beta_1, \ldots, \beta_m \ / \ w)$$

where $\alpha, \beta_1, \ldots, \beta_m, w$ are closed wffs (that is: wffs without free variables). α is called the prerequisite*, $\beta_1, \ldots, \beta_m$ the* justifications *and w the* consequent*.*

We sometimes abbreviate the prerequisite of a default d as $Pre(d)$, the justifications as $Jus(d)$ and the consequent as $Cons(d)$. Also notice that we use the notation $Cn(S)$ as the *consequence set* of a set of formulas S. That is: $Cn(S) \equiv_{def} \{\varphi \mid S \vDash \varphi\}$, where "$\vDash$" stands for classical entailment.

Definition 5.4. *A* default theory *is a pair $(\mathcal{W}, \mathcal{D})$ where $\mathcal{W}$ is a set of wffs and $\mathcal{D}$ is a set of defaults.*

Definition 5.5. *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory. For any set of wffs $S \subseteq L$ let $\Gamma(S)$ be the smallest set satisfying the following three properties:*

1. $\mathcal{W} \subseteq \Gamma(S)$
2. $Cn(\Gamma(S)) = \Gamma(S)$
3. *If $(\alpha : \beta_1, \ldots, \beta_m \,/\, w) \in \mathcal{D}$ and $\alpha \in \Gamma(S)$, and $\neg\beta_1, \ldots, \neg\beta_m \notin S$, then $w \in \Gamma(S)$*

A set of closed wffs $E \subseteq L$ is an extension *for T iff $\Gamma(E) = E$, i.e. iff E is a fixed point of the operator Γ.*

Theorem 5.2. *Let $E \subseteq L$ be a set of wffs, and let $T = (\mathcal{W}, \mathcal{D})$ be a default theory. Let $E_0 = W$ and for $i \geq 0$:*
$E_i = Cn(E_i) \cup \{w \mid (\alpha : \beta_1 \ldots \beta_m \,/\, w) \in \mathcal{D}$ where $\alpha \in E_i$ and $\neg\beta_1 \ldots \neg\beta_m \notin E\}$
Then, E is an extension of T iff $E = \bigcup_{i=0}^{\infty} E_i$

Proof. See [Reit80]. □

Theorem 5.3. *If E and F are extensions for a default theory $(\mathcal{W}, \mathcal{D})$ and $E \subseteq F$, then $E = F$.*

Proof. See [Reit80]. □

Reiter dedicates a special part of his work to the concept of a *normal default theory.*

Definition 5.6. *A* normal default *is a default of the form $(\alpha : w/w)$.*
A default theory $(\mathcal{W}, \mathcal{D})$ is normal iff every default in $\mathcal{D}$ is normal.

Theorem 5.4. *Every closed normal default theory has at least one extension.*

Proof. See [Reit80]. □

Theorem 5.5 (semi-monotonicity). *Suppose $\mathcal{D}$ and $\mathcal{D}'$ are sets of closed normal defaults with $\mathcal{D}' \subseteq \mathcal{D}$. Let E' be an extension for the closed normal default theory $T' = (\mathcal{W}, \mathcal{D}')$ and let $T = (\mathcal{W}, \mathcal{D})$. Then T has an extension E such that $E' \subseteq E$.*

Proof. See [Reit80]. □

Reiter's system by itself does not have the concept of a *justified conclusion*; only extensions are defined. It is, however, not so difficult to introduce this concept, as we may define a conclusion to be justified iff it is contained in every extension (we use the sceptical approach). For this criterion to be meaningful, it is required that there is at least one extension (otherwise it would for instance be possible to derive a contradiction).

Definition 5.7. *A conclusion c is* justified *given a default theory $(\mathcal{W}, \mathcal{D})$ iff:*

1. *$(\mathcal{W}, \mathcal{D})$ has at least one extension, and*
2. *c is an element of every extension of $(\mathcal{W}, \mathcal{D})$.*

Antoniou's process interpretation of RDL

One of the disadvantages of Reiter's original formulation of RDL is that the extensions are not defined in a way that allows them to be easily constructed.

In theorem 5.2, for instance, it seems that the content of the extension E should be known in advance, in order for it to be constructed in an inductive way (using $E_0, E_1, \ldots, E_n$). While theorem 5.2 allows one to *verify* whether a set of formulas E is or is not an extension (simply check whether $E = \cup_{i=0}^{\infty} E_i$), it cannot directly be applied in order to actually *construct* an extension when no suitable hypothesis is available.

In order to overcome these problems, Antoniou restated RDL in terms of *processes* [Anto97]. The general idea is to keep on applying different defaults until either every applicable default has been applied, or a violation of some justification has taken place, in which case one needs to backtrack over the defaults applied earlier.

Antoniou defines the applicability of a default as follows:

Definition 5.8. *A default* $(\varphi : \psi_1, \ldots, \psi_n \,/\, \chi)$ *is* applicable *to a deductively closed set of formulas* E *iff* $\varphi \in E$ *and* $\neg\psi_1 \notin E, \ldots, \neg\psi_n \notin E$.

Using the notion of applicability, it now becomes possible to define a *process*, which essentially consists of a sequence of defaults. With each process two sets are associated: a set of formulas derived by the process (In) and a set of assumptions of the process (Out).

Definition 5.9. *Let* $T = (\mathcal{W}, \mathcal{D})$ *and let* $\Pi = (d_0, \ldots, d_n)$ *be a sequence of defaults without multiple occurrences. Let* $\Pi[k]$ *be the initial segment of* Π *of length* k*, provided that the length of* Π *is at least* k*. With each sequence* Π*, two sets of first-order formulas are associated:*

- $In(\Pi) = Cn(\mathcal{W} \cup \{Cons(d) \mid d \text{ occurs in } \Pi\})$
- $Out(\Pi) = \{\neg\psi \mid \psi \in Jus(d) \text{ for some } d \text{ occurring in } \Pi\}$

Π *is called a* process *of* T *iff* d_k *is applicable to* $In(\Pi[k])$ *for every* k *such that* d_k *occurs in* Π.

Definition 5.10. *A process* Π *is* closed *iff every* $d \in \mathcal{D}$ *that is applicable to* $In(\Pi)$ *already occurs in* Π.
A process Π *is* successful *iff* $In(\Pi) \cap Out(\Pi) = \emptyset$.

Theorem 5.6. *A set of formulas* E *is an extension of the default theory* T *iff there is some closed and successful process* Π *of* T *such that* $E = In(\Pi)$.

Proof. See [Anto97]. □

Theorem 5.7. *Each process of a normal default theory is successful.*

Proof. See [Anto97]. □

Some examples in RDL

Now that we have defined our preliminaries, it is interesting to look at some of our previous examples from the perspective of Reiter's default logic.

example (shipment of goods)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{tma, \neg is\}$
$\mathcal{D} = \{(tma, a \ / \ a),\ (\neg is, \neg cd \ / \ \neg cd),\ (a, cd \ / \ cd)\}$

Here, there exist two extensions: $Cn(\{tma, \neg is, a, cd\})$ and $Cn(\{tma, \neg is, a, \neg cd\})$. So we see that here too, a is justified (just like in P&S's original system).

example (tuff-tuff-club)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{f, c\}$
$\mathcal{D} = \{(f \wedge c, s \ / \ s),\ (s, ttc \ / \ ttc),\ (ttc, p \ / \ p),\ (p \wedge c, \neg s \ / \ \neg s)\}$

Here, only one extension exists: $Cn(\{f, c, s, ttc, p\})$. So again, we obtain the same results as in P&S's system: ttc is justified.

In the examples above, we see that Reiter's default logic derives the same results as in P&S's original system. That is, we obtain the results that are suitable only in situations where HY-arguments are not allowed (see section 4.2 about epistemic and constitutive reasoning). There is, however, one example that is handled in a particularly remarkable way in RDL:

example (Ajax-Feijenoord)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{af\}$
$\mathcal{D} = \{(af : \neg p, t \ / \ t),\ (t : p \ / \ p)\}$

Here *no* extension exists, meaning we cannot infer any meaningful conclusions.

In the now following discussion, three different alternatives are provided for dealing with this problem: the implementation of rule-maximality (section 5.2.1), the direct addition of HY-arguments (section 5.2.2) and the usage of free defaults (section 5.2.3). Of these three solutions, two (rule-maximality and free defaults) are restricted to normal default theories. Normal default theories have the advantage that they generally behave in a more "regular" way than a general (non-normal) default theory.[2] The disadvantage of normal default theories is that, although Reiter originally thought all naturally occurring defaults can be represented as normal defaults,[3] it was soon discovered that there are situations that *do* require the use of non-normal defaults [ReCr81, Ethe87, Brew89]. The third solution (the direct addition of HY-arguments, section 5.2.2) is therefore defined for unrestricted default theories.

[2] Normal default theories for instance always have at least one extension.

[3] "In fact I know of no naturally occurring default which cannot be represented in this form" [Reit80][p. 95]

5.2.1 Implementing rule-maximality in RDL

In section 3.3.2 the principles of conclusion-maximality and rule-maximality were introduced. It was shown that the principle of conclusion-maximality provides an upper-bound for a (rebutting only) P&S theory without HY-arguments, whereas rule-maximality provides an upper-bound for a (rebutting only) P&S theory with HY-arguments.

In the current section, our aim is to apply the principle of rule-maximality to RDL. The idea is that by doing so, one obtains a logic that supports HY-style reasoning without the need to define HY-arguments themselves.

Standard RDL as conclusion maximality

The first thing to notice is that RDL, as far as one restricts oneself to normal default theories, is in accordance with the principle of conclusion-maximality. A conclusion-maximal set of conclusions, we recall, is a maximal set of wffs that has a coherent argument (read: successful process) that derives it.

Theorem 5.8. *Let E be a set of formulas and T a normal default theory. E is a Reiter-extension of T iff E is a maximal set of wffs that has a successful process Π such that $E = In(\Pi)$.*

Proof.
"$\Longrightarrow$":
Let E be a Reiter-extension of a normal default theory T. Then, according to theorem 5.6, there is a closed and successful process Π with $In(\Pi) = E$. Furthermore, there is no process Π' with $In(\Pi') \supsetneq In(\Pi)$. This is because otherwise there would be an extension $E' \supsetneq E$ and this is not possible due to theorem 5.3.

"$\Longleftarrow$":
Let Π be a process with $In(\Pi) = E$, and suppose there is no process Π' with $In(\Pi') \supsetneq E$. Then Π must be a complete process. Furthermore, Π is according to theorem 5.7 also a successful process. Thus, by means of theorem 5.6 E is a Reiter-extension. □

Adjusting RDL for rule maximality

Given that normal RDL-theories implement conclusion-maximality, the next question is how to adjust the RDL-formalism such that rule-maximality is obtained.

In order to be able to define rule-maximality in RDL, one needs to have the notion of a (maximal) set of defaults that is "conflict-free". For the system of P&S, we said that a set of defeasible rules is conflict-free iff it does not allow the construction of an incoherent (= self-defeating) argument (definition 3.32 on page 70). The idea is to define a similar condition for RDL. The only problem is that until so far, no explicit concept has been defined that can be used as an *incoherent* argument.

At first sight, it would seem that an obvious way to determine whether a set of defaults $\mathcal{D}'$ is "conflict-free" or not would be the examine whether or not an incoherent (= unsuccessful) process can be constructed with it. Unfortunately, theorem 5.7 states that for a normal default theory, every process is automatically successful; this is because definition 5.8 makes sure that during the construction of a process, the justification of its defaults will not be violated.

An alternative way to determine whether or not a set of defaults is "conflict-free" is whether or not there exists a closed process Π and a default d such that $Pre(d) \in In(\Pi)$ but $Cons(d) \notin In(\Pi)$. Because Π is a closed process it must be that $\neg Jus(d) \in In(\Pi)$. From this and the fact that d is a normal default, we can infer that $\neg Cons(d) \in In(\Pi)$. Thus, d is a default that would be applicable if one would purely look at the prerequisite, but whose application is undesirable, as its consequent conflicts with the conclusions of Π. Thus, if we were to concatenate Π and d, the result could be seen as an incoherent argument, although from a technical point of view, this result would not be a process anymore.

As closed and successful processes coincide with RDL-extensions, this leads to the following definition:

Definition 5.11. *Let $(\mathcal{W}, \mathcal{D})$ be a normal default theory. A rule-maximal set of defaults is a set $\mathcal{D}' \subseteq \mathcal{D}$ such that for every extension E in $(\mathcal{W}, \mathcal{D}')$ there does not exist a default $d \in \mathcal{D}'$ whose prerequisite is in E, but whose consequent is not in E.*

Theorem 5.9. *If D' is a rule-maximal set of defaults for the closed normal default theory $(\mathcal{W}, \mathcal{D})$ then $(\mathcal{W}, \mathcal{D}')$ has exactly one extension.*[4]

Proof. According to theorem 5.4, every closed normal default theory has *at least* one extension, so we only have to prove that $(\mathcal{W}, \mathcal{D}')$ has *at most* one extension. This, we prove *reductio ad absurdum*.

Suppose $(\mathcal{W}, \mathcal{D}')$ has at least two extensions E and F with $E \neq F$. Then, according to theorem 5.3, it holds that $E \not\subseteq F$. This means that apparently there are one or more conclusions that are in E, but not in F, so $E - F$ is not empty. We know that E equals $E_0 \cup E_1 \cup E_2 \cup \ldots$. This means that every element of $E - F$ should be an element of $E_0 \cup E_1 \cup E_2 \cup \ldots$. Let us take the lowest i such that $E_i \cap (E - F)$ is not empty. This E_i always exists, since $E - F$ is not empty. Furthermore, this i will be at least 1 (*proof*: Suppose $E_0 \cap (E - F)$ is not empty. We have that $E_0 = \mathcal{W}$ and $F_0 = \mathcal{W}$. So $E_0 \cap (E - F) = \mathcal{W} - (E - F)$. And since $\mathcal{W} = E_0 \subseteq E$ and $\mathcal{W} = F_0 \subseteq F$, we have that $\mathcal{W} - (E - F) = \emptyset$. Contradiction). Now, let's again look at the set $E_i \cap (E - F)$. This set contains conclusions that are in E_i (and therefore also in E) but not in F. As E_i is also the "first" set that contains these conclusions, these conclusions are not contained in E_{i-1} (and we may talk about E_{i-1} because we know that i is at least 1). This means that there is at least one default $(\alpha : w \ / \ w) \in D'$ with $\alpha \in E_{i-1}$ and $\neg w \notin E$ and $w \in E_i \cap (E - F)$. Recall that E_i is the "first" set for which $E_i \cap (E - F)$ is not empty. This means that $E_{i-1} \cap (E - F)$ is empty. This means that whatever is in E_{i-1} is also in F (that is: $E_{i-1} \subseteq F$). Thus, for our default $(\alpha : w \ / \ w)$ it holds that α is in F (since α is in E_{i-1}). The conclusion w, however, is in $E_i \cap (E - F)$. This means $w \notin F$. Thus, we have a default whose prerequisite is in F but whose consequent is not in F, where F is an extension. This means that D' is not a rule-maximal set of defaults. Contradiction. □

As we have now proved that $(\mathcal{W}, \mathcal{D}')$ has exactly one extension, we may talk about *the* extension of $(\mathcal{W}, \mathcal{D}')$. We are now ready to provide the definition of the entailment of justified conclusions under rule-maximization.

[4] As an aside, it can be noticed that the converse of this theorem does not hold. That is, if $(\mathcal{W}, \mathcal{D}')$ has exactly one extension, then $\mathcal{D}'$ does not need to be a rule-maximal set of defaults. A counterexample would be $\mathcal{W} = \emptyset$ and $\mathcal{D} = \{(true : tr \ / \ tr),\ (tr : bd \ / \ bd),\ (bd : fb \ / \ fb),\ (fb : \neg tr \ / \ \neg tr)\}$. Here, exactly one extension exists: $E = Cn(tr, bd, fb)$, but $\mathcal{D}$ is not a rule-maximal set of defaults, since $(fb : \neg tr \ / \ \neg tr)$ has its prerequisite in E, but its consequent not in E.

Definition 5.12. *A conclusion c is justified under rule-maximization in $(\mathcal{W}, \mathcal{D})$ iff for every rule-maximal set $\mathcal{D}' \subseteq \mathcal{D}$, it holds that the extension of $(\mathcal{W}, \mathcal{D}')$ contains c.*

One of the things that can be noticed about the thus defined notion of rule-maximality is that every extension under conclusion-maximization (read: every extension under standard normal RDL) is also an extension under rule-maximization. Thus, what rule-maximization does is taking the existing extensions under standard (normal) RDL and optionally adding one or more additional extensions[5]. This property is stated in the following theorem.

Theorem 5.10. *Let $T = (\mathcal{W}, \mathcal{D})$ be a normal default theory and E an extension in T. There exists a rule-maximal set of defaults $\mathcal{D}' \subseteq \mathcal{D}$ such that $(\mathcal{W}, \mathcal{D}')$ has extension E.*

Proof. Let $T = (\mathcal{W}, \mathcal{D})$ be a closed normal default theory and E be an extension in T. Then, according to theorem 5.6 there exists a closed and successful process Π with $In(\Pi) = E$. Let $\mathcal{G}$ be the set of defaults in Π. It trivially holds that $\mathcal{G} \subseteq \mathcal{D}$. Furthermore, it also holds that every extension F in $(\mathcal{W}, \mathcal{G})$ does not have a default d in $\mathcal{G}$ whose prerequisite is in F but whose consequent is not in F (this is because otherwise Π would not have been successful).

Because $\mathcal{G}$ is a set of defaults that satisfies "$\mathcal{G} \subseteq \mathcal{D}$ and every extension F in $(\mathcal{W}, \mathcal{G})$ does not have a default d in $\mathcal{G}$ whose prerequisite is in F but whose consequent is not in F", there is also a *maximal superset* $\mathcal{G}'$ of $\mathcal{G}$ which satisfies this property. According to definition 5.11, this $\mathcal{G}'$ is a rule-maximal set of defaults. According to theorem 5.9, $(\mathcal{W}, \mathcal{G}')$ has exactly one extension; call this extension E'. According to theorem 5.5 it holds that $E \subseteq E'$ (this is because $\mathcal{G} \subseteq \mathcal{G}'$ and E is an extension of $(\mathcal{W}, \mathcal{G})$). We know that $\mathcal{G}' \subseteq \mathcal{D}$, so it also holds that $E' \subseteq E$. Thus, it follows that $E' = E$. □

Examples

It is interesting to see how this definition can be applied to the examples treated earlier.

example (shipment of goods)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{tma, \neg is\}$
$\mathcal{D} = \{(tma, a \mathbin{/} a),\ (\neg is, \neg cd \mathbin{/} \neg cd),\ (a, cd \mathbin{/} cd)\}$

Here, there exist the following rule-maximal set of defaults:

- $\{(tma : a \mathbin{/} a),\ (\neg is : \neg cd \mathbin{/} \neg cd)\}$
  $E = Cn(\{tma, \neg is, a, \neg cd\})$
- $\{(tma : a \mathbin{/} a),\ (a : cd \mathbin{/} cd)\}$
  $E = Cn(\{tma, \neg is, a, cd\})$
- $\{(\neg is : \neg cd \mathbin{/} \neg cd), (a : cd \mathbin{/} cd)\}$
  $E = Cn(\{tma, \neg is, \neg cd\})$

This means that we have tma and $\neg is$ as conclusions that are justified under rule-maximization (and not a, cd or $\neg cd$).

[5] This is similar to what was shown regarding the logic of P&S in lemma 3.6 (page 74).

example (tuff-tuff-club)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{f, c\}$
$\mathcal{D} = \{(f \wedge c, s \ / \ s),\ (s, ttc \ / \ ttc),\ (ttc, p \ / \ p),\ (p \wedge c, \neg s \ / \ s)\}$

Here, there exist the following rule-maximal sets of defaults:

- $\{(f \wedge c : s \ / \ s),\ (s : ttc \ / \ ttc),\ (ttc : p \ / \ p)\}$
  $E = Cn(\{f, c, s, ttc, p\})$
- $\{(f \wedge c : s \ / \ s),\ (s : ttc \ / \ ttc),\ (p \wedge c : \neg s \ / \ s)\}$
  $E = Cn(\{f, c, s, ttc\})$
- $\{(f \wedge c : s \ / \ s),\ (ttc : p \ / \ p),\ (p \wedge c : \neg s \ / \ s)\}$
  $E = Cn(\{f, c, s\})$
- $\{(s : ttc \ / \ ttc),\ (ttc : p \ / \ p),\ (p \wedge c : \neg s \ / \ s)\}$
  $E = Cn(\{f, c\})$

This means that we have f and s as conclusions that are justified under rule-maximization (and not s, ttc or p).

As an aside, it may now seem that under rule-maximization, all possible conclusions of the defaults are blocked. This, however, is not true; conclusions that have nothing to do with any conflicts are still entailed, as we hope to make clear with the following example:

example

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{a\}$
$\mathcal{D} = \{(a : b \ / \ b),\ (b : \neg a \ / \ \neg a),\ (a : c \ / \ c)\}$

Here, there exist the following rule-maximal set of defaults:

- $\{(a : b \ / \ b),\ (a : c \ / \ c)\}$
  $E = Cn(\{a, b, c\})$
- $\{(b : \neg a \ / \ \neg a),\ (a : c \ / \ c)\}$
  $E = Cn(\{a, c\})$

This means that although b and $\neg a$ are not justified under rule-maximization, a and especially c *are* justified under rule-maximization.

Discussion

The difference between rule-maximization and conclusion-maximization can be seen as a possible design-choice when formalizing a certain system for nonmonotonic reasoning. Let us illustrate this with the following example:

"a's are usually b's"
"b's are usually c's"
"c's are usually d's"
"The object in question is an a"

"The object in question is not a d"

$\mathcal{W}$: a $\neg d$

$\mathcal{D}$: $a \Rightarrow b \Rightarrow c \Rightarrow d$

It is clear that in the above example, not all rules can be applied, as this would lead to a conflict. The question then is which of the defeasible rules or defaults should be blamed for entailing the contradiction, and subsequently be disabled.

An obvious approach would be that every defeasible rule is equally blamable. This would lead to the following three rule-maximal sets of defaults:

1. $\mathcal{D}' = \{(a : b \,/\, b),\ (b : c \,/\, c)\},\ E = Cn(\{a, b, c, \neg d\})$

2. $\mathcal{D}' = \{(a : b \,/\, b),\ (c : d \,/\, d)\},\ E = Cn(\{a, b, \neg d\})$

3. $\mathcal{D}' = \{(b : c \,/\, c),\ (c : d \,/\, d)\},\ E = Cn(\{a, \neg d\})$

This, however, is not the approach taken by RDL. Instead of blaming an *arbitrary* rule leading to the conflict, RDL blames the *last* rule leading to the conflict (this can be proven by taking the notion of a process into account). Thus, RDL deliberately leaves out extensions 2 and 3, while only retaining extension 1. The overall approach of RDL is to select a relatively small number of extensions, each of which entails a *maximal* amount of conclusions.

Although this approach may seem arbitrary at first sight, it does have certain advantages, when one for instance looks at it from the perspective of an intelligent agent. The relatively small number of extensions means the agent has relatively little uncertainty, while the maximal number of conclusions in each extension means that the agent is provided with relatively much information. It seems that the aim of RDL is to squeeze as much information (and as little uncertainty) as possible out of a default theory while preserving consistency. It must be mentioned that RDL achieves these properties by leaving out certain possibilities (such as the extensions 2 and 3 in the example above), and examples like *shipment of goods* and *ttc* can make one wonder whether RDL's eagerness does not come at the expense of its carefulness.

5.2.2 Implementing HY-arguments in RDL

Another possibility to achieve a HY-style entailment is to directly add HY-arguments to RDL, in a way that is similar to what was done in the logic of P&S. In order to be able to do so, RDL needs to be interpreted in terms of arguments of chained rules (in this case: arguments of chained defaults). The formalism below is such an interpretation. It combines Dung's assumption-based argument interpretation [Dung95] with Antoniou's process theory [Anto97].

The notion of a *process* is the same as in section 5.2.1, with the important exception that for a process $(d_0, \ldots, d_n)$ it is no longer required that $\neg Jus(d_i) \not\subseteq In(\Pi[i])$. That is, when adding a new default to a sequence, we do not look whether its justifications have already been violated. Instead, we only look whether the prerequisite has been satisfied. Obviously, by dropping the justification-check before applying a default, a much wider

range of unsuccessful (= self-defeating) processes becomes possible, including the processes that we need in order to define HY-arguments.[6]

The following definition takes a default theory and defines an argumentation theory based on it. The idea is to use processes as arguments and let a process defeat another process iff it derives the negation of one or more of the other process's justifications.

Definition 5.13. *For any default theory $T = (\mathcal{W}, \mathcal{D})$, the argumentation theory $AT(T) = (Args_T, defeat_T)$ is defined as follows:*

- $Args_T = \{\Pi \mid \Pi$ *is a finite process of* $T\}$
- Π $defeats_T$ Π' *iff* $\varphi \in In(\Pi)$ *for some* $\varphi \in Out(\Pi')$

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

As an example, take $T = (\mathcal{W}, \mathcal{D})$ with $\mathcal{W} = \{a, d\}$ and $\mathcal{D} = \{(a : b \,/\, b),\ (b : c \,/\, c),\ (d : \neg b \,/\, \neg b)\}$. Here, the processes $((d : \neg b \,/\, \neg b))$ and $((a : b \,/\, b),\ (b : c \,/\, c))$ defeat each other.

Under this process-based interpretation of RDL, a correspondence can be proved between extensions in RDL and stable extensions in the process-based interpretation. This has been taken from [Prak03].[7]

Definition 5.14. *Let T be a default theory.*

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

Theorem 5.11 ([Prak03]). *For any default theory T:*

1. *If S is a stable extension of $AT(T)$, then $Concs(S)$ is a Reiter-extension of T*
2. *If E is a Reiter-extension of T, the $Args(E)$ is a stable extension of $AT(T)$.*

Proof. To prove (1), we first append all processes in S into a sequence of defaults Π and delete each repeated occurrence of every default. Clearly, Π is a process. We claim that Π is a closed and successful process.

Firstly, since S is conflict-free, it follows by definition 5.13 of defeat that $In(\Pi) \cap Out(\Pi) = \emptyset$, so Π is successful.

Next, consider any default d not in Π and suppose that $Pre(d) \in In(\Pi)$. We claim that $In(\Pi) \vdash \neg k$ for some $k \in Jus(d)$. By compactness[8] of first-order logic, $Pre(d)$ is implied by some finite subset of $In(\Pi)$. With this subset a finite subprocess $\Pi[i]$ of Π can be associated. Since d is not an element of Π, we have that $\Pi[i]; d$ is not a subprocess of Π. So by construction of Π we have $\Pi[i]; d \notin S$. Then, since S is stable, S defeats $\Pi[i]; d$. This means that $In(\Pi) \vdash \neg k$ for some $k \in Jus(d)$. Hence Π is closed.

[6] The downside of dropping the justification-check is that theorem 5.7 no longer holds, and thus cannot be used like was done in section 5.2.1. Notice that theorem 5.6 is still unaffected, since the only effect of dropping the justification-check is that more unsuccessful processes come into existence.

[7] We applied a small correction to the proof of theorem 5.11.

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

Next, to prove (2), consider a closed process Π generating E and let $Args(\Pi)$ be the set of all processes whose defaults are a subset of Π (that is: $Args(\Pi) = \{\Pi' \mid defaults(\Pi') \subseteq defaults(\Pi)\}$). Clearly, since Π is successful and closed, we have that $Args(\Pi) = Args(E)$.

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

Until now, the reader may have wondered why we have not explicitly referred to the processes in the above proof as *arguments*. The reason is that we want to reserve the term "argument" for a construct in which a process is embedded. An argument will be a pair $(\mathcal{F}, \Pi)$ where $\mathcal{F}$ is subset of the consequences of the defaults of $\mathcal{D}$ and Π is a process. The idea is that $\mathcal{F}$ contains the foreign commitments of an argument. Classical arguments do not have foreign commitments, so in their case $\mathcal{F}$ is empty. Syntactically, this makes classical arguments a special case of HY-arguments.

Definition 5.15. *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory. An* argument *is a pair $(\mathcal{F}, \Pi)$ where $\mathcal{F} \subseteq Cons(\mathcal{D})$ and Π is a process in $(\mathcal{W} \cup \mathcal{F}, \mathcal{D})$. We say that c is a conclusion of $(\mathcal{F}, \Pi)$ iff $c \in In(\Pi)$.*

Definition 5.16 (classical defeat). *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory and $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments in T. We say that* $(\mathcal{F}_2, \Pi_2)$ classically defeats $(\mathcal{F}_1, \Pi_1)$ *iff $\mathcal{F}_2 = \mathcal{F}_1 = \emptyset$ and Π_2 $defeats_T$ Π_1.*

Adding HY-arguments

The next question is how the above-described argument-based interpretation of RDL can be used for the implementation of HY-arguments. From section 3.2 we recall that the idea of HY-arguments is to use the commitments of the other party against him. A rebutting HY-argument uses the foreign commitments to entail a contradiction, whereas an undercutting HY-argument uses the foreign commitments to entail an undercutter against the original argument. In both cases, the foreign commitments themselves are not allowed to be fc-based in the original arguments.

Before we continue with the formal definition of HY-arguments for RDL, we first need to define the notion of a conclusion (of a process) being *based* on a set of formulas. Once this is defined, we can determine whether a conclusion is fc-based or not.

Definition 5.17 (based on). *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory, $(\mathcal{F}, \Pi)$ be an argument under this default theory, and c be a conclusion of this argument. We say that c is* based on $\mathcal{F}$ *iff there is no process Π' under $(\mathcal{W} \backslash \mathcal{F}, \mathcal{D})$ that is a weak sublist of Π and has c as a conclusion.*

example

Let $T = (\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{a, g\}$ and
$\mathcal{D} = \{(a : b \,/\, b),\ (c : d \,/\, d),\ (b \wedge d : e \,/\, e),\ (g : e \,/\, e)\}$.
The argument $(\{c\},\ ((a : b \,/\, b),\ (c : d \,/\, d),\ (b \wedge d : e \,/\, e)))$

has conclusions b, d and e (as well as many others).
conclusion c is based on $\{c\}$
conclusion b is not based on $\{c\}$
conclusion d is based on $\{c\}$
conclusion e is based on $\{c\}$

The idea of "based on" is that the formulas in $\mathcal{F}$ are in some way necessary for the derivation of the conclusion in the process Π. Notice that if $\mathcal{F} = \emptyset$ (as is the case for classical arguments) then *no* conclusion of Π is based on $\mathcal{F}$.

There exists a clear connection between the notion of "based on" in definition 5.17 and the notion of "based on" as defined for the logic of P&S (definition 3.16 on page 60). It turns out that if one restricts the default theory so that all defaults conform to the somewhat restricted format of P&S defeasible rules, then these two notions coincide.

Theorem 5.12. *Let $T_{P\&S} = (\mathcal{S}, \mathcal{D}_{\mathcal{P}\&\mathcal{S}})$ be a defeasible theory in the sense of P&S, where $\mathcal{S}$ consists of premises only, and let $T_{RDL} = (\mathcal{W}, \mathcal{D}_{RDL})$ be a default theory in the sense of RDL, such that $\mathcal{W}$ is the smallest set that contains a literal L for every premise $\rightarrow L$ in $\mathcal{S}$, and $\mathcal{D}_{RDL}$ is the smallest set that contains a default $(L_1 \wedge \ldots \wedge L_k : L_{k+1}, \ldots, L_{n-1}, L_n \,/\, L_n)$ for every defeasible rule $L_1 \wedge \ldots \wedge L_k \wedge \sim L_{k+1} \wedge \ldots \wedge \sim L_{n-1} \Rightarrow L_n$ in $\mathcal{D}_{P\&S}$.*
We define a mapping from a P&S argument A to an RDL argument $(\mathcal{F}, \Pi)$ as follows:
Π is the sequence of all mapped defeasible rules in A and $\mathcal{F}$ is the set of all foreign commitments in A.
It now holds that:

1. *If $R_L(A)$ contains foreign commitments, then L is based on $\mathcal{F}$ in the mapped argument $(\mathcal{F}, \Pi)$*
2. *If $R_L(A)$ does not contain any foreign commitments, then L is not based on $\mathcal{F}$ in the mapped argument $(\mathcal{F}, \Pi)$.*

Proof.

1. Let A be an argument such that L is based on one or more foreign commitments, that is: $R_L(A)$ contains fc's. Then $R_L(A)$ actually defines a complete subargument of A. As A contains no different defaults with the same consequents (this is implied by the definition of a P&S-argument), there is no subargument of A with conclusion L not containing foreign commitments. This means there is also no process Π' under $(\mathcal{W}\backslash\mathcal{F}, \mathcal{D})$ that is a weak sublist of Π and has L as conclusion.
2. Let A be an argument such that L is not based on any foreign commitments, that is: $R_L(A)$ does not contain any fc's. Then $R_L(A)$ actually defines a complete subargument of A, a subargument without fc's. This means that there also exists a process Π' under $(\mathcal{W}\backslash\mathcal{F}, \mathcal{D})$ that is a weak sublist of Π and has L as conclusion.

□

The essence of theorem 5.12 is that the notion of "based on" in RDL can be seen as a broadening of the notion of "based on" in the system of P&S; this is because the defaults of RDL have a more general nature then the restricted rule-format of P&S.

Now that the notion of "based on" has been defined for RDL, we can proceed with defining the notion of HY-defeat. Notice that in this definition, the approach is to include *all* possible consequents of the attacked argument as foreign commitments, instead of taking just a subset of them.

Definition 5.18 (HY-defeat). *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory and $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments (classical or HY) in this theory. Argument $(\mathcal{F}_2, \Pi_2)$ HY-defeats argument $(\mathcal{F}_1, \Pi_1)$ iff:*

1. *$\mathcal{F}_2$ is the set of all consequents f of defaults in Π_1 such that f is not based on $\mathcal{F}_1$ in Π_1, and*

2. *Π_2 has a conclusion L that is based on $\mathcal{F}_2$ and Π_1 contains a default with a justification $\neg L$.*

At first sight, this definition somewhat looks like the definition of of HY-undercutting in P&S. The question can be raised whether definition 5.18 only captures the notion of undercutting and not the notion of rebutting. It turns out, however, that rebutting can be seen as a special case of definition 5.18.

Definition 5.19 (HY-rebut). *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory and $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments (classical or HY) under this theory. Argument $(\mathcal{F}_2, \Pi_2)$ HY-rebuts argument $(\mathcal{F}_1, \Pi_1)$ iff:*

1. *$\mathcal{F}_2$ is the set of all consequents f of defaults in Π_1 such that f is not based on $\mathcal{F}_1$ in Π_1, and*

2. *Π_2 has a conclusion L and a conclusion $-L$, where either L or $-L$ (or both) is based on $\mathcal{F}_2$.*

Theorem 5.13. *Let $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments. If $(\mathcal{F}_2, \Pi_2)$ HY-rebuts $(\mathcal{F}_1, \Pi_1)$ then $(\mathcal{F}_2, \Pi_2)$ HY-defeats $(\mathcal{F}_1, \Pi_1)$.*

Proof. Let $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments such that $(\mathcal{F}_2, \Pi_2)$ HY-rebuts $(\mathcal{F}_1, \Pi_1)$. Then condition (1) of the definition of HY-defeat is already fulfilled. As for condition (2), the following can be said. Π_2 has a conclusion that is based on $\mathcal{F}_2$ (this follows from definition 5.19). This means that $\mathcal{F}_2$ is not empty, which also means that Π_1 contains at least one default. Call this default d. Let $j \in Jus(d)$. It then holds that because $(\mathcal{F}_2, \Pi_2)$ has conclusion L and $\neg L$, it also has conclusion $\neg j$. And because either L or $\neg L$ (or both) is based on $\mathcal{F}_2$, $\neg j$ is also based on $\mathcal{F}_2$. □

In the system of P&S, it was argued that it is desirable for a HY-rebut to be symmetrical, just like a classical rebut is symmetrical. In order to achieve this, we define the notion of reverse HY-rebutting.

Definition 5.20 (reverse HY-rebut). *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory and $(\mathcal{F}_1, \Pi_1)$ and $(\mathcal{F}_2, \Pi_2)$ be two arguments. We say that $(\mathcal{F}_2, \Pi_2)$* reverse HY-rebuts *$(\mathcal{F}_1, \Pi_1)$ iff $(\mathcal{F}_1, \Pi_1)$ HY-rebuts $(\mathcal{F}_2, \Pi_2)$ and $\Pi_2 \neq \emptyset$.*

The complete definition of HY-enriched defeat contains three subcases:

Definition 5.21 (defeat HY-enriched RDL arguments). *An argument* $(\mathcal{F}_2, \Pi_2)$ *defeats an argument* $(\mathcal{F}_1, \Pi_1)$ *iff:*

- $(\mathcal{F}_2, \Pi_2)$ *classically defeats* $(\mathcal{F}_1, \Pi_1)$*, or*
- $(\mathcal{F}_2, \Pi_2)$ *HY-defeats* $(\mathcal{F}_1, \Pi_1)$*, or*
- $(\mathcal{F}_2, \Pi_2)$ *reverse HY-rebuts* $(\mathcal{F}_1, \Pi_1)$*.*

It is interesting to see that if one restricts oneself to classical defeat, and requires $\mathcal{F}$ to be empty, the result is equivalent to the argumentation framework for classical RDL. Our HY-enrichment therefore essentially boils down to broadening the set of possible arguments as well as broadening the defeat relation between these arguments.

The HY-enrichment is consistency-preserving. That is, if $\mathcal{W}$ is consistent then every stable extension of the HY-enriched system is also consistent, as far as classical arguments are concerned.

Theorem 5.14. *Let* $T = (\mathcal{W}, \mathcal{D})$ *be a default theory where* $\mathcal{W}$ *is consistent, and let* S *be a stable extension in the HY-enriched argumentation framework of* T*. It then holds that the union of all conclusions of the classical arguments of* S *is consistent.*

Proof. Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory where $\mathcal{W}$ is consistent, and S be a stable extension. Let S' be the set of processes of the classical arguments in S. Now, append all processes of S' into a sequence of defaults Π and delete each repeated occurrence of every default. Clearly, Π is a process. Also, since S' is conflict-free (this is because S is conflict-free) it follows from the definition of defeat that $In(\Pi) \cap Out(\Pi) = \emptyset$, so Π is successful. But if Π is successful, then it cannot entail an inconsistency (that is: $\bot \notin In(\Pi)$).

This can be seen as follows. Suppose Π *can* derive an inconsistency (that is: $\bot \in In(\Pi)$). We then distinguish two possible cases:

1. Π contains at least one default. Can this default d and let $j \in Jus(d)$. Then because $\bot \in In(\Pi)$ it also holds that $\neg j \in In(\Pi)$. Thus $In(\Pi) \cap Out(\Pi) \neq \emptyset$ so Π would not be successful. Contradiction.

2. Π does not contain any defaults (that is: Π is empty). The fact that $\bot \in In(\Pi)$ then means that $\mathcal{W}$ would be inconsistent, which is against our assumption. Contradiction.

The fact that $\bot \notin In(\Pi)$ also means that $\bot \notin Cn(\{In(\Pi_i) \mid \Pi_i$ is a process whose defaults are a subset of the defaults of $\Pi\})$. Because every element of S' is also a process whose defaults are a subset of the defaults of Π, it also holds that $\bot \notin Cn(\{In(\Pi_i) \mid \Pi_i \in S'\})$. □

The last remaining thing to do is that the concept of a justified conclusion is defined. For this, we take stable semantics, and a sceptical approach, just as was done for classical RDL.

Definition 5.22. *A conclusion* c *is* justified *iff:*

1. *there is at least one stable extension, and*
2. *every stable extension contains a classical argument with* c *as conclusion.*

Examples

It is interesting to see how the HY-enriched RDL can be applied to some of the examples treated earlier.

example (Ajax-Feijenoord)

$T = (\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{af\}$
$\mathcal{D} = \{(af : \neg p, t \ / \ t),\ (t : p \ / \ p)\}$

Argument A_1: $(\emptyset,\ ((af : \neg p, t \ / \ t)))$
Argument A_2: $(\{t\},\ ((t : p \ / \ p)))$

Here, argument A_2 HY-defeats argument A_1.

example (shipment of goods)

$T = (\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{tma, \neg is\}$
$\mathcal{D} = \{(tma : a \ / \ a),\ (\neg is : \neg cd \ / \ \neg cd),\ (a : cd \ / \ cd)\}$

Argument A_1: $(\emptyset,\ ((tma : a \ / \ a)))$
Argument A_2: $(\{a\},\ ((a : cd \ / \ cd),\ (\neg is : \neg cd \ / \ \neg cd)))$

Here, argument A_2 has conclusion $cd \wedge \neg cd$ (so also conclusion $\neg a$) which is based on $\{a\}$, so A_2 HY-defeats A_1.

example (tax relief)

$(\mathcal{W}, \mathcal{D})$ with
$\mathcal{W} = \{pmp\}$
$\mathcal{D} = \{(pmp : tr \ / \ tr),\ (tr : bd \ / \ bd),\ (bd : fb \ / \ fb),\ (fb : \neg tr \ / \ \neg tr)\}$

Argument A_1: $(\emptyset,\ ((pmp : tr \ / \ tr)))$
Argument A_2: $(\{tr\},\ ((tr : bd \ / \ bd),\ (bd : fb \ / \ fb),\ (fb : \neg tr \ / \ \neg tr)))$

Here, argument A_2 has conclusion $\neg tr$ which is based on $\{tr\}$, so A_2 defeats A_1.

On the existence of extensions

A well-known problem with respect to standard RDL is the possibility that for a specific default theory, no extensions exist. Take for instance the earlier mentioned Ajax-Feijenoord example: $\mathcal{W} = \{af\}$, $\mathcal{D} = \{(af : \neg p, t \ / \ t),\ (t : p \ / \ p)\}$. If only classical arguments are allowed, then no extensions exists. This can be seen as follows. Basically three arguments can be constructed using this default theory: $A_1 = (\emptyset,\ ())$, $A_2 = (\emptyset,\ ((af : \neg p, t \ / \ t)))$ and $A_3 = (\emptyset,\ ((af : \neg p, t \ / \ t),\ (t : p \ / \ p)))$. The last argument is self-defeating. This gives us

the defeat-relation of figure 5.3.

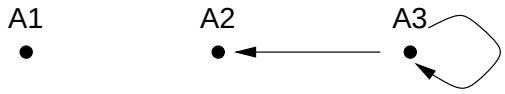


Figure 5.3: Ajax-Feijenoord example without HY-arguments

The problem with the defeat-relation illustrated in figure 5.3 is that there exists a self-defeating argument (A_3) that is not defeated by any argument other than itself. Obviously, for any stable extension of arguments S it holds that either $A_3 \in S$ or $A_3 \notin S$. If $A_3 \in S$ then S would not be conflict-free, since A_3 is self-defeating; hence S would not be a stable extension. If, on the other hand $A_3 \notin S$, then S does not contain any argument that defeats A_3; hence S would not be a stable extension either. As this holds for any S, it can easily be seen that argument A_3 causes the fact that no stable extension (and therefore also no Reiter-extension) exists at all.

With the introduction of HY-arguments, the situation becomes different. The reason is that now a counterargument against A_3 is available: $A_5 = (\{t, p\}, ())$. In addition, a counterargument against A_2 is available: $A_4 = (\{t\}, ((t : p \ / \ p)))$. The situation then becomes as in figure 5.4. Here a stable extension exists of the form $\{A_1, A_4, A_5\}$. The conclusions of this stable extension are limited to $Cn(\{af\})$.

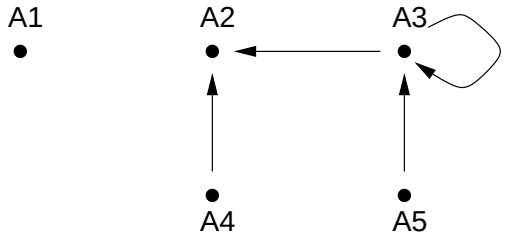


Figure 5.4: Ajax-Feijenoord example with HY-arguments

The Ajax-Feijenoord example illustrates that sometimes, the addition of HY-arguments results in the existence of an extension where without HY-arguments no extension would exist. In figure 5.3, the problem is caused because there is a self-defeating argument (A_3) that is defeated by no argument other than itself. The addition of HY-arguments makes sure that self-defeating arguments (like A_3) *are* defeated, by an argument (in this case A_5) that itself is undefeated. This observation is not limited to the Ajax-Feijenoord example. Instead, it is a general property, as stated in the lemma 5.3.

Lemma 5.3. *Let $T = (\mathcal{W}, \mathcal{D})$ be a default theory and let $(\emptyset, \Pi)$ be a classical argument that is self-defeating. There exists an undefeated HY-argument that defeats $(\emptyset, \Pi)$.*

Proof. Let $(\emptyset, \Pi)$ be a classical argument in T that is self-defeating. Then it holds that $In(\Pi) \cap Out(\Pi) \neq \emptyset$. Let $c \in In(\Pi) \cap Out(\Pi)$. Now take $(\mathcal{F}, ())$, with $\mathcal{F}$ as the set of all consequents of the defaults in Π. It then holds that $\mathcal{W} \cap \mathcal{F} \vDash c$, thus $c \in In(())$, which means that $(\mathcal{F}, ())$ HY-defeats $(\emptyset, \Pi)$. Furthermore, because the process of $(\mathcal{F}, ())$ is empty, $(\mathcal{F}, ())$ is not defeated by any argument. □

While classical arguments can be self-defeating, self-defeating HY-arguments do not exist.

Lemma 5.4. *Let* $(\mathcal{F}, \Pi)$ *be an argument with* $\mathcal{F} \neq \emptyset$. $(\mathcal{F}, \Pi)$ *does not defeat* $(\mathcal{F}, \Pi)$.

Proof. Let $(\mathcal{F}, \Pi)$ be an argument with $\mathcal{F} \neq \emptyset$. Suppose $(\mathcal{F}, \Pi)$ defeats $(\mathcal{F}, \Pi)$. Then according to the definition of HY-defeat (definition 5.18), every $f \in \mathcal{F}$ must be the consequent of some default in Π such that f is not based on $\mathcal{F}$ in Π. This means that $\mathcal{F}$ is in some sense "superfluous"; we don't need $\mathcal{F}$ in order to derive the conclusions of Π. That is, none of the conclusions of Π are actually based on $\mathcal{F}$. The fact that $(\mathcal{F}, \Pi)$ HY-defeats $(\mathcal{F}, \Pi)$, however, means (by means of definition 5.18) that Π has a conclusion L that is based on $\mathcal{F}$. Contradiction. □

Although HY-arguments provide a way of dealing with self-defeating arguments, and therefore sometimes can prevent the absence of extensions, HY-arguments by themselves are not sufficient to *guarantee* the existence of extensions. As an example: in the case of $\mathcal{W} = \emptyset$ and $\mathcal{D} = \{(true : \neg a, b \ / \ b),\ (true : \neg b, c \ / \ c),\ (true : \neg c, a \ / \ a)\}$, no stable extensions exist, either with or without HY-arguments.

In his original 1980 paper, Reiter shows that normal default theories always have at least one extension. It turns out that for normal default theories, the existence of extensions is also guaranteed when HY-arguments are allowed. The proof of this uses the following lemma.

Lemma 5.5. *Let* $T = (\mathcal{W}, \mathcal{D})$ *be a normal default theory. Let* $(\mathcal{F}_2, \Pi_2)$ *be an argument with* $\mathcal{F}_2 \neq \emptyset$ *and* $\Pi_2 \neq ()$ *that defeats* $(\mathcal{F}_1, \Pi_1)$ *in* T. *It holds that* $(\mathcal{F}_1, \Pi_1)$ *also defeats* $(\mathcal{F}_2, \Pi_2)$.

Proof. Let $(\mathcal{F}_2, \Pi_2)$ be an argument with $\mathcal{F}_2 \neq \emptyset$ that defeats $(\mathcal{F}_1, \Pi_1)$ in T. Because $\mathcal{F}_2 \neq \emptyset$ it holds that $(\mathcal{F}_2, \Pi_2)$ is not a classical argument and thus cannot classically defeat $(\mathcal{F}_1, \Pi_1)$. This leaves us with only two possible remaining types of defeat:

1. $(\mathcal{F}_2, \Pi_2)$ HY-defeats $(\mathcal{F}_1, \Pi_1)$. As T is a normal default theory, the only form of defeat is rebutting, and thus $(\mathcal{F}_2, \Pi_2)$ HY-rebuts $(\mathcal{F}_1, \Pi_1)$. This also means that $(\mathcal{F}_1, \Pi_1)$ reverse HY-rebuts $(\mathcal{F}_2, \Pi_2)$, and thus $(\mathcal{F}_1, \Pi_1)$ defeats $(\mathcal{F}_2, \Pi_2)$.

2. $(\mathcal{F}_2, \Pi_2)$ reverse HY-rebuts $(\mathcal{F}_1, \Pi_1)$. Then it holds that $(\mathcal{F}_1, \Pi_1)$ HY-rebuts $(\mathcal{F}_2, \Pi_2)$, and thus $(\mathcal{F}_1, \Pi_1)$ defeats $(\mathcal{F}_2, \Pi_2)$.

□

Theorem 5.15. *Let* $T = (\mathcal{W}, \mathcal{D})$ *be a normal default theory. In the HY-enriched argumentation framework* T *has at least one stable extension.*

Proof. Let $T = (\mathcal{W}, \mathcal{D})$ be a normal default theory. Without HY-arguments, there always exists at least one Reiter-extension E [Reit80, p. 95]. According to theorem 5.11, this means there also exists a stable extension S of arguments. Now do the following:

1. Take all "strict" HY-arguments (that is, HY-arguments not containing any defaults) and add them to S. The resulting S will still be non-conflicting (for otherwise one of the classical arguments in S would have been self-defeating, so S could not have been a stable extension).

2. For each "non-strict" HY-argument A (that is, an HY-argument that *does* contain at least one default) that is not attacked by S, add it to S. Because defeat is symmetric (due to lemma 5.5), A will also not attack S.

The result is a set that is still conflict-free and attacks all arguments not in it: a stable extension. □

5.2.3 Implementing free defaults in RDL

A third alternative way of incorporating HY-style reasoning into RDL is the use of *free defaults*. A free default (sometimes also also called a *supernormal default* [Anto97]) is a default of the form $(true : w \ / \ w)$. Therefore free defaults are a special form of normal defaults. The difference boils down to the way the defeasible rules are represented. A defeasible rule $p \Rightarrow q$ can be represented as $(p : q \ / \ q)$ using a normal default, but if the representation is limited to that of free defaults, it must be represented as $(true : p \supset q \ / \ p \supset q)$.

Free defaults and rule-maximization

Using free defaults instead of general normal defaults also boils down on the difference between rule-maximality and conclusion-maximality. This can be illustrated with the following example.

> Let $\mathcal{W} = \{a, \neg d\}$ and $\mathcal{D}_a$ be a set of abstract defeasible rules $\{a \Rightarrow b,\ b \Rightarrow c,\ c \Rightarrow d\}$. Let $\mathcal{D}_n$ be the representation of $\mathcal{D}_a$ using normal defaults, that is, $\mathcal{D}_n = \{(a : b \ / \ b),\ (b : c \ / \ c),\ (c : d \ / \ d)\}$. Let $\mathcal{D}_f$ be the representation of $\mathcal{D}_a$ using free defaults, that is, $\mathcal{D}_f = \{(true : a \supset b \ / \ a \supset b),\ (true : b \supset c \ / \ b \supset c),\ (true : c \supset d \ / \ c \supset d)\}$
>
> $(\mathcal{W}, \mathcal{D}_n)$ has one extension: $Cn(a, b, c, \neg d)$. This corresponds to the maximal argument $\rightarrow a,\ a \Rightarrow b,\ b \Rightarrow c,\ \rightarrow \neg d$.
>
> $(\mathcal{W}, \mathcal{D}_f)$ has three extensions: $Cn(a,\ a \supset b,\ b \supset c,\ \neg d)$, $Cn(a,\ b \supset c,\ c \supset d,\ \neg d)$ and $Cn(a,\ a \supset b,\ c \supset d,\ \neg d)$. These correspond to the rule-maximal set of rules $\{a \supset b,\ b \supset c\}$, $\{a \supset c,\ c \supset d\}$ and $\{a \supset b,\ c \supset d\}$. [9].

The next step is to state the correlation between normal defaults and conclusion maximization at one hand, and between free defaults and rule maximization at the other hand.

In section 5.2.1 it was shown that RDL is based upon the notion of conclusion maximality. It is not difficult to see that by using free defaults, one obtains a logic that is based on rule maximality.

Theorem 5.16. *Let E be a set of formulas and R a free default theory. E is a Reiter-extension of T iff $E = Cn(\mathcal{W} \cup \mathcal{D}_{max})$ for some $\mathcal{D}_{max}$ that is a maximal subset of propositions $p_i \supset q_i$ (corresponding to abstract rules $p_i \Rightarrow q_i$ in $\mathcal{D}_a$) such that $\mathcal{W} \cup \mathcal{D}_{max}$ is consistent.*

[9] Notice that the rules of the rule-maximal set of rules consist of material implications. This is another effect of using free defaults as a representation.

Proof.
"$\Longrightarrow$":
Let E be a Reiter-extension. Then, according to theorem 5.6 there exists a complete and successful process Π with $In(\Pi) = E$. Now take $Cons(\Pi)$ as the set of all consequents of defaults in Π. It now holds that $Cons(\Pi) = \mathcal{D}_{max}$.

"$\Longleftarrow$":
Let $\mathcal{D}_{max}$ be a maximal set of formulas $p_i \supset q_i$ (corresponding to abstract rules $p_i \Rightarrow q_i$) such that $\mathcal{W} \cup \mathcal{D}_{max}$ is consistent. Let $E = Cn(\mathcal{W} \cup \mathcal{D}_{max})$. It is not difficult to use this $\mathcal{D}_{max}$ to construct a closed process Π such that $In(\Pi) = E$. From theorem 5.7 it follows that this process is successful, and from theorem 5.6 it follows that E is a Reiter-extension. □

Simulation of HY-arguments by free default theories

The next thing to notice is that for normal default theories, the functionality of classical defeat, as well as of HY-defeat, is simulated by free default theories. That is, whenever a normal defaults argument A has a defeater (classical or HY), the associated free defaults argument A' also has a defeater.

In order to proof this, we first define a function that converts normal defaults to their associated free default counterparts.

Definition 5.23. *Let $f_{default}$ be a function such that*
$f_{default}(a : b \,/\, b) = (true : (a \supset b) \,/\, (a \supset b))$.
Let $f_{process}$ be a function such that
$f_{process}(d_0, d_1, \ldots, d_n) = (f_{default}(d_0), f_{default}(d_1), \ldots, f_{default}(d_n))$.
Let f_{theory} be a function such that
$f_{theory}(\mathcal{D}) = \{f_{default}(d) \mid d \in \mathcal{D}\}$.

When no confusion can arise, we simply write f for $f_{default}$, $f_{process}$ or f_{theory}.

It now holds that every classical argument A consisting of normal defaults has an associated classical argument A' consisting of free defaults that has exactly the same conclusions.

Lemma 5.6. *Let $A = (\emptyset, \Pi)$ be a classical argument in a normal default theory $(\mathcal{W}, \mathcal{D})$, let $A' = (\emptyset, f(\Pi))$, and let c be a formula. c is a conclusion of A in $(\mathcal{W}, \mathcal{D})$ iff c is a conclusion of A' in $(\mathcal{W}, f(\mathcal{D}))$.*

Proof.
"$\Longrightarrow$":
By induction over the length of the process:

Basis Every conclusion of $\Pi[0]$ is also a conclusion of $f(\Pi[0])$. This is because $In(\Pi[0]) = \mathcal{W} = In(f(\Pi[0]))$.

Step Suppose every conclusion of $\Pi[i]$ is also a conclusion of $f(\Pi[i])$. Let without loss of generality $d_i = (a : b \,/\, b)$. The fact that Π is a process means that a is a conclusion of $\Pi[i]$, and thus, by means of the induction hypothesis, that a is a conclusion of $f(\Pi[i])$. Now, $f(\Pi[i+1])$ contains as additional consequent (and therefore also as additional conclusion) $a \supset b$. Thus, $f(\Pi[i+1])$ also contains conclusion b. Thus, $f(\Pi[i+1])$ contains at least all conclusions of $\Pi[i+1]$.

"$\Longleftarrow$":
This follows from the fact that the consequents of $f(\Pi)$ (like $a \supset b$) are weaker than the consequents of Π (like b). □

Now, at first sight it might seem that, in a normal default theory, if process Π_2 defeats process Π_1, then $f(\Pi_2)$ also defeats $f(\Pi_1)$. This, however, is is not the case. Consider the following example.

$$\begin{array}{rcl} \mathcal{W} & = & \{a\} \\ \mathcal{D} & = & \{(a : b \,/\, b),\ (b : c \,/\, c),\ (a : d \,/\, d),\ (d : \neg c \,/\, \neg c)\} \\ \Pi_1 & = & ((a : b \,/\, b),\ (b : c \,/\, c)) \\ \Pi_2 & = & ((a : d \,/\, d),\ (d : \neg c \,/\, \neg c)) \end{array}$$

Here Π_2 defeats Π_1.
Now, take the free default versions of these processes.

$$\begin{array}{rcl} f(\Pi_1) & = & ((true : (a \supset b) \,/\, (a \supset b)),\ (true : (b \supset c) \,/\, (b \supset c))) \\ f(\Pi_2) & = & ((true : (a \supset d) \,/\, (a \supset d)),\ (true : (d \supset \neg c) \,/\, (d \supset \neg c))) \end{array}$$

Here, $f(\Pi_2)$ does not defeat $f(\Pi_1)$.
$f(\Pi_1)$, however, is defeated by another free defaults process:

$$\begin{array}{rcl} \Pi_3 & = & ((true : (a \supset b) \,/\, (a \supset b)),\ (true : (a \supset d) \,/\, (a \supset d)), \\ & & (true : (a \supset \neg c) \,/\, (a \supset \neg c))) \end{array}$$

This process has conclusions b, d and $\neg c$, so also $\neg(b \supset c)$. Thus, it defeats Π_1.

The idea that if an argument (A) has a defeater (B, classical or HY), then the free default version of this argument ($f(A)$) also has a defeater (B', with B' not necessarily equal to $f(B)$), is stated in the following two theorems.

Theorem 5.17. *Let $(\mathcal{W}, \mathcal{D})$ be a normal default theory and let $(\emptyset, \Pi)$ be a classical argument in $(\mathcal{W}, \mathcal{D})$ that has a classical coherent defeater in $(\mathcal{W}, \mathcal{D})$. Then $(\emptyset, f(\Pi))$ is a classical argument in $(\mathcal{W}, f(\mathcal{D}))$ that has a classical coherent defeater in $(\mathcal{W}, f(\mathcal{D}))$.*

Proof. Let $(\emptyset, \Pi_1)$ be a classical argument in $(\mathcal{W}, \mathcal{D})$ and let $(\emptyset, \Pi_2)$ be a minimal argument in $(\mathcal{W}, \mathcal{D})$ that defeats $(\emptyset, \Pi_1)$. This means that Π_2 has a conclusion c while Π_1 contains an assumption $\neg c$. Because all defaults are normal, Π_1 also has a conclusion $\neg c$. Thus, $\Pi_1; \Pi_2$ is inconsistent (it has conclusion $\bot$), and therefore it also holds that $f(\Pi_1; \Pi_2)$ is inconsistent (this follows from lemma 5.6). Now take Π_3 as a minimal subargument (weak sublist) of $f(\Pi_1; \Pi_2)$ that is still inconsistent. Π_3 contains at least one default (say d) from $f(\Pi_1)$ (otherwise Π_2 would have been incoherent). If one would leave out d from Π_3 then the result would not be inconsistent anymore, and thus also not incoherent. This means that $\Pi_3 - d$ has conclusion $\neg Cons(d)$, and therefore (free defaults are a special form of normal defaults) conclusion $\neg Jus(d)$. That is, $\Pi_3 - d$ is a coherent (classical) defeater of $f(\Pi_1)$. □

Theorem 5.18. *Let $(\mathcal{W}, \mathcal{D})$ be a normal default theory and let $(\emptyset, \Pi)$ be a classical argument in $(\mathcal{W}, \mathcal{D})$ that has a HY-defeater in $(\mathcal{W}, \mathcal{D})$. Then $(\emptyset, f(\Pi))$ is a classical argument in $(\mathcal{W}, f(\mathcal{D}))$ that has a classical coherent defeater in $(\mathcal{W}, f(\mathcal{D}))$.*

Proof. Let $(\emptyset, \Pi_1)$ be a classical argument in $(\mathcal{W}, \mathcal{D})$, and let $(\mathcal{F}, \Pi_2)$ be a minimal HY-defeater of $(\emptyset, \Pi_1)$. The fact that $(\mathcal{F}, \Pi_2)$ is a HY-defeater of $(\emptyset, \Pi_1)$ means that $\Pi_1; \Pi_2$ is inconsistent (it has conclusion $\bot$), and therefore that $f(\Pi_1; \Pi_2)$ is inconsistent (this follows from lemma 5.6). Now take Π_3 as a minimal subargument (weak sublist) of $f(\Pi_1; \Pi_2)$

that is still inconsistent. Π_3 contains at least one default (say d) from $f(\Pi_1)$ (otherwise the inconsistency shown by $(\mathcal{F}, \Pi_2)$ would not be due to A, so $(\mathcal{F}, \Pi_2)$ would not be a HY-argument). If one would leave out d from Π_3 then the result would not be inconsistent anymore, and thus also not incoherent. This means that $\Pi_3 - d$ has conclusion $\neg Cons(d)$, and therefore (free defaults are a special form of normal defaults) conclusion $\neg Jus(d)$. That is, $\Pi_3 - d$ is a coherent (classical) defeater of $f(\Pi_1)$. □

In addition to HY-arguments, many other kinds of arguments are also simulated by free defaults. Examples include arguments involving contraposition ($\mathcal{W} = \{\neg c\}, \mathcal{D}_a = \{a \Rightarrow b,\ b \Rightarrow c\}$) and arguments involving disjunction ($\mathcal{W} = \{a \vee b\}, \mathcal{D} = \{a \Rightarrow c,\ b \Rightarrow c\}$). An overview of the properties satisfied by free defaults, the reader is referred to the discussion of THEORIST in the Handbook of Logic in AI and Logic Programming [Maki94].

From the above discussion it follows that a particularly easy way to implement a HY-style of reasoning into RDL would be the use of free defaults, which results in a system that can be compared with Poole's THEORIST. The price of doing so, however, is that with the addition of HY-arguments many other styles of reasoning are also added (such as contraposition and disjunction) as these properties are inherently connected to the use of free defaults. The advantage of explicitly defining HY-arguments, as was done in section 3.3.1 (P&S) and section 5.2.2 (RDL) is that it allows us to study these arguments (and their effects) as an entity of its own, without necessarily involving other principles.

Another consideration is that the way in which explicit HY-arguments are used is quite similar to the way that people actually argue. Even if the effects of HY-arguments can also be reached by other formalisms, similarity between the process of formal reasoning and the process of human reasoning has certain advantages.

Free defaults and other systems

The concept of using free defaults is closely related to other systems for nonmonotonic reasoning. In THEORIST [Pool88] knowledge is represented in a fact base $\mathcal{F}$ and a set of hypotheses Δ. The idea is to take extensions as maximal subsets D of Δ that are consistent with $\mathcal{F}$. A proposition p is then called *explainable* iff it follows from an extension. Poole proves his system to be equivalent to RDL with free defaults — the proof is slightly different from what we have done. In addition, Poole uses a construction of *named hypotheses* that allows for the hypotheses to be undercut. THEORIST also has the notion of *constraints*, which prohibit the derivation of certain (undesirable) conclusions, while not entailing the negation of these conclusions by themselves. For a more elaborate discussion, we refer to Poole's original paper [Pool88].

Under some restrictions, similarity to free defaults can also be achieved for circumscription. If we have an abstract defeasible theory $(\mathcal{W}, \mathcal{D}_a)$ where every defeasible rule has the form $P_i(x) \Rightarrow Q_i(x)$, then this could be represented in circumscription as $P_i(x) \wedge \neg \mathtt{ab}_i(x) \supset Q_i(x)$, which is equivalent to $\neg \mathtt{ab}_i(x) \supset (P_i(x) \supset Q_i(x))$. If we assume that the $\mathtt{ab}_i$ predicates are circumscribed and the other predicates are allowed to vary, then on a semantical level, one only takes into account the models that *minimize* the $\mathtt{ab}_i$ predicates. This means that one selects models in which a *maximal* set of statements $P_i(x) \supset Q_i(x)$ is true, thus achieving rule-maximality for material implications, an effect that is similar to that of using free defaults.

Similar remarks can be made about Reiter's application of default logic for the purpose of diagnosis [Reit87]. Under this theory, a defeasible rule $p_i \Rightarrow q_i$ is represented by a

classical logic formula $p_i \wedge \neg \mathtt{ab}_i \supset q_i$ and a default $(true : \neg \mathtt{ab}_i / \neg \mathtt{ab}_i)$ which results in a solution similar to that of circumscription.

5.3 Pollock's system

The last remaining system to be treated in this thesis is that of John Pollock. Pollock has been studying defeasible reasoning since his Ph.D. thesis of 1965 and has published a significant number of papers on this subject. This has resulted in a formalism for defeasible reasoning, as well as a concrete implementation of this formalism, which Pollock calls OSCAR. In this thesis, however, we will limit ourselves to the formalism itself, leaving out the issue of implementation. During his years of research, Pollock has produced different versions of his formalism. In this thesis, we focus on two of these versions:

- the one based on grounded semantics [Poll87, Poll92] (section 5.3.1)
- the one resulting from a thorough analysis of self-defeating arguments [Poll95] (section 5.3.2)

In section 5.3.1 we start with a summary of Pollock's grounded semantics based system. Notice that this summary is partly based on the Handbook of Philosophical Logic [PrVr02].

5.3.1 Pollock's grounded semantics based system

In the system of Pollock, arguments are constructed by means of reasons. Just like P&S distinguish two kinds of rules (strict and defeasible), Pollock distinguishes two kinds of reasons: conclusive and prima facie.

Conclusive reasons are reasons that logically entail their conclusions. A conclusive reason is any valid form of first order deduction. The following are examples of conclusive reasons.

$\{p,\ p \supset q\}$ is a conclusive reason for q
$\{\exists x : Px\}$ is a conclusive reason for $\neg \forall x : \neg Px$

Prima facie reasons, at the other hand, are not necessarily valid in first order logic; they only create a presumption in favour of their conclusion. This presumption can be defeated by other reasons, depending on the strength of the conflicting reasons. Pollock distinguishes several kinds of prima facie reasons, for instance principles of perception, such as:

$\lceil x$ appears to me as $Y \rceil$ is a prima facie reason for believing $\lceil x$ is $Y \rceil$[10]

Another source of prima facie reasons is the statistical syllogism:

If $(r > 0.5)$ then $\lceil x$ is an F and $prob(G/F) = r \rceil$ is a prima facie reason of strength r for believing $\lceil x$ is a $G \rceil$.

[10]Here $\lceil . \rceil$ stands for the *objectification* operator. With this operator, expressions in the meta-language are translated into expressions in the object-language.

Other sources of prima facie reasons are also available [Poll95].

Although Pollock sometimes defines defeat in terms of inference graphs, we will instead use the equivalent argument interpretation of Prakken and Vreeswijk [PrVr02]. As an example of a possible argument, consider the following information. The object on one's plate appears as beef. Beef is a particular kind of meat. Furthermore, say, 80% of al meat comes from the bio-industry. In Pollock's system, this can be formalized as follows (assume that the perception principle has a strength of 0.9):

$$\texttt{INPUT} = \{\texttt{appears}(\texttt{beef}(\texttt{o})),\ \texttt{beef}(x) \supset \texttt{meat}(x),\ \lceil prob(\texttt{bio\_ind}(x)/\texttt{meat}(x)) = 0.8\rceil\}$$

1. $\langle \texttt{appears}(\texttt{beef}(o)),\ \infty\rangle$	($\texttt{appears}(\texttt{beef}(o))$ is in INPUT)
2. $\langle \texttt{beef}(o),\ 0.9\rangle$	(1 and the principle of perception)
3. $\langle \texttt{beef}(o) \supset \texttt{meat}(o),\ \infty\rangle$	($\texttt{beef}(o) \supset \texttt{meat}(o)$ is in INPUT)
4. $\langle \texttt{meat}(o),\ 0.9\rangle$	(2, 3 and $\{p,\ p \supset q\}$ is a conclusive reason for q)
5. $\langle \lceil prob(\texttt{bio\_ind}(x)/\texttt{meat}(x)) = 0.8\rceil,\ \infty\rangle$	($\lceil prob(\texttt{bio\_ind}(x)/\texttt{meat}(x)) = 0.8\rceil$ is in INPUT)
6. $\langle \texttt{meat}(o) \wedge \lceil prob(\texttt{bio\_ind}(x)/\texttt{meat}(x)) = 0.8\rceil,\ 0.9\rangle$	(4, 5 and $\{p, q\}$ is a conclusive reason of $p \wedge q$)
7. $\langle \texttt{bio\_ind}(o),\ 0.8\rangle$	(6 and the statistical syllogism)

In the above argument, which Pollock calls a *linear* argument, each line is a pair consisting of a proposition and a numerical value that indicates the strength, or degree of justification in the proposition. The lines 1, 3 and 5 have a strength ∞ because the propositions of these lines originate from INPUT; they are put forward as absolute facts. At line 7, the *weakest link* principle is applied, with the result that the strength of the argument line is the minimum of the values 0.8 and 0.9 taken from the previous line.

Besides linear arguments, Pollock also specifies *suppositional* arguments. The idea of suppositional reasoning is to "suppose" something that is not derived from other information, draw conclusions from it, and then "discharge" the supposition to obtain a conclusion that no longer depends on the supposition. The way in which Pollock's system deals with suppositional reasoning is very similar to the use of assumptions in natural deduction. As a result of this, each line of inference contains an associated set of suppositions.

Pollock's notion of an argument is made formal in the following definition (taken from [PrVr02]) which is essentially an argument-based interpretation of [Poll95].

Definition 5.24. *Let* INPUT *be a consistent set of first-order formulas. An* argument *based on* INPUT *is a finite sequence $\sigma_1, \ldots \sigma_n$, where each σ_i is a line of argument. A* line of argument *σ_i is a triple $\langle X_i, p_i, \nu_i\rangle$, where X_i, a set of propositions, is the set of* suppositions *of σ_i, p_i is a proposition, and ν_i is the* degree of justification *of σ_i. A new line of argument is obtained from the earlier lines of argument according to one of the following rules of argument formation.*

Input. *If p is in* INPUT *and σ is an argument, then for any X it holds that $\sigma, \langle X, p, \infty\rangle$ is an argument.*

Reason. *If σ is an argument, $\langle X_1, p_1, \eta_1\rangle, \ldots, \langle X_n, p_n, \eta_n\rangle$ are members of σ, and $\{p_1, \ldots, p_n\}$ is a reason of strength ν for q, and for each i, $X_i \subseteq X$, then $\sigma, \langle X, q, \min\{\eta_1, \ldots, \eta_n, \nu\}\rangle$ is an argument.*

Supposition. *If σ is an argument, X a set of propositions and $p \in X$, then $\sigma, \langle X, p, \infty\rangle$ is also an argument.*

Conditionalization. *If σ is an argument and some line of σ is $\langle X \cup \{p\}, q, \nu\rangle$, then $\sigma, \langle X, (p \supset q), \nu\rangle$ is also an argument.*

Dilemma *If σ is an argument and some line of σ is $\langle X, p \vee q, \nu\rangle$, and some line of σ is $\langle X \cup \{p\}, r, \mu\rangle$, and some line of σ is $\langle X \cup \{q\}, r, \xi\rangle$, then $\sigma, \langle X, r, \min\{\nu, \mu, \xi\}\rangle$ is also an argument.*

Pollock [Poll95] notes that other inference rules could be added as well.

The addition of argument formation rules like *supposition*, *conditionalization* and *dilemma* makes it possible to construct *suppositional arguments*, in addition to linear arguments. OSCAR is one of the very few nonmonotonic logics that allow for suppositional reasoning. An example of the usefulness of suppositional arguments is that it enables "reasoning by cases", which is left unsupported in various other logics for nonmonotonic reasoning. For instance, if Dutch people usually like ice-skating, Norwegian people usually like ice-skating, and Sven is either Dutch or Norwegian, then it seems a reasonable conclusion that, presumably, Sven likes ice-skating. In Pollock's system, this argument can be stated as follows.

Given the following reasons:

(1) $\texttt{Dutch}(x)$ is a prima facie reason of strength ν for $\texttt{likes\_skating}(x)$
(2) $\texttt{Norwegian}(x)$ is a prima facie reason of strength μ for $\texttt{likes\_skating}(x)$

Let $\texttt{INPUT} = \{\texttt{Dutch}(Sven) \vee \texttt{Norwegian}(Sven)\}$. The conclusion $\texttt{likes\_skating}(t)$ can be defeasibly derived as follows.

1.	$\langle \emptyset, \texttt{Dutch}(Sven) \vee \texttt{Norwegian}(Sven), \infty\rangle$	($\texttt{Dutch}(Sven) \vee \texttt{Norwegian}(Sven)$ is in $\texttt{INPUT}$)
2.	$\langle \{\texttt{Dutch}(Sven)\}, \texttt{Dutch}(Sven), \infty\rangle$	(Supposition)
3.	$\langle \{\texttt{Dutch}(Sven)\}, \texttt{likes\_skating}(Sven), \nu\rangle$	(2 and prima facie reason (1))
4.	$\langle \{\texttt{Norwegian}(Sven)\}, \texttt{Norwegian}(Sven), \infty\rangle$	(Supposition)
5.	$\langle \{\texttt{Norwegian}(Sven)\}, \texttt{likes\_skating}(Sven), \mu\rangle$	(4 and prima facie reason (2))
6.	$\langle \emptyset, \texttt{likes\_skating}(Sven), \min(\nu, \mu)\rangle$	(3, 5 and Dilemma)

At line 1, the proposition $\texttt{Dutch}(Sven) \vee \texttt{Norwegian}(Sven)$ is put forward as an absolute fact. At line 2, the proposition $\texttt{Dutch}(Sven)$ is temporarily assumed to be true. From this assumption, at line 3 the conclusion $\texttt{likes\_skating}(Sven)$ is defeasibly derived using the first prima facie reason. Line 4 is an alternative continuation of line 1, in which $\texttt{Norwegian}(Sven)$ is supposed to be true. At line 5 this is used to again defeasibly derive $\texttt{likes\_skating}(Sven)$, this time from the second prima facie reason. Finally, at line 6, the Dilemma rule is applied to 3 and 5, discharging the assumptions in the alternative suppositional arguments, and concluding to $\texttt{likes\_skating}(Sven)$ without any assumptions.

Now that the notion of an argument has been explained, we can continue with Pollock's definition of defeat. For now, we use Pollock's old version of defeat [Poll92].

Definition 5.25 (rebut). *An argument σ* rebuts *an argument η iff:*

1. *η contains a line of the form $\langle X, q, \alpha\rangle$ that is obtained by the argument formation rule* Reason *from some earlier lines $\langle X_1, p_1, \alpha_1\rangle, \ldots, \langle X_n, p_n, \alpha_n\rangle$ where $\{p_1, \ldots, p_n\}$ is a prima facie reason for q, and*

2. *σ contains a line of the form $\langle Y, \neg q, \beta\rangle$ where $X \subseteq Y$ and $\beta \geq \alpha$.*

Definition 5.26 (undercut). *An argument σ* undercuts *an argument η iff:*

1. *η contains a line of the form $\langle X, q, \alpha\rangle$ that is obtained by the argument formation rule* Reason *from some earlier lines $\langle X_1, p_1, \alpha_1\rangle, \ldots, \langle X_n, p_n, \alpha_n\rangle$ where $\{p_1, \ldots, p_n\}$ is a prima facie reason for q, and*

2. *σ contains a line of the form $\langle Y, \neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil, \beta\rangle$ where $Y \subseteq X$ and $\beta \geq \alpha$.*

Notice that $\lceil\{p_1, \ldots, p_n\} \gg q\rceil$ is a translation of "$\{p_1, \ldots p_n\}$ is a prima facie reason for q" into the object language.

Definition 5.27 (defeat). *An argument σ* defeats *an argument η iff σ rebuts or undercuts η.*

Regarding justified arguments, Pollock uses the following inductive definition [Poll87].

Definition 5.28 ([Poll87]).

- *All arguments are* in at level 0.
- *An argument is* in at level $n+1$ *($n > 0$) iff it is in at level 0 and it is not defeated by any argument that is in at level n.*
- *An argument is* justified *iff there is an m such that for every $n \geq m$ the argument is in at level n.*

This definition is shown by Dung to be (almost) equivalent to the definition of grounded semantics [Dung95].

5.3.2 Pollock and self-defeating arguments

Pollock has given special consideration to the issue of self-defeating arguments. One of his first remarks is that it may be undesirable for self-defeating arguments to prevent other arguments from becoming justified. Take the example of figure 5.5.

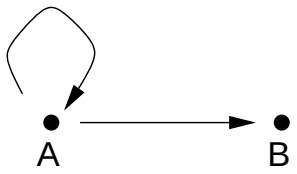


Figure 5.5: A simple example of self-defeat.

Intuitively, one could argue that B should be justified, since the only argument defeating it is self-defeating. Under definition 5.28, however, B is not justified. This can be seen as follows:

level 0: $\{A, B\}$
level 1: $\emptyset$
level 2: $\{A, B\}$
level 3: $\emptyset$
$\vdots$ $\vdots$

Thus, for any n: if n is even then A and B are in at level n, but if n is odd then A and B are not in at level n. Therefore, neither A nor B is justified.

The situation becomes more serious when one considers that self-defeating arguments can be extended to an argument that can defeat an arbitrary (other) argument. Take the following example.

$\texttt{INPUT} = \{A, B, C\}$

A is a prima facie reason for D.
B is a prima facie reason for $\neg D$.
C is a prima facie reason for E.

There now exists an argument that defeats the argument for E.

1. $\langle \emptyset, A, \infty \rangle$ ($A \in \texttt{INPUT}$)
2. $\langle \emptyset, D, \alpha \rangle$ (A is a prima facie reason for D)
3. $\langle \emptyset, B, \infty \rangle$ ($B \in \texttt{INPUT}$)
4. $\langle \emptyset, \neg D, \alpha \rangle$ (B is a prima facie reason for $\neg D$)
5. $\langle \emptyset, \neg E, \alpha \rangle$ ($\{D, \neg D\}$ is a conclusive reason for $\neg E$)

The interaction between the arguments for D, $\neg D$, $\neg E$ and E is shown in figure 5.6. The argument for $\neg E$ now prevents E from becoming justified, even though intuitively E has nothing to do with the conflict between D and $\neg D$.[11]

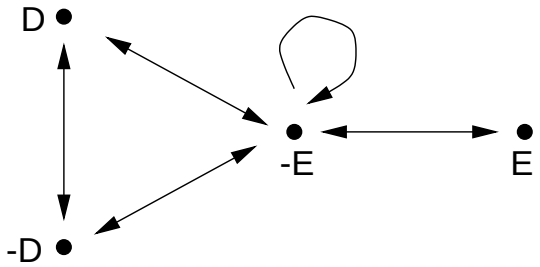


Figure 5.6: A more complex example of self-defeat.

In order to deal with these kinds of problems, Pollock proposes the following alternative to definition 5.28 [Poll87], which aims to prevent self-defeating arguments from competition with other arguments.

Definition 5.29.

- *All non-self-defeating arguments are* in at level 0.
- *An argument is* in at level $n+1$ *($n \geq 0$) iff it is in at level 0 and it is not defeated by any argument that is in at level n.*
- *An argument is* justified *iff there is an m such that for every $n \geq m$ the argument is in at level n.*

[11] With stable semantics, there would be two extensions: $\{D, E\}$ and $\{\neg D, E\}$, both containing E.

Pollock's solution boils down to preventing self-defeating arguments from defeating any other argument, essentially by ignoring them. An alternative approach for neutralizing self-defeating arguments would be to define a special argument (like the empty argument of P&S) that defeats all self-defeating arguments, and that has no arguments defeating it. In either way, the effect is that in figure 5.5, argument B becomes justified and argument A does not become justified.

Pollock's new system

Although the problem of self-defeating arguments seemed to be solved by definition 5.29 [Poll87], Pollock later discovers that the situation is actually more complicated [Poll91a, Poll92, Poll95]. An an illustration, he specifies the following example [Poll91a]:

> Robert says the elephant besides him looks pink (RSELP).
> The fact that Robert says the elephant looks pink is a reason to believe that it is looks pink (ELP).
> The fact that the elephant looks pink is a reason to believe that it is pink (EIP).
> Robert becomes unreliable in the presence of pink elephants (RUPPE).
>
> We can then construct the following argument, assuming that all prima facie reasons have the same strength α.
>
> 1. $\langle \emptyset, \text{RSELP}, \infty \rangle$
> (RSELP $\in$ INPUT)
> 2. $\langle \emptyset, \text{ELP}, \alpha \rangle$
> (RSELP is a prima facie reason for ELP)
> 3. $\langle \emptyset, \text{EIP}, \alpha \rangle$
> (ELP is a prima facie reason for EIP)
> 4. $\langle \emptyset, \text{RUPPE}, \infty \rangle$
> (RUPPE $\in$ INPUT)
> 5. $\langle \emptyset, \text{EIP} \wedge \text{RUPPE}, \alpha \rangle$
> ($\{\text{EIP}, \text{RUPPE}\}$ is a conclusive reason for EIP $\wedge$ RUPPE)
> 6. $\langle \emptyset, \neg\lceil\{\text{RSELP}\} \gg \text{ELP}\rceil, \alpha \rangle$
> (EIP $\wedge$ RUPPE is a prima facie reason for $\neg\lceil\{\text{RSELP}\} \gg \text{ELP}\rceil$)

The key point to notice about the above example is that it concerns a self-defeating argument that has an undercutter that undercuts one of the conclusions it is based on (one could say that "the argument's head bites its tail"). In this respect, Pollock's pink elephant example is similar to our Ajax-Feijenoord example.[12] Pollock argues that in the pink elephant example, not only EIP but also ELP should be prevented from becoming justified. In this respect, Pollock shares the same intuition as us. The problem, however, is that the only classical argument defeating subargument $(1, 2)$ is self-defeating, and as we showed earlier, self-defeating arguments should in general not prevent other arguments from becoming justified. How does one deal with this problem?

At first, Pollock tries to find the solution by generalizing the notion of self-defeat [Poll91a], but later he retreats from this approach and instead tries to solve it using a multiple status assignment [Poll95]. Although Pollock states his new system in terms of

[12] We think, however, that the Ajax-Feijenoord example is more natural.

inference graphs, we will follow the argument-based interpretation of [PrVr02]. In definition 5.30, defeat is defined based on the *last* lines of the respective arguments. The overall defeat status of arguments is then defined by evaluating the status of the relevant subarguments (definition 5.31 and 5.32).

Definition 5.30. *An argument σ defeats another argument η iff:*

1. *η's last line is $\langle X, q, \alpha\rangle$ and is obtained by the argument formation rule* Reason *from some earlier lines $\langle X_1, p_1, \alpha_1\rangle, \ldots \langle X_n, p_n, \alpha_n\rangle$ where $\{p_1, \ldots p_n\}$ is a prima facie reason for q, and*

2. *σ's last line is $\langle Y, r, \beta\rangle$ where $Y \subseteq X$ and either*

   (a) *r is $\neg q$ and $\beta \geq \alpha$, or*

   (b) *r is $\neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil$ and $\beta \geq \alpha$.*

Definition 5.31 (subarguments). *An argument A is a* subargument *of an argument B iff A is a subsequence of B and there exists a tree T of argument lines such that:*

1. *T contains all and only lines from A, and*

2. *T's root is A's last element, and*

3. *l is a child of l' iff l was inferred from a set of lines one of which was l'.*

A proper *subargument of A is any subargument of A unequal to A.*

Definition 5.32. *An assignment of "defeated" or "undefeated" to a closed set S of arguments is a* partial defeat status assignment *iff it satisfies the following conditions:*

1. *All arguments in S with only lines obtained by the* input *argument formation rule are assigned "undefeated"*

2. *$A \in S$ is assigned "undefeated" iff:*

   (a) *All proper sub-arguments of A are assigned "undefeated", and*

   (b) *All arguments in S defeating A are assigned "defeated"*

3. *$A \in S$ is assigned "defeated" iff:*

   (a) *One of A's proper sub-arguments is assigned "defeated", or*

   (b) *A is defeated by an argument in S that is assigned "undefeated".*

A defeat status assignment *is a maximal (with respect to set inclusion) partial defeat status assignment.*

Notice that the conditions (a2) and (3a) on the sub-arguments of A make the weakest link principle hold by definition. The similarity of defeat status assignments to Dung's preferred extensions can be seen as follows: the conditions (2b) and (3b) on the defeaters of A are the analogues of Dung's notion of acceptability, which make a defeat status assignment an admissible set. Then the fact that a defeat status assignment is a maximal partial assignment induces the similarity with preferred extensions.

It is easy to verify that when two arguments defeat each other, an input has more than one status assignment. Since Pollock wants to define a sceptical consequence notion, he therefore has to consider the intersection of all assignments. This leads to the following definition.

Definition 5.33. *Let S be a closed set or arguments based on* `INPUT`*. Then, relative to S, an argument is* undefeated *iff every status assignment to S assigns "undefeated" to it; it is* defeated outright *iff no status assignment to S assigns "undefeated" to it; otherwise it is* provisionally defeated.

It is interesting to see how Pollock's new system deals with the problem of self-defeating arguments.

$\texttt{INPUT} = \{A, B, C\}$

A is a prima facie reason for D.
B is a prima facie reason for $\neg D$.
C is a prima facie reason for E.

Now, the following holds:

1. There is no status assignment that assigns "undefeated" to the argument for $\neg E$. This can be seen as follows. Suppose there is one. Then, according to definition 5.32 all proper subarguments of the argument for $\neg E$ are assigned "undefeated". Thus, the argument for D and the argument for $\neg D$ are undefeated. But from the fact that the argument for D is undefeated, it follows that the argument for $\neg D$ is defeated. Contradiction.
2. Any status assignment that assigns nothing ("defeated" nor "undefeated") to the argument for $\neg E$ also assigns nothing to the argument for E and nothing to the arguments for D and for $\neg D$.
3. There is a status assignment that assigns "defeated" to the argument for $\neg E$, "undefeated" to the argument for E, "defeated" to the argument for D, and "undefeated" to the argument for $\neg D$.
4. There is a status assignment that assigns "defeated" to the argument for $\neg E$, "undefeated" to the argument for E, "undefeated" to the argument for D, and "defeated" to the argument for $\neg D$.

The status assignments 3 and 4 are maximal status assignments. Thus, it overall holds that E is undefeated, $\neg E$ is defeated outright, and D and $\neg D$ are provisionally defeated. This is in line with what one might expect.

As for the pink elephant problem, Pollock's new system deals with it in the following way.

$\texttt{INPUT} = \{\text{RSELP}, \text{RUPPE}\}$

RSELP is a prima facie reason for ELP.
ELP is a prima facie reason for EIP.
$\text{EIP} \wedge \text{RUPPE}$ is a prima facie reason for $\neg\lceil\{\text{RSELP}\} \gg \text{ELP}\rceil$.

Now, suppose there is a status assignment that assigns "undefeated" to the argument for ELP. Then, the argument for EIP is also assigned "undefeated", as well as the argument for EIP $\wedge$ RUPPE. But then the argument for ELP should be assigned "defeated". Contradiction.

Alternatively, suppose that there is a status assignment that assigns "defeated" to ELP. Then the argument for EIP is also assigned "defeated", just as the argument for EIP $\wedge$ RUPPE. But then the argument for ELP would have to be assigned "undefeated". Contradiction.

Thus, nothing ("defeated" nor "undefeated") is assigned to ELP, and therefore also nothing is assigned to EIP and EIP $\wedge$ RUPPE. This means that both ELP, EIP and EIP $\wedge$ RUPPE are defeated outright.

There is some controversy about whether ELP should or should not be defeated outright. Prakken and Vreeswijk argue that although EIP should be defeated, ELP could also be undefeated, because its only defeater is a self-defeating argument, and EIP is not a deductive consequence of ELP [PrVr02]. Perhaps the best way to see why ELP should be defeated is by means of a dialogue.

P: The elephant besides Robert looks pink, because Robert says so.
O: But if the elephant looks pink, then it probably also *is* pink, don't you think?
P: (cannot give any good reason for denying this inference) Euhh, yes...
O: But you know that Robert becomes unreliable in the presence of pink elephants, so how can you maintain that the elephant looks pink in the first place?
P: (understands that his statement ELP has lost grounds) Euhh...

The point is that a rational agent that beliefs ELP and allows for its beliefs to be critically questioned, will soon find out that its belief ELP is based on quicksand. As this holds for any rational agent with this belief, ELP should not be justified.

Unfortunately, there also exists an example that is handled in a somewhat less intuitive way by Pollock's new system.

$\texttt{INPUT} = \{pmp\}$

pmp is a prima facie reason for tr.
tr is a prima facie reason for bd.
bd is a prima facie reason for fb.
fb is a prima facie reason for $\neg tr$.

Here, there exists the following argument.

1. $\langle \emptyset, pmp, \infty \rangle$
2. $\langle \emptyset, tr, \alpha \rangle$
3. $\langle \emptyset, bd, \alpha \rangle$
4. $\langle \emptyset, fb, \alpha \rangle$
5. $\langle \emptyset, \neg tr, \alpha \rangle$

Here, argument $(1, 2, 3, 4, 5)$ defeats $(1, 2)$, and $(1, 2)$ defeats $(1, 2, 3, 4, 5)$. Only one maximal status assignment exists: $(1, 2, 3, 4)$ and all of its subarguments are

assigned "undefeated" and $(1, 2, 3, 4, 5)$ is assigned "defeated". Thus, overall pmp, tr, bd, fb are undefeated and $\neg tr$ is defeated outright.

The fact that in Pollock's new system tr, bd and fb are justified means that the tax-relief problem is dealt with in a structurally different way than the pink elephant problem, even though both of them are based on self-defeating arguments of the type "head bites tail". Furthermore, one could describe the same kind of small conversation in which an agent is confronted with the untenability of its standpoint. Therefore, we believe both examples should be dealt with in a uniform way.

In section 5.3.3, we will provide a solution. It will come as no surprise that this solution is based on the concept of HY-arguments. With these arguments, we also do not need the more complex semantics of definition 5.30, 5.31, 5.32 and 5.33. As far as self-defeating arguments are concerned, Pollock's old grounded semantics based system will do fine.

A different analysis on self-defeating arguments

Pollock identified (some of) the problems related to self-defeating arguments and sought the solution to a great extent in adjusting the semantics of his system. Before continuing with our own solution (the implementation of HY-arguments) it is interesting to discuss somewhat more thoroughly the approach of adjusting the semantics as a solution to the difficulties of self-defeating arguments.

It should be noticed that the issue of how to deal with self-defeating arguments is not exclusively related to Pollock's system. Comparable problems play a role in RDL and in the system of P&S.

Consider the following examples, which are for simplicity stated in the logic of P&S.

I: $\mathcal{S} = \{\rightarrow a\}$
$\mathcal{D} = \{a \wedge \sim u \Rightarrow b,\ b \Rightarrow u\}$
$A_1 :\ \rightarrow a,\ a \wedge \sim u \Rightarrow b,\ b \Rightarrow u$
$A_2 :\ \rightarrow a,\ a \wedge \sim u \Rightarrow b$

II: $\mathcal{S} = \{\rightarrow a,\ \rightarrow b\}$
$\mathcal{D} = \{a \wedge \sim u \Rightarrow u,\ b \wedge \sim u \Rightarrow c\}$
$B_1 :\ \rightarrow a,\ a \wedge \sim u \Rightarrow u$
$B_2 :\ \rightarrow b,\ b \wedge \sim u \Rightarrow c$

On a semantical level, these examples result in the following argumentation frameworks, assuming a logic with classical arguments only.

Figure 5.7: A semantical perspective of the examples I and II.

Figure 5.7 shows that example I and II share the same argumentation framework.[13] Still,

[13]This is not completely true, since some arguments like $\rightarrow a$ and $\rightarrow a,\ \rightarrow b$ have been left out. These arguments, however, are not relevant for the nature of the problem.

there is an important difference between the examples I and II.

Example I is basically of the same type as Pollock's pink elephant example and the Ajax-Feijenoord example (in fact, apart from the syntactical sugar, it *is* the Ajax-Feijenoord example). It was argued earlier that A_2 should not be justified.

Also example II is a case of a self-defeating argument that defeats another argument. In this case, however, it is undesirable that the self-defeating argument defeats the other argument. B_2 is simply a legitimate argument and any agent that wishes to defeat it will discover that its own position is untenable.

If one purely regards it from a semantical perspective, then we have two examples that share the same formalization, while having a different intuitively desirable outcome. Such a pair of examples can be called a *mirror example*, as was described in section 2.2.5. As there is no reason to consider one of the formalizations as "wrong", the only remaining possibility is that, apparently, the purely semantical representation of figure 5.7 is not rich enough; it is too abstract. The point is that on the level of an argumentation framework, one has abstracted from the internal structure of the arguments, so one has lost the information *why* in the left-hand side A_2 should be defeated, while in the right-hand side B_2 should not be defeated.

The solution, therefore, cannot simply come by altering the semantics, because every possible semantics will take the same argumentation-framework as input and therefore derive the same outcome for both examples.

It must be mentioned that Pollock, who searches for a mainly (though not totally) semantical solution to the issue of self-defeat, also does not make the distinction between examples I and II in his new system. This can be seen as follows:

1. the left-hand side example goes similar to the pink elephant example discussed earlier. the result is no status assignment exists that assigns something to A_1 and A_2 and that therefore A_2 is defeated outright.

2. as for the right-hand side example, basically the same holds. Because in every possible status assignment B_2 will not have a status, B_2 will also not have a status. Thus, the outcome in the right-hand side example is the same as in the left-hand side example: B_2 is defeated outright.

With HY-arguments, at the other hand, it *is* possible to distinguish between the examples I and II. This results in the HY-enriched argumentation frameworks of figure 5.8.

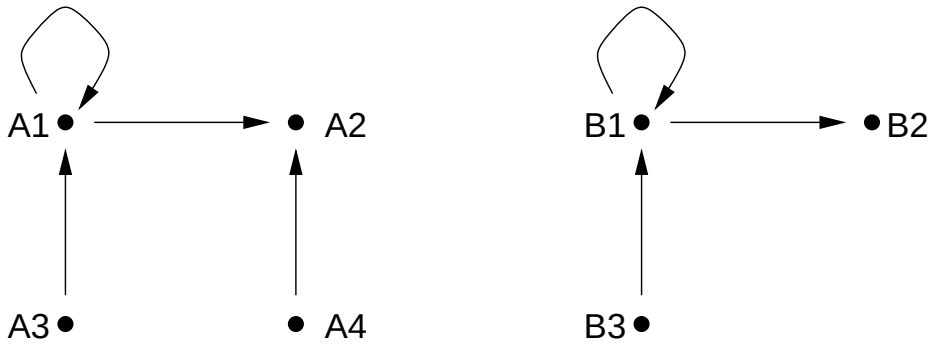


Figure 5.8: Examples I and II with HY-arguments.

Here, the desired outcome is achieved. In the left-hand side example, A_2 is *out* because it is defeated by the undefeated HY-argument A_4 ($= \rightsquigarrow b,\ b \Rightarrow u$), while in the right-hand side example B_2 is reinstated by the HY-argument B_3 ($= \rightsquigarrow u$). All of this is independent of which particular Dung-semantics is being applied. Thus, we see that where a purely semantics based approach can fail to make the necessary fine nuances, HY-arguments provide a workable solution.

5.3.3 Implementing HY-arguments in Pollock's system

The next question is whether HY-arguments can be implemented in Pollock's system, and, if yes, how such can be done. The first thing to notice is that HY-arguments are in fact based on a specific kind of suppositional reasoning. With a HY-argument one supposes one or more of the commitments of the other party in order to derive something that undermines the other party's position (either a contradiction or an undercutter of the other party's original argument). Unfortunately, this particular form of suppositional reasoning is not supported in Pollock's framework for defeasible reasoning. This can be illustrated using the following example.

> $\texttt{INPUT} = \{p, \neg r\}$
>
> p is a prima facie reason for q.
> q is a prima facie reason for r.
>
> There now exists an argument for q (argument I):
> 1. $\langle \emptyset, p, \infty \rangle$ ($p \in \texttt{INPUT}$)
> 2. $\langle \emptyset, q, \alpha \rangle$ (p is a prima facie reason for q)
>
> An HY-argument would then suppose q and derive a contradiction (argument II).
> 1. $\langle \{q\}, q, \infty \rangle$ (supposition)
> 2. $\langle \{q\}, r, \alpha \rangle$ (q is a prima facie reason for r)
> 3. $\langle \emptyset, \neg r, \infty \rangle$ ($\neg r \in \texttt{INPUT}$)
>
> Argument II, however, is self-defeating and does not prevent argument I from becoming justified (at least, not in Pollock's new system).

So, even though Pollock's system supports suppositional reasoning, it does not support the specific suppositional reasoning required for HY-arguments.

Another approach to defeat argument I would be to apply modus tollens to construct an argument for $\neg q$, based on $\neg r$ and the fact that q is a prima facie reason for r. This would go as follows.

1. $\langle \{q\}, q, \infty \rangle$ (supposition)
2. $\langle \{q\}, r, \alpha \rangle$ (q is a prima facie reason for r)
3. $\langle \emptyset,\ q \supset r,\ \alpha \rangle$ (conditionalization)
4. $\langle \emptyset, \neg r, \infty \rangle$ ($\neg r \in \texttt{INPUT}$)
5. $\langle \emptyset, \neg q, \alpha \rangle$ ($\{q \supset r,\ \neg r\}$ is a conclusive reason for $\neg q$)

Unfortunately, this argument is (strictly) defeated on line 2 by the one-line argument

$\langle \emptyset, \neg r, \infty \rangle$. So although it is possible to use supposition and conditionalization for the purpose of modus tollens, the resulting argument is of little use, since it is directly defeated by an argument with strength ∞. Thus, Pollock's system does not have modus tollens as a general property.[14]

It is clear that in order to allow for HY-arguments, Pollock's framework needs to be adjusted. Our proposed adjustment is similar to the way P&S and RDL were adjusted, as the design considerations (as layed out in section 3.2) are the same.

First, what is needed is a way to determine which conclusions are fc-based and which are not. This difference is important, as for instance the possibilities of attacking a conclusion depend on whether it is fc-based or not. In the logic of P&S, finding out whether a conclusion L in argument A is fc-based or not can be done by calculating $R_L(A)$ and inspecting whether it contains foreign commitments. The approach, therefore, is looking at the "start" of the argument. In Pollock's framework, on the other hand, suppositions are included in every node. Therefore, it would seem that there is no need to look back at the "start" of the argument, since the relevant suppositions are already connected with each conclusion in the argument. This, however, is not completely true, since it is possible to make suppositions that are not entirely necessary for at least part of the inference graph.

As an example, take $\texttt{INPUT} = \{a\}$, "a is a prima facie reason for b" and "$b \wedge c$ is a prima facie reason for d." In Pollock's original system, the following is an argument:

1.	$\langle \{c\}, a, \infty \rangle$	($a \in \texttt{INPUT}$)
2.	$\langle \{c\}, b, \alpha \rangle$	(a is a prima facie reason for b)
3.	$\langle \{c\}, c, \infty \rangle$	(supposition)
4.	$\langle \{c\},\ b \wedge c,\ \alpha \rangle$	($\{b, c\}$ is a conclusive reason for $b \wedge c$)
5.	$\langle \{c\}, d, \alpha \rangle$	($b \wedge c$ is a prima facie reason for d)

Although every line contains supposition $\{c\}$, the derivation of b does in fact not depend on this supposition at all. What is needed, therefore, is a formalization that only allows for suppositions that are *relevant* in the associated line. This means that each line should be assigned a minimal set of suppositions. Therefore, the definitions *input*, *supposition* and *reason* need to be adjusted; *input* needs to produce a line with an empty supposition, *supposition* (which will be called *foreign commitment* in definition 5.34) should produce a line whose supposition is a singleton, and *reason* should take the union of the suppositions of the lines it uses.

Another design consideration is that the foreign commitments are conclusions of the argument being attacked. Furthermore, the conclusions on which the foreign commitments are based should themselves not be based on foreign commitments.

The now following formalization allows for the use of HY-arguments. In order to keep things simple and focus on our main point, we assume that all classical (non-HY) arguments are linear; thus, suppositional reasoning is limited to HY-arguments only.

Definition 5.34. *An* argument *based on* INPUT *is a finite sequence $\sigma_1, \ldots, \sigma_n$, where each σ_i is a line of argument. A line of argument is a triple $\langle X_i, p_i, \nu_i \rangle$, where X_i is the set of* foreign commitments *of σ_i, p_i is a proposition, and ν_i is the degree of justification of σ_i. A new line of argument is obtained from earlier lines of argument according to the following rules of argument formation.*

[14] We thank Henry Prakken for this observation.

Input. *If p is in* INPUT *then it holds that $\sigma, \langle \emptyset, p, \infty \rangle$ is an argument.*

Reason. *If σ is an argument $\langle X_1, p_1, \eta_1 \rangle, \ldots, \langle X_n, p_n, \eta_n \rangle$ are members of σ and $\{p_1, \ldots, p_n\}$ is a reason of strength ν for q, then $\sigma, \langle X_1 \cup \ldots \cup X_n, q, \min\{\eta_1, \ldots \eta_n, \nu\} \rangle$ is an argument.*

Foreign commitment. *If σ is an argument and p is a proposition, then $\sigma, \langle \{p\}, p, \infty \rangle$ is also an argument.*

To avoid redundancy, we require that different lines always have different propositions.

Definition 5.35. *An argument σ* classically rebuts *an argument η iff:*

1. *η contains a line of the form $\langle \emptyset, q, \alpha \rangle$ that is obtained by the argument formation rule* reason *from some earlier lines $\langle \emptyset, p_1, \alpha_1 \rangle, \ldots, \langle \emptyset, p_n, \alpha_n \rangle$ where $\{p_1, \ldots p_n\}$ is a prima facie reason for q, and*

2. *σ contains a line of the form $\langle \emptyset, \neg q, \beta \rangle$ where $\beta \geq \alpha$.*

Definition 5.36. *An argument σ* classically undercuts *an argument η iff:*

1. *η contains a line of the form $\langle X, q, \alpha \rangle$ that is obtained by the argument formation rule* reason *from some earlier lines $\langle X_1, p_1, \alpha_1 \rangle, \ldots, \langle X_n, p_n, \alpha_n \rangle$ where $\{p_1, \ldots p_n\}$ is a prima facie reason for q, and*

2. *σ contains a line of the form $\langle \emptyset, \neg \ulcorner \{p_1, \ldots, p_n\} \gg q \urcorner, \beta \rangle$ where $\beta \geq \alpha$.*

Definition 5.37. *An argument σ* HY-rebuts *an argument η iff:*

1. *σ contains a line of the form $\langle X_1, L, \beta_1 \rangle$ and a line of the form $\langle X_2, \neg L, \beta_2 \rangle$ where $X_1 \neq \emptyset$ or $X_2 \neq \emptyset$, and*

2. *for each $f_i \in X_1 \cup X_2$ it holds that there is a line in η of the form $\langle \emptyset, f_i, \alpha_i \rangle$, and*

3. *it holds that* $\min\{\beta_1, \beta_2\} \geq \min\{\alpha_1, \ldots, \alpha_n\}$.

If argument σ HY-rebuts argument η and $\min\{\beta_1, \beta_n\} = \min\{\alpha_1, \ldots, \alpha_n\}$ *then η* reverse HY-rebuts *σ.*

Definition 5.38. *An argument σ* HY-undercuts *an argument η iff:*

1. *η contains a line of the form $\langle X, q, \alpha \rangle$ that is obtained by the argument formation rule* reason *from some earlier lines $\langle X_1, p_1, \alpha_1 \rangle, \ldots, \langle X_n, p_n, \alpha_n \rangle$ where $\{p_1, \ldots p_n\}$ is a prima facie reason for q, and*

2. *σ contains a line of the form $\langle Y, \neg \ulcorner \{p_1, \ldots, p_n\} \gg q \urcorner, \beta \rangle$ with $Y \neq \emptyset$, and*

3. *for each $f_i \in Y$ it holds that there is a line in η of the form $\langle \emptyset, f_i, \alpha_i \rangle$, and*

4. *it holds that* $\beta \geq \min\{\alpha_1, \ldots, \alpha_n\}$.

Definition 5.39. *An argument σ* defeats *an argument η iff:*

1. *σ classically rebuts η, or*

2. σ classically undercuts η, or

3. σ HY-rebuts η, or

4. σ reverse HY-rebuts η, or

5. σ HY-undercuts η.

The formalism of definition 5.34 until 5.39 will be referred to as the HY-enriched Pollock system.

It must be noticed that if one restricts oneself to linear arguments without foreign commitments, the HY-enriched Pollock system is equivalent to Pollock's old system with linear arguments only. This can be seen as follows. First, if no foreign commitments are allowed then the remaining arguments are all linear (see definition 5.34). Also, without foreign commitments, arguments cannot HY-rebut, reverse HY-rebut or HY-undercut each other. Thus, on the level of a Dung-style argumentation framework, what the above definitions do is that they take the existing set of arguments and the defeat-relation, and add new arguments (HY-arguments) and extend the defeat-relation.

In order to see how the HY-enriched system works, consider the following examples.

Ajax-Feijenoord

$\texttt{INPUT} = \{af\}$

af is a prima facie reason for t.
t is a prima facie reason for p.
p is a prima facie reason for $\neg\lceil\{af\} \gg t\rceil$.

argument pro:
1. $\langle\emptyset, af, \infty\rangle$ ($af \in \texttt{INPUT}$)
2. $\langle\emptyset, t, \alpha\rangle$ (af is a prima facie reason for t)

argument con:
1. $\langle\{t\}, t, \infty\rangle$ (foreign commitment)
2. $\langle\{t\}, p, \alpha$ (t is a prima facie reason for p)
3. $\langle\{t\}, \neg\lceil\{af\} \gg t\rceil, \alpha\rangle$ (t is a prima facie reason for $\neg\lceil\{af\} \gg t\rceil$)

Here, argument con is an undercutting HY-argument against argument pro.

Tax-relief

$\texttt{INPUT} = \{pmp\}$

pmp is a prima facie reason for tr.
tr is a prima facie reason for bd.
bd is a prima facie reason for fb.
fb is a prima facie reason for $\neg tr$.

argument pro:

1. $\langle \emptyset, pmp, \infty \rangle$ ($pmp \in$ INPUT)
2. $\langle \emptyset, tr, \alpha \rangle$ (pmp is a prima facie reason for tr)

argument con:

1. $\langle \{tr\}, tr, \infty \rangle$ (foreign commitment)
2. $\langle \{tr\}, bd, \alpha \rangle$ (tr is a prima facie reason for bd)
3. $\langle \{tr\}, fb, \alpha \rangle$ (bd is a prima facie reason for fb)
4. $\langle \{tr\}, \neg tr, \alpha \rangle$ (fb is a prima facie reason for $\neg tr$)

Here, the argument con is a rebutting HY-argument against the argument pro.

With respect to the issue of self-defeating arguments, the following can be said. First of all, the issue of self-defeating arguments applies only to classical rebutting and classical undercutting.

Lemma 5.7. *For each argument σ in the HY-enriched Pollock system such that σ defeats itself, it holds that σ classically rebuts itself or σ classically undercuts itself.*

Proof. The fact that σ defeats σ means (definition 5.39) that either σ classically rebuts itself, σ classically undercuts itself, σ HY-rebuts itself, σ reverse HY-rebuts itself, or σ HY-undercuts itself.

Suppose σ HY-undercuts itself. Then (definition 5.38 (1)) σ contains a line of the form $\langle Y, \neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil, \beta\rangle$ with $Y \neq \emptyset$. Then (definition 5.38 (2)), σ also contains at least one line $\langle \{f_i\}, f_i, \infty \rangle$ for some $f_i \in Y$. But definition 5.38 (3) also requires that σ contains a line $\langle \emptyset, f_i, \alpha_i \rangle$. This, however, conflicts with the definition of an argument (definition 5.34), where it is required that different lines have different propositions. Contradiction.

Thus, σ cannot HY-undercut itself. For similar reasons, σ also cannot HY-rebut itself (which implies that σ also cannot reverse HY-rebut itself). Thus, the only remaining possibilities are that σ classically rebuts itself, or that σ classically undercuts itself. □

The fact that self-defeat is limited to classical arguments means that every self-defeating argument is defeated by an argument that is itself undefeated.

Theorem 5.19. *If an argument σ in the HY-enriched Pollock system defeats itself, then there exists an argument η that defeats σ and is not defeated by any argument.*

Proof. Let σ be an argument such that σ defeats itself. Then, according to lemma 5.7, σ either classically rebuts itself or classically undercuts itself. If σ classically rebuts itself then (definition 5.35) σ contains a line $\langle \emptyset, q, \alpha \rangle$ and a line $\langle \emptyset, \neg q, \beta \rangle$. Then the argument $\eta = (\langle \{q\}, q, \infty \rangle, \langle \{\neg q\}, \neg q, \infty \rangle)$ HY-rebuts σ and is not defeated by any argument. If σ classically undercuts itself, then (definition 5.36) σ contains a line of the form $\langle \emptyset, \neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil, \beta\rangle$. Then the one-line argument $\eta = (\langle \{\neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil, \beta\}, \neg\lceil\{p_1, \ldots, p_n\} \gg q\rceil, \alpha\rangle)$ HY-undercuts σ and is not defeated by any argument. □

Now again consider the issue of semantics (definition 5.28 and 5.29). In definition 5.29, self-defeating arguments were explicitly ruled out, so that they do not prevent other arguments from becoming justified. It is interesting to see that with HY-arguments, this behavior also emerges using the "unrepaired" definition of grounded semantics (definition 5.28). This can be seen as follows. Suppose σ is a self-defeating argument. Then, according to theorem 5.19, there exists an argument η that defeats σ and is itself undefeated. The fact that η is not defeated by any argument means that η is *in* at every level. This, however,

also means that σ is *out* and stays *out* at every level starting from level 1, and will thus not keep other arguments from becoming justified.

The point is that with HY-arguments, there is no need for a "hacked" version of grounded semantics (like definition 5.29) or for Pollock's new system, at least not in order to neutralize the effects of self-defeating arguments. With HY-arguments, ordinary grounded semantics will do fine.

Chapter 6

Summary and conclusions

In this section, the most important findings of this thesis are summarized. Overall, we can distinguish nine main conclusions.

1. HY-style informal reasoning has been known since antiquity and is still widely in use today.

A well-known example of HY-style reasoning can be found in Socrates's elenchus. When applying the elenchus, one keeps confronting the opponent with the consequences of his or her own points of view, until it becomes clear that the opponent's view is simply untenable.

The idea of confronting one's opponent with the (defeasible) consequences of his or her own point of view is still common in today's world. Examples hereof can be found in critical interviews, as well as in modern philosophy.

2. Despite the ubiquity of informal HY-reasoning, HY-arguments are often not supported by existing formalisms for defeasible reasoning.

The system of Prakken and Sartor, as well as Reiter's default logic and Pollock's system do not support HY-arguments by themselves, although an implicit form of HY-style reasoning can be implemented in RDL by using free defaults. The effects of the lacking support of HY-arguments has been illustrated by the various running examples in this thesis.[1]

3. HY-style reasoning can be understood by taking into account the concept of nested commitments.

In dialogue, when one confronts the opponent with what seems to be the consequence of his or her own standpoints, one does not become committed to this consequence itself. Instead, one claims that the *opponent* is in fact committed to this consequence. One then has to combat any denials on the side of the opponent that this consequence indeed follows from opponent's standpoints.

4. It is possible to implement HY-style reasoning in the logic of Prakken and Sartor, in Reiter's default logic and in Pollock's system.

[1] like "Ajax-Feijenoord", "shipment of goods", ...

The implementation of HY-arguments can be done in accordance to the general principles that have been stated in section 3.2.2. These have been applied to the logic of P&S, to RDL and to Pollock's system, in order to achieve HY-enriched versions of these respective formalisms. For the logic of P&S this resulted in a formalism that allows for a uniform treatment of strict and defeasible rules. For RDL it was shown that the presence of HY-arguments can result in the existence of extensions, where otherwise no extensions would exist. For Pollock's system, it was argued that HY-arguments actually involve a particular form of suppositional reasoning not implemented in Pollock's original system.

5. The proposed formalizations of HY-arguments complies with a couple of intuitive examples, but does not depend on them.

In section 2.2 it was argued that possible methods for defining formal reasoning based on intuitive reasoning include the use of examples, postulates and formal semantics.

As for the method of examples, several running examples have been specified for which it was shown that the HY-enriched formalisms correctly deal with them, whereas the original formalisms fail to derive the desired outcome.

As for the use of postulates, although no postulates were defined directly on the overall entailment relation, general principles have been defined for the application of HY-arguments themselves; this was done at the end of section 3.2.2. The specification of the HY-enriched formalisms has been carried out in accordance with these principles.

As for the use of formal semantics, it has been shown that the notion of HY-arguments complies with the concept of a Dung-style argumentation framework. This makes the application of various argumentation-based semantics possible. Argumentation can also be studied from the perspective of dialogue, in which HY-arguments can serve as an additional form of counterarguments, at points in the debate where the use of such an argument would be intuitively warranted.

Overall, one can say that although a number of small examples have been used as a starting point, the overall research methodology has been broader than simply defining a formalism that complies with a limited number of examples. Therefore, the pitfalls mentioned in section 2.2.1 (like illustrated by Vreeswijk's interpolation theorem) have to a large extent been avoided.

6. The difference between argumentation formalisms with HY-arguments and argumentation formalisms without HY-arguments is related to the difference between rule maximality and conclusion maximality.

Nonmonotonic logic is specially designed to handle the possibility of conflicting conclusions. In general, there are two overall ways of dealing with potential conflicts. With conclusion maximality, one blames a *last* rule leading to the conflict, whereas with rule maximality, one blames an *arbitrary* rule that is involved in the conflict. In a simple defeasible theory — that is, one without priorities and undercutting — it holds that a justified conclusion of DS_{HY} is also a justified conclusion under rule maximality as well as in $DS_{classic}$. Furthermore, a justified conclusion under rule maximality, or in $DS_{classic}$ is also a justified conclusion under conclusion maximality (see figure 3.2 on page 76).

7. The application of HY-arguments (and contraposition) is in general compatible with

epistemical reasoning but needs not to be compatible with constitutive reasoning.

Conceptually, one can distinguish two kinds of reasoning: epistemical reasoning and constitutive reasoning [Hage97]. The difference is that the epistemical reasoning process can be considered as a perfect or imperfect procedure. There exists an objective reality and the aim of the reasoning process is to derive valid facts about it, with a varying level of success. Constitutive reasoning, at the other hand, can be seen as a pure procedure in that the reasoning *itself* determines the validity of the statements being derived.

As for epistemical reasoning, a possible way to interpret defaults or defeasible rules is the statistical one. Here, a default $A \Rightarrow B$ is interpreted as "most A's are B's" (unrestricted statistical interpretation), or as "nearly all A's are B's" (ε-semantics). In order to allow nonmonotonic entailment, however, what is needed is a "normality assumption" under the statistical interpretation. It was argued the the principle of maximum entropy provides such a normality assumption. Using maximum entropy, one can validate simple forms of contraposition and HY. It was argued that some well-known counterexamples against contraposition (as well against HY) are based on implicit additional information that should have been formalized explicitly, in order to avoid undesirable derivations.

As for constitutive reasoning, examples were given to illustrate that contraposition and HY are in general not valid. For legal reasoning, part of the fact that contraposition and HY do not hold is related to the fact that legal reasoning requires the effects of conflicts to remain as local as possible, in order not to prevent other inferences from becoming valid. The rationale behind this is that one should always try to interpret the law as consistent as possible.

8. HY-arguments provide a subtle way of dealing with the issue of self-defeating arguments

One particular tricky issue in formal argumentation is that of self-defeating arguments. In general, there are good reasons for making sure that self-defeating arguments do not keep other arguments from becoming justified. In the logic of P&S, for example, this is achieved by self-defeating arguments being defeated by the empty argument, whereas in Pollock's old system it was proposed to explicitly exclude self-defeating arguments from the inductive definition of justified arguments. With HY-arguments, there is no need for special constructs like these, as for every self-defeating classical argument there exists an undefeated HY-argument that defeats it. Furthermore, no HY-argument is itself self-defeating.

Another advantage of HY-arguments is that they allow different situations of self-defeat to be distinguished (see figure 5.8 on page 141) whereas without HY-arguments, essentially the same argumentation frameworks would result, forcing one to make a decision that is suitable to only one of the two given situations.

9. Mirror examples can provide a useful addition as a research method for the construction of logical formalisms.

Mirror examples have been used twice in our analysis of HY-arguments. In section 4.2.2, a mirror example was the starting point for the discussion about epistemical versus constitutive reasoning.

In section 5.3.2 a mirror example was given in order to argue that for a proper treatment of self-defeating arguments, what is needed is more than just an argumentation framework based on classical arguments only.

In general, the research method of stating and solving mirror-examples is useful because it forces the researcher to carefully think about the concepts or forms of reasoning that are implemented by the logic in question, as well as about how the translation of an informal example to the logical formalism should take place. Applying mirror examples is similar to the dialectical method of thesis, antithesis and synthesis. The thesis would be "in formal example FE, conclusion c should be derived because of intuitive interpretation IE_1". The antithesis would be "in formal example FE, conclusion c should not be derived because of intuitive interpretation IE_2". The synthesis would then be the resolvement of this dilemma, either by (1) providing additional guidelines about how information should be formally represented, or by (2) constructing a logic that is rich enough to distinguish between IE_1 and IE_2, or by (3) the construction of two separate logical formalisms, one that is applicable to the the kind of reasoning required by IE_1 and one that is applicable to the kind of reasoning required by IE_2.[2]

Related research

Despite the fact that the idea of using the other party's statements against him has been known since antiquity, surprisingly little research is carried out to include this phenomenon in formal models for dialogue or argumentation.

As for the somewhat related issue of self-defeating arguments, John Pollock has given the matter some consideration. His approach, however, is largely based on changing the semantics, not on directly implementing HY-arguments. In section 5.3.2 it was argued that Pollock's approach has important drawbacks.

In legal procedure, one feature that is particularly close to the notion of a Socratic dialogue is the cross-examination of a witness. Some thoughts about the structure of a legal cross-examination have been stated by Joseph Fulda [Fuld00]. Fulda describes cross-examination as "an opportunity to impeach evidence given by the witness during direct examination". The idea is that by asking a carefully selected series of questions, the witness can be led to commit himself to an inconsistency[3], thereby undermining his credibility. Fulda's treatment, however, is quite brief, and no attention is being paid to the specific features of defeasible reasoning.

Related issues

The difference between constitutive reasoning and epistemical reasoning also has consequences for, say, the construction a (defeasible) deontic logic. One possible approach would be to take an existing general purpose logic, and enrich it with the necessary deontic concepts. One example of this is SDL, which basically consists of classical logic, with

[2] An example of this is the discussion of epistemical versus constitutive reasoning, where it was mentioned that an argumentation formalism with HY-arguments is suitable for epistemical reasoning, but that constitutive reasoning is better modeled by an argumentation formalism without HY-arguments.

[3] The contradiction can either be immediate — in case the witness directly contradicts something he claimed earlier — or indirectly [Fuld00, p. 338]: " (...) the testimony can be impeached by contradicting not just a proposition in the testimony space, but by contradicting *any logical consequence of any proposition in the testimony space*."

embedded deontic concepts.

The example of the snoring professor, however, shows that sometimes the desirable outcome does not merely depend on the properties of the embedded deontic concepts, but also on the properties of the "host logic" (the logic in which the deontic concepts are embedded). Let us restate the example in question.

$$\begin{array}{rl} \textit{facts}: & \{SN,\ P\} \\ \textit{defaults}: & \{SN \Rightarrow M,\ M \Rightarrow Perm(R),\ P \Rightarrow \neg Perm(R)\} \end{array}$$

As indicated in section 4.2.2, one of the desirable outcomes of this example is M. In order to obtain this outcome, contraposition should not be sanctioned. This is a requirement that can only be fulfilled by the host logic; it cannot be achieved by tuning the properties of the embedded deontic concepts. In order to obtain a proper formalism for deontic reasoning, the host logic should satisfy the specific requirements for constitutive reasoning. One should be aware that these are not necessarily the same as for epistemical reasoning. Hence, a careful examination of the host logic is advisable.

Future research

This thesis should be seen as a first exploration on formalizing the principle of using the standpoints of the other party against him in argumentation and dialogue. Despite the elaborate treatment many questions are still open. For instance, in Pollock's system, HY was defined for linear arguments only. An interesting question would be how HY can be combined with Pollock's general suppositional reasoning. Another technical issue is whether the relationship between DS_{HY}, $DS_{classic}$, rule maximality and conclusion maximality is changed when priorities are defined between rules.

One particular interesting issue is that of an HY-enriched speech act based dialogue system. In section 3.2.1 it was suggested that such a dialogue system could be implemented by using the *but-then* statement, but no complete formalization was given; in particular we did not specify when certain moves are applicable and when not. Part of the reason why such a formal dialogue system was not specified has to do with the problem of *relevance* of dialogue moves. Consider the following defeasible theory.

$$\begin{array}{rcl} \mathcal{S} & = & \{\rightarrow a,\ \rightarrow c,\ \rightarrow \neg e\} \\ \mathcal{D} & = & \{a \Rightarrow b\ c \Rightarrow d,\ b \wedge d \Rightarrow e\} \\ < & = & \emptyset \end{array}$$

Now, if the proponent puts forward an argument for b, the HY-counterargument would be $\rightarrow c,\ c \Rightarrow d,\ \rightsquigarrow b,\ b \wedge d \Rightarrow e,\ \rightarrow \neg e$, so the argument-based dialogue would be:

$$\begin{array}{ll} \text{P:} & \rightarrow a,\ a \Rightarrow b \\ \text{O:} & \rightarrow c,\ c \Rightarrow d,\ \rightsquigarrow b,\ b \wedge d \Rightarrow e,\ \rightarrow \neg e \end{array}$$

Now consider the same example using a speech act based dialogue:

P:	claim b	$C_P(b)$
O:	claim d	$C_O(d)$
P:	why d	[unchanged]
O:	because $c \Rightarrow d$	$C_O(c, d)$
P:	concede c, d	$C_P(b, c, d)$
O:	but-then $b \wedge d \Rightarrow e$	$C_O(c, d, C_P(e))$
P:	concede e	$C_P(b, c, d, e)$
O:	claim $\neg e$	$C_O(c, d, \neg e)$
P:	concede $\neg e$	$C_P(b, c, d, e, \neg e)$ (inconsistent)

Now, the main difficulty is in the second step, where P's claim of b is responded by O's claim of d. At the moment O claims d it is not clear *why* this is a relevant countermove against the claim of b, since this claim appears to be unrelated; its relation only becomes clear after the dialogue has ended. However, if one cannot determine in advance when a certain move is relevant then this allows parties to stretch the dialogue indefinitely by putting forward all kinds of irrelevant moves.[4] The problem of relevance would be solved if, in the above example, O would immediately provide the entire HY-argument ($\rightarrow c,\ c \Rightarrow d,\ \rightsquigarrow b,\ b \wedge d \Rightarrow e,\ \rightarrow \neg e$) instead of just gradually "rolling out" this argument. This approach, however, has the downside that it does not correspond to how people actually argue in dialogues like cross-examinations.

The problem of relevance also plays a role in informal dialogue. When, in an Anglo-Saxon legal system, a lawyer cross-examines a witness it is sometimes not immediately clear what the point is that he or she wants to make. In that case, the counterparty can object at which the lawyer may have to explain privately to the judge his question technique as well as the relevance of it. As the problem of relevance plays a role in real-life dialogue and argumentation it should not come as a surprise to encounter the same problem when attempting to formalize this kind of dialogue. Joseph Fulda encounters basically the same problem when trying to determine a criterion that allows one to distinguish between those cross-examinations that are "to the point" and those that are not [Fuld00]. Fulda concludes that this is only possible if the future part of the line of questioning is also taken into account. It thus seems that one possible way to deal with this issue is to have a third party to whom such a future line of questions (or a future line of but-then statements) could be entrusted, and who would then determine whether the questioner's dialogue steps can still be considered as relevant. The role of this third party would be comparable to that of a judge in informal cross-examinations. The question of how such should be concretely implemented is left open for future research.

One final open issue to be mentioned has to do with the difference between constitutive reasoning and epistemical reasoning. As explained in section 4.2.2, contraposition and HY can be seen as valid principles for epistemical reasoning, but not necessarily for constitutive reasoning. The approach that was taken in this thesis is to go towards two seperate logics, one for epistemical reasoning (including HY) and one for constitutive reasoning (without HY). The underlying assumption is that the reasoning that takes place is either entirely epistemical or entirely constitutive. The situation in which part of the reasoning is epistemical and part of the reasoning is constitutive would require a hybrid logic, which is until now an issue that is still open for future research.

[4] This is the reason that in [MBPW02] relevance is stated as one of the necessary properties of an agent communication protocol. The problem of relevance in dialogues for defeasible reasoning has also been studied by Prakken [Prak00], but his solution is not applicable to the specific problems of the *but-then* statement.

Epilogue

In this epilogue, some possible applications of formal dialogue and argumentation are sketched. In particular, we ask ourselves the question what role formal dialogue and argumentation can play in the field of agent-mediated electronic commerce. For this, two stages are distinguished: ex-ante and ex-post. Both have different characteristics and require a different approach.

A particular relevant question is according to which criteria a certain formalism can be relied upon for an actual implementation in an e-commerce environment. For this, we mention four relevant conditions and discuss how an (HY-enriched) argumentation formalism would perform on these points.

Electronic commerce: ex-ante versus ex-post

Electronic commerce can be defined as doing business by means of electronic networks, of which the Internet is the most well-known. Examples of e-commerce include on-line retailing, auction sites and the use of EDI.

Overall, one can distinguish two phases of an (e-)commerce transaction: ex-ante and ex-post (see figure 6.1). In the ex-ante phase, the aim is for parties to come to an agreement. This means that parties first have to get in touch with each other, exchange the necessary information and possibly enter into a negotiation. In the ex-post phase, the agreement is carried out. If, during the execution of the agreement, unforeseen circumstances or disagreements arise, then some mechanism is needed to deal with these.

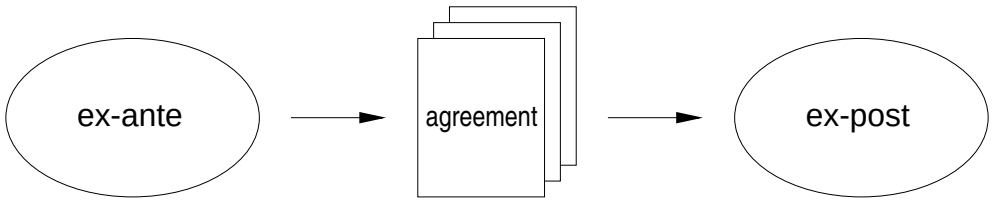


Figure 6.1: Ex-ante versus ex-post.

With respect to the ex-ante phase, much research has been carried out about how software agents can play a role in the process of negotiation. A relevant topic is the specification of formal protocols in which agents can try to persuade each other to perform certain behaviour. An agent can, for instance, persuade another agent to give up a certain resource by explaining that the other agent can still reach its goal without the resource [PaSJ98]. Persuation can also be brought about by enriching the negotiation protocol with features like threats and promises of rewards [SJNP98, KrSE98]. Another topic of research related to

the ex-ante phase is that of quantitative negotiation, which is concerned with questions like how much should one concede in each step of the negotiation [SiFJ97, FaSJ98, GoHO00].

In the ex-post phase, the agreement that has been settled in the ex-ante phase is meant to be carried out. When all goes according to plan, then the e-commerce support can be limited to the electronic exchange of the necessary documents (such as the despatch advice or invoice). Despite the relative simplicity of these messages, electronic support can yield significant cost advantages, and reduce the number of errors [VJKN94].

A more complex situaton arises when disagreements or unforeseen circumstances occur. An example would be if the price on the invoice is higher that previously agreed, due to additional surcharges (like taxes and costs of transport). The buyer may argue that such additional charges should have been indicated explicitly before it agreed the purchase. The seller may argue that in its field of operation, it is common to leave the charges implicit, and that this is a well-known practice.

In general, there are various ways in which disputes among parties can be settled. An obvious way would be to go to court. A disadvantege of doing so, apart from aspects as time and costs, would be that traditional legal action is most likely to disturb the relationship between the two parties, which can seriously hamper any future business transactions. An alternative to going to court would be *Alternative Dispute Resolution* (ADR). ADR can be divided into three options: negotiation, mediation and arbitration [Lodd02]. In the case of negotition, parties try to resolve the conflict or work out a compromise without a third party being involved. In the case of mediation, an independent third party (the mediator) moderates the dispute between parties, but it does not have the power to impose a particular settlement. In the case of arbitration, however, the third party *does* have the power to decide the case.

A recent trend is for ADR to be supported by electronic means, for instance, over the Internet. The thus implemented ADR is also referred to as ODR (online dispute resolution) or eADR. An example of ODR is the settlement of disputes regarding the use of Internet domain names, where arbitration is performed by the World Intellectual Property Organization (WIPO).[5] The rise of ODR has inspired researchers to come up with tools aiming to support it; an example is the system of Lodder [Lodd02].

On a more abstract level, one can ask the question how the issue of dispute resolution could be handled by software agents. Would it be possible for software agents to settle disputes among themselves, on behalf of their respective owners?

Safe protocols for electronic commerce

Given the idea of letting software agents settle any disputes among themselves on behalf of their owners, an interesting question is under which circumstances the involved agents and their communication protocols could be considered as "safe". That is, under what circumstances can an agent-based implementation for dispute resolution be trusted?

Apart from traditional computer security issues — which can also be problematic, see [Cami98, CRZD98a, CRZD98b] — additional properties are desirable in order for an agent-based implementation to be rightfully trusted. Four of these properties to be discussed here are verifyability, acceptability, minimal disclosure, and fairness under bounded rationality and private information.

[5] See arbiter.wipo.int/domains

With *verifyability*, we mean that the results obtained by the agents can easily be verified by their human owners. Argumentation and dialogue systems, for instance, have the advantage that the outcome can easily be understood by inspecting the dialogue that led to the outcome. A traditional defeasible logic, where justified arguments are based on a fixed-point definition, does not necessarily have the advantage of an intuitive explanation. In order for humans to trust their artificial agents, a clear explanation of why a certain outcome was derived is likely to be an important condition.

Another important property for agent-based dispute resolution is *acceptability*. It is not enough that an agent is able to explain how a certain outcome was derived or argued, the explanation should also be accepted by its human owner. It should not be the case that the owner goes through the dialogue and wonders why a particular argument was not put forward by its agent, for otherwise it would have won the dispute. That is, for an agent to be trusted, it should be able to put forward at least the same arguments as its owner would. One example would be HY-arguments. As indicated in chapter 1, HY-arguments are commonly used among humans, and it would therefore make sense also to have them implemented by artificial agents. If an agent repeatedly loses a dispute by not being able to put forward a certain type of argument, its owner might no longer regard it as a safe alternative to settling disputes himself.

A third desirable property is the *minimal disclosure* of information. A party may not want to provide more information than is necessary to defend its case. For instance, when a buyer disputes the quality of a certain shipment, then the seller may defend itself by revealing that part of the very same shipment has been received and approved by another buyer. The seller, however, may only be willing to reveal information, like its existing customer base, when absolutely necessary. One of the properties of argumentation or dialogue systems is that it can be determined on the spot what information is to be provided, whereas with "traditional" (nonmonotonic) logics, the idea is often that all information is available beforehand.[6]

The fourth and last desirable property to be discussed is that of *fairness under bounded rationality and limited information*. When an agent does not have complete information and unlimited computational resources, then it may not be able to put forward an argument that would otherwise have helped to win the debate. Thus, the result of the dialogue may be different from under ideal circumstances. Bur even so, it should *never* be the case that a party *loses* the debate due to incomplete information or bounded rationality at the side of the *other* party. Consider the following example, which is specified in the formalism of P&S [PrSa97].

$$\begin{array}{lcll} \mathcal{S} & = & \{\rightarrow A,\ C \rightarrow \neg A\} & \\ \mathcal{D} & = & \{A \Rightarrow B,\ B \Rightarrow C\} & \\ < & = & \emptyset & \\ \text{P:} & & \rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C & (A_1) \\ \text{O:} & & \emptyset & (A_2) \end{array}$$

Here, P constructs an argument that is inconsistent, so O gives an empty argument in

[6]It must be said, however, that this disadvantage of traditional (nonmonotonic) logics can be overcome by defining a dialectical protocol where parties take turns in supplying new information, and after each turn the (defeasible) consequences are calculated using the logic in question. In that way, a traditional (nonmonotonic) logic can be applied while still allowing parties to decide on the spot what information is to be released. This approach, however, appears less natural than the dialogue approach. See also [Brew01].

response.

Now, imagine a situation where P simply does not know about the existence of the rule $C \rightarrow \neg A$ (although O does). Then P would have no way of interpreting O's counterargument (the empty argument). P would only accept the fact that its argument A_1 is inconsistent after O explicitly points out *why* it is inconsistent.

Another possibility is that the rule $C \rightarrow \neg A$ *is* present to agent P, but that it was not *aware* of it, or that it simply did not have the necessary computational resources to apply this rule and find out that argument A_1 is inconsistent. In this case, the fact that agent O claims that A_1 is inconsistent may again be not enough since agent P may want to know the additional reasoning-steps that make it inconsistent. Hence, it would be desirable if agent O were able to provide these steps in an explicit way.

If one would want to explicitly provide this information in P&S's original framework, this would result in the following dialogue:

P: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C$
O: $\rightarrow A,\ A \Rightarrow B,\ B \Rightarrow C,\ C \rightarrow \neg A$
P: $\emptyset$

Here O makes an incoherent argument itself, which can be defeated by the empty argument, so instead of winning the argument, O *looses* it!

If circumstances of limited information or bounded rationality make the outcome of the dialogue different than the outcome in an ideal situation, then the outcome should be in the disadvantage of the agent where the unideal situation applies to, and not the agent that is still capable of ideal reasoning. The above problem can be solved by allowing HY-arguments. In that case, A_2 would be $\rightarrow C,\ C \rightarrow \neg A,\ \rightarrow A$, which contains all relevant information.

Interaction between artificial agents and humans

A different possible application of formal argumentation and dialogue theory can be found in the interaction between humans and artificial agents. An artificial agent, working among humans, will often have to persuade its human companions in order to be able to achieve its goals. Also, in order to gain acceptance, it will need to respond intelligently to any arguments coming from its human counterparts. This requires a full-blown theory on human-style argumentation, and it is our view that HY-arguments should be part of it.

The prospect of artificial agents interacting with humans in a human-style manner still has a significant way to go, but several organizations are already investing significant resources into implementing this ideal [CaNe02]. The prospect of humans arguing with artificial agents may be closer than it would seem.

Bibliography

[Adam75] E. Adams, *The Logic of Conditionals.* Reidel, Dordrecht, The Netherlands. (1975)

[AlGM85] C. Alchourrón, P. Gärdenfors and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50:510-530 (1985)

[AlMa82] C. Alchourrón and D. Makinson, On the logic of theory change: Contraction functions and their associated revision functions. *Theoria* 48:14-37 (1982)

[Ansc57] G.E.M. Anscombe, *Intentions.* Blackwell, Oxford (1957)

[Anto97] G. Antoniou, *Nonmonotonic Reasoning.* The MIT Press, Cambridge, Massachusetts (1997)

[BaGG00] P. Baroni, M. Giacomin and G. Guida, Extending abstract argumentation systems theory. *Artificial Intelligence* 120:251-270 (2000)

[BDKT97] A. Bondarenko, P.M. Dung, R.A. Kowalski and F. Toni, An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence* 93:63-101 (1997)

[BoPa99] R.A. Bourne and S. Parsons, Maximum entropy and variable strength defaults. *Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Stockholm, pp. 50-55 (1999)

[BoPa00] R.A. Bourne and S. Parsons, Connecting lexicographic with maximum entropy entailment. *Proceedings of the Fifth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (LNAI 1638)*, pp. 80-91 (2000)

[Bour99] R.A. Bourne, *Default Reasoning Using Maximum Entropy and Variable Strength Defaults.* PhD thesis, University of London (1999)

[Brew89] G. Brewka. *Nonmonotonic Reasoning: From Theoretical Foundations Towards Efficient Computation.* PhD thesis, Universität Hamburg. (1989)

[Brew91] G. Brewka, Assertional default theories. *Symbolic and Quantitative Approaches to Uncertainty, Proceedings of the European Conference ECSQAU. (LNCS 548)* pp. 120-124 (1991)

[Brew01] G. Brewka, Dynamic argument systems: A formal model of argumentation processes based on situation calculus. *Journal of Logic and Computation* 11(2):257-282 (2001)

[Buck91] B. Buck and V.A. Macaulay (eds.) *Maximum Entropy in Action.* Clarandon Press, Oxford (1991)

[CaJo97] J. Carmo and A.J.I. Jones, A new approach to contary-to-duty obligations. In: *Defeasible Deontic Logic.* Donald Nute (ed.), Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 317-344 (1997)

[Cami98] M.W.A. Caminada, *Onderzoek naar Internet-gerelateerde Beveiligingsincidenten binnen Nederlandse Organisaties.* KPMG EDP Auditors (afstudeerscriptie) (1998)

[CaNe02] M.W.A. Caminada and E. van Neer, Land van de rijzende robot; Japanse bedrijven willen een kunstmatige vriend in elk huishouden. *Computable* nr. 11 (2002)

[ChHo85] P.W. Cheng and K.J. Holyoak, Pragmatic reasoning schemas. *Cognitive Psychology* 17:391-406 (1985)

[Chis63] R.M. Chisholm, Contrary-to-duty imperatives and deontic logic. *Analysis* 24:33-36 (1963)

[ChWo94] M.K. Chang and C.C. Woo, A speech-act-based negotiation protocol: design, implementation, and test use. *ACM Transactions on Information Systems* 12(4):360-382 (1994)

[Clar96] H. Clark, *Using Language.* Cambridge University Press (1996)

[ClSc89] H. Clark and E. Schaefer. Contributing to discourse. *Cognitive science* 13:259-294 (1989)

[CRZD98a] M.W.A. Caminada, R.P. van de Riet, A. van Zanten and L. van Doorn, Internet security incidents, a survey within Dutch organisations. *Proceedings of WebNet'98* (1998)

[CRZD98b] M.W.A. Caminada, R.P. van de Riet, A. van Zanten and L. van Doorn, Internet security incidents, a survey within Dutch organisations. *Computers & Security* 17(5):417-433 (1998)

[Cupp94] F. Cuppens, Roles and deontic logic. *Proceedings of the Second International Workshop on Deontic Logic in Computer Science (ΔEON'94).* Andrew J.I. Jones and Marek Sergot (eds.), pp. 86-106 (1994)

[DiMW94] F. Dignum, J.-J. Ch. Meyer and R.J. Wieringa, Contextual permission: A solution to the free choice paradox. *Proceedings of the Second International Workshop on Deontic Logic in Computer Science (ΔEON'94).* Andrew J.I. Jones and Marek Sergot (eds.), pp. 107-144 (1994)

[Dung95] P.M. Dung, On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence* 77:321-357 (1995)

[EeGK87] F.H. van Eemeren, R. Grootendorst and T. Kruiger, *Handbook of Argumentation Theory; A Critical Survey of Classical Backgrounds and Modern Studies.* Foris publications, Dordrecht, The Netherlands (1987)

[Ethe87] D.W. Etherington. Reasoning with incomplete information. In: *Research Notes in Artificial Intelligence.* Pitman, London (1987)

[Ethe89] D.W. Etherington, K.D. Forbus, M.L. Ginsberg, D. Israel and V. Lifschitz. Critical issues in nonmonotonic reasoning. *Proceedings of the 1st International Conference on Knowledge Representation and Reasoning.* Morgan Kaufmann Publishers, pp. 500-504 (1989)

[FaSJ98] P. Faratin, C. Sierra and N. R. Jennings, Negotiation decision functions for autonomous agents. *Int. Journal of Robotics and Autonomous Systems.* 24(3-4):159-182 (1998)

[Fels86] W. Felscher, Dialogues as a foundation for intuitionistic logic. In: *Handbook of Philosophical Logic, volume III: Alternatives to Classical Logic.* pp. 341-372. D. Gabbay and F. Guenthner (eds.), D. Reidel publishing company, Dordrecht, The Netherlands (1986)

[Fuld00] J.S. Fulda, The logic of "improper cross". *Artificial Intelligence and Law* 8:337-341 (2000)

[GaGu02] D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic, Second Edition, Vol 4* Kluwer Academic Publishers, Dordrecht etc. (2002)

[Gärd82] P. Gärdenfors. Rules for rational changes of beliefs. In: *Philosophical Essays Dedicated to Lennart Åqvist on His Fiftieth Birthday.* pp. 88-101. T. Pauli (ed.), University of Uppsala Philosophical Studies 34 (1982)

[GePe92] H. Geffner and J. Pearl, Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence* 53:209-244 (1992)

[GHRN94] Dov M Gabbay, C.J. Hogger, J.A. Robinson, D. Nute (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming; Volume 3; Nonmonotonic Reasoning and Uncertain Reasoning.* Clarendon Press, Oxford (1994)

[Gins94] M.L. Ginsberg, AI and nonmonotonic reasoning. In: [GHRN94] pp. 1-33 (1994)

[GoHO00] G. Governatori, A.H.M. ter Hofstede, Arthur H.M. and P. Oaks, Defeasible logic for automated negotiation. *Proceedings of CollECTeR.* P. Swatman and P.M. Swatman (eds.), Deakin University (2000)

[Gold93] M. Goldszmidt, P. Morris and J. Pearl, A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15:220-232 (1993)

[Gord95] T.F. Gordon, *The Pleadings Game – An Artificial Intelligence Model of Procedural Justice*, Kluwer, Dordrecht (1995)

[Haac78] S. Haack, *Philosophy of Logics.* Cambridge University Press. (1978)

[Hage96] J.C. Hage, A theory of legal reasoning and a logic to match. *Artificial Intelligence and Law* 4:199-273 (1996)

[Hage97] J.C. Hage, *Reasoning with Rules, An Essay on Legal Reasoning and Its Underlying Logic.* Kluwer Academic Publishers, Dordrecht (1997)

[Hage00] J.C. Hage, Dialectical models in artificial intelligence and law. *Artificial Intelligence and Law* 8:137-172 (2000)

[Hamb70] C.L. Hamblin, *Fallacies*, Richard Clay (The Chaucer press) Ltd., Bungay, Suffolk (1970)

[Hamb87] C.L. Hamblin, *Imperatives.* Basil Blackwell (1987)

[Harm86] G. Harman, *Change in View, Principles of Reasoning.* The MIT Press, Cambridge, Massachusetts (1986)

[Hart61] H.L.A. Hart, *The Concept of Law.* Clarendon Press, Oxford (1961)

[HaVe94] J.C. Hage and B. Verheij, Reason-based logic: A logic for reasoning with rules and reasons. *Law, Computers & Artificial Intelligence* 3(2/3):171-209 (1994)

[Hort01] J.F. Horty, Argument construction and reinstatement in logics for defeasible reasoning. *Artificial Intelligence and Law* 9:1-28 (2001)

[Jayn79] E.T. Jaynes, Where do we stand on maximum entropy? In: *The Maximum Entropy Formalism* (R.D. Levine and M. Tribus, eds.), MIT Press, Cambridge, Massachusetts (1979)

[Kono88] K. Konolige, Defeasible argumentation in reasoning about events. *Methodologies for Intelligent Systems* 3:380-390 (1988)

[KrSE98] S. Kraus, K. Sycara and A. Evenchik, Reaching argreements through argumentation: a logical model and implementation. *Artificial Intelligence journal*, 104(1-2):1-69 (1998)

[Lifs94] V. Lifschitz, Circumscription. In: [GHRN94] pp. 297-352 (1994)

[Lodd98] A.R. Lodder, *Dialaw, On Legal Justification and Dialog Games.* Ph.D. thesis, University of Maastricht (1998)

[Lodd99a] A.R. Lodder, *Dialaw, On Legal Justification and Dialogical Models of Argumentation.* Kluwer Academic Publishers, Dordrecht (1999)

[Lodd99b] A.R. Lodder, DiaLaw: levels, dialog trees, convincing arguments. *Legal Knowledge Based Systems, JURIX 1999, The Twelfth Conference.* Nijmegen, The Netherlands (1999)

[Lodd02] A.R. Lodder, Online negotiation and mediation: Is there room for argument support tools? - Some reflections on argumentation and eADR. *BILETA 17th Annual Conference* (2002)

[LoLo78] P. Lorenzen and K. Lorenz, *Dialogische Logic.* Wissenschaftliche Buchgesellschaft, Darmstadt (1978)

[Loui90] R.P. Loui, Ampliative inference, computation, and dialectic. In: *AI and Philosophy.* J.L. Pollock (ed.), MIT Press (1990)

[Loui98] R.P. Loui, Process and policy: Resource-bounded non-demonstrative reasoning. *Computational Intelligence* 14(1):1-38 (1998)

[MacK79] J.D. MacKenzie, Question-begging in non-cumulative systems. *Journal of Philosophical Logic*, 8, pp. 117-133 (1979)

[MacK90] J.D. MacKenzie, Four dialogue systems. *Studia Logica 4/90*, pp. 567-583 (1990)

[Maki94] D. Makinson, General Patterns in Nonmonotonic Reasoning. In: [GHRN94] pp. 35-110 (1994)

[MaTo01] D. Makinson and L. van der Torre, Constraints for input/output locics. *Journal of Philosophical Logic* 30(2): 155-185 (2001)

[MBPW02] P. McBurney, S. Parsons and M. Wooldridge, Desiderata for agent argumentation protocols. *Proceedings of the First International Joint Conference on Autonomous Agents & Multi-Agent systems (AAMAS)* (2002)

[McCa86] J. McCarthy. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence* 26(3):89-116 (1986)

[MeWi93] J.-J. Ch. Meyer and R.J. Wieringa, Deontic logic: A concise overview. In: *Deontic Logic in Computer Science; Normative System Specification.* J.-J. Ch. Meyer and R.J. Wieringa (eds.), John Wiley & Sons, Chichester, England (1993)

[Meyd90] R. van der Meyden, The dynamic logic of permission. *Proceedings of the 5th Conference on Logic in Computer Science (LICS'90)*, pp. 72-78 (1990)

[Meye88] J.-J. Ch. Meyer, A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic* 29(1):109-136 (1988)

[Nels94] L. Nelson, *De Socratische Methode.* Uitgeverij Boom, Amsterdam (1994)

[NuEr98] D. Nute and K. Erk, Defeasible logic graphs; I. Theory. *Decision Support Systems* 22:277-293 (1998)

[NuHH98] D. Nute, Z. Hunter and Ch. Henderson, Defeasible logic graphs; II. Implementation. *Decision Support Systems* 22:295-306 (1998)

[Nute80] D. Nute, *Topics in Conditional Logic.* D. Reidel publishing company, Dordrecht, The Netherlands (1980)

[Pari98] J.B. Paris, Common sense and maximum entropy. *Synthese* 117(1):75-93 (1998)

[PaSJ98] S.D. Parsons, C.A. Sierra and N.R. Jennings, Agents that reason and negotiate by arguing. *Journal of Logic and Computation* 8(3), 261-292 (1998)

[PaVe90] J.B. Paris and A. Vencovská, A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning* 4:183-223 (1990)

[PaVe97] J.B. Paris and A. Vencovská, In defense of the maximum entropy inference process. *International Journal of Approximate Reasoning* 17:77-103 (1997)

[Pear88] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA (1988)

[Pear92] J. Pearl, Epsilon-semantics. In: *Encyclopedia of Artificial Intelligence.* Editor-in-chief S.C. Shapiro. Wiley. pp. 468-475 (1992)

[Pear99] J. Pearl, Reasoning with cause and effect. *Proceedings IJCAI'99.* pp. 1437-1449 (1999)

[Pere82] C.H. Perelman, *The Realm of Rhetoric.* translated by William Kluback. University of Notre Dame Press, Notre Dame, Indiana. (1982)

[Plato1] Plato, Lysis. In: *Socratic Discourses by Plato and Xenophon*, E. Rhys (ed.), J.M. Dent & Sons ltd. London (1910)

[Plato2] Plato, *Sophist.* translated by Benjamin Jowett. (360 BC)

[Poll87] J.L. Pollock. Defeasible reasoning. *Cognitive Science* 11:481-518 (1987)

[Poll91a] J.L. Pollock. Self-defeating arguments. *Minds and Machines* 1:367-392 (1991)

[Poll91b] J.L. Pollock. A theory of defeasible reasoning. *International Journal of Intelligent Systems* 6:33-54 (1991)

[Poll92] J.L. Pollock. How to reason defeasibly. *Artificial Intelligence* 57:1-42 (1992)

[Poll95] J.L. Pollock, *Cognitive Carpentry: A Blueprint for How to Build a Person.* The MIT Press, Cambridge, Massachusetts (1995)

[Pool88] D. Poole, A logical framework for default reasoning. *Artificial Intelligence* 36:27-47 (1988)

[Prak93] H. Prakken, *Logical Tools for Modeling Legal Argument.* Ph.D. thesis Vrije Universiteit, Amsterdam (1993)

[Prak97] H. Prakken, *Logical Tools for Modeling Legal Argument; A Study of Defeasible Reasoning in Law.* Kluwer Academic Publishers, Dordrecht, The Netherlands (1997)

[Prak00] H. Prakken, On dialogue systems with speech acts, arguments, and counterarguments. *Proceedings of JELIA'2000, The 7th European Workshop on Logic for Artificial Intelligence (LNAI 1919).* pp. 239-253 (2000)

[Prak01] H. Prakken, Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese* 127:187-219 (2001) (special issue on New Perspectives in Dialogical Logic. S. Rahman and H. Rückert (eds.))

[Prak02] H. Prakken, Intuitions and the modelling of defeasible reasoning: some case studies. *Proceedings of the Ninth International Workshop on Nonmonotonic Reasoning.* pp. 91-99 (2002)

[Prak03] H. Prakken, *Commonsense Reasoning.* Institute of Information and Computing Sciences. Universiteit Utrecht.

[PrSa96] H. Prakken and G. Sartor, A dialectical model of assessing conflicting arguments in legal reasoning. In: *Logical Models of Legal Argument.* Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 175-212 (1996)

[PrSa97] H. Prakken and G. Sartor, Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics* 7: 25-75 (1997) (special issue on 'Handling inconsistency in knowledge systems')

[PrSa98] H. Prakken and G. Sartor, Modelling reasoning with precedents in a formal dialogue game. *Artificial Intelligence and Law* 6:231-287 (1998)

[PrSe97] H. Prakken and M. Sergot, Dyadic deontic logic and contrary-to-duty obligations. In: *Defeasible Deontic Logic.* Donald Nute (ed.) Kluwer Academic Publishers, Dordrecht, The Netherlands. pp. 223-262 (1997)

[PrVr00] H. Prakken and G.A.W. Vreeswijk, Credulous and sceptical argument games for preferred semantics. *Proceedings of JELIA'2000, The 7th European Workshop on Logic for Artificial Intelligence (LNAI 1919).* pp. 239-253 (2000)

[PrVr02] H. Prakken and G. Vreeswijk, Logical systems for defeasible argumentation. In: [GaGu02] pp. 219-318 (2002)

[Rawl00] J. Rawls, *A Theory of Justice* (revised edition). Oxford University Press, Oxford (2000)

[ReCr81] R. Reiter and G. Criscuolo, On interacting defaults. *Proceedings IJCAI-81.* pp. 270-276 (1981)

[Reit80] R. Reiter, A logic for default reasoning. *Artificial Intelligence* 13:81-132 (1980)

[Reit87] R. Reiter, A theory of diagnosis from first principles. *Artificial Intelligence* 32:57-95 (1987)

[RoAn82] L. Ross and C.A. Anderson, Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments. In: *Judgment under Certainly: Heuristics and Biases.* Daniel Kahneman, Paul Slovic and Amos Tversky (eds.) Cambridge University Press. pp. 129-152 (1982)

[Ross41] A. Ross, Imperatives and logic. *Theoria* 7:53-71 (1941)

[Rsch77] N. Rescher, *Dialectics, A Controversy-Oriented Approach to the theory of Knowledge.* State University of New York Press, Albany (1977)

[Sear69] J.R. Searle, *Speech Acts, An Essay in the Philosophy of Language.* Cambridge University Press (1969)

[Sear79] J.R. Searle, A taxonomy of illocutionary acts. In: *Expression and Meaning.* Cambridge University Press. pp. 1-29 (1979)

[Sear83] J.R. Searle, *Intentionality, An Essay in the Philosophy of Mind.* Cambridge University Press (1983)

[Shoh87] Y. Shoham, A semantical approach to non-monotonic logics. In: *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI)* (1987)

[Shoh88] Y. Shoham, *Reasoning about Change.* MIT Press, Cambridge, USA (1988)

[SiFJ97] C. Sierra, P. Faratin and N.R. Jennings, A service-oriented negotiation model between autonomous agents. *Proceedings of the 8th European Workshop on Modeling Autonomous Agents in a Multi-Agent World (MAAMAW-97).* Ronneby, Sweden. pp. 17-35 (1997)

[SiLo92] G.R. Simari and R.P. Loui, A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence* 53:125-157 (1992)

[Skid02] J. Skidmore, Skepticism about practical reasoning: transcendental arguments and their limits. *Philosophical Studies* 109:121-141 (2002)

[SJNP98] C. Sierra, N.R. Jennings, P. Noriega and S. Parsons, A framework for argumentation-based negotiation. *Intelligent Agents IV (LNAI 1365)*, M. P. Singh, A. Rao and M. J. Wooldridge (eds.). (1998)

[Tars56] A. Tarski, *Logic, Semantics, Metamathematics.* Oxford University Press. J.H. Woodger (ed.) (1956)

[Tars86] A. Tarski, What are logical notions? *History and Philosophy of Logic* 7:143-154. John Corcoran (ed.) (1986)

[TaTo96] Y.-H. Tan and L.W.N. van der Torre, How to combine ordering and minimizing in a deontic logic based on preferences. *Deontic Logic, Agency and Normative Systems; ΔEON'96: Third International Workshop on Deontic Logic in Computer Science*, Sesimbra, Portugal. pp. 216-232 (1996)

[Torr97] L.W.N. van der Torre, *Reasoning about Obligations; Defeasibility in Preference-Based Deontic Logic.* Ph.D. thesis, Erasmus Universiteit Rotterdam (1997)

[ToTa97] L.W.N. van der Torre and Y.-H. Tan, The many faces of defeasibility in defeasible deontic logic. In: *Defeasible Deontic Logic.* Donald Nute (ed.) Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 79-121 (1997)

[Toul58] S.E. Toulmin, *The Uses of Argument.* Cambridge University Press (1958)

[VJKN94] P. van der Vlist, W.J. de Jong, A.E. Kolff, F.J. van der Net, A. van Overbeek, A.T.C. Siebbeles, *EDI in de Handel.* Samson Bedrijfsinformatie, Alphen aan de Rijn (1997)

[Vree91] G.A.W. Vreeswijk, The feasibility of defeat in defeasible reasoning. *The Proceedings of the 2nd Int. Conf. on Knowledge Representation and Reasoning (KR'91).* pp. 526-534 (1991)

[Vree92] G.A.W. Vreeswijk, Nonmonotonicity and partiality in defeasible argumentation. In: *Nonmonotonic Reasoning and Partial Semantics.* W. van der Hoek, J-J.Ch. Meyer, Y.H. Tan and C. Witteveen (eds.) Prentice Hall, New York. pp. 157-180 (1992)

[Vree93] G.A.W. Vreeswijk, *Studies in Defeasible Reasoning.* Ph.D. thesis Vrije Universiteit, Amsterdam

[Vree94] G.A.W. Vreeswijk, *IACAS: An Interactive Argumentation System.* Technical report CS 95-01, Department of Computer Science, FdAW, University of Limburg, The Netherlands (1994)

[Vree95a] G.A.W. Vreeswijk, Interpolation of benchmark problems in defeasible reasoning. *Proceedings of 2nd World Conference on the Fundamentals in AI (WOCFAI'95).* pp. 453-468 (1995)

[Vree95b] G.A.W. Vreeswijk. The computational value of debate in defeasible reasoning. *Argumentation: An International Scientific Journal* 9(2):305-342 (1995)

[Vree95c] G.A.W. Vreeswijk, *Interpolation of Benchmark Problems in Defeasible Reasoning.* Technical report CS 95-05. University of Limburg, Maastricht (1995)

[Vree97] G.A.W. Vreeswijk, Abstract argumentation systems. *Artificial Intelligence* 90:225-279 (1997)

[WaKr95] D.N. Walton and E.C.W. Krabbe. *Commitment in Dialogue, Basic Concepts of Interpersonal Reasoning.* State University of New York Press. (1995)

[Waso66] P. Wason, Reasoning. In: *New Horizons in Psychology.* B. Foss (ed.) Penguin Books Ltd., Harmondsworth, England (1966)

[Webe08] A. Weber, *History of Philosophy.* translated by Frank Thilly. Charles Scribner's Sons, New York (1908)

Samenvatting

Dit proefschrift, met als ondertitel "verkenningen in argument-gebaseerd redeneren", gaat over wiskundige modellen van redeneren en argumentatie. Hoewel op het gebied van formele argumentatie al vrij veel werk is verzet gaan vele formalismen uit van argumenten als ketens van regels, beginnend met de premissen. Wat men echter al sinds de klassieke oudheid in discussies ziet zoals deze daadwerkelijk door mensen worden gevoerd is dat een tegenargument soms niet begint met premissen, maar dat juist het standpunt van de tegenstander als uitgangspunt wordt genomen. Dit soort argumenten, welke in dit proefschrift worden aangeduid als HY ("hang yourself") argumenten, vormt het uitgangspunt van dit proefschrift. Ons doel is dan ook een analyse van deze argumentvorm en te laten zien hoe enkele reeds bestaande argumentatieformalismen hiermee kunnen worden uitgebreid.

Een belangrijke vraag is volgens welke criteria een redeneerformalisme dient te worden opgezet. Met andere woorden, hoe kan men vaststellen of een formalisme "correct" is. Een redeneermechanisme is immers wiskundig van aard. Wiskunde voorziet welliswaar in de instrumenten om verschillende vormen van redeneren te specificeren, maar geeft op zichzelf geen criterium welke van deze vormen als "redelijk" of "correct" kunnen worden beschouwd. De criteria hiervoor zullen dus minstens voor een deel buiten de wiskunde moeten worden gezocht, op basis van "intuitief", informeel redeneren.

Een relatief eenvoudige manier om een redeneerformalisme te rechtvaardigen op basis van intuitief redeneren is om gebruik te maken van kleine voorbeelden. Geschetst wordt een zekere situatie waarin een bepaalde conclusie wel of niet voor de hand ligt. Vervolgens wordt gekeken of met behulp van het formalisme deze conclusie al dan niet kan worden afgeleid. Een formalisme wordt correct bevonden wanneer het uit een aantal standaard voorbeeldjes de gewenste "intuitieve" uitkomsten haalt. Het probleem is dat het gebruik van intuitieve voorbeeelden misleidend kan zijn wanneer niet alle informatie expliciet mee wordt gemodelleerd — een probleem dat voor non-monotone redeneerformalismes ernstiger kan zijn dan voor monotone redeneerformalismes. Daarnaast heeft het gebruik van voorbeelden ook een zeker ad-hoc karakter. Gegeven een verzameling van voorbeelden die elkaar niet uitsluiten is het altijd mogelijk om een formalisme te definieren dat alle voorbeelden uit de verzameling correct afhandelt, zonder dat de garantie bestaat dat ook een voorbeeld dat niet tot de verzameling behoort correct wordt afgehandeld.

Een ander criterium om te bepalen of een redeneerformalisme als correct kan worden aangemerkt is om uit te gaan van postulaten. Het idee is dat het redeneerformalisme bepaalde algemene, intuitief klinkende eigenschappen dient te ondersteunen. Net als bij het gebruik van voorbeelden is ook hier het probleem dat het praktisch niet mogelijk is om correct redeneren geheel te definieren in termen van een bepaalde verzameling postulaten. Daarnaast is het voor nonmonotoon redeneren vaak een moeilijke opgave om een formalisme te specificeren dat voldoet aan bepaalde postulaten.

Een veel gebruikte methode om de betekenis van een logica te omschrijven is om gebruik te maken van een formele semantiek. Traditionele semantiek is vaak model-georienteerd. Voor defeasible logica, waar argumentatie als onderdeel van kan worden gezien, resulteert dit veelal in een semantiek van geprefereerde modellen, waarmee niet altijd even makkelijk gewerkt kan worden. Een alternatieve semantiek voor defeasible logica bestaat uit Dung's argument-interpretaties. Een algemeen probleem is dat het bestaan van een model-georienteerde semantiek niet betekent dat het formalisme ook intuitief gerechtvaardigd is. In sommige gevallen kan een semantiek gebaseerd op een dialoogspel een uitkomst bieden.

Een algemeen probleem wanneer men wil nagaan of een bepaald redeneerformalisme intuitief is, is dat men moet bepalen wat correcte intuities zijn. De intuities van een in logica ongeschoolde persoon zijn veelal niet correct, zoals verschillende studies hebben aangetoond. De intuities van iemand die zich in formeel redeneren heeft verdiept, daarentegen, zijn vaak beinvloed door de specifieke formalismen waar de betreffende persoon mee in aanraking is gekomen.

Een aanvullende methode om inzicht te verwerven in zowel informeel als formeel redeneren is dat van een *spiegelvoorbeeld*. Een spiegelvoorbeeld bestaat uit twee informele logische voorbeeldjes die, afgezien van syntax, dezelfde formalisatie delen, hoewel ze verschillende conclusies hebben. Het bestaan van een spiegelvoorbeeld dwingt de onderzoeker na te denken over het precieze verschil tussen de twee intuitieve voorbeelden en hoe dit tot uiting moet komen in de formalisatie of in de toegepaste logica(s).

Er zijn in totaal drie argumentatieformalismes gebruikt om het concept van HY-argumenten in te illustreren. Als eerste is gekozen voor het formalisme van Prakken en Sartor. Een van de redenen hiervoor is dat dit ondanks zijn relatieve eenvoud de twee bekende manieren waarop argumenten elkaar kunnen aanvallen (rebutting en undercutting) ondersteunt; daarnaast kan het tevens omgaan met prioriteiten. Aan de hand van enkele intuïtieve voorbeelden worden de negatieve gevolgen van het ontbreken van HY-argumenten geschetst. Vervolgens wordt er een informele analyse gemaakt van hoe HY-argumenten klassieke argumenten en andere HY-argumenten kunnen aanvallen. Deze analyse wordt gedaan aan de hand van het begrip *commitments*, de standpunten die een partij in een dialoog heeft ingenomen en welke hij tegen mogelijke kritiek moet verdedigen.

Gebruik makend van deze analyse wordt vervolgens het formalisme van Prakken en Sartor uitgebreid met HY-argumenten en wordt getoond dat de intuitieve voorbeelden die zonder HY nog verkeerd gingen nu correct worden opgelost. Daarnaast wordt getoond hoe HY samenhangt met de principes van conclusiemaximalisatie en regelmaximalisatie. Voor regelmaximalisatie geldt het postulaat *cautious monotonity*, iets wat voor conclusiemaximalisatie niet geldt. Een speciale eigenschap van het met HY-argumenten uitgebreide formalisme van Prakken en Sartor is dat het mogelijk wordt om een defeasible theory uitsluitend met defeasible regels te beschrijven; stricte (non-defeasible) regels zijn niet nodig.

Hoewel HY en contrapositie verschillende fenomenen zijn met verschillende effecten bestaan er tevens belangrijke overeenkomsten. Beide vormen van redeneren kunnen als geldige principes worden beschouwd in epistemische redeneerprocessen, waarbij het doel is het zo goed mogelijk redeneren over een objectieve werkelijkheid. Voor constitutief redeneren, waarbij feiten worden geschapen door het redeneerproces zelf, hoeven contrapositie en HY geen geldige principes te zijn. Veel vormen van constitutief redeneren zijn immers van menselijke aard en er zijn duidelijke indicaties dat mensen veelal niet goed overweg kunnen met contrapositie en HY. Het feit dat epistemisch en constitutief redeneren verschillende eisen stellen kan worden geïllustreerd met behulp van een spiegelvoorbeeld.

HY-argumenten zijn in principe volledig verenigbaar met de argument-gebaseerde semantieken zoals gedefinieerd door Dung. Het enige waarop gelet moet worden is dat het mogelijk is dat HY-argumenten gerechtvaardigd worden, een begrip dat eigenlijk alleen een zinnige betekenis heeft voor klassieke argumenten. Wanneer men alleen geïnteresseerd is in de gerechtvaardigde conclusies, dan kan dit probleem vrij eenvoudig worden ondervangen door uitsluitend te kijken naar de *klassieke* gerechtvaardigde argumenten; een alternatieve oplossing is overigens ook mogelijk.

HY-argumenten kunnen niet alleen worden toegevoegd aan het formalisme van Prakken en Sartor, maar ook aan default logic en aan het formalisme van Pollock. Wat betreft default logic bestaan er grofweg drie manieren om een HY-stijl van redeneren te bewerkstelligen. Om te beginnen kan men uitgaan van een argument-interpretatie van default logic en hier direct HY-argumenten aan toevoegen. De bekende en wenselijke eigenschap dat een normal default theory altijd minstens één extensie heeft blijft gelden, ook na toevoeging van HY-argumenten. Daarnaast kan voor normal defaults ook de afleidingsrelatie worden aangepast om zo het principe van regelmaximalisatie te verkrijgen. Een derde mogelijkheid is om uit te gaan van free defaults, waarbij overigens diverse andere effecten op de koop toe moeten worden genomen.

John Pollock is één van de weinige onderzoekers die hetzelfde soort problemen heeft onderzocht als waar HY een aanpak voor heeft. Pollock's oplossing is echter grotendeels van semantische aard. Met behulp van een spiegelvoorbeeld valt in te zien dat er situaties zijn waarin een puur semantische aanpak geen uitkomst biedt. Als in Pollock's formalisme suppositioneel redeneren wordt beperkt is het echter zeer wel mogelijk om ook hier HY-argumenten te definiëren, waarmee de problemen die Pollock heeft aangestipt op een nette manier worden opgelost.

SIKS Dissertation Series

1998

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-4 Dennis Breuker (UM)
Memory versus Search in Games

1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

1999

1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-3 Don Beal (UM)
The Nature of Minimax Search

1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-5 Aldo de Moor (KUB) *Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems*

1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-7 David Spelt (UT)
Verification support for object database design

1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

2000

2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-2 Koen Holtman (TUE) *Prototyping of CMS Storage Management*

2000-3 Carolien M.T. Metselaar (UvA)
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectie

2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-8 Veerle Coupé (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

2001

2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-3 Maarten van Someren (UvA)
Learning as problem solving

2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-6 Martijn van Welie (VU)
Task-based User Interface Design

2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design

2001-11 Tom M. van Engers (VU)
Knowledge Management: The Role of Mental Models in Business Systems Design

2002

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel (KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (UvA)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UvA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UvA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek (UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise