Reasoning with Defeasible Reasons

Pandzic, Stipe

DOI:
10.33612/diss.136479932

Stipe Pandžić

Reasoning with Defeasible Reasons

Printing: Ridderprint B.V.

Colophon: This thesis was typeset with LaTeX, using Diego Puga's Pazo math fonts.

university of groningen

Reasoning with Defeasible Reasons

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Thursday 29 October 2020 at 9.00 hours

by

Stipe Pandžić

born on 5 July 1987
in Mostar, Bosnia and Herzegovina

Supervisors

Prof. B.P. Kooi
Prof. L.C. Verbrugge
Prof. A.M. Tamminga

Assessment Committee

Prof. J.M. Broersen
Prof. T. Studer
Prof. H.B. Verheij

The child's toys and the old man's reasons
Are the fruits of the two seasons.

—William Blake, *Auguries of Innocence*

Contents

Introduction

What is this thesis about?

This thesis grew out of interests in understanding the principles of ordinary or commonsense reasoning. This type of reasoning is easily performed by most human reasoners, thus deserving the title of "ordinary" and "commonsense". Imagine that you were to see a picture of cherry blossoms from Tokyo in an October newspaper edition. Knowing that Japanese cherry normally blossoms in March or April, you reasonably conclude that the photo must be at least half-a-year old. But were you to further learn that the Tokyo temperatures this autumn are similar to those of spring's, you would be inclined to discard your original conclusion that the photo is old. This phenomenon of withdrawing conclusions upon considering additional information is known as "non-monotonicity" of inference. A good deal of what commonsense reasoning is about is connected to non-monotonic inferences. Although humans seem to easily engage in commonsense reasoning, it is notoriously difficult to systematically explain its underlying workings. This problem came to the attention of AI researchers who realized that the design of intelligent computer programs requires understanding of and ability to engineer common sense.

One distinctive feature of commonsense reasoners, as opposed to ideal reasoners, is that they make *errors*. Reasoning errors often do not result from obtuseness or irrational behavior, but rather from a *need* to draw conclusions despite having only incomplete information about a relevant subject matter. If an agent has complete information about a situation and it is able to reason deductively, then its inferences are monotonic and any addition of new information will not question previous conclusions. Ordinary reasoners seldom (if ever) have complete information about any contingent fact and they are "forced" to draw conclusions that can turn

out to be wrong. From the 1970s onward, researchers in AI have noticed the importance of reasoning errors. This has led to the later development of formal systems with inference rules that hold *other things being equal*, but fall short of deductive validity.

Recognizing that monotonic logics may not offer sufficient tools for modeling ordinary reasoning has led some researchers to a more skeptical stance toward formal logics. While AI researchers accepted non-monotonicity as one of the staples of the new types of logical systems, some trends in the 20^{th} century philosophy saw reasoning errors and limitations of human reasoning as an indication that formal logics and ordinary reasoning are not as closely connected as the philosophical tradition has it.[1] For example, Harman (1984, p. 112) defends a view according to which logic has no "special role in reasoning". He thinks that logic is neither a descriptive theory of how humans reason nor a prescriptive theory of how humans humans ought to reason.[2]

In contrast to such trends, the unifying idea behind this thesis is that there are both non-monotonic logics that adequately describe ordinary reasoning and those that show how logical norms are prescriptive in ordinary reasoning. As it becomes clear throughout the thesis, we do take reasoning errors and logical limitations of ordinary reasoning as constituents of the systems we develop. However, we do not accept skepticism regarding the role of logical norms in ordinary reasoning and we do not accept skepticism regarding the role of logic in modeling ordinary reasoning. In that sense, this thesis is an attempt to advance the optimistic view of the connections between formal logic and ordinary reasoning that currently has more proponents among AI researchers. A long-term goal, however, is to advocate that understanding the logical principles of commonsense reasoning should also be in the focus of philosophical theories of reasoning.

One of the main steps that we plan to undertake in this direction is to reinstate arguments as a subject matter of formal logic. The 20^{th} century witnessed formal logic and argumentation theory parting their ways,

[1] Notably, Kant (1781/1998, p. 194 A52/B76) claimed that logic is "the science of the rules of understanding in general" and Frege (1893/1964, p. 12) saw logic as prescribing "the way in which one ought to think if one is to think at all".

[2] Traditionally, the view that logic is not a descriptive theory of human reasoning has had many proponents, among them also Frege (1893/1964, p. 12). The view that logic is not a prescriptive theory either is a more recent one, gaining its popularity throughout the second part of the twentieth century. For example, one recent defense of such view is given by Russell (2017).

most famously in the seminal work of Toulmin (1958/2003, p. 111), who believed that deciding on the tenability of (most) arguments requires more than looking at their logical form. This trend gave rise to the field of informal logic which aims to analyze those features of arguments that are deemed out of the scope of formal methods.[3]

This trend has not curbed the development of formal methods that deal with arguments. In the 1980s, researchers in the field of artificial intelligence became interested in developing systems that formalize argumentation. In this respect, Pollock's work (1987, 1992, 1995) on defeasible reasoning is a pioneering attempt to find a formal system for argument-based inference (Prakken, 2017, p. 2186). However, the most influential formal account of arguments has been Dung's (1995) theory of abstract argumentation frameworks. Although Dung's frameworks do not represent arguments with the richness of their internal structures, they offer an elegant mathematical account of oppositions or attacks among arguments. A good amount of later research strove to find a comparably elegant formal system that would also include the structure of arguments in argumentation frameworks.

This thesis gives an answer to the problem of modelling structured arguments from a formal logic perspective. To enable logical representation of arguments, we first define a logical system with **defeasible reasons** represented in its object language. The logic proposed here has two main basic components. The first component is Artemov's (2001) justification logic. Justification logic is an extension of standard epistemic logic in which we replace the 'modal box' operator preceding some proposition formula, e.g. $\Box F$ for "F is known", with a justification term or reason term t that gives information on the source of epistemic justification for F. The resulting expression $t : F$, called *justification assertion*, reads as "F is known because of the reason t". The format of justification assertions alone is already suggestive of the paradigmatic pairs of reasons and conclusions that are typically associated with structures of arguments. In this thesis, we want to pin down the logical workings behind this intuition.

The second basic component to our new system is non-monotonic

[3]Recently, Hample (2007) proposed that no symbolic representation of arguments, be it in a formal or in a natural language, should be the primary object of argument analysis. Instead, Hample suggests that the attention should be on arguers and that any symbolic and textual form of argument is just an artifact of the process of arguing (Hample, 2007, p. 164).

reasoning or, more specifically, defeasible reasoning. To obtain the desired connection between justification assertions and arguments, we need to be able to model such reasons that are able to conflict and defeat each other. This is the idea of defeasibility of reasons that permeates the formal study of arguments in AI. What makes *defeasibility* central to the study of arguments? The answer is that (most) arguments rely on reasons that, in principle, cannot eliminate every possibility of encountering reasons that would oppose them — this holds at least of those reasons that are not as strong as mathematical proofs. Therefore, to develop a logical theory of arguments, the logic needs to be able to deal with defeasible reasons.

Pollock (1987, p. 482) was the first to notice out that what philosophers study as "defeasibile reasoning" had already been studied in AI through what is known as "non-monotonic reasoning". This connection is important for the system that we present here. Our method to formalize defeasible reasons is to define a non-monotonic logic with explicit representation of defeasible reasons based on the language of justification assertions. The AI tradition offers Reiter's default logic (Reiter, 1980) as a standard way to deal with the type of non-monotonicity that is induced by allowing defeasible inferences.[4] Reiter proposed inference rules called "defaults", which permit drawing defeasible conclusions that hold normally, but not without exceptions, as long as drawing such conclusions does not lead to inconsistency.

We adapt Reiter's idea of inference rules with defeasible conclusions to the mentioned calculus of reason terms in justification logic. The resulting logic of default justifications fulfills the goal of representing arguments in justification logic. The new logic brings value to justification logic, which can now be considered as a general theory of reasons. By extending the calculus of reason terms to the case of defeasible reasons, justification logics can be fully integrated with the philosophical study of (non-mathematical) reasons justifying contingent statements. This is the key step to enable a formal account of the Platonic definition of knowledge as *justified* and true belief.[5]

Default justification logic also brings value to argumentation theory. Most importantly, it shows that tenability of arguments is a subject-matter

[4]Non-monotonic approaches in AI offer a variety of alternatives to formalize the ideas of defeasible reasoning, including *circumscription* (McCarthy, 1980) and *autoepistemic logic* (Moore, 1985).

[5]The idea of modelling justified true beliefs has been one of the focal points of Artemov's (2008) program for bridging justification logic and mainstream epistemology.

of formal logic. One of the results is that our logic of default justifications determines whether an argument is acceptable or not at a purely symbolic level through a normative system with logical consequence. This is one of the features that distinguishes our logic with structured arguments from the existing structured argumentation *frameworks*.[6] These frameworks are less-abstract formal accounts of arguments compared to Dung's abstract argumentation frameworks, since they do attempt to represent internal structures of arguments. What is new in our logic is that we represent arguments as primary objects of the logical language and we decide on their acceptability through a definition of logical consequence. Thus the aim of the thesis is not only to use formal logic notions to model arguments, but to define a full-fledged logic of arguments that manipulates structured arguments at a purely symbolic level.

The way in which we interpret default assumptions in justification logic provides a way to model the basic types of argumentative attacks called *rebuttal* and *undercut* (Pollock, 1987, p. 485). The two concepts play an important role in the semantics of justification logic default theories and we want to introduce them here informally. Given a *prima facie* reason[7] and some conclusion justified by that reason, a rebutting defeater is a reason for the opposite conclusion. An undercutting defeater for that *prima facie* reason is a reason that attacks the connection between the *prima facie* reason and the conclusion it supports. The logic we develop in this thesis especially aims to advance the study of undercutting or exclusionary defeat, which has been notoriously difficult to model by logical means. To see why, consider that modelling rebuttals can be done in a more straightforward manner, since rebuttals can be translated into inconsistency among statements. However, to model undercutters, we need a more expressive language that represents or "reifies" (Horty, 2007) reasons in its object-level formulas. This is so because we cannot simply reduce undercut to inconsistency. What we need instead is a way to say that a default conclusion is *normally* acceptable when supported by a given *prima facie* reason, but not under some exclusionary circumstances.[8]

In addition to undercutting and rebutting defeat, AI researchers have

[6]Some well-known structured argumentation frameworks are ABA (Bondarenko et al., 1997), deductive argumentation (Besnard and Hunter, 2001), DeLP (García and Simari, 2004) and ASPIC+ (Prakken, 2010).

[7]A reason that provisionally holds, unless disproved by new information.

[8]This challenge is recognized by Horty (2012). See (Horty, 2012, Ch. 5) for his variant of default logic-based representation of undercutting or exclusionary defeaters.

investigated an additional standard type of argument defeat called "undermining" (van Eemeren et al., 2014, p. 626). Intuitively, an argument is undermined when one of its premises is denied. This thesis also provides a logical account of undermining in default justification logic. Notice that undermining does not target default inference, as undercutting, or default conclusion, as rebutting, but rather attacks an argument's premise as a starting point for default reasoning. This motivated the distinction between default and plausible reasoning in formal argumentation that we adopt in this thesis. In the plausible reasoning paradigm, fallibility of reasoning results from adding new information that questions old information and, thereby, it might question old conclusions.[9] In contrast, in the default reasoning paradigm, fallibility results from adding some further true information on top of existing information and this new information in turn gives reasons to question old conclusions, but they do not question old information.[10]

In our default logic, reasoning starts from a set of facts (also called "axioms" and "premises"), which is then extended by conclusions that hold by default. We argue that modeling plausible reasoning and undermining defeaters in the view of default theories requires changing the set of starting premises upon receiving new information. Thus we give a dynamic aspect to our default justification logic and model changes to premises using the techniques from the logic of belief revision (Hansson, 1999a). More specifically, undermining is modeled with belief revision operations that include *contracting* the set of starting premises, that is, removing some information from a set of facts.

Besides the logical system of default reasons, this thesis uses the idea of defeasibility to shed a new light on the problem of normativity of logical rules. This problem has its roots in Harman's (1986) criticism of

[9]Rescher's (1976, 1977) work is the landmark reference for the study of plausible reasoning. Rescher (1977, p. 39) claims that "a thesis is more or less plausible depending on the reliability of the sources that vouch for it".

[10]Note here that Prakken (2017, p. 2198) refers to the difference between *defeasible* and *plausible* reasoning, instead of default and plausible reasoning. To be clear about the terminology, we use the etymologically close terms "defeasible" and "defeat" in a more general sense, so that, e.g., undermining is normally also considered as a type of defeat. This conforms to the standard usage of "defeat" and "defeasibility" which simply mean that something is annulled. The term "default", on the other hand, has a more specific meaning related to default assumptions introduced by Reiter (1980, p. 82). Vreeswijk (1993) introduced the distinction between the two kinds of non-monotonicity to argumentation theory (using the term "defeasible").

the relevance of formal logic for human reasoning. Harman argues that classical logic has neither a normative role nor an explanatory role in human reasoning. His position on the role of logic in human reasoning is known as "anti-psychologism". According to Harman, if logical rules had a normative role in human reasoning, we would be able to come up with a normative principle that connects formal logic and human reasoning.[11]

Harman considers multiple candidate principles to bridge logic and human reasoning, only to reject each of them and to skeptically conclude (Harman, 1986, p. 20) that "there is no clearly significant way in which logic is specially relevant" for human reasoning. The idea of such principle spurred the "Bridge Principle" debate in the philosophy of logic, with an aim to find a principle that fulfills Harman's requirements. We argue that Harman's conclusion does not follow, once we take into account that normative rules in human reasoning, just like normative rules in general, are defeasible rules only. We offer a system that interprets logical rules as default norms to show that Harman's counterexamples to the normative role of logic in human reasoning do not hold. Moreover, we argue that it is not necessary to "bridge" logic and reasoning by coming up with a bridge principle so as to claim that classical logic is normative for human reasoning.

We stated in the introduction that we focus on fallible agents who, unlike ideal agents, are prone to making reasoning errors. It seems that such agents need to be aware of their fallibility and adopt a modest attitude toward their ability to form true beliefs. This issue is known as "doxastic modesty". As a final topic of this thesis, we investigate the limits on how far could a fallible and modest agent go in acknowledging its fallibility. The phenomenon of doxastic modesty statements came into prominence after Makinson (1965) published the paradox of the preface. According to the paradox, an author of a non-fictional book is justified to believe each assertions in one such book. However, being aware of one's own fallibility, the author is justified to disbelieve the conjunction of all assertions in the book and to acknowledge so in the book preface with an appropriate statement of doxastic modesty. It seems that doxastic modesty requires the author to entertain justified inconsistent beliefs. Moreover and more generally, it seems that doxastic modesty requires all

[11]One might, for example, think that deductive closure can bridge logic and reasoning by means of the following principle:"If some statement is classically entailed by one's set of beliefs, then that statement should be added to the set of beliefs".

fallible agents to believe the doxastic modesty statement "At least one of my beliefs is false".

We analyze the process by which an agent could learn that statement. Instead of focusing on inconsistency of beliefs, we highlight the connection between doxastic modesty statements and Moorean statements. We argue that agents cannot in principle learn any of the straightforward versions of doxastic modesty statements. Similar results are already known in the case of Moorean statements. This weakens arguments in support of the claim that doxastic modesty requires agents to believe that one of their beliefs is false. What is needed to save those arguments is to employ some *ad hoc* assumptions on agents' beliefs that give special treatment to their beliefs in doxastic modesty statements.

Outline of the chapters

The rest of this thesis is structured as follows.

- In **Chapter 1**, we present technical requirements for reading the rest of the thesis. We first present justification logics, which give the basic language for the logic of defeasible arguments. Then we describe the basics of standard default logic. Finally, we briefly familiarize readers with abstract argumentation frameworks. The order of presentation follows the order of use of these systems throughout the thesis.

- In **Chapter 2**, we develop a logic of structured defeasible arguments using the language of justification logic. In this logic, we introduce defeasible justification assertions of the type $t : F$ that read as "t is a defeasible reason that justifies F". Such formulas are then interpreted as arguments and their acceptance semantics is given in analogy to Dung's abstract argumentation framework semantics.

  We first define a new justification logic that relies on operational semantics for default logic. One of the key features that is absent in standard justification logics is the possibility to weigh different epistemic reasons or pieces of evidence that might conflict with one another. To amend this, we develop a semantics for "defeaters": conflicting reasons to doubt the original conclusion or to believe an opposite statement. In our logic, reasons are non-monotonic and their acceptability status can be revised in the course of reasoning.

Then we present our logic as a system for abstract argumentation with structured arguments. The format of conflicting reasons overlaps with the idea of attacks between arguments to the extent that it is possible to define all the standard notions of extensions of argumentation frameworks.

- In **Chapter 3**, we establish a formal correspondence between Dung's original argumentation semantics and our operational semantics for default theories. We show that a large subclass of Dung's frameworks that we call "warranted" frameworks is a special case of our logic: (1) Dung's frameworks can be obtained from justification logic-based theories by focusing on a single aspect of attacks among justification logic arguments and (2) Dung's warranted frameworks always have multiple justification logic instantiations, called "realizations", in the sense of multiple corresponding default theories.

  In the same chapter, we compare our logic to Reiter's default logic interpreted as an argumentation framework. The comparison is done by analyzing differences in the ways in which process trees are built for the two logics. The aim is to show that our logic solves the problem of modeling undercut and exclusionary reasons in default logic.

- **Chapter 4** covers information changes in default justification logic with argumentation semantics. We introduce dynamic operators that combine belief revision and default theory tools to define both prioritized and non-prioritized operations of contraction, expansion and revision for justification logic-based default theories. We argue that the combination enriches both default logics and belief revision techniques. We model the kind of attack called "undermining" with those operations that contract a knowledge base by an attacked formula.

- In **Chapter 5**, we argue for weak psychologism — the claim that logical rules are normative for human reasoning — by offering a new, default logic perspective on the normativity of logic. First we discuss Harman's proposed counterexamples to the normativity of classical logic. We show that Harman's argument hinges on the claim that there is no exceptionless normative principle that requires human agents to follow the rules of classical logic. This is right, but, contrary to what Harman claims, we argue that this does

not suffice to refute weak psychologism. Instead, we argue that Harmanian bridge principles presuppose two requirements that a normative principle cannot meet, namely the non-defeasibility requirement and the relevance requirement. We show that both requirements are unnecessary. Moreover, we define a new variant of default logic for ordinary reasoning as an alternative framework for normative rules. Using this default logic, we present a picture of how logic is normative for human reasoning.

- In **Chapter 6**, we argue that an agent cannot in principle form a belief in the statement "At least one of my beliefs is false", without having to revise it immediately after. Once this statement has been learned, it should not be believed any more. Agents encounter a problem of the similar kind when learning Moorean statements. To avoid this problem, agents can refer to their totality of beliefs slightly differently and, thereby, avoid the change of the believed statement. We argue that each of the two *ad hoc* solutions that we discuss cannot be convincingly defended. Finally, we suggest that doxastic modesty justifies suspension of the belief in the conjunction of one's beliefs and it also justifies believing doxastic modesty statements that do not claim that one in fact believes falsely.

Chapter 1

Preliminaries

1.1 Introduction

This chapter introduces the basic formal ingredients used throughout the thesis: justification logics, default logic, and abstract argumentation frameworks. Since each of these systems has yielded a field of research with a rich tradition, the chapter focuses on the standard aspects of the three systems that contribute to a better understanding of the system developed in the rest of this thesis. Since the role of the language of justification logic is central to the development of the logic of defeasible argumentation in Chapter 2, the most extensive part of this chapter is given to a systematic exposition of justification logics.

1.2 Justification logic

Informally, justification logics are systems that enable mathematically rigorous representation of reasons or justifications. The terms "reason" and "justification" are usually understood as reasons to believe or know, but, in general, the language supports other non-doxastic and non-epistemic interpretations. However, justification logic grew out of a more specific interest in formalizing the idea from constructive mathematics that truth can be identified with provability. Thus, the original intention was not to deal with reasons in their broadest capacity, but only with a specific group of reasons: formal mathematical proofs. In this thesis, we adopt the usual interpretation of justification logics as logics that model reasons to *believe*, to *know*, or, in general, to *accept* claims.

Typical for justification logics is their use of the format of labelled formulas:

$$term : formula,$$

representing pairs of reasons and claims. In the object language, they are written as the so-called "justification assertions" $t : F$ that read as "t is a reason that justifies formula F". The first justification logic was developed as a logic of proofs in arithmetic (logic of proofs, **LP**) by Artemov (2001).[1] On the original reading of pairs $t : F$, the term t encodes some Peano arithmetic derivation for the statement F.

Soon after Artemov introduced the logic of proofs (**LP**) in (2001), Fitting (2005a, 2005b) proposed a possible worlds semantics for this logic in order to incorporate justification logics in the family of modal logics. Syntactic objects that represent mathematical proofs in the logic of proofs **LP** are then more broadly interpreted as epistemic or doxastic reasons by Fitting (2005a, 2005b) and Artemov and Nogina (2005). A distinctive feature of justification logic taken as epistemic logic is replacing belief and knowledge modal operators that precede propositions ($\Box F$ for "F is known") by proof terms or, in a generalized epistemic context, justification terms. Next to the usual possible world condition for the truth of $t : F$ that F is true in all accessible alternatives, Fitting's semantics requires that the reason t is admissible for formula F.

The language of justification logic builds on the language of propositional logic, which is augmented by formulas labelled with reason terms ($t : F$) and a grammar of operations on such terms. Reason terms are built from constants and variables, using operations on terms. Intuitively, constants justify logical postulates and variables justify contingent facts or inputs outside the structure. The basic operation of standard justification logics is *application*. Intuitively, application produces a reason term $(u \cdot t)$ for a formula G which is a syntactic "imprint" of the *modus ponens* step from $F \to G$ and F to G for some labelled formulas $u : (F \to G)$ and $t : F$. We say that the term u has been applied to the term t to obtain the term $(u \cdot t)$. The *Application* axiom is present in all standard justification logics:

$$u : (F \to G) \to (t : F \to (u \cdot t) : G).$$

[1] The idea of explicit proof terms as a way to find the semantics for the provability calculus **S4** dates back to Gödel's 1938 lecture published in (Gödel, 1995). For a more encompassing overview of standard justification logics see (Artemov and Fitting, 2019) or (Kuznets and Studer, 2019).

The axiom displays a distinctive feature of justification terms by which the history of reasoning steps taken in producing such terms is recorded in their structure.

Another common operation on justification terms is *sum*. Intuitively, if a reason term t justifies some formula F, then, by sum, we can add any other reason term u so that the new reason term $(t + u)$ still justifies F. On an epistemic interpretation, this operation can be informally motivated as follows (Artemov and Fitting, 2016, Section 2.2): t and u might be thought of as two volumes of an encyclopedia that are used as evidence for some statement F. If one volume justifies F, then adding the other volume to the corpus of evidence does not compromise the justification for F. This intuition is captured by the *Sum* axioms:

$$t : F \to (t + u) : F \quad \& \quad u : F \to (t + u) : F.$$

These axioms represent the requirement of monotonicity on reasons and prevent that adding new information compromises already accepted reasons. The axioms regulating the sum and application operations are formally described in this section, following the definition of the language. In relation to monotonicity of reasons, it is worth noting here that this thesis seeks to meet what Artemov (2001, p. 482) considers to be "an intriguing challenge to develop a theory of nonmonotonic justifications which prompt belief revision".

Additionally, standard justification logics may include unary operators '!' and '?' on terms that occur in axioms about agents' introspective abilities. The *Positive Introspection* axiom

$$t : F \to !t : t : F$$

is a justification logic variant of the modal logic axiom **4**: $\Box F \to \Box\Box F$. On an epistemic reading of the modal logic "box", the axiom says that "if an agent knows F, then the agent knows that it knows F". The operation '!' does not simply iterate the reason t for F, but gives a "meta-evidence" (Artemov, 2008, p. 494) that t is a correct reason for F. An example motivated by the original provability reading of justification terms could be that the output term $!t$ is taken to be a justification of each line in a natural deduction proof t for a proposition F. Therefore, the operation '!' is known under the name *Proof Checker*.

Historically, the first justification logic (logic of proofs **LP**) consisted of the above Application, Sum and Positive Introspection axioms, together

with the *Factivity* axiom: $t : F \rightarrow F$. This axiom is an explicit counterpart to the modal *Truth axiom*: $\Box F \rightarrow F$ read as "If F is known, then F". Together with Sum, Factivity is an "embodiment" of the requirement of non-defeasibilty for reasons: "there can be no other truths such that, had I believed them, would have destroyed my justification for believing F". The ramifications of non-defeasibility requirements on reasons will be among the main topics of this thesis. In particular, we search for a logical theory of reasons that do not necessarily persist as acceptable reasons after new information has been added.

In contrast to Positive Introspection, the *Negative Introspection* axiom

$$\neg t : F \rightarrow ?t : \neg t : F$$

is not accepted for a logic of arithmetic proofs. The type of operation that '?' represents "does not exist for formal mathematical proofs since $?t$ should be a single proof of infinitely many propositions $\neg t : F$, which is impossible" (Artemov, 2008, p. 495). Consider that, in order to be suitable for the context of formal proofs, '?' would need to take t as its only input to justify that $\neg t : F$ holds for infinitely many propositions F that a proof represented by t does not prove. Throughout the rest of the thesis, we do not consider the introspection axioms. In fact, we will build our logic starting with a system of non-defeasible reasons that includes only propositional axioms, Application, Sum and Factivity. However, for the purposes of this preliminaries chapter, we describe the most well-known justification logic: the logic of proofs **LP**.

The following grammar summarizes the informal discussion of the available operations and describes a way to build the formulas from the language of **LP** starting from the propositional base:

- a countable set $\mathcal{P}$ of propositional atoms: $P_1, \ldots, P_n, \ldots$
- connectives: $\neg, \wedge, \vee, \rightarrow$
- parentheses: $(,)$
- the 'top' symbol denoting an arbitrary tautology: $\top$
- reason terms (polynomials) $t_1, \ldots, t_n, \ldots$ built from:
  1. justification variables $x_1, \ldots, x_n, \ldots$
  2. justification constants $c_1, \ldots, c_n, \ldots$

using binary ('$+$' and '$\cdot$') and unary ('!') operators

- operator symbol of the type $\langle term \rangle : \langle formula \rangle$

On the basis of the alphabet above, we define the set of all reason terms Tm and the set of all formulas Fm. We first say that each term from the set of all terms Tm has to be built according the following grammar:

1. Any constant c is a reason term and any variable x is a reason term.
2. If t is a reason term, then $(t \cdot t)$, $(t + t)$ and $!t$ are reason terms.

Using Tm, we give the following grammar of **LP** formulas from Tm:

1. Any propositional atom $P \in \mathcal{P}$ is a formula and $\top$ is a formula.
2. If F is a formula, then $\neg F, F \to F, F \vee F$ and $F \wedge F$ are formulas.
3. If t is a reason term from Tm and F is a formula, then the combination $t : F$ is also a formula.

The selection of axioms for **LP**, which were all introduced above, is given by the following list:

A0 *All the instances of propositional logic tautologies from Fm*

A1 $t : (F \to G) \to (u : F \to (t \cdot u) : G)$ (Application)

A2 $t : F \to (t + u) : F$; $\qquad u : F \to (t + u) : F$ (Sum)

A3 $t : F \to F$ (Factivity)

A4 $t : F \to !t : t : F$ (Positive Introspection)

Combined with the following two rules, we described the logic **LP**:

R0 *From F and* $F \to G$ *infer G* (Modus ponens)

R1 *If F is an axiom instance of A0-A4 and c a proof constant, then infer* $c : F$ (Axiom necessitation)

The formula F is **LP**-provable ($\mathbf{LP} \vdash F$) if F can be derived using the axioms A0-A4 and rules R0 and R1. The following is an example derivation of a formula in **LP**:

LP $\vdash x : (F \wedge G) \rightarrow ((c \cdot x) : F \wedge (d \cdot x) : G)$.

1	$(F \wedge G) \rightarrow F, (F \wedge G) \rightarrow G$	(A0)
2	$c : ((F \wedge G) \rightarrow F), d : ((F \wedge G) \rightarrow G)$	(1 R1)
3	$c : ((F \wedge G) \rightarrow F) \rightarrow (x : (F \wedge G) \rightarrow (c \cdot x) : G)$	(A1)
4	$d : ((F \wedge G) \rightarrow G) \rightarrow (x : (F \wedge G) \rightarrow (d \cdot x) : G)$	(A1)
5	$x : (F \wedge G) \rightarrow (c \cdot x) : F$	(2,3 R0)
6	$x : (F \wedge G) \rightarrow (d \cdot x) : G$	(2,4 R0)
7	$(x : (F \wedge G) \rightarrow (c \cdot x) : F) \rightarrow ((x : (F \wedge G) \rightarrow (d \cdot x) : G) \rightarrow$ $(x : (F \wedge G) \rightarrow ((c \cdot x) : F \wedge (d \cdot x) : G)))$	(A0)
8	$(x : (F \wedge G) \rightarrow (d \cdot x) : G) \rightarrow (x : (F \wedge G) \rightarrow$ $((c \cdot x) : F \wedge (d \cdot x) : G))$	(5,7 R0)
9	$x : (F \wedge G) \rightarrow ((c \cdot x) : F \wedge (d \cdot x) : G)$	(7,8 R0)

□

The theorem above is an explicit version of the formula $\Box(F \wedge G) \rightarrow (\Box F \wedge \Box G)$, which is a theorem of the modal logic **K**.

Notice that our use of the constants c and d in this proof is arbitrary in the sense that R1 does not restrict our choice of proof constants used in line 2. In justification logics, basic logic axioms are taken to be justified by virtue of their status within a system and their justifications are not further analyzed. Moreover, we may also treat any such formula $c : F$ as an axiom in the system and postulate that some proof constant d justifies $c : F$. A set of instances of all such canonical formulas in justification logic is called a *Constant Specification* ($\mathcal{CS}$) set. The following is the general definition of constant specification sets, which subsumes the set produced as the set of instances of rule R1 above:

Definition 1 (Constant Specification)**.**

$$\mathcal{CS} = \{c_n : c_{n-1} : \cdots : c_1 : F \mid F \textit{ is an axiom instance of A0-A4,} \\ c_n, c_{n-1}, \ldots, c_1 \textit{ are proof constants and } n \in \mathbb{N}\}$$

Rule R1 generates a set of formulas in which any constant justifies any instance of $A0 - A4$. This defines only one possible constant specification set. One could require, for example, that every axiom instance comes with a unique constant.

The choice of a constant specification set may be included as a parameter of logical awareness for a justification logic. This is done by relativizing the Axiom necessitation rule to a constant specification as follows:

R1* *If F is an axiom instance of A0-A4 and $c_n, c_{n-1}, \ldots, c_1$ are proof constants such that $c_n : c_{n-1} : \cdots : c_1 : F \in \mathcal{CS}$, then infer $c_n : c_{n-1} : \cdots : c_1 : F$* (Iterated axiom necessitation)

For example, the simplest standard justification logic $\mathbf{J}_\emptyset$ is defined by axioms A1, A2, rules R0, R1 and an empty constant specification, which means that $\mathbf{J}_\emptyset$ does not support any form of axiom necessitation rules.

Next to the *Empty* constant specification ($\mathcal{CS} = \emptyset$), other standard choices of constant specification sets include (Artemov and Fitting, 2019, pp. 17-18):

- *Total* ($\mathcal{TCS}$): any axiom instance can be labelled with any sequence of proof constants;
- *Finite*: $\mathcal{CS}$ is a finite set of formulas;
- *Axiomatically Appropriate*: for each axiom instance A, there is a constant c such that $c : A \in \mathcal{CS}$ and for each formula $c_n : c_{n-1} : \cdots : c_1 : A \in \mathcal{CS}$ such that $n \geq 1$, there is a constant c_{n+1} such that $c_{n+1} : c_n : c_{n-1} : \cdots : c_1 : A \in \mathcal{CS}$;
- *Injective*: each proof constant c justifies at most one formula.

Replacing rule R1 with R1* relative to a choice of $\mathcal{CS}$ gives the logic $\mathbf{LP_{CS}}$. Notice that the necessitation rules in justification logics regulate only logical awareness of axioms, unlike their modal logic counterpart "If F is provable, then infer $\Box F$". In justification logics with an axiomatically appropriate $\mathcal{CS}$, theorem necessitation turns into a constructive property of derivations for which the following theorem holds:[2]

(Strong) Internalization 2. *Given an axiomatically appropriate $\mathcal{CS}$ and the corresponding rule R1*, if a formula F is provable in a justification logic system*

[2]However, for any logic that contains axiom A4, an axiomatically appropriate $\mathcal{CS}$ is not necessary to ensure that the formula $c : F$ is justified. With A4, the proof checker operation ensures that $!c : c : F$ is derivable. Therefore, the logic **LP** above fulfills the requirement of internalizing each formula $c : F$ with the constant specification set generated with R1. This is the original approach taken by Artemov (2001).

with $\mathcal{CS}$ and R1, then $t : F$ is also provable for some term t built from proof constants using only '$\cdot$'.*

Proof. See (Artemov and Fitting, 2019, p. 21). □

The choice of a constant specification is thus an important parameter and not least so because it could affect complexity results, as discussed by, e.g., Milnikel (2007).[3] However, it will not be central to the development of our system of defeasible reasons in Chapter 2. Because of that, we simply assume axiomatically appropriate and injective constant specifications in which each axiom instance and each formula inferred through necessitation has its own proof constant. An intuitive class of such constant specifications (Artemov, 2018, p. 31) are $\mathcal{CS}$ sets produced by assigning Gödel numbers to axioms.

As mentioned before, on the original semantics of the first justification logic **LP**, justifications are interpreted as codes of proofs of arithmetical statements. Possible worlds semantics for justifications of generalized statements are introduced by Fitting (2005a,b). *Fitting models* made it possible to extend interpretations of syntactic objects that represent mathematical proofs as epistemic reasons (Fitting, 2005a,b, Artemov and Nogina, 2005, Artemov, 2008). As mentioned above, justification logics interpreted as doxastic or epistemic logics replace belief and knowledge modal operators that precede propositions ($\Box F$ for "F is known") by justification terms. For the truth of the justification assertion $t : F$, Fitting's semantics requires F to be true in all accessible alternatives, as familiar from standard epistemic logic, and that the reason t is admissible for formula F in the current state. In Fitting semantics, admissibility of reasons is a given determined by the admissibility function in the $\mathbf{LP_{CS}}$ model (Definition 3). In the semantics of default reasons presented in Chapter 2, admissibility is not taken to be a primitive notion. To determine whether a default reason is among admissible reasons for a formula, it is necessary to establish that its admissibility is not overridden by a conflicting reason.

Definition 3 ($\mathbf{LP_{CS}}$ model)**.** *A frame $\mathcal{F}$ is defined as a pair $< \mathcal{S}, \mathcal{R} >$ such that $\mathcal{S}$ is a non-empty set of states and $\mathcal{R}$ a binary accessibility relation on states.*

[3] Consider also epistemic implications of this choice. If we define an empty $\mathcal{CS}$, we eliminate logical awareness for an agent, while any infinite axiomatically appropriate $\mathcal{CS}$ imposes logical omniscience.

We define a function reason assignment *based on* $\mathcal{CS}$, $*(\cdot) : \mathcal{S} \times Tm \to 2^{Fm}$, *a function mapping each pair of states and terms to a set of formulas from* Fm. *We assume that it satisfies the following conditions:*

1. *If* $F \to G \in *(\mathbf{w}, t)$ *and* $F \in *(\mathbf{w}, u)$, *then* $G \in *(\mathbf{w}, t \cdot u)$
2. $*(\mathbf{w}, t) \cup *(\mathbf{w}, u) \subseteq *(\mathbf{w}, t + u)$
3. *If* $c : F \in \mathcal{CS}$, *then* $F \in *(\mathbf{w}, c)$
4. *If* $F \in (\mathbf{w}, t)$, *then* $t : F \in (\mathbf{w}, !t)$

A truth assignment $v : \mathcal{P} \to 2^{\mathcal{S}}$ *is a function assigning a set of states to each propositional formula. We define the interpretation* $\mathcal{I}$ *as a quadruple* $(\mathcal{S}, \mathcal{R}, v, *)$. *For an interpretation* $\mathcal{I}$, $\models$ *is a truth relation on the set of formulas of* $\mathbf{LP_{CS}}$.

For any formula $F \in Fm$, $\mathcal{I}, \mathbf{w} \models F$ iff

- For any $P \in \mathcal{P}$, $\mathcal{I}, \mathbf{w} \models P$ iff $\mathbf{w} \in v(P)$
- $\mathcal{I}, \mathbf{w} \models \neg F$ iff $\mathcal{I}, \mathbf{w} \not\models F$
- $\mathcal{I}, \mathbf{w} \models F \to G$ iff $\mathcal{I}, \mathbf{w} \not\models F$ or $\mathcal{I}, \mathbf{w} \models G$
- $\mathcal{I}, \mathbf{w} \models F \vee G$ iff $\mathcal{I}, \mathbf{w} \models F$ or $\mathcal{I}, \mathbf{w} \models G$
- $\mathcal{I}, \mathbf{w} \models F \wedge G$ iff $\mathcal{I}, \mathbf{w} \models F$ and $\mathcal{I}, \mathbf{w} \models G$
- $\mathcal{I}, \mathbf{w} \models t : F$ iff $F \in *(\mathbf{w}, t)$ and for each $\mathbf{w}' \in \mathcal{S}$ such that $\mathbf{w}\mathcal{R}\mathbf{w}'$, it holds that $\mathcal{I}, \mathbf{w} \models F$

In (Fitting, 2005b), axiomatic soundness and completeness of $\mathbf{LP_{CS}}$ with respect to Fitting models are proved for axiomatically appropriate constant specifications.

1.3 Default logic

The second formal ingredient in this thesis is Reiter's default logic (Reiter, 1980). Default logic is a non-monotonic logic that extends classical reasoning by introducing conclusions that hold normally, but not without exceptions. Conclusions of this type are introduced by default rules such as the following:

$$\frac{bird(Tweety) : flies(Tweety)}{flies(Tweety)}.$$

The default reads as follows: "If Tweety is a bird and if it is consistent with the current theory to assume that Tweety flies, then conclude that Tweety flies". The reasoning behind this default tells us that, *normally*, if we know that something is a bird and if it is consistent with what we already believe that it flies, then we may also believe that it indeed flies.

The idea of logic built around such rules is to take some incomplete set of facts and use default rules to extend the set of facts with defeasible conclusions as much as possible without introducing contradictory conclusions. Default reasoning of this type is formalized with Reiter's default theories:

Definition 4 (Reiter's Default Theory)**.** *A default theory Δ is defined as a pair (W, D), where the set W is a finite set of first-order logic formulas and D is a countable set of default rules.*

The set W contains facts or known information. The general form of a default rule from D in Reiter's theory is

$$\delta = \frac{\varphi : \psi_1, \ldots, \psi_n}{\chi},$$

for predicate logic formulas φ, $\psi_1, \ldots, \psi_n$ and χ.[4] By $pre(\delta)$ we denote the prerequisite φ of δ, by $just(\delta)$ we denote the set $\{\psi_1, \ldots, \psi_n\}$ of the justifications of δ and by $cons(\delta)$ we denote the consequent χ of δ.

How exactly to extend an initial set of facts with default conclusions? To give a clear formal answer, we will need a definition of default *applicability*. A default rule $\delta = \frac{\varphi:\psi_1,\ldots,\psi_n}{\chi}$ is applicable to a deductively closed set of first-order formulas S iff

- $\varphi \in S$ and
- $\neg\psi_i \notin S$ for all $\psi_i \in \{\psi_1, \ldots, \psi_n\}$.

Starting from the definition of applicability, there are two standard ways to define Reiter's theory extensions. Reiter's (1980) original approach uses fixed-point equations such that, if a set S is chosen as an extension of a theory Δ, then S corresponds to the outcome of applying all S-applicable defaults with respect to the set W. Another standard way,

[4]Note that there are also *open* defaults of the form $\frac{bird(X):flies(X)}{flies(X)}$, where X is a free variable. Such rules are default schemes and they are dealt with by using a ground substitution which assigns ground terms to variables. Open defaults thus represent *sets* of defaults.

that of Antoniou (1997), relies on an operational procedure of applying defaults to build extensions in a step-by-step manner. In this thesis, we focus on Antoniou's operational semantics that also serves as an inspiration for the operational semantics of the default justification logic from Chapter 2.

The details of operational semantics for building Reiter's logic extensions will be given shortly. Here are some desiderata for an extension set E proposed by Antoniou (1997, pp. 27-28):

- The set of facts W is included in E ($W \subseteq E$);
- E is closed under classical logical consequence ($Th(E) = E$);
- E is closed under the application of defaults in D, that is, if E is an extension, all applicable defaults have been applied.

In building extensions, we consider possible orders in which defaults from D could be applied without repetitions or possible *sequences*: $\Pi = (\delta_1, \delta_2, \ldots)$, were $\delta_1, \delta_2, \ldots \in D$. The initial segment containing the first k elements of Π is denoted with $\Pi[k]$. Any segment $\Pi[k]$ is also a sequence. In particular, $\Pi[0]$ is the empty list $(\ )$, $\Pi[1]$ is the list with the first element of Π, and for $k \geq 2$, $\Pi[k]$ is the list k elements of Π. With any sequence Π we associate the following two sets:

- $In(\Pi) = Th(W \cup \{cons(\delta) \mid \delta \in \Pi\})$;
- $Out(\Pi) = \{\neg\psi \mid \psi \in just(\delta) \text{ for some } \delta \in \Pi\}$.

Intuitively $In(\Pi)$ represents a knowledge base resulting from default application and $Out(\Pi)$ collects formulas that are supposed not to become a part of it after defaults have been applied.

Whether a sequence $\Pi = (\delta_1, \delta_2, \ldots, \delta_n)$ can be executed in the proposed order or not depends on the applicability of each rule δ_{k+1} from Π to the closed set of formulas $In(\Pi)[k] = Th(W \cup cons(\delta_1, \delta_2, \ldots, \delta_k))$. This observation is central for Antoniou's definition (1997, p. 32) of default *processes* which he uses for defining Reiter's extensions:

Definition 5 (Process)**.** *A sequence of default rules* Π *is a* process *of a default theory* $\Delta = (W, D)$ *iff every* $\delta_{k+1} \in \Pi$ *is* applicable *to the set* $In(\Pi[k])$*, where* $\Pi[k] = (\delta_1, \ldots \delta_k)$.

As mentioned before, extensions of default theories should be closed under the application of defaults. We say that a process Π is *closed* iff every $\delta \in D$ that is applicable to $In(\Pi)$ belongs to Π.

Besides closure, extension-producing processes fulfill an additional condition called *success* (Antoniou, 1997, p. 32). A process Π is *successful* if for each default rule $\frac{\varphi:\psi_1,\ldots,\psi_m}{\chi}$ from Π, justifications $\psi_1,\ldots,\psi_n$ are consistent with the consequents added to an *In*-set after all defaults have been applied. In other words, none of the formulas from an *Out*-set should become a part of an *In*-set for the same process. Intuitively, assumptions made in the process of extending the set of facts should not be invalidated with the addition of further conclusions.

We give an example of both a process that is closed and not successful and a process that is successful and not closed, using propositional logic. Let $W_0 = \emptyset$ and let

$$D_0 = \left\{ \delta_1 = \frac{\top : \neg a}{b}, \delta_2 = \frac{\top : a}{a} \right\}.$$

We define the Reiter's default theory $\Delta_0 = (W_0, D_0)$. Take the sequence $\Pi_1 = (\delta_1)$. This sequence is a process, since δ_1 is applicable to $In(\Pi[0])$. Moreover, this is a successful process because the intersection of $In(\Pi_1)$ and $Out(\Pi_1)$ is empty. However, Π_1 is not closed. The reason is that the rule δ_2 is applicable to $In(\Pi_1)$ and it is not included in Π_1.

It is easy to check that the sequence $\Pi_2 = (\delta_1, \delta_2)$ is also a process and that it is closed. Notice, however, that Π_2 is a failed process. After applying the rule δ_2, the intersection of $In(\Pi_2)$ and $Out(\Pi_2)$ both contain the formula $\neg a$. Intuitively, $cons(\delta_2)$ invalidates the assumption made to draw the conclusion $cons(\delta_1)$. Moreover, notice that the sequence $\Pi_3 = (\delta_2, \delta_1)$ is not a process and that the sequence $\Pi_4 = (\delta_2)$ is both closed and successful. The latter type of sequences is used to define Reiter's extensions:

Definition 6 (Reiter's Theory Extension)**.** *A set of first-order formulas E is an extension of a default theory $\Delta = (W, D)$ iff there is a closed and successful process Π of Δ such that $E = In(\Pi)$.*

For the theory Δ_0, our analysis implies that its only extension is the set $In(\Pi_4)$. For more complex default theories, Antoniou (1997, p. 34) introduces a convenient method of finding default theory extensions through drawing process trees that we use in Chapter 2 and Chapter 3.

At last, we can define the notion of validity for Reiter's default logic. Using the definition of extensions, there are two different notions of entailment for a Reiter's theory Δ:

Skeptical entailment $\Delta \mid\sim_s \varphi$ iff φ is in all extensions of Δ.

Credulous entailment $\Delta \mid\sim_c \varphi$ iff φ is in at least one extension of Δ.

Notice that the set of formulas S that consists of all credulous consequences for a theory Δ may be inconsistent.

For an illustration of default reasoning with inconsistent conclusions in Reiter's logic, consider the "Nixon diamond" scenario in which applying defaults leads to the existence of inconsistent extensions for a theory. The scenario concerns the following dilemma: we assume that, usually, Quakers are pacifists and Republicans are not, but which of the two properties holds of Nixon who is both a Quaker and a Republican?

Formally, we start from the facts that $quaker(Nixon)$ and $republican(Nixon)$, together with the following default schemes

$$\left\{ \frac{quaker(X) : pacifist(X)}{pacifist(X)}, \frac{republican(X) : \neg pacifist(X)}{\neg pacifist(X)} \right\}.$$

Using ground substitution, we obtain the Reiter's theory $\Delta_N = (W_N, D_N)$ with $W_N = \{quaker(Nixon), republican(Nixon)\}$ and $D_N = \{\delta_3, \delta_4\}$ with

$$\left\{ \delta_3 = \frac{quaker(Nixon) : pacifist(Nixon)}{pacifist(Nixon)}, \delta_4 = \frac{republican(Nixon) : \neg pacifist(Nixon)}{\neg pacifist(Nixon)} \right\}.$$

The theory Δ_N has two extensions:

$$E_1 = Th(W \cup \{pacifist(Nixon)\}) \text{ and}$$
$$E_2 = Th(W \cup \{\neg pacifist(Nixon)\}).$$

Multiple extensions mean that neither $cons(\delta_3)$ nor $cons(\delta_4)$ is valid on the definition of skeptical entailment and both $cons(\delta_3)$ and $cons(\delta_4)$ are valid on the definition of credulous entailment. Notice that, if a theory has no extensions, then any first-order formula follows according to the skeptical entailment and no formula follows according to the credulous entailment. As a limiting case, a theory that has an inconsistent set of facts W always has a closed and successful process corresponding to the sequence $\Pi[0]$. To see why, consider that the set $Out(\Pi[0])$ is empty. This means that the set $In(\Pi[0]) = Th(W)$ defines the extension of that theory.

1.4 Abstract argumentation frameworks

The last formal ingredient in this thesis are abstract argumentation frameworks (henceforth AF). They offer answers to the problem of the acceptability of arguments based exclusively on the information about the attacks from one argument to another. An argumentation framework is a pair of a set of arguments, and a binary relation representing the attack-relationship (defeat) between arguments. More formally, $AF = (Arg, Att)$, where Arg is a set of arguments $A_1, A_2, \ldots$ and Att is a relation on $Arg \times Arg$ such that A_i attacks A_j if and only if $(A_i, A_j) \in Att$. These frameworks are abstract in at least two ways. First, it is immediately observable that the structure of arguments does not enter the formal workings of AFs. Secondly, and less obviously, the exact nature of attacks between arguments is not specified. As a result of their abstract nature, the mathematical structure of AFs is simply the structure of directed graphs, where nodes represent arguments and arrows represent attacks.

The study of arguments at this level of abstraction was initiated by Dung (1995). The generality of abstract argumentation enabled Dung to establish connections between argumentation frameworks on one side and logic programming, Reiter's default logic, Pollock's inductive logic, game theory (n-person games) on the other side, among others. From then on, there have been various attempts to develop frameworks where both the structure of arguments and the exact nature of attacks is specified, most notably in Prakken's (2010) ASPIC+ framework.

The generality of AFs turned out to be an asset, at least according to the amount of research originating from the simple idea of arguments modeled as graphs. The importance of Dung's theory of arguments for this thesis lies in the semantics of arguments acceptance in AFs. These semantics mediate between the language of justification logic and the operational methods for default theories. The concepts developed in (Dung, 1995, Section 2) are thus used as an additional level to the operational semantics that is inherited from default theories. We now present the basics of the AF semantics.

Starting from a framework $AF = (Arg, Att)$, the following can be said about collective acceptance of arguments from Arg. For the following definitions, it holds that a set of arguments S attacks an argument A_1 if $(A_2, A_1) \in Att$ for some A_2 from S.

Definition 7 (Conflict-free sets)**.** *A set of arguments S is said to be conflict-free*

if there are no arguments A_1 and A_2 in S such that $(A_1, A_2) \in Att$.

Definition 8 (Acceptability)**.** *An argument A_1 from Arg is acceptable with respect to a set of arguments S iff, for each argument A_2 from Arg it holds that, if $(A_2, A_1) \in Att$, then S attacks A_2.*

Using the definitions of conflict-free sets and acceptability, we can define a variety of standard semantics. Each of the semantics defined below represent a different way to answer the problem of determining those arguments that are considered to be the winning arguments.

Definition 9 (AF Extensions)**.** *For an abstract argumentation framework $AF = (Arg, Att)$, the following extensions are defined:*

Admissible Extension *A conflict-free set of arguments S is an admissible extension iff each argument in S is acceptable with respect to S.*

Preferred Extension *If S is a maximal admissible extension with respect to set inclusion, then S is a preferred extension.*

Complete Extension *An admissible extension S is a complete extension iff each argument that is acceptable with respect to S belongs to S.*

Grounded Extension *A complete extension S is the grounded extension if it is the least complete extension with respect to set inclusion.*

Stable Extension *A conflict-free set of formulas S is a stable extension if S attacks each argument that is not in S.*

We will use again the Nixon diamond example, this time to illustrate the semantics of AF's. Let A and B to be argument abstractions representing the claims "Nixon is not a pacifist because he is a Republican" and "Nixon is a pacifist because he is a Quaker", respectively. Additionally, we include an argument C that resolves the conflict of A and B such that C represents the claim "Nixon never used the right to exempt himself from the military draft, although the right is granted to all birthright Quakers". Thus the winning argument becomes the argument for the claim that Nixon is not a pacifist.

We can now define an abstract argumentation framework for the Nixon diamond: $AF^N = (Arg, Att)$, where $Arg = \{A, B, C\}$ and $Att = \{(A, B), (B, A), (C, B)\}$. The structure of attacks from AF^N can be conveniently represented by way of a directed graph. In Figure 1.1, we show the graph that corresponds to the framework AF^N.

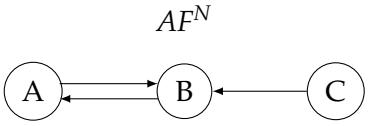

Figure 1.1: AF example

The nodes represent the arguments from *Arg* and the edges represent the direction of attacks obtained form *Att*. The graph shows that the argument C resolves the dilemma of Nixon diamond by deciding that A is the winning argument. The arguments C and A are contained in the preferred extension of AF^N, but also in its grounded extension. In the context of AF semantics, one can think of preferred and grounded extension as representing credulous and skeptical approach, respectively. In fact, for the framework AF^N, we find the coincidence of preferred, complete, grounded and stable extensions.

If all the types of semantics are uniformly defined, as in the case of AF^N, it is easy to determine which arguments need to be accepted. This is one of the motivations behind specifying conditions under which we have a unique answer to the problem of selecting a group of winning arguments. Dung (1995, p. 331) specifies a subclass of *well-founded* AFs for which we can establish a unique answer to this problem. An argumentation framework is well-founded iff there are no infinite sequences of arguments $A_1, A_2, \ldots, A_n, \ldots$ such that for each i, A_{i+1} attacks A_i. In Chapter 3, we specify the well-foundedness criteria for justification logic default theories.

Chapter 2

Default justification logic

> *As such, every great degree of caution in inferring, every skeptical disposition, is a great danger to life. No living being would be preserved had not the opposite disposition — to affirm rather than suspend judgement, to err and make things up rather than wait, to agree rather than deny, to pass judgement rather than be just — been bred to become extraordinarily strong.*
>
> —Nietzsche (1882/2001, p. 112,§ 111)

2.1 Introduction

In this chapter, we introduce default justification logic. We start from a variant of justification logic, namely **JT**, that models non-defeasible reasons. We use **JT** as the basic logic for default theories with default rules containing justification assertions. Then we introduce an operational semantics for justification logic default theories. The combination of default theories and justification logic enables us to interpret justification assertions as defeasible arguments. Finally, we define conditions of argument acceptance of justification assertions that we then use to define all the standard notions of extensions from abstract argumentation systems.

2.2 Justification logic and formal theories of defeasible arguments

Default reasoning is a key concept in the development of computational models of argument. Default reasons became a topic of interest for AI researchers largely due to Pollock's (1987) work, which brought closer together the ideas of non-monotonic reasoning from AI and defeasible reasoning from philosophy. To highlight the importance of defeasibility for the study of reasoning, we use a variant of Pollock's (1987) "red-looking table" vignette, previously discussed by Chisholm (1966): Suppose you are standing in a room where you see red objects in front of you. This can lead you to infer that a red-looking table in front of you is in fact red. However, the reason that you have for your conclusion is defeasible. For a typical defeat scenario, suppose you learn that the room you are standing in is illuminated with red light. This gives you a reason to doubt your initial reason to conclude that the table is red, though it would not give you a reason to believe that it is not red. However, if you were to learn, instead, that the original factory color of the table is white, then you would also have a reason to believe the denial of the claim that the table is red.

The example specifies two different ways in which reasons defeat other reasons: the former is known as *undercut* and the latter as *rebuttal*, in Pollock's (1987) terminology. If you obtain additional information about the light conditions, this will incur your suspension of the applicability of your initial reason to believe that the table is red. In contrast, if you learn that there is a separate reason to consider that the table is not red, this will not directly compromise your initial reason itself. The differences between undercutting and rebutting reasons are illustrated in Figure 2.1.

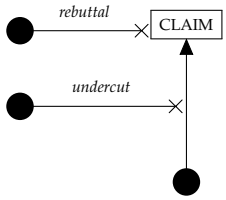

Figure 2.1: Two types of defeat: undercut and rebuttal

An argument relying on default reasons is itself regarded as defeasi-

ble. The formal study of defeasible arguments is already well-developed, most prominently in the frameworks for structured argumentation represented in the 2014 special issue of the *Argument and Computation* journal (vol. 5, issue 1): ABA (Toni, 2014), ASPIC+ (Modgil and Prakken, 2014), DeLP (García and Simari, 2014) and deductive argumentation (Besnard and Hunter, 2014).[1] These frameworks differ in the way they formalize argument structures and their defeasibility. Importantly, although all these frameworks use logic as a part of their formalization, none of them is a *logic of defeasible arguments*. The current chapter introduces a logic of defeasible arguments using the language of justification logic introduced by Artemov (2001). Among the many advantages of formalizing arguments in a logical system, for now we will point out only a couple of the more obvious ones. First, our logic of arguments is a full-fledged normative system with definition(s) of logical consequence that satisfies structured argumentation postulates by relying only on the definitions of logical consequence. We will show this in Section 3.3. Secondly, our logic is not a framework for specifying other systems and it does not use any meta-level rules from an unspecified system. Instead, we formalize arguments using only object-level formulas and inference rules. From a computational perspective, such a system is desirable as a way to manipulate arguments at a purely symbolic level.

The idea of finding a logical system with arguments as object-level formulas has already influenced the formal argumentation community. One especially interesting contribution in this direction is the logic of argumentation (LA) by Krause, Ambler, Elvang-Gøransson, and Fox (1995). These authors present a system in which inference rules manipulate labelled formulas interpreted as pairs of arguments and formulas[2]

$$arg : formula.$$

Our logic advances the search for the logic of arguments and builds on the take-away message from (Krause et al., 1995, p. 129) that we should take arguments "to be first-class objects themselves". By refining the way

[1] The acronyms ABA, ASPIC and DeLP refer to "Assumption-Based Argumentation", "Argumentation Service Platform with Integrated Components" and "Defeasible Logic Programming", respectively.

[2] The system has been used to develop applications that support medical diagnosis (Elvang-Gøransson et al., 1993, Fox et al., 2001). In LA, labels *arg* are interpreted as terms in the typed λ-calculus (Barendregt et al., 2013). Thanks to Artemov (2001, p. 7), we know that justification logic advances typed combinatory logic and typed λ-calculus.

in which we handle defeat among arguments, we make it possible to determine argument acceptance at a purely symbolic level and without using any measures of acceptability extraneous to the logic itself. This is one of the desiderata that the LA authors left open (Krause et al., 1995, Section 6).

In order to formalize arguments, we embrace the strategy of using a formal language with labelled formulas. In justification logic, such labelled formulas represent pairs of reasons and claims. They are written as the so-called "*justification assertions*" $t : F$ that read as "t is a reason that justifies formula F". The first justification logic was developed as a logic of proofs in arithmetic (logic of proofs, LP) by Artemov (2001).[3] On the original reading of pairs $t : F$, the term t encodes some derivation of the statement F in Peano arithmetic. Thus, the original logic of proofs does in fact give one particular formalization of arguments, namely a formalization of non-defeasible arguments. Accordingly, subsequent epistemic interpretations of justification logics provided a formal framework to deal with justifications and reasons, albeit non-defeasible ones. Even so, the underlying language of justification logic offers a powerful formal tool to model reasons as objects with operations. In this chapter, the language of justifications is used to study defeasible reasons.

The language of justifications is expressive enough to combine desirable features of the four mentioned structured argumentation frameworks in a single system. In Section 2.7, we will present how to use this logical language to provide justification assertions with argumentation semantics. Here are some outcomes that a reader can expect from our novel default justification logic:

- We show that default justification logic fulfills Pollock's project of defining a single formal system with strict and defeasible rules reified through deductive and default reasons. The four mentioned approaches dealing with structured argumentation are useful generalizations on how to understand arguments, but the problem we address here is how to unify their meta-analysis into a logical theory of undercut and rebuttal.

- Our system abstracts from the content of arguments, but, unlike ASPIC+ or ABA, represents arguments in the object language with

[3]The idea of explicit proof terms as a way to find the semantics for the provability calculus **S4** dates to 1938 and Gödel's lecture published in 1995.

default reasons. Compared to the level of abstraction in our logic, frameworks like ASPIC+ and ABA could be justly considered as meta-approaches to argumentation.

- Although ABA, ASPIC+ and deductive argumentation can generate Dung's frameworks, they cannot be said to provide a logical realization of Dung's frameworks because they do not define a specific logical system. In default justification logic, Dung's attack graphs, whose nodes can be interpreted as existential statements of the type "There is some argument", are realized with an explicit logical formula $t : F$ ascribed to each node of an attack graph. Determining acceptability of arguments through a normative system with logical consequence promises improvements in the area of computational argumentation.

- The logic we present here is capable of capturing all the components of Toulmin's six-fold argumentation scheme, with the exception of what he calls "qualifiers". The presence of elements like "warrant" and "backing" leads to a multi-layered understanding of an argument.[4] None of the mentioned structured argumentation frameworks gives a formalization of the added components of arguments such as warrants and backings. In contrast, our logic represents three layers of arguments which are codified in reason terms t justifying formulas F that are not necessarily explicitly represented at every stage of manipulating the formula $t : F$ in the semantics. [5]

- Justification logic enables us to integrate default logic and argumentation theory. Our logic remedies an important limitation of constructing arguments as Reiter's defaults (Verheij, 2009, p. 227):

[4]Toulmin's book *The Uses of Argument* (Toulmin, 1958/2003) is an acclaimed anti-formalistic argumentation monograph that separates logical methods and argumentation theory (Verheij, 2009, p. 219). Toulmin (1958/2003, p. vii) himself stated that the aim of his book was "to criticize the assumption, made by most Anglo-American academic philosophers, that any significant argument can be put in formal terms". One of the aims of this thesis is to reunite logical methods and argumentation theory.

[5]With the help of these distinctions, we are able to verify apparently conflicting claims about the nature of defeat in the literature. For example, ASPIC+ correctly models undercut by referring to the exclusion of a rule that does not apply in a given context. However, at the "lower" level of the argument backing, undercut eliminates an assumption made in justifying that rule — which suggests that this type of attack might be reduced to an assumption attack, as claimed in e.g. ABA. Such meta-disagreements on the nature of defeasibility can be reconciled in a fine-grained account of arguments.

> Reiter's defaults are givens and it is not possible to provide reasons for why they hold. Introducing justification logic as the basic language of default rules supplies them with a formal version of Toulmin's warrants and provides a way to further reason about the acceptability of rules. In this way, default logic with warrants is able to subsume formal argumentation semantics.[6]

The rest of this chapter is structured as follows. The next section introduces the basic justification logic system for reasoning with certain information. Then we use this formal system to introduce default justifications based on default rules with justification formulas. The "red table" example will be used as a running example that illustrates the use of such default rules. A preliminary survey of this system was carried out in (Pandžić, 2018). The system enables us to interpret formulas of the type $t : F$ as structured arguments with mutual attacks and to define the extension notions of Dung's argumentation frameworks in justification logic.

2.3 The logic of non-defeasible reasons JT

In Chapter 1, we saw how justification logic introduced the notions of justification and reason into epistemic logic. However, standard justification logics do not formally study ways of *defeat* among reasons and they takes admissibility of reasons as a primitive notion in the definition of semantics. Given the pervasiveness of commonsense reasoning, we know that only a restricted group of epistemic reasons may be treated as completely immune to undercutting and rebutting defeaters: mathematical proofs.[7] But mathematical reasons form only a small part of possible reasons to accept a statement and, being a highly-idealized group of reasons, they have rarely been referred to as reasons. Fitting's possible worlds semantics for justification logics was meant to model not only mathematical and logical truths, but also facts of the world or "inputs from outside the structure" (Fitting, 2009, p. 111). Yet the original intent of the first justification logic LP to deal with mathematical proofs, together

[6] Relations between Reiter's default logic (Reiter, 1980) and argumentation are explored in Dung's seminal paper (Dung, 1995), but the idea of modelling arguments in default logic has been initiated earlier in the AI literature e.g. by Prakken (1993).

[7] It is, however, possible to defeat a premise in a mathematical proof. This means that mathematical proofs are not immune to undermining defeaters.

with the fact that mathematics is cumulative, is reflected in its epistemic generalizations. Accordingly, reasons that justify facts of the world were left encapsulated within a framework for non-defeasible mathematical proofs.[8]

Non-mathematical reasons and justifications are commonly held to depend on each other in acquiring their status of "good" reasons and justifications. Still, the questions related to non-ideal reasons have only recently been raised in the justification logic literature.[9] In the present chapter we develop a non-monotonic justification logic with justification terms such that (1) their defeasibility can be tracked from the term structure and (2) other justifications can defeat them by means of an undercut or a rebuttal. Our logic combines techniques from default logic, justification logic and formal argumentation to represent conflicts of reasons produced in less-than-ideal ways.

The strategy we take in defining default justification logic is to start from a logic of non-defeasible reasons and then to extend this logic with inference rules that produce defeasible reasons. In this respect, our strategy is analogous to the standard default logic approach (Antoniou, 1997, Reiter, 1980) where agents reason from a background theory containing certain or non-defeasible information. The next section presents a variant of justification logic with non-defeasible reasons. In the context of justification logics, this means that the logic for non-defeasible reasoning includes the *Sum* and *Factivity* axioms, which were tagged as non-defeasibility axioms in Chapter 1. Each of the two axiom schemes requires that the existing reasons cannot be questioned by any incoming information.

While we do want our underlying logic to represent "non-defeasible" and "truth-inducing" reasons, we do not need additional constraints on the system to introduce the system of default reasons. For the sake of

[8]See (Bench-Capon and Dunne, 2007, p. 620) for a discussion on the difference between mathematical proofs and persuasive arguments.

[9]The first proposed formalism that includes the idea of evidence elimination specific to a multi-agent setting is by Renne (2012). Baltag, Renne, and Smets (2012, 2014) bring together ideas from belief revision and dynamic epistemic logic and offer an account of good and conclusive evidence. Several approaches (Milnikel (2014), Kokkinis et al. (2015, 2016), Ognjanović et al. (2017)) start from the idea of merging probabilistic degrees of belief with justification logic, while Fan and Liau (2015) and Su et al. (2017) develop a possibilistic justification logic. Fitting (2017) introduces a paraconsistent formal system with justification assertions where contradictions can be interpreted as conflicting evidence.

formal clarity, we leave out standard axioms and operations that ensure positive or negative introspection, although these can easily be added. Accordingly, an adequate logical account of factive justifications is the logic **JT**, a justification logic with the axiom schemes that are explicit analogues of the axiom schemes for the modal logic **T**.[10]

Intuitively, a reader can think of the **JT** logic as modelling an idealized arguer whose arguments fully exhaust all the possible information regarding claims and who, therefore, gives indisputable reasons for those claims. Note that there are also weaker variants of justification logic that do not assume factivity of reasons. These systems are not adequate for our purposes since we want to build defeasible arguments from a base of fact-inducing reasons — just as in standard default logic where reasoning starts from non-defeasible information or facts (Antoniou, 1997, p. 19). After we define the underlying logic that represents non-defeasible argumentation, we develop our novel non-monotonic approach to reasons and provide this logic with the semantics for defeasible argumentation.

Syntax

In Chapter 1, we defined the syntax for the justification logic **LP**. We use the formal ingredients given in Chapter 1 to define the logic **JT** in this chapter. Recall that, given that "t" is a justification term and that "F" is a formula, "$t : F$" is a justification assertion, where t is informally interpreted as a reason or justification for F. For a countable set of variables $x_1, \ldots, x_n, \ldots$ and a countable set of proof constants $c_1, \ldots, c_n, \ldots$, we define the set Tm that consists of exactly all justification terms, constructed from variables and proof constants by means of operations $\cdot$ and $+$. The grammar of justification terms for the logic **JT** is restricted to the application and sum operations, as defined by the following Backus-Naur form:

$$t ::= x \mid c \mid (t \cdot t) \mid (t + t)$$

where x is a variable and c is a proof constant. Proof constants are atomic within the system. For a justification term t, a set of subterms $Sub(t)$ is defined by induction on the construction of t. Formulas of **JT** are defined

[10] Justification logic **JT** was first introduced by Brezhnev (2001). Justification logics with axiom schemes equivalent to the logic we define in this section are also defined and investigated by Kuznets (2000) and Fitting (2008).

by the following Backus-Naur form:

$$F ::= \top \mid P \mid (F \to F) \mid (F \vee F) \mid (F \wedge F) \mid \neg F \mid t : F$$

where $P \in \mathcal{P}$ and $\mathcal{P}$ is a countable set of atomic propositional formulas and $t \in Tm$. The set Fm consists of exactly all formulas.

Axioms and rules of JT

We can now define the logic of non-defeasible reasons **JT**. The logic **JT** is the weakest logic with "truth inducing" justifications containing axiom schemes for the two basic operations $\cdot$ and $+$. We will use t and u to denote some elements from Tm and F and G to denote some elements from Fm. These are the axioms and rules of **JT**:

A0 *All the instances of propositional logic tautologies from Fm*

A1 $t : (F \to G) \to (u : F \to (t \cdot u) : G)$ (Application)

A2 $t : F \to (t + u) : F$; $\qquad u : F \to (t + u) : F$ (Sum)

A3 $t : F \to F$ (Factivity)

R0 *From F and* $F \to G$ *infer G* (Modus ponens)

R1* *If F is an axiom instance of A0-A3 and* $c_n, c_{n-1}, \ldots, c_1$ *are proof constants such that* $c_n : c_{n-1} : \cdots : c_1 : F \in \mathcal{CS}$, *then infer* $c_n : c_{n-1} : \cdots : c_1 : F$ (Iterated axiom necessitation)

The *Constant Specification* ($\mathcal{CS}$) set used in rule R1* is defined as follows:

Definition 2.1 (Constant Specification)**.** *The Constant Specification set is required to be Axiomatically appropriate and Injective as defined on p. 17.*

$$\mathcal{CS} = \{c_n : c_{n-1} : \cdots : c_1 : F \mid F \text{ is an axiom instance of A0-A3, } c_n, c_{n-1}, \ldots, c_1 \text{ are proof constants and } n \in \mathbb{N}\}$$

We consider a limited class of $\mathcal{CS}$-sets where each axiom instance has its own proof constant.[11] The logic **JT** relativized to an axiomatically appropriate and injective $\mathcal{CS}$ in R1* will be referred to by its full name $\mathbf{JT_{CS}}$. We say that the formula F is $\mathbf{JT_{CS}}$-provable ($\mathbf{JT_{CS}} \vdash F$) if F can be derived using the axioms A0-A3 and rules R0 and R1*.

[11] We require that $\mathcal{CS}$ is axiomatically appropriate to ensure that standard properties such as *Internalization* (Artemov, 2001) hold and we take each constant to justify at most one axiom instance, so $\mathcal{CS}$ is also required to be injective. See Section 1.2.

Semantics

The semantics for $\mathbf{JT_{CS}}$ is an adapted version of the semantics for the logic of proofs (**LP**) given by Mkrtychev (1997).[12] Intuitively, the semantics extends that of propositional logic with a function that ascribes reason terms to formulas in such a way that it respects the sum and application axioms and some axiomatically appropriate and injective constant specification $\mathcal{CS}$.

Definition 2.2 ($\mathbf{JT_{CS}}$ model)**.** *We define a function* reason assignment *based on $\mathcal{CS}$, $*(\cdot) : Tm \to 2^{Fm}$, a function mapping each term to a set of formulas from Fm. We assume that it satisfies the following conditions:*

1. *If $F \to G \in *(t)$ and $F \in *(u)$, then $G \in *(t \cdot u)$*
2. *$*(t) \cup *(u) \subseteq *(t + u)$*
3. *If $c : F \in \mathcal{CS}$, then $F \in *(c)$.*

A truth assignment $v : \mathcal{P} \to \{True, False\}$ is a function assigning truth values to propositional formulas in $\mathcal{P}$. We define the interpretation $\mathcal{I}$ as a pair $(v, *)$. For an interpretation $\mathcal{I}$, $\models$ is a truth relation on the set of formulas of $\mathbf{JT_{CS}}$.

For any formula $F \in Fm$, $\mathcal{I} \models F$ iff

- For any $P \in \mathcal{P}$, $\mathcal{I} \models P$ iff $v(P) = True$
- $\mathcal{I} \models \neg F$ iff $\mathcal{I} \not\models F$
- $\mathcal{I} \models F \to G$ iff $\mathcal{I} \not\models F$ or $\mathcal{I} \models G$
- $\mathcal{I} \models F \vee G$ iff $\mathcal{I} \models F$ or $\mathcal{I} \models G$
- $\mathcal{I} \models F \wedge G$ iff $\mathcal{I} \models F$ and $\mathcal{I} \models G$
- $\mathcal{I} \models t : F$ iff $F \in *(t)$.

[12] The condition for justifications of the type '!t' are not needed in the $\mathbf{JT_{CS}}$ semantics. Note that Mkrtychev's model does not make use of different states and accessibility relations among them. This type of model can be thought of as a single world justification model. Since the notion of defeasibility introduced in the next section turns on the incompleteness of available reasons, our system eliminates worries about the trivialization of justification assertions that otherwise arise from considering justifications as modalities in a single-world model.

Moreover, an interpretation $\mathcal{I}$ is *reflexive* iff the truth relation for $\mathcal{I}$ fulfills the following condition:

- For any term t and any formula F, if $F \in *(t)$, then $\mathcal{I} \models F$.

In the absence of the reflexivity condition, it is possible that $F \in *(t)$, hence $\mathcal{I} \models t : F$, but also that $\mathcal{I} \models \neg F$. While reasons in reflexive models can be taken as *conclusive* or factive, without the reflexivity condition reasons are interpreted as being only *admissible*. In possible worlds semantics, the admissibility condition $F \in *(t)$ for the truth of $t : F$ is supplemented with the condition that F holds in all accessible alternatives (Fitting, 2005b, p. 4). The consequence relation of the logic of factive reasons $\mathbf{JT_{CS}}$ is defined in terms of reflexive interpretations:

Definition 2.3 ($\mathbf{JT_{CS}}$ consequence relation)**.** $\Sigma \models F$ *iff for all reflexive interpretations* $\mathcal{I}$*, if* $\mathcal{I} \models B$ *for all* $B \in \Sigma$*, then* $\mathcal{I} \models F$*.*

In the next section, we will use deductively closed set of $\mathbf{JT_{CS}}$ formulas:

Definition 2.4 ($\mathbf{JT_{CS}}$ closure)**.** JT_{CS} *closure is given by* $Th^{JT_{CS}}(\Gamma) = \{F | \Gamma \models F\}$*, for a set of formulas* $\Gamma \subseteq Fm$ *and the* JT_{CS} *consequence relation* $\models$ *defined above.*

For any closure $Th^{JT_{CS}}(\Gamma)$, it follows that $\mathcal{CS} \subseteq Th^{JT_{CS}}(\Gamma)$.

We can prove that the compactness theorem holds for the $\mathbf{JT_{CS}}$ semantics.[13] Compactness turns out to be a useful result in defining the operational semantics of default reason terms. We first say that a set of formulas Γ is $\mathbf{JT_{CS}}$*-satisfiable* iff there is a reflexive interpretation $\mathcal{I}$ that meets $\mathcal{CS}$ (via the third condition of Def. 2.2) for which all the members of Γ are true. A set Γ is $\mathbf{JT_{CS}}$-finitely satisfiable if every finite subset Γ' of Γ is $\mathbf{JT_{CS}}$-satisfiable.

Theorem 2.5 (Compactness)**.** *A set of formulas is* $\mathbf{JT_{CS}}$*-satisfiable iff it is* $\mathbf{JT_{CS}}$*-finitely satisfiable.*

Proof. See the Appendix A. □

[13] A compactness proof for **LP** satisfiability in possible world semantics is given by Fitting (2005b). A similar proof is given for $\mathbf{JT_{CS}}$ in the Appendix A to provide a self-contained introduction to $\mathbf{JT_{CS}}$ in this thesis.

2.4 Justification logic default theories

Throughout the rest of this chapter, we develop a system based on $\mathbf{JT_{CS}}$, in which an agent forms default justifications reasoning from incomplete information. Justification logic $\mathbf{JT_{CS}}$ is capable of representing the construction of a new piece of evidence out of existing ones by application ('$\cdot$') or sum ('$+$') operation. However, to extend an incomplete $\mathbf{JT_{CS}}$ theory, we need to import reasons that are defeasible. We come up with both a way in which such reasons are imported and a way in which they might get defeated. These possibilities are opened up by introducing concepts familiar from defeasible reasoning literature into justification logic.

We start from the above-defined language of the logic $\mathbf{JT_{CS}}$ and develop a new variant of justification logic $\mathbf{JT_{CS}}$ that enables us to formalize the import of reasons outside the structure as well as to formalize *defeaters* or reasons that question the plausibility of other reasons.

Our logical framework of defeasible reasons represents both factive reasons produced via the axioms and rules of $\mathbf{JT_{CS}}$ and plausible reasons based on default assumptions that "usually" or "typically" hold for a restricted context. We follow the standard way (Reiter, 1980) of formalizing default reasoning through default theories to extend the logic of factive reasons with defeasible reasons. Building on the syntax of $\mathbf{JT_{CS}}$, we introduce the definition of the *default theory*:

Definition 2.6 (Default Theory)**.** *A default theory T is defined as a pair (W, D), where the set W is a finite set of $\mathbf{JT_{CS}}$ formulas and D is a countable set of default rules. Each default rule is of the following form:*

$$\delta = \frac{t : F :: (u \cdot t) : G}{(u \cdot t) : G}.$$

The informal reading of the default δ is: "If t is a reason justifying F, and it is consistent to assume that $(u \cdot t)$ is a reason justifying G, then $(u \cdot t)$ is a defeasible reason justifying G". The formula $t : F$ is called the prerequisite and $(u \cdot t) : G$ is both the consistency requirement[14] and the consequent of the default rule δ. We refer to each of the respective formulas as $pre(\delta)$, $req(\delta)$ and $cons(\delta)$. For the set of all consequents from the entire set of defaults D, we use

[14]In order to avoid any misunderstanding, we avoid the name *justification* for the formula $req(\delta)$ since justification logic terms are commonly known as justifications.

$cons(D) = \{cons(\delta) \mid \delta \in D\}$. The default rule δ introduces a unique reason term u, which means that, for a default theory T, the following holds:

1. *For any formula $v : H \in Th^{JT_{CS}}(W)$, $u \neq v$;*
2. *For any formula $H \in W$, $u : (F \to G)$ is not a subformula of H and*
3. *For any default rule $\delta' \in D$ such that $\delta' = \frac{t':F'::(u'\cdot t'):G'}{(u'\cdot t'):G'}$, if $u = u'$, then $F = F'$ and $G = G'$.*

Note that the term u does not need to be fresh in the sense that it cannot appear in two different defaults' consequents.[15] Default reasons may refer to other default reasons and this possibility is crucial to represent interactions among defaults. The unique reason term u witnesses the defeasibility of the *prima facie* reason $(u \cdot t)$ for G. Whether a reason actually becomes defeated or not depends on other default-reason formulas from $cons(D)$. Other defaults might question both the plausibility of the reasoning that u codifies and the plausibility of the proposition G. Section 2.5 gives an example of a concrete $\mathbf{JT_{CS}}$ derivation that instantiates unique reason terms.

A formal way of looking at a default reason of this kind is that $(u \cdot t)$ codifies the default step we apply on the basis of the known reason t. A distinctive feature of such rules is generating justification terms as if it were the case that $cons(\delta)$ was inferred by using an instance of the application axiom: $u : (F \to G) \to (t : F \to (u \cdot t) : G)$. The difference is that an agent cannot ascertain that an available reason justifies applying the conditional $F \to G$ without restrictions. Still, sometimes a conclusion must be drawn without being able to remove all of the uncertainty as to whether the relevant conditional actually applies or not. In such cases, an agent turns to a plausible assumption of a justified "defeasible" conditional $F \to G$ that holds only in the absence of any information to the contrary. While the internal structure of the default reason[16] $(u \cdot t)$ indicates that it is formed on the basis of the formula $u : (F \to G)$, the

[15]Compare Artemov's (2018, p. 30) introduces "single-conclusion" (or "pointed") justifications that enable handling "justifications as objects rather than as justification assertions".

[16]In this chapter and in Chapter 3, we use "defeasible reasons" and "default reasons" interchangeably. In Chapter 4, an additional type of defeaters is introduced, namely undermining defeaters. Undermining defeaters do not target arguments $t : F$ that are introduced through default rules. Therefore, Chapter 4 extends the class of defeasible reasons beyond default reasons.

defeasibility of $(u \cdot t)$ lies in the fact that the formula $u : (F \to G)$ is not a part of the same evidence base as $(u \cdot t) : G$.

One can think of our use of the operation "$\cdot$" in default rules as the same operation that is used in the axiom A1, only being applied on an incomplete $\mathbf{JT_{CS}}$ theory. Similarly, we can follow Reiter (1980, p. 82) and Antoniou (1997, p. 21) in thinking of a standard default rule such as $\frac{A:B}{B}$ as merely saying that an implication $A \wedge \neg C \wedge \neg D \cdots \to B$ holds, provided that we can establish that a number of exceptions $C, D, \ldots$ does not hold. However, if the rule application context is defined sufficiently narrowly, the rule is classically represented as an implication $A \to B$. Generalizing on such interpretation of defeasibility, our defaults with justification assertions can be represented as instantiations of the axiom A1 applied in a sufficiently narrow application context.

Analogous to standard default theories, we take the set of facts W to be underspecified with respect to a number of facts that would otherwise be specified for a complete $\mathbf{JT_{CS}}$ interpretation. Besides simple facts, our underlying logic contains justification assertions. To deal with justification assertions, a complete $\mathbf{JT_{CS}}$ interpretation would also further specify whether a reason is acceptable as a justification for some formula. Therefore, except the usual incomplete specification of known propositions, default justification theories are also incomplete with respect to the actual specification of the reason assignment function. For our default theory, this means that, except the valuation v, default rules need to approximate an actual reason-assignment function $*(\cdot)$.

Let us again consider the red-looking-table example from the Introduction to see how *prima facie* reasons and their defeaters are imported through default rules.

Example 2.7. *Let R be the proposition "the table is red-looking" and let T be the proposition "the table is red". Take t_a and u_a to be some specific individual justifications. The reasoning whereby one accepts the default reason $(u_a \cdot t_a)$ might be described by the following default rule:*

$$\delta_a = \frac{t_a : R :: (u_a \cdot t_a) : T}{(u_a \cdot t_a) : T}.$$

We can informally read the default as follows: "If t_a is a reason justifying that a table is red looking and it is consistent for you to assume that this gives you a reason $(u_a \cdot t_a)$ justifying that the table is red, then you have a defeasible reason $(u_a \cdot t_a)$ justifying that the table is red". Suppose you then get to a belief

that "the room you are standing in is illuminated with red light", a proposition denoted by L. For some specific justifications t_b and u_b, the following rule gives you an undercutting reason for $(u_a \cdot t_a)$:

$$\delta_b = \frac{t_b : L :: (u_b \cdot t_b) : \neg[u_a : (R \to T)]}{(u_b \cdot t_b) : \neg[u_a : (R \to T)]},$$

where the rule is read as "If t_b is a reason justifying that the lighting is red and it is consistent for you to assume that this gives you a reason $(u_b \cdot t_b)$ denying that the reason u_a justifies that if the table is red-looking, then it is red, then you have a defeasible reason $(u_b \cdot t_b)$ denying that the reason u_a justifies that if the table is red-looking, then it is red". The formula $cons(\delta_b)$ denies your reason to conclude $cons(\delta_a)$, although note that it is not directly inconsistent with $cons(\delta_a)$. In Section 2.6, we define what undercutting defeaters are semantically.

Suppose that instead of learning about the light conditions in the room as in δ_b, you learn that the original factory color of the table is white. This would also prompt a rebutting defeater - a separate reason to believe the contradicting proposition $\neg T$. Let W denote the proposition "the table is originally white" and let t_c and u_c be some specific justifications. We have the following rule:

$$\delta_c = \frac{t_c : W :: (u_c \cdot t_c) : \neg T}{(u_c \cdot t_c) : \neg T}.$$

The rule reads as "If t_c is a reason justifying that the table is originally white and it is consistent for you to assume that this gives you a reason $(u_c \cdot t_c)$ justifying that the table is not red, then you have a defeasible reason $(u_c \cdot t_c)$ justifying that the table is not red". Note that the formula $cons(\delta_c)$ does not directly mention any of the subterms of $(u_a \cdot t_a)$. The defeat among the reasons $(u_a \cdot t_a)$ and $(u_c \cdot t_c)$ comes from the fact that they cannot together consistently extend an incomplete $\mathbf{JT_{CS}}$ *theory.*

The entire example can be described by the following default theory $T_0 = (W_0, D_0)$, where $W_0 = \{t_a : R, t_b : L, t_c : W\}$ and $D_0 = \{\delta_a, \delta_b, \delta_c\}$.

Each defeater above is itself defeasible and considered to be a *prima facie* reason. The way in which *prima facie* reasons interact is further specified through their role in the operational and argumentative semantics for default theories. By the end of this chapter, we explain the workings of the operational semantics and different ways to determine the sets of acceptable reasons given a definition of a default theory.

2.5 Operational semantics of default justifications

The logic of default justifications we develop here relies on the idea of operational semantics for standard default logics presented by Antoniou (1997).[17] Here is an informal description of the role of operational semantics steps in determining acceptable reasons. First, in the operational part of the semantics, default reasons are taken into consideration at face value. After the default reasons have been taken together, we check dependencies among them in order to find out what are the non-defeated reasons. Finally, a rational agent includes in its knowledge base only acceptable pieces of information that are based on those reasons that are ultimately non-defeated. An important part of the latter step is an acceptance semantics analogous to the argument acceptance semantics of formal argumentation frameworks.

The basis of operational semantics for a default theory $T = (W, D)$ is the procedure of collecting new, reason-based information from the available defaults. A *sequence* of default rules $\Pi = (\delta_0, \delta_1, \ldots)$ is a possible order in which a list of default rules without multiple occurrences from D is applied (Π is possibly empty). Applicability of defaults is determined in the following way:

Definition 2.8 (Applicability of Default Rules)**.** *For a* $\mathbf{JT_{CS}}$*-closed set of formulas* Γ *we say that a default rule* $\delta = \frac{t:F::(u\cdot t):G}{(u\cdot t):G}$ *is applicable to* Γ *iff*

- $t : F \in \Gamma$ *and*
- $\neg(u \cdot t) : G \notin \Gamma$.[18]

Reasons are brought together in the set of $\mathbf{JT_{CS}}$ formulas that represents the current *evidence base*:

Definition 2.9. $In(\Pi) = Th^{JT_{CS}}(W \cup \{cons(\delta) \mid \delta \text{ occurs in } \Pi\})$, *where* Π *is a sequence of defaults.*

[17] Following Antoniou (1997, p. 31), we think of operational definitions as those that "give a procedure that can be applied to examples". However, to determine extensions for justification logic default theories, Section 2.7 specifies further semantic conditions on sets of justification logic formulas, in addition to the procedural conditions discussed in this section.

[18] We follow the convention of omitting parentheses around the expression $(u \cdot t) : G$ and interpret the negation as binding the entire expression $(u \cdot t) : G$. The convention is also familiar from modal logics.

The set $In(\Pi)$ collects reason-based information that is yet to be determined as acceptable or unacceptable depending on the acceptability of reasons and counter-reasons for formulas.

We need to further specify sequences of defaults that are significant for a default theory T: default processes. For a sequence Π, the initial segment of the sequence is denoted as $\Pi[k]$, where k stands for the number of elements contained in that segment of the sequence and where k is a minimal number of defaults for the sequence Π. Any segment $\Pi[k]$ is also a sequence. Intuitively, the set of formulas $In(\Pi)$ represents an update of the incomplete evidence base W where the new information is not yet taken to be granted. Using the notions defined above, we can now get clear on what a default process is:

Definition 2.10 (Process)**.** *A sequence of default rules* Π *is a* process *of a default theory* $T = (W, D)$ *iff every* k *such that* $\delta_k \in \Pi$ *is* applicable *to the set* $In(\Pi[k])$*, where* $\Pi[k] = (\delta_0, \ldots \delta_{k-1})$*.*

The kind of process that we are focusing on here is called *closed* process and we say that a process Π is closed iff every $\delta \in D$ that is applicable to $In(\Pi)$ is already in Π. For default theories with a finite number of defaults, closure for any process Π is obviously guaranteed by the applicability conditions. However, if a set of defaults is infinite, then this is less-obvious.

Lemma 2.11 (Infinite Closed Process)**.** *For a theory* $T = (W, D)$ *and infinitely many* k*'s, an infinite process* Π *is closed iff for every default rule* δ_k applicable *to the set* $In(\Pi[k])$*,* $\delta_k \in \Pi$*.*

Proof. From the compactness of $\mathbf{JT_{CS}}$ semantics we have that if a set $In(\Pi[k]) \cup \{req(\delta)\}$ is satisfiable for all the finite k's, it is also satisfiable for infinitely many k's. Therefore the applicability conditions for a rule δ are equivalent to the finite case. □

To illustrate how the basic notions of the operational semantics work, Figure 2.2 shows the process tree for the default theory T_0 from our running Example 2.7. The figure shows that T_0 has four closed processes: $\Pi_1 = (\delta_a, \delta_b)$, $\Pi_2 = (\delta_b, \delta_a)$, $\Pi_3 = (\delta_b, \delta_c)$ and $\Pi_4 = (\delta_c, \delta_b)$. The *In*-sets $In(\Pi_1)$ and $In(\Pi_2)$ are equal and $\mathbf{JT_{CS}}$-inconsistent with $In(\Pi_3)$ and $In(\Pi_4)$, which are also equal. Whenever two sets $In(\Pi)$ and $In(\Pi')$ are not equal, they are $\mathbf{JT_{CS}}$-inconsistent. We can already see that $\mathbf{JT_{CS}}$-inconsistent *In*-sets capture the idea of rebuttal in our semantics, as

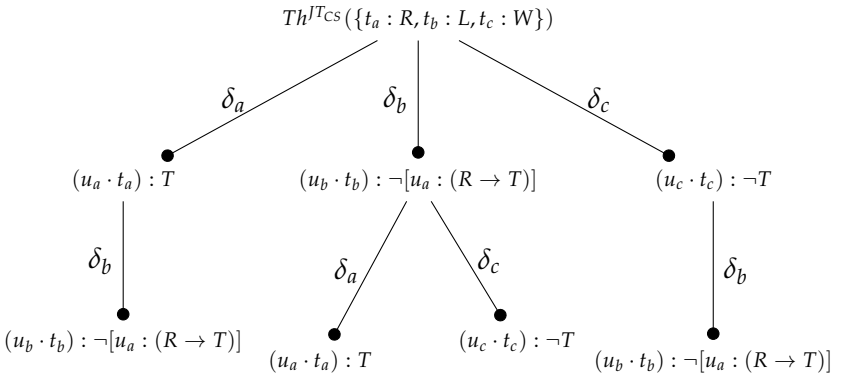

Figure 2.2: The process tree of T_0 from Example 2.7.

introduced informally in Example 2.7. For example, $\mathbf{JT_{CS}}$-inconsistent In-sets $In(\Pi_1)$ and $In(\Pi_4)$ reflect the opposition between the reasons $(u_a \cdot t_a)$ and $(u_c \cdot t_c)$. At the level of process trees, however, we are not yet able to explain the attack on $(u_a \cdot t_a)$ by the undercutting reason $(u_a \cdot t_a)$. To do so, we need to move further from the semantics of collecting new information.

We have already discussed the key components of our operational semantics that bear some similarity to standard default theories. Now we develop our new argument semantics that builds on the expressivity of the justification logic language. We show that the default variant of the application operation is essential to the way in which we represent arguments and their mutual attacks in justification logic.

2.6 Argumentative schemes and argumentative attacks in justification logic

In a complete specification of $\mathcal{I}$, acceptability of reasons for a formula is determined *ex officio* by assigning formulas to reasons through the function $*(\cdot)$. In contrast, in reasoning from an incomplete evidence base W, a closure $Th^{JT_{CS}}(W)$ is typically underspecified as to whether a reason t is acceptable for a formula F. In "guessing" what a true interpretation is, every default rule introduces a reason term whose structure codifies an application operation step from an unknown justified conditional. For

example, in rule δ above, we rely on the justified conditional $u : (F \to G)$. Even though this justified conditional is not a part of the rule δ itself, it is the underlying assumption on the basis of which we are able to extend an incomplete evidence base. The propositions of this kind are in one sense taken as rules allowing for default steps, but they are also specific justification logic formulas. They will be referred to as "warrants", because their twofold role in our system corresponds to Toulmin's concept of argument warrants.[19] Justification logic defaults give a formal meaning to Toulmin's philosophical idea that warrants are formulated as statements, even though they function as rules of inference within arguments. Each underlying formula of this kind can be made explicit by means of a function *warrant assignment*: $\#(\cdot) : D \to Fm$. The function maps each default rule to a specific justified conditional as follows:

$$\#(\delta) = u : (F \to G),$$

where $\delta \in D$ and

$$\delta = \frac{t : F :: (u \cdot t) : G}{(u \cdot t) : G},$$

for some reason term t, a unique reason term u and some formulas F and G.

A set of all such underlying warrants of default rules is called *Warrant Specification* ($\mathcal{WS}$) set.

Definition 2.12 (Warrant specification)**.** *For a default theory $T = (W, D)$, justified defeasible conditionals are given by the Warrant Specification set:*

$$\mathcal{WS}^T = \#(D) = \{\#(\delta) \mid \delta \in D\}.$$

We will use warrant specification sets that are relativized to default processes:

$$\mathcal{WS}^{\Pi} = \{\#(\delta) \mid \delta \in \Pi\}.$$

In reasoning from incomplete information, defeasible justification assertions from $\mathcal{WS}^T$ are the only available resource to approximate a reason

[19] Toulmin explains (1958/2003, p. 91) inference-licensing warrants as follows: "...taking these data as a starting point, the step to the original claim or conclusion is an appropriate and legitimate one. At this point, therefore, what are needed are general, hypothetical statements, which can act as bridges, and authorise the sort of step to which our particular argument commits us."

assignment function that actually holds. Moreover, the use of underlying assumptions from $\mathcal{WS}^T$ is responsible for the non-monotonic character of default reasons. Thus our default rules are in contrast with the standard application operation represented by the axiom A1. The extended meaning of the application operation via default rules will be referred to as **default application**. Importantly, default application extends the standard idea of "proof terms" in justification logic so as to include reason terms that codify inference steps from assumptions to warrant formulas as conclusions dependent on those assumptions. We briefly explain this idea after we specify how warrants and default application are decisive for the semantics of attacks between arguments.

The extension of the application operation to its defeasible variant opens new possibilities for a semantics of justifications. In particular, it enables reasoning that is not regimented by the standard axioms A1 and A2 of basic justification logic (Artemov, 2008, p. 482). For instance, if a set of $\mathbf{JT_{CS}}$ formulas contains both a *prima facie* reason t and its defeater u, then the set containing a conflict of justifications does not support concatenation of reasons by which $t : F \rightarrow (t + u) : F$ holds for any two terms t and u. In other words, the possibility of a conflict between reasons requires an adaptation that eliminates the monotonicity property of justifications assumed in the sum axioms (A2).

In explaining the basics of the operational semantics, we qualified the semantics of rebutting attacks as being straightforward. Rebuttal is already captured in the mechanism of multiple extensions known from standard default theories. What requires additional explanation is the semantics of undercutting defeaters. Notice that each formula $\#(\delta)$ has the format of a justified material conditional. This formula is not a part of a default inference δ itself, but the default application described by δ depends on a conjecture that the conditional holds and the justification assertion $cons(\delta)$ encodes this conjecture in the internal structure of the resulting reason term. This brings to attention the following possibility: an evidence base may at the same time contain justified formulas of the type $t : F$, $(u \cdot t) : G$ and $v : \neg[u : (F \rightarrow G)]$, without the evidence base being $\mathbf{JT_{CS}}$-inconsistent.

Although the application axiom A1 does not say that $t : F$ and $(u \cdot t) : G$ together entail the formula $u : (F \rightarrow G)$, there is, intuitively, something wrong with the reason $(u \cdot t)$ justifying the formula G, taken together with t justifying F and v justifying $\neg u : (F \rightarrow G)$. This new type of opposition among reasons explains why we need to refer to

warrant formulas. The co-occurrence of the formulas $t : F$, $(u \cdot t) : G$ and $v : \neg[u : (F \to G)]$ together is not significant in standard justification logic where reasoning is exclusively regulated by the standard axioms for idealized reasons, such as the axioms of the basic $\mathbf{JT_{CS}}$ logic. It only becomes significant with default application.[20] We will now use the presented "reverse engineering" of axiom A1 to model undercut.[21]

We have already discussed why the semantics of undercut cannot be reduced to the existence of multiple inconsistent extensions. Nevertheless, $\mathbf{JT_{CS}}$ inconsistency is important for undercutting attacks.[22] Notice that adding arbitrary warrants from $\mathcal{WS}^T$ to an evidence base $In(\Pi)$ could lead to an inconsistent set of $\mathbf{JT_{CS}}$ formulas. In Example 2.7, if we start from any evidence base of T_0 and add the warrant $u_a : (R \to T)$ of δ_a to it, the union becomes $\mathbf{JT_{CS}}$-inconsistent with both the warrant $u_b : (L \to \neg[u_a : (R \to T)])$ of δ_b and the warrant $u_c : (W \to \neg T)$ of δ_c. This means that the three warrants are jointly incompatible in the context of default reasoning defined by T_0. An agent needs to find out which warrants and, thereby, which reasons prevail in a conflicting set of warrants. This procedure relies on the following definition that captures the above discussed intuition behind undercut:

Definition 2.13 (Undercut)**.** *A reason u undercuts reason t being a reason for a formula F in a set of $\mathbf{JT_{CS}}$ formulas $\Gamma \subseteq In(\Pi[k])$ iff $\bigvee_{(v)\in Sub(t)} u : \neg[v : (G \to H)] \in Th^{JT_{CS}}(\Gamma)$ and there is a process Π' of T such that $v : (G \to H) \in \mathcal{WS}^{\Pi'}$.*

The definition says that the reason u undercuts the reason t, if there is a subterm v of t such that u denies the justified conditional $v : (G \to H)$ as a warrant that supports one of the default steps made in building the

[20] Notice that a ($\mathbf{JT_{CS}}$-closed) evidence base that contains the formulas $t : F$ and $(u \cdot t) : G$, also contains the formula $((c \cdot t) \cdot (u \cdot t)) : (F \to G)$, assuming that the constant c justifies the axiom $F \to (G \to (F \to G))$. This is so regardless of whether $u : (F \to G)$ is also in the evidence base or not.

[21] One way to model exclusionary reasons and undercutters in default logic is to use non-normal defaults. However, with the use of non-normal defaults, many desirable features of default logics are lost, and this holds already for semi-normal defaults (Antoniou, 1997, Chapter 6). Besides that, the use of justification logic warrants provides an elegant way to subsume argumentation semantics in default logic. For a more extensive discussion on the benefits of warrants over non-normal defaults see the next chapter based on the work from (Pandžić, 2019).

[22] Later, in Lemma 2.7, we characterize the relation between rebuttal and undercut formally.

argument $t : F$. Intuitively, if we think of each warrant of a default rule introduces a default argument, we can say that u attacks t on one its sub-arguments.

We will also specify the way in which sets of $\mathbf{JT_{CS}}$ formulas undercut some default reason. This definition will be used in defining different variants of default theory extensions. Sets of justification logic formulas are said to undercut reasons according to the following definition:

Definition 2.14. *A set of* $\mathbf{JT_{CS}}$ *formulas* $\Gamma \subseteq In(\Pi[k])$ *undercuts reason* t *being a reason for a formula* F *iff* $\bigvee_{(v)\in Sub(t)} \neg[v : (G \to H)] \in Th^{JT_{CS}}(\Gamma)$ *and there is a process* Π' *of* T *such that* $v : (G \to H) \in \mathcal{WS}^{\Pi'}$.

One can think of Γ as a set of reasons against which the reason t is tested as a reason that justifies the formula F. This is further elaborated in the semantics of acceptability of reasons. By introducing default reasons through default application and considering rebuttal and undercut among such reasons, it is possible to take an argumentation perspective to justification logic formulas. For example, Figure 2.3 provides an intuitive Toulminian interpretation of the default reasoning steps discussed in Example 2.7, where each step can be associated with a corresponding step in the Toulminian argument scheme.[23]

Note that the formula $(u_c \cdot t_c) : \neg T$ is captioned as a rebuttal of the formula $(u_a \cdot t_a) : T$, but $(u_a \cdot t_a) : T$ also rebuts $(u_c \cdot t_c) : \neg T$. Their rebuttal relation is symmetric because the two conclusions T and $\neg T$ of the default reasons $(u_a \cdot t_a)$ and $(u_c \cdot t_c)$ are contradictory, which means that applying either of the default rules δ_a and δ_c blocks the application of the other default rule.[24] Moreover, in Toulmin's scheme of argumentation, backing is understood as a certification or evidence for the use of a warrant to introduce some conclusion. In justification logic, backing naturally translates into a $\mathbf{JT_{CS}}$ derivation (with undischarged assumptions) of a

[23] A reader should take the following two provisos into account here. Firstly, Toulmin does not use the term "undercutter". Instead, Toulmin uses rebuttal as an ambiguous concept that, among other kinds of defeat, covers for circumstances in which the general authority of the warrant would have to be set aside (Verheij, 2009, p. 235). Secondly, our scheme does not include "qualifiers" (Toulmin, 1958/2003, p. 94) that indicate the strength of the step from grounds to claim.

[24] In Example 2.7, rebuttal results from the straightforward JT_{CS}-inconsistency of the two formulas $(u_c \cdot t_c) : \neg T$ and $(u_a \cdot t_a) : T$. However, it is generally possible that two default rules block each other's applicability for some, but not all default processes. This is a result of a more intricate structure of joint inconsistencies among sets of default reasons.

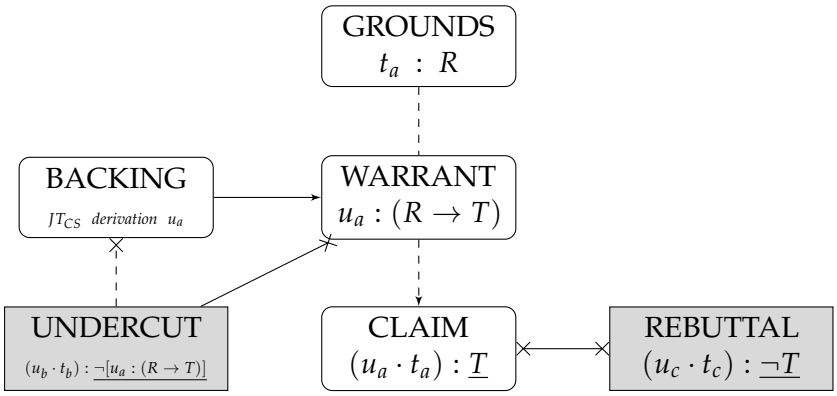

Figure 2.3: Toulminian layout of the arguments in Example 2.7

default conditional, and the steps of that derivation are codified in the reason term u_a that justifies the conditional $R \to T$. Consider a simple backing for δ_a from Example 2.7:

1	$x : (\neg T \to \neg R)$	(Assumption)
2	$(\neg T \to \neg R) \to (R \to T)$	(A0)
3	$c : [(\neg T \to \neg R) \to (R \to T)]$	(R1)
4	$(c \cdot x) : (R \to T)$	(1,3 A1)

By taking that $u_a = (c \cdot x)$, one can recover the underlying structure of reasoning for the warrant $u_a : (R \to T)$, which corresponds to the idea of backing. Informally, the backing $(c \cdot x)$ describes reasoning when one assumes that if the table was not actually red, then it would not look red. This is a simple backing example, but, in general, such reasoning structures can become more complex. For example, assumptions made in deriving a warrant formula may include literals that are not subformulas of the warrant itself, as Example 2.25 later illustrates. In general, representative cases of warrants cannot be derived from a knowledge base W, without using (undischarged) assumptions. This is also the case in our Example 2.7, where the warrant formula $u_a : (R \to T)$ is not contained in the knowledge base closure $Th^{JT_{CS}}(W)$. Clearly, proof terms are thus interpreted more broadly than in standard justification logics.

A reader may notice here that the self-referential mechanism in which the language of justification logic treats its own reasoning steps within the language gives a three-layered understanding of arguments. The first

layer is an argument seen as a pair of reason terms and formulas, e.g. the formula $(u_a \cdot t_a) : T$, resulting from the default $\delta_a = \frac{t_a:R::(u_a \cdot t_a):T}{(u_a \cdot t_a):T}$. In argumentative terms, this layer includes the formula $t_a : R$ that represents Toulminian grounds or data. Since the term $(u_a \cdot t_a)$ formally realizes the default application step of δ_a, the formula $(u_a \cdot t_a) : T$ will always be explicitly featured in the semantic treatment of the acceptability of the reasoning steps codified by the term $(u_a \cdot t_a)$. Argumentation semantics for such formulas will be presented in the next section. The second layer gives a wider understanding of the argument. It includes the rule δ_a together with its warrant formula $u_a : (R \to T)$. This layer explains the reasoning step from the grounds $t_a : R$ to the claim T. It provides an answer to Toulmin's question (1958/2003, p. 90) "How did you get there?", that is, how to justify that some claim follows from the available data or grounds. Finally, the third layer of the argument for T additionally includes the backing or the unfolded formal structure of the reasoning steps represented by u_a that are given in support of the use of the warrant $u_a : (R \to T)$. Analogously to Toulmin's argument scheme (1958/2003, p. 92), the warrant makes explicit the connection between the grounds and the claim, while the backing explains why the warrant counts as a justified one. Argument warrant can themselves become a part of the reasoning process, especially upon questioning their authority. This is illustrated by the default rule δ_b in the running example.

2.7 Argument acceptance in justification logic

By introducing default reasons in justification logic it becomes possible not only to use argumentation terminology in talking about formulas of the type $t : F$ but also to give standard abstract argumentation theory conditions of argument acceptance of such formulas. The idea of conflicting default reasons overlaps with abstract argumentation frameworks that treat conflicts between arguments. This section shows that all the formal conditions of argument acceptance as defined in Dung's framework (1995) can be defined for default justifications introduced here. In Section 3.2, this is used to prove that the logic of default justifications generalizes Dung's frameworks.

The semantics of reason acceptance starts from characterizing conflict-free sets of $\mathbf{JT_{CS}}$ formulas. Notice that by introducing reason terms through default application, conflicts are not only defined in terms of

$\mathbf{JT_{CS}}$ inconsistency, but also in terms of undercut from Definition 2.13. The following definition specifies conditions for conflict-free sets with respect to undercut:

Definition 2.15 (Conflict-free Sets)**.** *A set of* $\mathbf{JT_{CS}}$*-closed formulas* Γ *is conflict-free iff* Γ *does not undercut a formula* $t : F$ *such that* $t : F \in \Gamma$*.*

Note that, if a set of formulas $In(\Pi)$ for any process Π is conflict-free according to Def. 2.15, then it is also free from rebuttal for a consistent set of formulas W. To see why, first consider that rebuttal occurs between formulas that are contained in jointly $\mathbf{JT_{CS}}$-inconsistent evidence bases. Since we know that the conditions under which a default can be applied to an evidence base $In(\Pi)$ preserve consistency of each segment $In(\Pi[k])$ of $In(\Pi)$, we also know that $In(\Pi)$ is rebuttal-free. Consistency preservation of extended evidence bases is established in the following theorem:

Theorem 2.16. *For a theory* $T = (W, D)$ *and a process* Π *of* T*, if the set of formulas* W *is* $\mathbf{JT_{CS}}$*-consistent, then any conflict-free set of formulas* $In(\Pi)$ *is also* $\mathbf{JT_{CS}}$ *consistent.*

Proof. The property of $\mathbf{JT_{CS}}$ consistency for a set of formulas $In(\Pi)$ follows from the applicability conditions for any default rule $\delta \in \Pi$ of the form $\frac{t:F::(u\cdot t):G}{(u\cdot t):G}$ and the fact that W is $\mathbf{JT_{CS}}$ consistent. □

The theorem ensures that, for any non-empty process Π, a set of conflict-free formulas $In(\Pi)$ that an agent could eventually accept is free from any possible conflict.

As stated before, the set W contains certain information and this means that any information from W is always acceptable regardless of what has been collected later on. Therefore, any set of formulas Γ that extends the initial information contains W. To decide whether a consequent of a default δ is acceptable, an agent looks at those sets of reasons that can be defended against all the available counter-reasons. For any set of $\mathbf{JT_{CS}}$ formulas Γ, we define the notion of acceptability of a justified formula $t : F$:

Definition 2.17 (Acceptability)**.** *For a default theory* $T = (W, D)$*, a formula* $t : F \in cons(\Pi)$ *is acceptable w.r.t. a set of* $\mathbf{JT_{CS}}$ *formulas* $\Gamma \subset In(\Pi[k])$ *iff for each undercutting reason* u *for* t *being a reason for* F *such that* $u : G \in In(\Pi[k])$*,* Γ *undercuts* u *being a reason for* G*.*

In the definition, we use the process segment $\Pi[k]$, because we want to determine acceptability for each stage of building a process Π. Intuitively, an agent looks at finding a defensible set of arguments in the space of all possible arguments defined by all certain information taken together with the consequents of applicable defaults. Accordingly, for a default theory $T = (W, D)$, an agent considers *potential extension sets* of $\mathbf{JT_{CS}}$ formulas Γ that meet the following conditions:

1. $W \subseteq \Gamma$ and
2. $\Gamma \subseteq W \cup \{cons(\Pi) \mid \Pi \text{ is some process of } T\}$.

Informally, an agent has yet to test any potential extension against all the other available reasons before it can be considered as an admissible extension of the evidence base.

Definition 2.18 ($\mathbf{JT_{CS}}$-Admissible Extension). *A potential extension set of $\mathbf{JT_{CS}}$ formulas $\Gamma \subset In(\Pi)$ is a $\mathbf{JT_{CS}}$-admissible extension of a default theory $T = (W, D)$ iff $Th^{JT_{CS}}(\Gamma)$ is conflict-free, each formula $t : F \in \Gamma$ is acceptable w.r.t. Γ and Π is closed.*

After considering all the available reasons, an agent accepts only those defeasible statements that can be defended against all the available reasons against these statements.[25]

The two latter definitions introduce the idea of "external stability" of knowledge bases (Dung, 1995, p. 323) into default logic by taking into account that only those reasons that are able to defend themselves against the reasons that question their plausibility eventually become accepted. In addition to that, our operational semantics prompts an implicit revision procedure. Any new default rule that is applicable to the set of formulas $In(\Pi[k])$ potentially makes changes to what an agent considered to be acceptable relying on the set of formulas $In(\Pi[k-1])$. Before we show this on the formalized example from the beginning of this section, we introduce the idea of default extension for a default theory T. Extension

[25] We do not require $\mathbf{JT_{CS}}$-admissible extension sets to be closed under $\mathbf{JT_{CS}}$ consequence. This is required for $\mathbf{JT_{CS}}$ variants of preferred, complete, grounded and stable extensions that give a definitive answer to the problem of which arguments are entailed according to a the criteria defined by the semantics that those extensions represent. Preferred, complete, grounded and stable semantics are required to give comprehensive answers about those arguments that are entailed for a default theory and, therefore, the corresponding extensions need to be $\mathbf{JT_{CS}}$-closed.

is the fundamental concept in defining logical consequence in standard default theories. We think of preferred extensions as maximal plausible world views based on the acceptability of reasons:

Definition 2.19 (**JT$_{CS}$**-Preferred Extension)**.** *For a default theory $T = (W, D)$, a closure $Th^{JT_{CS}}(\Gamma)$ of a* **JT$_{CS}$***-admissible extension Γ is a* **JT$_{CS}$***-preferred extension of T iff for any other* **JT$_{CS}$***-admissible extension Γ', $\Gamma \not\subset \Gamma'$.*

In other words, **JT$_{CS}$**-preferred extensions are maximal **JT$_{CS}$**-admissible extensions with respect to set inclusion. The existence of **JT$_{CS}$**-preferred extensions is universally defined for default theories. To ensure that this result also holds for the case of an infinite number of default rules and infinite closed processes, we make use of Zorn's lemma and restate it as follows:

Lemma 2.20 (Zorn, 1935)**.** *For every partially ordered set A, if every chain of (totally ordered subset of) B has an upper bound, then A has a maximal element.*

Theorem 2.21 (Existence of **JT$_{CS}$**-Preferred Extension)**.** *Every default theory $T = (W, D)$ has at least one* **JT$_{CS}$***-preferred extension.*

Proof. If W is inconsistent, then for any default δ, negation of the consistency requirement $req(\delta)$ is contained in $Th^{JT_{CS}}(W)$ and the only closed process Π is the empty sequence. Therefore, the only potential and **JT$_{CS}$**-admissible extension is W itself and T has a unique **JT$_{CS}$**-preferred extension $Th^{JT_{CS}}(W)$ containing all the formulas of **JT$_{CS}$**.

Assume that W is consistent. In general, if there is a finite number of default rules in D, any closed process Π of T is also finite. **JT$_{CS}$**-admissible extensions obtained from closed processes form a complete partial order with respect to $\subseteq$. Since there are only finitely many **JT$_{CS}$**-admissible sets, any **JT$_{CS}$**-admissible set Γ has a maximum Γ' within a totally ordered subset of a set of all **JT$_{CS}$**-admissible sets. Therefore, $\Gamma \subseteq \Gamma'$ and $Th^{JT_{CS}}(\Gamma')$ is a **JT$_{CS}$**-preferred extension of T.

For the case where D is infinite and closed processes $\Pi_1, \Pi_2, \ldots$ are infinite, there is again a complete partial order formed from a set of all **JT$_{CS}$**-admissible sets. The argument for finite processes does not account for the case where Γ', the union of **JT$_{CS}$**-admissible sets $\Gamma_1, \Gamma_2, \ldots$, could be contained in some Γ'' for an ever increasing sequence $\Gamma_1, \Gamma_2, \ldots$. We first state that Γ', the union of an ever increasing sequence of **JT$_{CS}$**-admissible sets $\Gamma_1, \Gamma_2, \ldots$, is also a **JT$_{CS}$**-admissible set. To ensure this, we turn to its subsets. That is, if Γ' was not admissible, then some of

its subsets Γ_n for $n \geq 1$ would not be conflict-free or would contain a formula that is not acceptable, but this contradicts the assumption that Γ_n is $\mathbf{JT_{CS}}$-admissible. Now, for the set of all $\mathbf{JT_{CS}}$-admissible sets ordered by $\subseteq$, any chain (totally ordered subset) has an upper bound, that is, the union of its members $\Gamma' = \bigcup_{n=1}^{\infty} \Gamma_n$. According to Lemma 2.20, there exists a maximal element and, therefore a $\mathbf{JT_{CS}}$-preferred extension of T. □

The semantics of defeasible reasons enables us to define additional types of extensions that are not necessarily based on the admissibility of reasons. One of them is the stable extension familiar from formal argumentation theory (Dung, 1995):

Definition 2.22 ($\mathbf{JT_{CS}}$-Stable Extension)**.** *For a default theory $T = (W, D)$, a conflict-free closure $Th^{JT_{CS}}(\Gamma)$ of a potential extension Γ is a $\mathbf{JT_{CS}}$-stable extension of T iff for any process Π of T, Γ undercuts all the formulas $t : F \in cons(\Pi)$ outside $Th^{JT_{CS}}(\Gamma)$.*

The intuition behind the definition is that every reason left outside the accepted set of reasons is attacked. To understand the process semantics workings of the stable extension definition, we can parse this definition into two components. First, it is clear that a stable extension $Th^{JT_{CS}}(\Gamma)$ undercuts each default reason t for every $cons(\delta) = t : F$ such that $t : F$ is not contained in Γ, but δ occurs in a closed process Π of T for which it holds that Γ is a subset of the evidence base $In(\Pi)$. Intuitively, from those reasons that are applicable within a closed default process, only the reasons that are undercut by $Th^{JT_{CS}}(\Gamma)$ are left outside. But notice that, secondly, for each default reason u and a formula $cons(\delta') = u : G$ such that Γ and $u : G$ do not co-occur in any potential extension of T, but $u : G$ is included in some potential extension of T, it holds that u also has to be undercut. This means that if δ' cannot be applied to the default process Π and δ' occurs in some other closed process Π', then Γ undercuts u. To see why, take for example the justification assertions $t : F = cons(\delta)$ and $v : \neg F = cons(\delta'')$. For any potential extension $\Gamma \subset In(\Pi)$ such that $\delta \in \Gamma$ and δ'' is not applicable to Π due to the inconsistency of the formula $req(\delta'')$, $Th^{JT_{CS}}(\Gamma)$ contains an undercutter for the reason v. In fact, if $t : F \in \Gamma$, then $Th^{JT_{CS}}(\Gamma)$ entails a formula $\neg r : (J \rightarrow \neg F)$ for any formula $J \in In(\Pi)$ and any reason r. Therefore, it also contains some reason term s that undercuts the warrant of the default rule $cons(\delta'')$. This means that inconsistent justification assertions responsible for rebuttal

indirectly undercut rebutted reasons. This undercut is further inherited by all the potential default reasons that are inferred from inconsistent default reasons, even if these are not involved in any rebuttal induced by $\mathbf{JT_{CS}}$ inconsistency. The following lemma generalizes this observation on the dependence between rebuttal and undercut:

Lemma 2.23. *For a default theory $T = (W, D)$ and its closed processes Π and Π', if some rule $\delta = \frac{t:F::(u\cdot t):G}{(u\cdot t):G}$ from Π' is inapplicable to $In(\Pi)$ and $t : F \in In(\Pi)$, then there is a potential extension $\Gamma \subset In(\Pi)$ that undercuts $(u \cdot t)$ being a reason for G.*

Proof. By Theorem 2.5, we know that there is some segment $In(\Pi[k])$ that contains the formula $t : F$ and, by assumption, that δ is inapplicable to $In(\Pi[k])$. Therefore, $In(\Pi[k])$ contains the formula $\neg(u \cdot t) : G$. According to axiom A1 and propositional reasoning, if the $\mathbf{JT_{CS}}$ closure $In(\Pi[k])$ contains $t : F$ and $\neg(u \cdot t) : G$, then it also contains the formula $\neg[u : (F \rightarrow G)]$. By the definition of an *In*-set (Def. 2.9) and the way in which potential extensions are built for T, there is some potential extension $\Gamma \subset In(\Pi[k])$ such that $Th^{JT_{CS}}(\Gamma)$ contains $\neg[u : (F \rightarrow G)]$. Since $\#(\delta) = u : (F \rightarrow G)$ and $u : (F \rightarrow G) \in \mathcal{WS}^{\Pi'}$, Γ undercuts $(u \cdot t)$ being a reason for G by Definition 2.14. □

If a potential extension Γ of T undercuts all the formulas left outside, then Γ also has to maximize admissibility with respect to set inclusion. This straightforwardly leads to the following lemma:

Lemma 2.24. *Every $\mathbf{JT_{CS}}$-stable extension of a default theory $T = (W, D)$ is also a $\mathbf{JT_{CS}}$-preferred extension of T.*

We can check that in the red-looking-table example, $\mathbf{JT_{CS}}$-stable and $\mathbf{JT_{CS}}$-preferred extension coincide. Formally, theory T_0 has a unique $\mathbf{JT_{CS}}$-stable and $\mathbf{JT_{CS}}$-preferred extension $Th^{JT_{CS}}(W_0 \cup \{cons(\delta_b), cons(\delta_c)\})$. Moreover, note that the process $\Pi_1 = (\delta_a, \delta_b)$ includes revising the resulting set of acceptable reasons, since the reason $(u_b \cdot t_b)$ undercuts $(u_a \cdot t_a)$ being a reason for formula T.

However, $\mathbf{JT_{CS}}$-stable extensions are not universally defined for any default theory T. To show this, we will formalize Pollock's "pink elephant" example (1995, pp. 119-120, 2009, pp. 181-182). This example is an instance of defeasible reasoning with a self-defeating argument. The concept of self-defeat is notorious in argumentation theory. Firstly, suppose that

Robert says that the elephant beside him looks pink. Normally, we would take Robert's testimony to support the conclusion that the elephant is pink. However, Robert suffers from what is known as "pink-elephant phobia". People in this condition "become strangely disoriented so that their statements about their surroundings cease to be reliable" (Pollock, 2009, p. 181). Therefore, Pollock concludes that it seems that "if it were true that the elephant beside Robert is pink, we could not rely upon his report to conclude that it is".

Example 2.25. *Let P be the proposition "The elephant looks pink", let E be the proposition "The elephant is pink", and let H be the proposition "Robert suffers from pink-elephant phobia". The pink elephant example is then described by the default theory $T_1 = (W_1, D_1)$, where $W_1 = \{k : H, l : P\}$ and D_1 consists of the default rules*[26]

$$\delta_1 = \frac{l : P :: (m \cdot l) : E}{(m \cdot l) : E} \text{ and}$$

$$\delta_2 = \frac{(m \cdot l) : E :: (n \cdot (m \cdot l)) : \neg[m : (P \to E)]}{(n \cdot (m \cdot l)) : \neg[m : (P \to E)]}.$$

While the structure of the backing for δ_1 resembles that of δ_a from Example 2.7, the backing for the default rule δ_2 has a more intricate structure:

1	$x : [m : (P \to E) \to \neg(E \wedge H)]$	(Assumption)
2	$k : H$	(Assumption)

[26]Notice that in the original formulation of his pink elephant example, Pollock introduces (1995, p. 120) an intermediate inference between the rules δ_1 and δ_2. Namely, he thinks that there is an inference from Robert's saying (reason term l) that the elephant looks pink to him, to the conclusion that it *does* look pink. We follow a version of the example that does not take the intermediate step as a separate inference, taken from (Koons, 2017, § 4.1). There are two reasons for this decision. Firstly, Pollock's red table example that we formalized in Example 2.7 has the same structure of inference that starts from seeing a red-looking table to conclude that the table is red. There is no mention of the table looking red independently of an agent's report that it does. It is not clear why to think that Robert's unreliability in the presence of pink elephants would question the fact that the elephant does look pink, even if Robert himself realizes that he suffers from the phobia. It is also not clear what would it mean for an object to look pink, regardless of being perceived as pink by some agent. Secondly, a report of another agent to whom the elephant does not look pink would be treated differently in justification logic. Such report would undermine Robert's own report and the subject matter of undermining attacks is dealt with in Chapter 4, together with the topic of how to model testimonies. In any case, an intermediate default rule could formally be added without affecting the significance of the example for the discussion.

3 $[m : (P \to E) \to \neg(E \wedge H)] \to$
$[(H \to (E \to \neg[m : (P \to E)])]$ (A0)

4 $c : ([m : (P \to E) \to \neg(E \wedge H)] \to$
$[(H \to (E \to \neg[m : (P \to E)])])$ (R1)

5 $(c \cdot x) : [(H \to (E \to \neg[m : (P \to E)])]$ (1,4 A1)

6 $((c \cdot x) \cdot k) : (E \to \neg[m : (P \to E)])$ (2,5 A1)

Let $n = ((c \cdot x) \cdot k)$. The above inference steps in $\mathbf{JT_{CS}}$ *formalize the backing for the warrant $n : (E \to \neg[m : (P \to E)])$ of δ_2. Notice that, in the formalization of its backing, the warrant of δ_2 is supported by appeal to the presupposed information about the phobia that Robert suffers from, that is, to the justification assertion $k : H$.*

The theory T_1 has a $\mathbf{JT_{CS}}$-preferred extension $Th^{JT_{CS}}(W_1)$. However, it has no $\mathbf{JT_{CS}}$-stable extension, because the available reasons cannot form a conflict-free set that attacks all the reasons outside that set. This result conforms to similar results about preferred and stable semantics in abstract argumentation frameworks (Dung, 1995, p. 328). By the end of the section, we define the theory T_3 that shows the same type of a self-defeating argument alongside other arguments. In our default theories, self-defeating arguments do not influence other independent arguments, except in the above-illustrated sense of affecting the existence of stable semantics.

In addition, we can easily define other significant notions of extensions in formal argumentation. In particular, we can define variants of Dung's (1995, p. 329) *complete* and *grounded* extension:

Definition 2.26 ($\mathbf{JT_{CS}}$-Complete Extension)**.** *For a default theory $T = (W, D)$, a closure $Th^{JT_{CS}}(\Gamma)$ of a* $\mathbf{JT_{CS}}$*-admissible extension Γ is a* $\mathbf{JT_{CS}}$*-complete extension of T iff for each closed process Π of T such that there is a* $\mathbf{JT_{CS}}$*-admissible extension Γ' in $In(\Pi)$ and $\Gamma \subset \Gamma'$, if a formula $t : F \in cons(D)$ is acceptable w.r.t. Γ in $In(\Pi)$, then $t : F$ belongs to Γ.*

Definition 2.27 ($\mathbf{JT_{CS}}$-Grounded Extension)**.** *For a default theory $T = (W, D)$, a* $\mathbf{JT_{CS}}$*-complete extension $Th^{JT_{CS}}(\Gamma)$ is the unique* $\mathbf{JT_{CS}}$*-grounded extension if Γ is the smallest potential extension with respect to set inclusion such that $Th^{JT_{CS}}(\Gamma)$ is a* $\mathbf{JT_{CS}}$*-complete extension of T.*[27]

[27]Note here that the we know that there is the smallest potential extension which is

Unsurprisingly, the results for different types of extensions defined by Dung (1995) are valid for our default theory extensions.

Lemma 2.28. *Every* $\mathbf{JT_{CS}}$*-preferred extension of a default theory* $T = (W, D)$ *is also a* $\mathbf{JT_{CS}}$*-complete extension of* T*.*

Proof. Assume that $Th^{JT_{CS}}(\Gamma)$ is a $\mathbf{JT_{CS}}$-preferred extension of T for some potential extension Γ. Assume towards contradiction that for some closed process Π such that $\Gamma \subset In(\Pi)$ and Γ is $\mathbf{JT_{CS}}$-admissible there exists a formula $cons(\delta)$, where $\delta \in \Pi$, acceptable with respect to Γ, but not included in Γ. According to Def. 2.18, there is a $\mathbf{JT_{CS}}$-admissible extension Γ' for which it holds that $\Gamma \subset \Gamma'$. But this contradicts the assumption that $Th^{JT_{CS}}(\Gamma)$ is a $\mathbf{JT_{CS}}$-preferred extension. Therefore, for any closed process Π' for which Γ is $\mathbf{JT_{CS}}$-admissible and for any formula $cons(\delta')$ such that $\delta' \in \Pi'$, if $cons(\delta')$ is acceptable with respect to Γ, then $cons(\delta')$ is included in Γ. □

It does not hold, however, that every $\mathbf{JT_{CS}}$-complete extension is also $\mathbf{JT_{CS}}$-preferred. The following theory T_2 is a counterexample. Let the theory be defined as $T_2 = (W_2, D_2)$, where $W_2 = \{p : K, q : L\}$ and D_2 consists of the default rules

$$\delta_3 = \frac{p : K :: (r \cdot p) : M}{(r \cdot p) : M} \text{ and}$$

$$\delta_4 = \frac{q : L :: (s \cdot q) : \neg M}{(s \cdot q) : \neg M}.$$

One of the $\mathbf{JT_{CS}}$-complete extensions of T_2 is $Th^{JT_{CS}}(W_2)$, as a result of the fact that none of the available default reasons is acceptable with respect to the potential extension W_2. However, $Th^{JT_{CS}}(W_2)$ is not one of $\mathbf{JT_{CS}}$-preferred extensions for T_2. The theory has two $\mathbf{JT_{CS}}$-preferred extensions such that one of them contains $cons(\delta_3)$, while the other contains $cons(\delta_4)$.

Considering some proposition as justified might be seen as a function of interacting reasons. Each of the presented $\mathbf{JT_{CS}}$ extensions is a method to compute extensions with justified formulas. Moreover, each

$\mathbf{JT_{CS}}$-complete since we can represent $\mathbf{JT_{CS}}$-admissible extensions as forming a complete partial order w.r.t. set inclusion. Ordered extensions lend themselves to a fixed-point reformulation of all admissibility-based extensions and a possibility of guaranteeing the existence of the smallest potential extension by the application of the Knaster-Tarski theorem (Tarski, 1955).

of the $\mathbf{JT_{CS}}$ extension definitions can be used as a way to define a corresponding characterization of logical consequence. Given a particular $\mathbf{JT_{CS}}$ extension of a theory T, the formulas contained in that extension are valid formulas for T under that specific $\mathbf{JT_{CS}}$ semantics. There are some analogies with the traditional notions of non-monotonic consequence relations. For example, $\mathbf{JT_{CS}}$-grounded extensions correspond to *cautious* consequence relations describing what a skeptical reasoner would accept for some default theory. In a similar way, $\mathbf{JT_{CS}}$-preferred semantics describes a *credulous* inference relation. The consequence relation defined by $\mathbf{JT_{CS}}$-stable extension is an interesting case in this context. Although for many default theories $\mathbf{JT_{CS}}$-stable and $\mathbf{JT_{CS}}$-preferred semantics coincides, there are some intuitive grounds to consider $\mathbf{JT_{CS}}$-stable extensions as skeptical in nature. This specifically relates to the demand that the existence of $\mathbf{JT_{CS}}$-stable extensions depends on whether a set of $\mathbf{JT_{CS}}$ formulas is able to defeat all other reasons outside that set or not. Such excessive demands on the validity of formulas do not comply to our ordinary intuitions about credulous consequence relations.

To illustrate the differences among the above defined semantics, we will elaborate on an example of a single default theory whose $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-complete, $\mathbf{JT_{CS}}$-preferred and $\mathbf{JT_{CS}}$-stable extensions do not coincide, although each of them exists. We define the default theory $T_3 = (W_3, D_3)$ with $W_3 = \{t_1 : F, t_2 : H, t_3 : I\}$ and $D_3 = \{\delta_5, \delta_6, \delta_7, \delta_8\}$, where δ_5, δ_6, δ_7 and δ_8 are defined as follows:

$$\delta_5 = \frac{t_1 : F :: (u_1 \cdot t_1) : G}{(u_1 \cdot t_1) : G},$$

$$\delta_6 = \frac{(u_1 \cdot t_1) : G :: (u_2 \cdot (u_1 \cdot t_1)) : \neg[u_1 : (F \to G)]}{(u_2 \cdot (u_1 \cdot t_1)) : \neg[u_1 : (F \to G)]},$$

$$\delta_7 = \frac{t_2 : H :: (u_3 \cdot t_2) : (J \wedge \neg[u_2 : (G \to \neg[u_1 : (F \to G)])])}{(u_3 \cdot t_2) : (J \wedge \neg[u_2 : (G \to \neg[u_1 : (F \to G)])])} \text{ and}$$

$$\delta_8 = \frac{t_3 : I :: (u_4 \cdot t_3) : \neg J}{(u_4 \cdot t_3) : \neg J}.$$

Any evidence base $In(\Pi)$ of T_3 containing the formula $cons(\delta_7)$ will also contain the formula $(c_1 \cdot (u_3 \cdot t_2)) : \neg[u_2 : (G \to \neg[u_1 : (F \to G)])]$, which represents the reasoning behind an argument that questions the warrant of the self-defeating argument given in δ_2 by undercutting $(u_2 \cdot (u_1 \cdot t_1))$. The undercutter $(c_1 \cdot (u_3 \cdot t_2))$ can be derived from $cons(\delta_7)$ with some propositional reasoning combined with the use of axiom A1 and rule R1*.

Moreover, default δ_5 provides an argument that rebuts the reason $(u_4 \cdot t_3)$ for $\neg J$, for any extension that contains $cons(\delta_7)$. This argument is codified within the term $(c_2 \cdot (u_3 \cdot t_2))$ justifying the formula J, again assuming some propositional reasoning, axiom A1 and rule R1*. Accordingly, the rules δ_7 and δ_8 cannot occur together in any default process of T_3.

In total, the theory T_3 has six closed processes, as shown in the process tree of T_3 displayed in Figure 2.7. Building a process tree for our default theories proceeds in the following way: each node of the process tree is labeled with an *In*-set after a default rule (connecting edges) has been applied. Note that, for each node of the process tree in Figure 2.7 and a closed process Π of T_3, if a node corresponds to some segment $\Pi[k]$ of Π we indicate only the formula that has been added to $In(\Pi[k])$ as a result of applying an available default rule to $In(\Pi([k-1])$. The process tree helps us to check the status of $\mathbf{JT_{CS}}$ extensions for T_3. The theory has two preferred extensions, namely $Th^{JT_{CS}}(W_3 \cup \{cons(\delta_5), cons(\delta_7)\})$ and $Th^{JT_{CS}}(W_3 \cup \{cons(\delta_8)\}$. Of the two $\mathbf{JT_{CS}}$-preferred extensions, only $Th^{JT_{CS}}(W_3 \cup \{cons(\delta_5), cons(\delta_7)\})$ is also $\mathbf{JT_{CS}}$-stable. A skeptical reasoner will only accept $Th^{JT_{CS}}(W_3)$, the unique $\mathbf{JT_{CS}}$-grounded extension of T_3. Finally, all the three mentioned $\mathbf{JT_{CS}}$ closures are $\mathbf{JT_{CS}}$-complete for T_3.

It is possible to specify conditions under which different $\mathbf{JT_{CS}}$ extension notions above coincide. Sufficient conditions need to eliminate the possibility of attack cycles. We first define the cycle of asymmetrical attacks:

Definition 2.29 (Undercut Cycle)**.** *A cycle of undercuts is an infinite periodic sequence of* $\mathbf{JT_{CS}}$ *formulas* $t_1 : F_1, \ldots, t_n : F_n, t_1 : F_1, \ldots, t_n : F_n, t_1 : F_1, \ldots$, *for some number of formulas* $n \geq 1$*, such that each reason* t_i *undercuts* t_k *being a reason for the formula* F_k *according to Def. 2.13 and* $t_i : F_i$ *is the predecessor of the formula* $t_k : F_k$ *in the sequence.*

Rebuttals among formulas ultimately derive from the property of $\mathbf{JT_{CS}}$ inconsistency. They are thus symmetric and can be traced through the process semantics and existence of different evidence bases $In(\Pi')$ and $In(\Pi'')$ for some closed processes Π and Π'. Therefore, we do not need to define rebuttal separately, but only provide a condition that excludes attacks induced by $\mathbf{JT_{CS}}$ inconsistency.

We are ready now to give the conditions for the coincidence of $\mathbf{JT_{CS}}$ extensions in *well-founded* default theories. A default theory $T = (W, D)$ is called well-founded if for all closed processes Π and Π' of T it holds that:

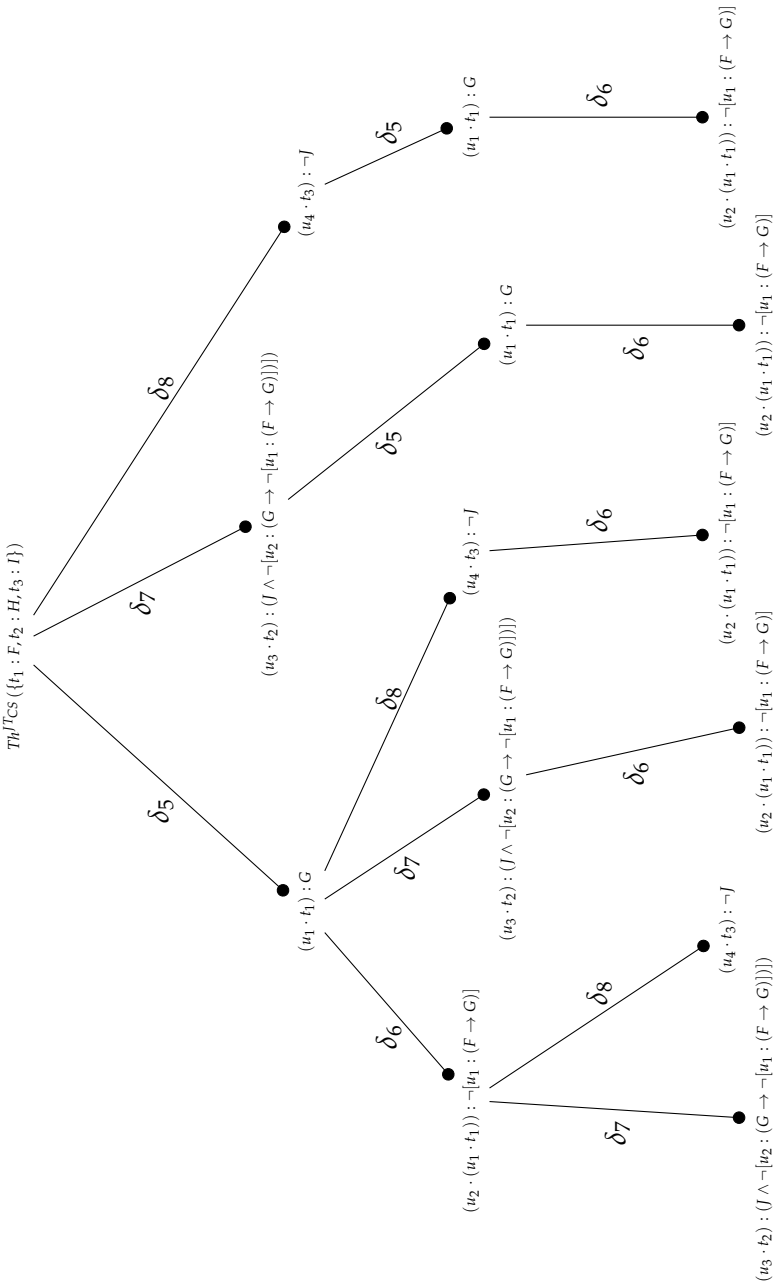

Figure 2.4: The process tree of T_3

1. $In(\Pi) = In(\Pi')$ and
2. There are no sets of $\mathbf{JT_{CS}}$ formulas $\Gamma \in In(\Pi)$ forming a cycle of undercuts.

The following theorem shows that $\mathbf{JT_{CS}}$ extensions of well-founded default theories coincide.[28]

Theorem 2.30. *Every well-founded default theory $T = (W, D)$ has a unique $\mathbf{JT_{CS}}$-complete extension $Th^{JT_{CS}}(\Gamma)$ which is $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred and $\mathbf{JT_{CS}}$-stable.*

Proof. Firstly, if a $\mathbf{JT_{CS}}$-grounded extension is also a $\mathbf{JT_{CS}}$-stable extension of a default theory $T = (W, D)$, then it is also $\mathbf{JT_{CS}}$-preferred and the unique $\mathbf{JT_{CS}}$-complete extension of T. Therefore, it is sufficient to focus on the proof that each $\mathbf{JT_{CS}}$-grounded extension is $\mathbf{JT_{CS}}$-stable for a well-founded theory.

Assume that a well-founded theory T has a $\mathbf{JT_{CS}}$-grounded extension $Th^{JT_{CS}}(\Gamma)$ that is not $\mathbf{JT_{CS}}$-stable. The set $\Gamma \subset In(\Pi)$ is the smallest potential extension such that $Th^{JT_{CS}}(\Gamma)$ is a $\mathbf{JT_{CS}}$-complete extension of T. Moreover, there is at least one formula $t : F \in cons(\delta)$ from the set of consequents $cons(D)$ such that $t : F \notin Th^{JT_{CS}}(\Gamma)$, but, since $Th^{JT_{CS}}(\Gamma)$ is not $\mathbf{JT_{CS}}$-stable, $Th^{JT_{CS}}(\Gamma)$ does not undercut t being a reason for F. Now we have to show that unless $Th^{JT_{CS}}(\Gamma)$ undercuts t being a reason for F, at least one of the following statements has to hold about T:

(1) $t : F$ is acceptable w.r.t. Γ in $In(\Pi)$, but Γ is a subset of $In(\Pi')$ for some other closed process Π' and $t : F$ is not acceptable w.r.t. Γ in $In(\Pi')$. But this means that the sets $In(\Pi)$ and $In(\Pi')$, which, in turn, means that T is not well-founded according to condition (1) on well-founded default theories;

(2) $t : F$ is not acceptable w.r.t. Γ in $In(\Pi)$ and there is some formula $v : G \in In(\Pi)$ such that v undercuts t being a reason for F, but $Th^{JT_{CS}}(\Gamma)$ does not undercut v being a reason for G. However, $v : G$ is not contained in Γ, since we assumed that Γ is not $\mathbf{JT_{CS}}$-stable and that Γ does not undercut t being a reason for F. But this means that there exists an infinite periodic sequence of $\mathbf{JT_{CS}}$ formulas $t_1 : F_1, \ldots, t_n : F_n, t_1 : F_1, \ldots, t_n : F_n, t_1 : F_1, \ldots$ forming an undercut cycle according to Def. 2.29. This means that T is not well-founded according to condition (2).

[28] Compare (Dung, 1995, p. 331) for well-foundedness of abstract argumentation frameworks. Here we adapt the proof idea for the coincidence of extensions of well-founded abstract argumentation frameworks that can be found there.

Therefore, since T is well-founded, it has a unique $\mathbf{JT_{CS}}$-complete extension $Th^{JT_{CS}}(\Gamma)$ which is $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred and $\mathbf{JT_{CS}}$-stable. □

2.8 Conclusions

In this chapter we defined default justification logic and interpreted justification logic formulas as arguments. we generalized justifications to represent defeasible reasons, whose acceptability depends on other available reasons. To this end, we generalized the application operation in such a way that default rules may introduce reason terms that are not proofs, but only provide support for contingent conditional statements to a certain extent. Formally such reason terms represent derivations with non-empty sets of assumptions.

The properties of our system are still to be thoroughly investigated. In the context of non-monotonic reasoning, our logic introduces some technical possibilities already for normal default theories. Among them are revision of extensions and interaction of different defaults without relying on their preference orderings, as commonly done in default logic (Delgrande and Schaub, 2000). An extensive account of default reasons that makes use of preference orderings on defaults is developed by Horty (2012). Horty's logic is based on a propositional language and develops from a different formal account of reasons, where reasons are not explicitly featured as object level terms. Horty uses the idea of preferences to represent undercutters or exclusionary reasons.

Our work provides a complementary addition to the study of less-than-ideal reasons in justification logic. Among related approaches, the logic of conditional probabilities developed by Ognjanović, Savić, and Studer (2017) introduces a way to model non-monotonic reasoning with justification assertions. Their proposal is based on defining operators for approximate probabilities of a justified formula given some condition formula. Using conditional probabilities, the logic models certain aspects of defeasible inferences with justification terms. Yet the system can neither encode the defeasibility of justification terms in their internal structure nor model defeat among reasons, to mention only some differences from our initial desiderata.

Baltag, Renne, and Smets (2012) define a justification logic in which an agent may hold a justified belief that can be compromised in the face

of newly received information. The logic builds on the ideas from belief revision and dynamic epistemic logic to model examples where epistemic actions cause changes to an agent's evidence. Concerning the possibility of modelling defeaters, the logic offers two dynamic operations that change the availability of evidence in a model, namely "updates" and "upgrades" (Baltag et al., 2012, p. 183). Evidence obtained by updates counts as "hard" or infallible, while upgrades bring about "soft" or fallible evidence. With the use of these actions, epistemic models can represent justified beliefs being defeated, for example, by means of an epistemic action of update with hard evidence. In this way, however, the mechanism by which reasons may conflict with one another is simply being "outsourced" to an extra-logical notion of fallibility and, therefore, the logic does not directly address the ways of defeat that we formalize in this chapter.

Further developments are possible starting from the basic form of default rules with justified formulas. We indicate some of the possibilities to extend the basic logic. On the technical side of the logic, we used only the expressiveness of normal default rules and we still need to investigate how it could be extended with non-normal default rules. Since all processes are successful for normal default theories, it is interesting to see whether the logic has some further desirable properties such as, for example, goal-driven query evaluation.

It is also possible to use the first-order variant of justification logic (Fitting, 2014), instead of the propositional justification logic used here. This is an intriguing direction because of the possibilities it opens. To mention one of them, a first-order warrant of a default rule would enable expressing default schemes with variables as placeholders for objects. Such rules would fully capture the informal idea of Toulminian warrants, which are meant to be schematic generalizations. Defining default rules on such rich language would be one step closer to a full logical account of structured arguments.

Finally, the logic of default justifications has a potential to link the logical analysis of justifications with the philosophical study of defeasibility and knowledge. Ever since the concept of justification entered into epistemic logics, there has been a tendency to model mainstream epistemology examples, proposed by e.g. Russell (1912), Gettier (1963) and Dretske (2005), with the use of justification logic (Artemov, 2008, 2018). With the introduction of default justifications, however, we gain flexibility for a more full-blooded integration of the formal theory of jus-

tification with the study of knowledge in philosophy, since paradigmatic examples include both incomplete specification of reasons and defeated reasons. Potential benefits of a non-monotonic system of justifications in this context were anticipated by Artemov (2008, p. 482), who proposes that "to develop a theory of non-monotonic justifications which prompt belief revision" stands as an "intriguing challenge".

Chapter 3

Relations of default justification logic to formal argumentation and Reiter's default logic

3.1 Introduction

In this chapter, we relate default justification logic to formal argumentation and Reiter's default logic. We first show that, by abstracting from the structure of arguments and focusing only on the direction of attacks, we can obtain Dung's frameworks from the default justification logic and, *vice versa*, our logic provides realization procedures for Dung's frameworks that assign justification assertions to Dung's arguments. We then discuss how our logic complies with the rationality postulates for structured argumentation frameworks proposed by Amgoud (2014). We conclude the chapter by discussing the benefits of modelling default reasons with our logic over standard default logic.

3.2 Realizing Dung's frameworks in justification logic

In Chapter 1, we briefly described Dung's abstract argumentation frameworks (AF) that deal with the problem of the acceptability of arguments

based on their mutual conflicts. An argumentation framework is defined as a pair of a set of arguments, and a binary relation representing the attack-relationship (defeat) between arguments. These frameworks were characterized as abstract because they neither represent the structure of arguments nor do they specify the exact nature of attacks between them. In Chapter 2 we presented a default justification logic approach that both represents the structure of arguments and spells out their mutual attacks in terms of undercut and rebuttal.

In this section we examine connections between abstract argumentation frameworks and our default justification logic. Our semantics of justification formulas $t : F$ can be naturally related to the concepts of argumentation semantics. Any justification formula can be plausibly regarded as an argument where t codifies premises and F is a conclusion of an argument.[1] However, the expressiveness of the language $\mathbf{JT_{CS}}$ enables us to construct the complex argument structures that result from logical operations on formulas. As expected, abstract argumentation frameworks are not able to capture all the subtleties of more complex default reasons. Interestingly, it turns out that there are also AF structures that cannot be translated into default theories.

We first focus on the possibility of mapping from default theories to AFs. To establish the connection between default reasons semantics and AF semantics, we need to restrict our attention to a subclass of our default theories. Since our logic is more expressive with respect to attack relations, we focus on non-complex default theories where attack relations are defined only by looking at the union of logical consequences of each consequent of a default rule. In this way, each default rule is taken separately as a self-contained argument. To achieve this, we first specify what it means for two default rules to *block* each other's applicability. For a process Π of $T = (W, D)$, the rules δ and δ' from D block each other in Π iff for some segment $\Pi[k]$ such that both δ and δ' are applicable to $In(\Pi[k])$, if either of the two defaults has been applied, the other default becomes inapplicable to $In(\Pi[k+1])$. A default theory $T = (W, D)$ is *non-complex* if it fulfills the following two conditions:

1. If two defaults δ and δ' from D block each other in a process Π of T, then for each process Π' with a segment $\Pi'[k]$ such that either δ or δ' has been applied to $In(\Pi'[k])$ it holds that the default that

[1] We can say this also about the formula $c : F$, where c is a proof constant, but in this case the attack relation will be empty.

has not been applied to $In(\Pi'[k])$ is inapplicable to $In(\Pi'[k+n])$ for any segment $\Pi'[k+n]$ of Π';

2. For a process Π of T, a reason t such that $t : F \in In(\Pi)$ and any undercutter u for t such that $u : \neg[v : (G \rightarrow H)] \in In(\Pi)$ for some $v \in Sub(t)$, there exists a reason $w \in Sub(u)$ such that $w : \neg[v : (G \rightarrow H)] \in Th^{JT_{CS}}(cons(\delta))$ for a default rule $\delta \in \Pi$.

In other words, we require for any defeat that occurs in a theory T that it be derivable only from a consequent of a default rule, because joint attacks cannot be represented in Dung's (1995) framework.

Using default justifications, one can look into the details of the structure of the arguments, including grounds, warrants, backings and different ways of attack, while Dung's framework treats arguments abstracting from their contents. This means that any translation from default theories with justification terms to Dung's framework has to "forget" information about the structure of the arguments. Having restricted our target theories to non-complex theories, we can now describe a mapping "$\Longrightarrow$" called *forgetful projection*. Forgetful projection converts each formula $cons(\delta)$ such that δ occurs in some process of a given default theory into a corresponding argument of an AF and it converts each attack among default reasons into a corresponding attack relation between arguments in an AF. A mapping $\Longrightarrow$ from a non-complex default theory $T = (W, D)$ to an abstract argumentation framework $AF = (Arg, Att)$, where Arg is a set of arguments $A_1, A_2, \ldots$ and Att is a a binary attack relation, is defined as follows:

- $\delta_n \in \Pi$ for a process $\Pi \quad \Longrightarrow \quad A_n \in Arg$
- $\delta_m \in \Pi'$ & $\delta_n \in \Pi''$ for some processes Π' & Π'' such that δ_m & δ_n do not occur together in any process $\Pi \quad \Longrightarrow \quad (A_m, A_n) \in Att$ & $(A_n, A_m) \in Att$
- $t : \neg[u : (F \rightarrow G)] \in Th^{JT_{CS}}(cons(\delta_m))$, $v : H = cons(\delta_n)$ such that $u \in Sub(v)$ & $u : (F \rightarrow G) \in \mathcal{WS}^{\Pi}$ and $\delta_m \in \Pi$ & $\delta_n \in \Pi \Longrightarrow \quad (A_m, A_n) \in Att$

Recall the theory T_0 from the Example 2.7, Chapter 2. The theory T_0 has its forgetful projection AF_0 that preserves the direction of the attacks from the original example. Consider that each of the rules δ_a, δ_b and δ_c is applicable to at least one process. This means that we can map all

three defaults to the arguments A_a, A_b and A_c in Arg_0. Given that δ_a and δ_c cannot be applied to the same process of T_0 and given the fact that they are applicable to some processes, both $(A_a, A_c) \in Att_0$ and (A_c, A_a) are in Att_0. Finally, notice that the rules δ_b and δ_a can be applied together in a default process and that the reason $(u_b \cdot t_b)$ undercuts $(u_a \cdot t_a)$ via justifying the denial of the warrant $u_a : (R \rightarrow T)$ of δ_a. Forgetful projection maps this relation between $cons(\delta_b)$ and $cons(\delta_a)$ into an additional attack (A_b, A_a) in Att_0.

Since forgetful projection does preserve the structure of conflicts among groups of arguments, it is possible to compare $\mathbf{JT_{CS}}$ extensions of default theories with extensions of the obtained AFs. It is not difficult to check that the following extension-correspondence statement holds:

Proposition 3.1. *For a formula $t : F = cons(\delta_n)$ such that $\delta_n \in D$ for a non-complex default theory $T = (W, D)$ and its $\mathbf{JT_{CS}}$-complete, $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred or $\mathbf{JT_{CS}}$-stable extension $Th^{JT_{CS}}(\Gamma)$, it holds that $t : F \in Th^{JT_{CS}}(\Gamma)$ iff an argument A_n is contained in the corresponding complete, grounded, preferred or stable extension sets for a forgetful projection $AF = (Arg, Att)$ of T.*

Proof. The proof is by induction on the acceptance conditions for a formula $t : F = cons(\delta)$ given by the definitions of $\mathbf{JT_{CS}}$-complete, $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred and $\mathbf{JT_{CS}}$-stable extension definitions for default theories.

The following is an argument for the $\mathbf{JT_{CS}}$-preferred extension case. Take as the induction base default theory $T = (W, D)$ such that T has a non-empty process $\Pi = (\delta)$ and $t : F = cons(\delta)$. The theory has a $\mathbf{JT_{CS}}$-preferred extension $Th^{JT_{CS}}(\Gamma)$, where $\Gamma = W \cup \{cons(\delta)\}$. The forgetful projection of T is defined as $AF = (Arg, Att)$, where $Arg = \{A\}$ and Att is empty. The only preferred extension of AF is A.

For the inductive step, assume that if $t : F = cons(\delta_k)$ is in a $\mathbf{JT_{CS}}$-preferred extension of a default theory $T_n = (W_n, D_n)$ such that $t : F$ occurs in a closed process $\Pi = (\delta_1, \ldots, \delta_m)$ of T, then it is also in a preferred extension of its forgetful projection $Af_n = (Arg_n, Att_n)$. By the induction hypothesis and the definition of $\mathbf{JT_{CS}}$-preferred extensions, it holds that $t : F \in \Gamma$ such that Γ is a $\mathbf{JT_{CS}}$-admissible extension and for no other $\mathbf{JT_{CS}}$-admissible extension Γ' it holds that $\Gamma \subset \Gamma'$. By the definition of a $\mathbf{JT_{CS}}$-admissible set, it holds that Γ is conflict-free and each formula in Γ is acceptable w.r.t. Γ in Π. For a non-complex default theory and the formula $t : F$, this means that for any undercutting

reason $u : G \in Th^{JT_{CS}}(W \cup \{cons(\delta_j)\})$ for t being a reason for F in Π, Γ undercuts u being a reason for G. The forgetful projection maps all the formulas $cons(\Pi')$ for any process Π' into arguments $A_1, \ldots, A_n$ of the framework Af_n and for the undercutter $u : G$ ascribes an attack relation (A_j, A_k), and analogously for any other possible undercutter. Moreover, any conflict-free set is also $\mathbf{JT_{CS}}$-consistent and for each formula $v : H = cons(\delta_p)$ such that $v : H \in \Gamma'$ for a $\mathbf{JT_{CS}}$-admissible extension Γ' of T and $v : H \notin cons(\Pi)$, it holds that $v : H$ and $t : F$ do not occur together in any process Π' because T is non-complex. According to the forgetful projection, (A_k, A_p) and (A_p, A_k) are both in Att_n. It is easy to check that, by the definition of Dung's preferred extension, the forgetful projection maps Γ into a preferred extension S of Af_n such that A_k is in S.

□

Intuitively, forgetful projections of justification logic arguments outline a single perspective on argumentation, namely that of opposition among arguments. Note that there are extensions of Dung's framework that formalize joint attacks from sets of arguments such as (Nielsen and Parsons, 2006). In this chapter, we on relating standard Dung's argumentation frameworks to our default justification logic. However, we assume that Proposition 3.1 can be generalized to any default theory with justification formulas for a richer abstract argumentation framework with joint attacks.

One may also ask whether the other direction of translating from argumentation frameworks to default theories always works. Since the content of arguments is not specified in Dung's framework, it is only possible to retrieve incomplete information about justification logic counterparts of Dung's frameworks. For any argument in Dung's framework, there are many justification logic realizations. Starting from a directed graph obtained from a framework $AF = (Arg, Att)$, each node A_i is paired with a corresponding formula $t_i : F_i$, where each $t_i : F_i$ is a consequent of some rule δ_i such that δ_i occurs in at least one process of a theory $T = (W, D)$ that realizes AF. Moreover, each node A_i is paired with a warrant $u_i : (G_i \to F_i)$.

The above defined procedure treats every single arrow in Dung's graph as a specification of $\mathbf{JT_{CS}}$ entailments from justification assertions paired with the nodes of a graph. Accordingly, we determine the structure of attacks among the obtained formulas. More specifically, a pointed arrow without an inverted arrow specifies that a default consequent

formula, which realizes a direct predecessor for the arrow, entails an undercut formula for the consequent formula via entailing the negation of a warrant that realizes the successor node. An arrow with an inverted arrow specifies inconsistency for consequent formulas paired with the connected nodes, that is, a rebuttal between the two formulas.[2] Using this procedure, we would get information on which formulas should a default consequent formula entail with respect to other default consequents, provided the definition of attack relations among arguments in Arg. However, the procedure fails as the following example shows. Take a simple framework $Af^* = (Arg, Att)$ with A as its only argument and $(A, A) \in Att$. It turns out that it is not possible to realize A as a single consequent of a default rule.

In fact, there are other cycles of attacks that cannot be realized in default justification logic using the proposed method to pair each argument in an AF with a consequent of a default rule from a theory T that realizes that AF. This problem can be generalized to a class of *unwarranted* argumentation frameworks featuring such attack cycles. An argumentation framework $AF = (Arg, Att)$ is said to be unwarranted iff:

1. There is an infinite sequence $A_1, A_2, \ldots, A_n, \ldots$ s. t. for each i, A_{i+1} attacks A_i;

2. For any two distinct arguments $A = A_k$ and $B = A_{k+1}$ s. t. A_k and A_{k+1} are adjacent members of the $A_1, A_2, \ldots, A_n, \ldots$ sequence, it does not hold that $(A, B) \in Att$ and $(B, A) \in Att$;

3. There exists no argument C outside the sequence s. t.:

    a) for some A from the sequence $A_1, A_2, \ldots, A_n, \ldots$ it holds that $(A, C) \in Att$;

    b) C is not a member of an infinite sequence $B_1, B_2, \ldots, B_n, \ldots$ s. t. for each i, B_{i+1} attacks B_i;

    c) for no two distinct arguments D and E from Arg it holds that $(D, C) \in Att$ and $(E, C) \in Att$.

We refer to all argumentation frameworks that are not unwarranted as *warranted* argumentation frameworks. The conditions above eliminate realizations of a small subclass of graphs with "floating" cycles, but they do

[2] A $\mathbf{JT_{CS}}$ formula can also entail both a rebutting and an undercutting reason for some default reason.

not eliminate the possibility to realize cycles of attacks in general. In fact, most cycles of attacks happen in warranted argumentation frameworks.

In the abstract argumentation (Baroni and Giacomin, 2003) and defeasible reasoning (Pollock, 2001) literature, only the semantics of odd-length cycles of attacks (or of defeats) is notorious for undesirable properties that odd-length cycles entail for different types of extensions. In our default reason theory, both odd- and even-length "floating-attack" cycles have no direct counterparts. This will be explained below in detail.

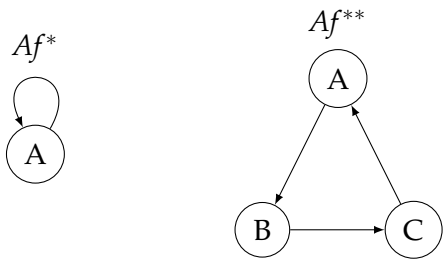

Figure 3.1: Unwarranted argumentation framework examples

Informally, we can say that such unwarranted frameworks violate the following postulate for structured argumentation frameworks:

- Prior to any undercutting attack there must be at least one reasoned claim.

Here by "reasoned claim" we understand a claim that is introduced with the use of a warrant of a default rule that is different from the warrant(s) whereby a cycle of undercutting attacks is introduced. From the perspective of our default theory, the frameworks Af^* and Af^{**} represented in Figure 3.1 are impossible. If precisely assessed, their status of *argumentation* frameworks can be attributed to the possibility in Dung's model to abstract from argument structure.

Once additional argument features are considered, and in particular arguments' warrants, the structures from Figure 3.1 can be proved to be impossible. The following theorem shows that, in our default theories, floating-attack cycles without at least one outgoing edge to an argument outside the cycle are not possible.

Theorem 3.2. *For a sequence $In(\Pi)[k]$ of a default theory $T = (W, D)$ and a set of formulas $\{t_1 : F_1, \ldots, t_n : F_n\} \in In(\Pi)[k]$, a cycle of undercuts among*

the reasons $t_1, \ldots, t_n$ is possible only if (1) there is a reason t_i for a formula F_i, where $1 \leq i \leq n$, such that one of its subterms $p \in Sub(t_i)$ for a warrant $p : (B \rightarrow C) \in \mathcal{WS}^{\Pi[k]}$ is not undercut by any of the reasons from the cycle $t_1, \ldots t_n$ and (2) there is a warrant $r : (D \rightarrow E) \in \mathcal{WS}^{\Pi[k]}$, such that $r \in Sub(t_i)$ and r is undercut by some reason from the cycle $t_1, \ldots t_n$, but none of the other warrants from $\mathcal{WS}^{\Pi[k]}$ is a subformula of E.

Proof. Assume that there is a cycle of undercuts in a set of formulas $In(\Pi)[k]$ among reasons $t_1, \ldots, t_n$, such that each t_i, where $2 \leq i \leq n$, is undercut by t_{i-1} as a reason for F_i and that t_1 is undercut by t_n as a reason for F_1. By Definition 2.13, for each reason t_i and each formula $t_i : F_i$ from a set of formulas $\{t_2 : F_1, \ldots, t_n : F_n\} \in In(\Pi)[k]$, there is a subterm $s \in Sub(t_i)$ such that $t_{i-1} : \neg[s : (G \rightarrow H)]$ and for the formula $t_1 : F_1 \in In(\Pi)[k]$ and a subterm $u \in Sub(t_1)$, it holds that $t_n : \neg[u : (I \rightarrow J)]$. Then assume that each reason term from the set $\{v \mid v \in \bigcup_{j=1}^{n} Sub(t_j)$ and $v : (K \rightarrow L) \in \mathcal{WS}^{\Pi[k]}\}$, is undercut in the cycle of undercuts $t_1, \ldots t_n$. This means that each warrant $v : (K \rightarrow L)$ for $v \in Sub(t_k)$ and $2 \leq k \leq n$ would have to be a proper subformula of a formula $t_{k-1} : F_{k-1}$ from the cycle such that $t_{k-1} : \neg[v : (K \rightarrow L)]$ and, thereby, t_{k-1} undercuts t_k being a reason for formula F_k. Additionally, the warrant $w : (M \rightarrow N)$ for $w \in Sub(t_1)$ would have to be a proper subformula of the formula $t_n : F_n$ such that $t_n : \neg[w : (M \rightarrow N)]$ and, thereby, t_n undercuts t_1 being a reason for formula F_1. But this is not possible since no formula is a proper subformula of itself. Therefore, at least one reason t_k from the cycle of undercuts $t_1, \ldots, t_n$ has to attack a warrant $r : (O \rightarrow P) \in \mathcal{WS}^{\Pi[k]}$, where $r \in Sub(t_k)$ and $1 \leq k \leq n$, such that none of the other warrants from $\mathcal{WS}^{\Pi[k]}$ is a subformula of P. □

The theorem ensures that cycles of asymmetrical attacks among arguments are possible only if there is an outlying argument and this argument is attacked by an argument in the cycle. Although our justification logic cannot realize the subclass of unwarranted frameworks, this result does not exclude circular argumentation from it in general. However, the result does show that there are constraints on interpreting directed graphs as argumentation frameworks and these constraints are due to the inclusion of additional argument features into our system.

In the literature about abstract argumentation frameworks, there are attempts to provide frameworks Af^* and Af^{**} with intuitive interpretations. For example, van Eemeren et al. (2014, p. 630) give the following

sports situation as an informal interpretation of Af^{**}. Imagine that Ajax has recently won matches against Feyenoord. We have a reason to think that Ajax is the best Dutch football club (argument A). But assume that it is also the case that Feyenoord has won recent matches against PSV and that PSV has won recent matches against Ajax. Then we have a reason to think that Feyenoord is the best Dutch club (argument B) and that PSV is the best Dutch club (argument C). The available arguments leave us with no answer to the question which football club is the best.

By fleshing out the content of these arguments in our default theory, it becomes clear that there is more to this example than the cycle of three attacks is able to show. There are two kinds of arguments involved in resolving the conflict among the claims to the status of the best club. First, the fact that Ajax has won recent matches against Feyenoord, provides a reason to claim that Ajax is the best club. Secondly, the same fact provides grounds to question the claim that Feyenoord is the best club. The first argument can be an attacker only as a rebuttal, while the second argument is an undercutter. Analogously, arguments can be provided with reference to Feyenoord and PSV, as we will formalize below.

Example 3.3. *Let $T_1 = (W_1, D_1)$, be the default theory describing the conflict of football clubs. The set of facts is defined by $W_1 = \{t_1 : A_1, t_2 : F_1, t_3 : P_1, t_4 : [\neg(A_2 \wedge F_2) \wedge \neg(A_2 \wedge P_2) \wedge \neg(F_2 \wedge P_2)]\}$. Let A_1, F_1 and P_1 be the propositions "Ajax/Feyenoord/PSV has won recent matches against Feyenoord/PSV/Ajax" and A_2, F_2 and P_2 are the propositions "Ajax/Feyenoord/PSV is the best Dutch football club". Notice that the set of facts contains a formula which corresponds to the background knowledge that only one club can be the best club. Finally, $D_1 = \{\delta_1, \delta_2, \delta_2, \delta_4, \delta_5, \delta_6\}$ is the set of defaults, where*

$$\delta_1 = \frac{t_1 : A_1 :: (u_1 \cdot t_1) : A_2}{(u_1 \cdot t_1) : A_2}, \delta_2 = \frac{t_2 : F_1 :: (u_2 \cdot t_2) : F_2}{(u_2 \cdot t_2) : F_2},$$

$$\delta_3 = \frac{t_3 : P_1 :: (u_3 \cdot t_3) : P_2}{(u_3 \cdot t_3) : P_2}, \delta_4 = \frac{t_1 : A_1 :: (u_4 \cdot t_1) : \neg[u_2 : (F_1 \rightarrow F_2)]}{(u_4 \cdot t_1) : \neg[u_2 : (F_1 \rightarrow F_2)]},$$

$$\delta_5 = \frac{t_2 : F_1 :: (u_5 \cdot t_2) : \neg[u_3 : (P_1 \rightarrow P_2)]}{(u_5 \cdot t_2) : \neg[u_3 : (P_1 \rightarrow P_2)]} \text{ and}$$

$$\delta_6 = \frac{t_3 : P_1 :: (u_6 \cdot t_3) : \neg[u_1 : (A_1 \rightarrow A_2)]}{(u_6 \cdot t_3) : \neg[u_1 : (A_1 \rightarrow A_2)]}.$$

It is easy to check that theory T_4 has a unique $\mathbf{JT_{CS}}$-stable and $\mathbf{JT_{CS}}$-preferred extension $Th^{JT_{CS}}(W_1 \cup \{cons(\delta_4), cons(\delta_5), cons(\delta_6)\})$. Therefore, the conflict

between Dutch football clubs results in accepting that the available reasons do not sanction any of the three Dutch football clubs to claim the title of the best club.

Theory T_4 shows that Af^{**} at best gives an incomplete representation of the conflict of Dutch football clubs. A more faithful abstract argumentation framework should include additional arguments and attack relations as Figure 3.2 shows. The only accepted arguments are the ad-

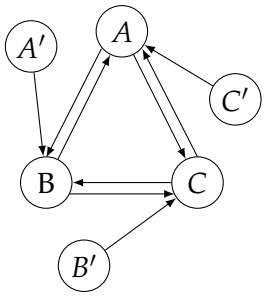

Figure 3.2: Abstract attack structure of Example 3.3

ditional arguments A', B' and C' that are not featured in Af^{**}. These three arguments ensure that none of the unjustified claims to the title of the best Dutch football club goes through. Note how the arguments that are eventually accepted as winning are those indicating the inability of reasons and warrants to justify claims – this layer of argumentation has been so far elusive to a strict logical formalization.

By excluding all the unwarranted Dung's frameworks as defined above, it is possible to formalize the *Realization* procedure ("$\longrightarrow$") of *warranted* Dung's frameworks in justification logic. For a warranted abstract argumentation framework $AF = (Arg, Att)$, there is a default theory $T = (W, D)$ such that:

- $A \in Arg \quad \longrightarrow \quad t : F = cons(\delta)$ and $u : (G \to F) \in \mathcal{WS}^{\Pi}$ such that $\delta \in \Pi$ for a process Π of T
- $(A_m, A_n) \in Att$ & $(A_n, A_m) \in Att \quad \longrightarrow \quad t : P \in Th^{JT_{CS}}(cons(\delta_m))$ and $u : \neg P \in Th^{JT_{CS}}(cons(\delta_n))$ such that P is a fresh propositional variable and, for some processes Π' and Π'' of T, $\delta_m \in \Pi'$ and $\delta_n \in \Pi''$
- $(A_m, A_n) \in Att$ & $(A_n, A_m) \notin Att \quad \longrightarrow \quad \neg[u : (G \to F)] \in Th^{JT_{CS}}(cons(\delta_m))$ and $u : (G \to F) \in \mathcal{WS}^{\Pi}$ for a term $u \in Sub(t)$

such that $t : F = cons(\delta_n)$ and, for a process Π of T, $\delta_m \in \Pi$ and $\delta_n \in \Pi$

The following proposition characterizes realizations of warranted *AF*'s:

Proposition 3.4. *An argument A_n is contained in a complete, grounded, preferred or stable extension of a warranted Dung's framework $AF = (Arg, Att)$ iff a formula $t : F = cons(\delta_n)$ such that $\delta_n \in D$ for a default theory $T = (W, D)$ is contained in the corresponding $\mathbf{JT_{CS}}$-complete, $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred or $\mathbf{JT_{CS}}$-stable extension $Th^{JT_{CS}}(\Gamma)$ for a realization T of AF.*

Proof. The proof is by induction on the acceptance conditions for an argument A given by the definitions of complete, grounded, preferred and stable extension definitions for Dung's abstract argumentation frameworks restricted to the subclass of warranted frameworks.

The proof of the inductive step relies on the fact that the realization procedure $\longrightarrow$ preserves the direction of attacks specified by Dung's attack relation. The direction of argument attacks in our operational semantics is defined exactly as semantics in abstract argumentation, where realized extensions can be instantiated with the corresponding consistent models from $\mathbf{JT_{CS}}$, modulo specifying the logical structure of attacks and closing the realized extensions under the $\mathbf{JT_{CS}}$ consequence relation.

The realization procedure is straightforward for *AF*'s that amount to directed acyclic graphs as well for (most) *AF*'s whose cycles include two-node-cycle components, which translate into rebuttal between formulas. For example, for the existence of a directed path, the realization assigns an undercutter formula $u : \neg[t : (F \rightarrow G)]$ as an argument that realizes the starting node such that any warrant of a subsequent node is a subformula of G. With the presence of other types of cycles, the realization forces the existence of sub-arguments for at least one argument $t : F$ corresponding to a node from a realized cycle. This follows from Theorem 3.2.

□

In Dung's framework arguments are only implicit and one can consider each argument A as a statement of the following type "There is an argument A". When realized in justification logic, each of this existential statements can be instantiated with an explicit argument structure $t : F$.

One may wonder what is the significance of (un)warranted abstract argumentation frameworks for formal argumentation in general. We will

conclude this section by pointing out what could the realization results from justification logic contribute to our understanding of arguments. The most important insight given by the justification logic realization of AFs is that once we include reason terms into our formal language, we bring forward the requirements on the logical language that are only implicit in representing arguments as graph nodes. One of these requirements is that the reasoning structure that we call "backing" in this paper has to be built according to the axioms and rules of the underlying calculus of reason terms. According to it, it is impossible to build a "proof term" or a reason term that would support and undercut one and the same conclusion, which is the result obtained in Theorem 3.2. However, an isolated AF cycle requires a possibility to have a default reason term, without other default reasons as its subterms, that supports and undercuts a conclusion introduced in a single consequent formula of a default rule.

Such loops and cycles that correspond to attack relations in AFs cannot easily be exemplified in natural language either. Even self-defeating argument that require more than a single default inference are difficult to exemplify, as Pollock's pink elephant example witnesses. Starting with the work of Caminada (2005), it has been argued that attack loops could be exemplified with the statement "I am unreliable" or, using the third-person perspective, "An agent says that the agent is unreliable". In the same vein, Caminada (Summer 2008) argues that the above discussed three-node cycle of attacks can be exemplified by a scenario featuring agents who question one another's reliability in the following way.[3] Suppose that there are three agents, namely Bert, Ernie and Elmo. If Bert says that Ernie is unreliable, then everything that Ernie says cannot be relied on. If Ernie says that Elmo is unreliable, then everything that Elmo says cannot be relied on. Finally, If Elmo says that Bert is unreliable, then everything that Bert says cannot be relied on. This creates a cycle of attacks among Bert, Ernie and Elmo.

It is in such borderline examples of arguments that we can value the precision of the justification logic language. Natural language allows the type of self-referentiality featured in the sentence "I am unreliable". With the use of the justification logic language, we can see that such examples belong to a special group of statements that require the logical machinery of propositional quantification or that of quantification into

[3]Similar examples are discussed by Pollock (2009) and Prakken and Horty (2012).

sentential position (Uzquiano, 2020, § 3.5). In an extended language with propositonal quantifiers, we could represent the statement "I am unreliable" with the following formula

$$t : \forall p(\neg t : p),$$

where p is a propositional variable. The Bert-Ernie-Elmo attack cycle would be just an extended version of the loop example. Take

$$t_1 : \forall p(\neg t_2 : p), \qquad t_2 : \forall q(\neg t_3 : q) \qquad \text{and} \quad t_3 : \forall r(\neg t_1 : r)$$

to realize the attack structure of the conflicting testimonies of unreliability among the three agents, where p, q and r are propositional variables.

There are two findings related to the above examples that deserve our attention in the context of modelling arguments. Firstly, if the above examples are to be taken as arguments on a par with arguments that do not require such strong logical machinery, they should not be considered as a part of the *default* reasoning paradigm of argumentation. Such examples of argumentative attack belong to the *plausible* reasoning paradigm. In the default reasoning paradigm, which is the paradigm we investigate in Chapters 2 and 3, argumentative attacks result from attacking defeasible inferences, as illustrated by rebutting and undercutting attacks form Example 2.7. In the plausible reasoning paradigm, argumentative attacks result from adding new information that questions old information and, thereby, it might question old conclusions. Notice that undercutting and rebutting attacks do not question the reliability of old information. For example, concluding that the table is white, rather than red, cannot question the fact that the table looks red under the red lighting. On the other hand, if you question the old information that the table is red looking, then you compromise both old information and any default conclusions that may follow from old information. This type of attack is called "undermining" and it is defined as an attack on premises of an argument (van Eemeren et al., 2014, p. 626). In the Bert-Ernie-Elmo attack cycle, the three sources of information are undermined in such a way that the testimonies of the three agents question one another in the proposed order. This differs from the attacks induced by default inferences, where some default step is being questioned, rather than the credibility of information sources.

Secondly, default justification logic shows that argumentation frameworks that include such arguments on a par with other defeasible arguments do not consider the paradoxical nature of the mentioned example.

In an important sense, ASPIC+ is still too abstract to capture the intensional paradox created by adding propositional quantification in the justification logic representation of attack cycles.[4] Notice that the reason term t in the statement $t : \forall p(\neg t : p)$ justifies that it cannot justify any proposition. To assess if such reason terms could ever be acceptable, we would first need to resolve what it is that t justifies. Following, for example, Prior's explanation (1961) of the intensional version of the Liar paradox, there have to be at least two statements justified by the operator t. The issues that $t : \forall p(\neg t : p)$ raises are fundamental to our understanding of arguments, but they cannot be further developed here. For now, we are able to conclude that AFs, as well as structured argumentation frameworks, are not capable of capturing the paradoxical nature and the exact logical structure of the examples discussed above. Structured argumentation frameworks may be more expressive in the sense that they allow such arguments with their definitions of arguments. However, their expressivity rests on the fact that they do not specify a logical system expressive enough neither to logically represent reasons nor to logically represent arguments. Once we have a precise language with reason terms, we are able to talk about the issues of whether to include paradoxical propositions of the type $t : \forall p(\neg t : p)$ on a par with other arguments or not. More importantly, we are in a position to discuss what is required to logically represent notorious cycles of attacks in formal argumentation.

3.3 Rationality postulates for structured argumentation

Chapter 2 shows that default rules with justification formulas are expressive enough to model elements of arguments that are traditionally seen as extra-logical, such as warrants and backings. The results from this chapter, Section 3.2, establish the logic of default justifications as a system that explicitly features the structure of arguments and uses Dung's methods for argument evaluation. The $\mathbf{JT}_{\mathbf{CS}}$ variants of admissible, complete, grounded, preferred and stable extensions preserve reasonable outputs of the corresponding Dung's extensions. An additional question that may be asked is whether our logic also behaves reasonably with respect to "ratio-

[4] See (Priest, 1991) for a discussion about intensional paradoxes. Intensional paradoxes belong to a "class of paradoxes of self-reference whose members involve intensional notions such as *knowing that*, *saying that*, etc." (Priest, 1991, p. 193).

nality postulates" that are set for structured argumentation frameworks in the literature (Amgoud, 2014, Caminada and Amgoud, 2007).

According to Amgoud (2014), the exact formulation of rationality postulates for structured argumentation frameworks depends on the family of a logical language that they use: rule-based or classical. In frameworks with rule-based languages, a distinction is made between strict rules (rules without exceptions) and defeasible rules (rules that may have exceptions). Arguments are built according to the available strict and defeasible rules. Examples of such systems are ASPIC+ (Prakken, 2010) and DeLP (García and Simari, 2004). In frameworks with classical languages, arguments are built from a knowledge base using an underlying monotonic logic. Examples of frameworks that use classical languages are (Besnard and Hunter, 2001) and (Besnard and Hunter, 2005). The framework described in (Besnard and Hunter, 2001) is based on a propositional language, while that of (Besnard and Hunter, 2005) is based on a first-order language.

Following Amgoud (2014), we will consider five postulates, originally formulated for argumentation frameworks built on classical languages. In general, classical logic-based argumentation frameworks start from the idea that there is some knowledge base with classical logic formulas. We define arguments from that (possibly inconsistent) knowledge base as pairs of sets of formulas and conclusion formulas such that a conclusion formula is classically entailed by a set of formulas. We will here present the five postulates without committing to Amgoud's definition of an argument. We do so deliberately, because the definition of an argument for *classical-logic based* frameworks cannot be applied to our logic. The reasons will be given shortly after we present the postulates. We give their "framework-neutral" formulation, leaving the exact definitions of framework extensions, arguments, sub-arguments, strict rules, premises and conclusions unspecified:

Closure The set of conclusions for each extension is closed under strict rules.

Sub-arguments If an argument is contained in an extension, then all the sub-arguments of the argument are contained in the extension.

Consistency The set of conclusions for each extension is consistent.

Exhaustiveness If each premise and the conclusions of an argument are

conclusions of an extension, then the argument is contained in the extension.

Free precedence If an argument is not involved in any conflict, then the argument contained in each extension.

Although we mentioned that the five postulates provide criteria to evaluate classical logic-based argumentation frameworks, they are also relevant for rule-based argumentation frameworks. In fact, their rule-based framework variants can be found in Amgoud and Besnard (2013).

Delimiting the notion of *argument* in default justification logic

To discuss whether rationality postulates hold for a system, it is required to have a precise definition of an argument. Note that default justification logic offers both a narrower and more broader understandings of an argument, which may include implicit components. The narrower understanding simply takes every formula of the type $t : F$ to be a structured argument such that t represents premises of an argument and F represents its conclusion. However, as Figure 2.3 shows, t codifies a more complex structure that involves implicit features of an argument, such as an argument's warrant and its backing. This offers a broader perspective whereby an argument can be rather seen as an argument schema, which is inclusive of its implicit elements. For the discussion on rationality postulates Closure, Consistency and Free-precedence, it suffices to focus on arguments' explicit features in the sense of the narrower understanding. To discuss Sub-arguments and Exhaustiveness, we will use additional elements from the broader understanding of arguments.

Although the idea of a classical logic-based system is closer to our default logic, the postulates given by Amgoud (2014) are not directly applicable to our logic. While the logic of default justifications uses $\mathbf{JT_{CS}}$ consequence to build arguments, it also allows for defeasible rules by means of extending the application operation $\cdot$ for default rules. It is not the case that all arguments are built from a knowledge base using only the monotonic consequence for the underlying language, as required in (Amgoud, 2014, p. 2030). Default reasons are built using warrants of the type $u : (F \to G)$ and warrants are functioning as defeasible rules, but they are not initially known and they do not need to become a part of the knowledge base, although they potentially could. Finally, and most importantly, arguments in the narrower sense are featured in the $\mathbf{JT_{CS}}$

language itself, which means that a pair $(\textit{premises}, \textit{conclusion})$ is also an object-level formula, unlike, for example, in (Besnard and Hunter, 2001).

On the other hand, rule-based languages introduce the differentiation between strict and defeasible rules, but these rules are not a part of the base language. In contrast with, for example, (Prakken, 2010), arguments in justification logic are built via the operations in the $\mathbf{JT_{CS}}$ language, where the strict rules are simply the rules of the logic $\mathbf{JT_{CS}}$ and defeasible rules are in the object language due to the fact that both warrants are a part of the $\mathbf{JT_{CS}}$ and the operation $\cdot$ is a part of the language. This makes any argument logically dependent on other strict and defeasible conclusions within the system. For example, since warrants are formulas of the $\mathbf{JT_{CS}}$ language, a consequent of a default rule may refer to the underlying warrant of another default rule in the way of an undercut attack.

Our logic takes the middle way between rule-based systems and classical logic-based systems by combining the distinction between strict rules and defeasible rules with logical dependency of arguments via $\mathbf{JT_{CS}}$ consequence. This middle way is epitomized by the two roles that warrants have in the system: they function as both implicit rules as well as statements.[5] Warrants in the role of rules enable default conclusions and warrants in the role of logical statements enable other formulas to refer to warrants within the logical system. However, this also means that our logic cannot be aligned with only one of the two families of logic-based argumentation systems identified in Amgoud (2014).

Postulates for default justification logic

Even without directly applying the postulates for classical logic-based argumentation, we can check whether the desiderata on which Amgoud (2014) builds the rationality postulates hold for our logic. We first examine three postulates from Amgoud (2014, pp. 2032-2035) that are easily adaptable for our logic. For any $\mathbf{JT_{CS}}$-complete, $\mathbf{JT_{CS}}$-grounded, $\mathbf{JT_{CS}}$-preferred or $\mathbf{JT_{CS}}$-stable extension Γ of a default theory $T = (W, D)$, the following postulates are required to hold:

$\mathbf{JT_{CS}}$ closure The set of conclusions for each $\mathbf{JT_{CS}}$ extension Γ is closed under strict rules;

[5] Note that this corresponds to Toulmin's ambiguous use of the term "warrant". For example, Toulmin (1958/2003, p. 91) refers to warrants as both rules and statements in a single paragraph.

$\mathbf{JT_{CS}}$ consistency The set of conclusions for each $\mathbf{JT_{CS}}$ extension Γ is $\mathbf{JT_{CS}}$-consistent;

$\mathbf{JT_{CS}}$ free precedence If some argument $t : F$ is not involved in any conflict, then $t : F \in \Gamma$ for each $\mathbf{JT_{CS}}$ extension Γ.

In our logic, strict rules are simply the rules of $\mathbf{JT_{CS}}$ logic. By Definitions 2.19, 2.22, 2.26, 2.27, extensions are closed under $\mathbf{JT_{CS}}$ consequence and, therefore, closed under strict rules.

The satisfaction of the consistency postulate is guaranteed for each default theory $T = (W, D)$ with a consistent set of facts W. For such default theories, it can be easily shown that $\mathbf{JT_{CS}}$ consistency of each extension is preserved by the conditions of application for each default rule. This follows from Theorem 2.16 and the fact that each $\mathbf{JT_{CS}}$ extension is conflict-free. Exceptions to the consistency postulate are theories with an inconsistent set of facts W. This reflects the way in which our logic deals with inconsistent information. Firstly, an agent starts with known facts represented by justified formulas that do not conflict with one another. Conflicts arise only after an agent needs to extend an incomplete knowledge base by default assumptions. Resolving such meaningful conflicts always leads to $\mathbf{JT_{CS}}$-consistent extensions.

The free precedence postulate requires that the system infers all the arguments and, in general, formulas that do not conflict with any other argument. As stated above, we take arguments in the narrower sense of formulas $t : F$ and these arguments may be based either on strict or on defeasible rules. This postulate follows trivially for all $\mathbf{JT_{CS}}$ extensions, except for $\mathbf{JT_{CS}}$-admissible extensions that do not maximize inclusion of arguments by their definition. Notice that for other $\mathbf{JT_{CS}}$ extensions, no formula $t : F = cons(\delta)$ is excluded from a $\mathbf{JT_{CS}}$ extension Γ, unless δ is inapplicable to the respective process containing Γ or one of the subterms of t is undercut by Γ. The inclusion of all free formulas and arguments built on conflict-free grounds is then ensured by the closure under $\mathbf{JT_{CS}}$ consequence.

For the two additional postulates from (Amgoud, 2014), the notion of a sub-arguments of an argument needs to be defined. We will start again from the narrower understanding of an argument in the sense of any formula $t : F$. The concept of a *sub-argument* for default application will be taken to mean the following:

- If a formula $(u \cdot t) : G$ is obtained by means of application (axiom

A1) or default application from the formulas $t : F$ and $u : (F \to G)$, then $t : F$ and $u : (F \to G)$ are *sub-arguments* of $(u \cdot t) : G$;

- If a formula $(t + u) : F$ is obtained by means of sum (axiom A2) from either the formula $t : F$ or the formula $u : F$, then at least one of the formulas $t : F$ and $u : F$ is a *sub-argument* of $(t + u) : F$.

If an argument $t : F$ is a sub-argument of $(u \cdot t) : G$ and a sub-argument of $(t + u) : F$, then any sub-argument of $t : F$ is also a sub-argument of $(u \cdot t) : G$ and $(t + u) : F$. Notice that if an argument $t : F$ is a sub-argument of $(t + u) : F$, it is not necessary that there is some formula $u : G$ which is also a sub-argument of $(t + u) : F$. It is possible that some justification term u does not justify any formula G. For a trivial example, take some justification constant c and any formula F, $(c \cdot c)$ is not a justification for F in $\mathbf{JT_{CS}}$ logic because the application operation that gives $(c \cdot c)$ is not meaningful for an injective constant specification $\mathcal{CS}$.

The following two postulates require rational acceptance of an argument with respect to its substructure:

$\mathbf{JT_{CS}}$ sub-arguments If an argument $t : F$ is in a $\mathbf{JT_{CS}}$ extension Γ, then any sub-argument of $t : F$ is also in Γ;

$\mathbf{JT_{CS}}$ exhaustiveness If each sub-argument and the formula F for some argument $t : F$ are conclusion of $\mathbf{JT_{CS}}$ extension Γ, then $t : F$ is in Γ.

In contrast to the Exhaustiveness postulate on page 81, notice that the $\mathbf{JT_{CS}}$ variant of exhaustiveness does not mention premises of an argument as conclusions of $\mathbf{JT_{CS}}$ extensions. On the narrower understanding of the arguments as justification assertions, the premises of an argument are reason terms, but not well-formed formulas. The postulate is reinterpreted to track conclusion formulas that are sub-arguments for an argument, because reason terms codify reasoning steps from those formulas such as, for example, the warrants of arguments. The $\mathbf{JT_{CS}}$ exhaustiveness postulate obviously holds for all $\mathbf{JT_{CS}}$ extensions closed under $\mathbf{JT_{CS}}$ consequence by axioms A1 and A2. Thus, informally, if the steps of an argument are contained in an extension, then the argument itself is.

The sub-arguments postulate can be seen as a dual version of exhaustiveness, in the sense that it requires that all the steps of an accepted

argument should also be accepted (Amgoud, 2014, p. 2029). This postulate is not directly satisfied by our logic. Take, for example, an argument $(u \cdot t) : G$ obtained by default application. According to default application, one of the sub-arguments of $(u \cdot t) : G$ is some formula $u : (F \to G)$ which is neither a part of a knowledge base W for a default theory T nor is it required for that formula to become a part of an extended knowledge base, which results from applying the available defaults.

Does that mean that arguments introduced by default rules are based on unjustified reasoning steps? We can show that this is not the case. Although the sub-arguments postulate is not directly satisfied, the basic idea behind the postulate is: "an argument cannot be accepted if at least one of its sub-parts are bad" (Amgoud, 2014, p. 2033). This desideratum holds because, even if the sub-argument $u : (F \to G)$ of an argument $(u \cdot t) : G$ does not become a part of a knowledge base, the system ensures that the warrant $u : (F \to G)$ has not been compromised by other available arguments in the knowledge base. For any argument $(u \cdot t) : G$ and its warrant $u : (F \to G)$, if $(u \cdot t) : G$ is in a $\mathbf{JT_{CS}}$ extension, then that extension contains the formula $((c \cdot t) \cdot (u \cdot t)) : (F \to G)$, assuming that the constant c justifies the axiom $F \to (G \to (F \to G))$ and that the sub-argument $t : F$ of $(u \cdot t) : G$ is also contained in the extension. Therefore, it is possible to ascertain that none of the steps in building the argument $(u \cdot t) : G$ has turned out to be bad, if the argument $(u \cdot t) : G$ is actually accepted in a $\mathbf{JT_{CS}}$ extension.

3.4 Undercutting in justification logic and Reiter's logic

In this section, we compare default theories based on justification logic to standard default theories based on first-order logic that were first defined by Reiter (1980). Our primary goal is not to address the question of the possibility to establish correspondence results between the two. Instead, we focus on their conceptual differences in modeling default reasoning. We start by showing how to represent exclusionary reasons and undercutting defeat in process trees for justification logic-based default theories. Then we show how to translate undercut into Reiter's default logic by interpreting its default processes as arguments. We argue that our logic conforms better to the idea of making a default inference without having to anticipate numerous exceptions to the inference.

Reiter's default logic (Reiter, 1980, Antoniou, 1997) is one of the most notable logics for non-monotonic reasoning with rules that enable "jumping" to conclusions. However, the question of what precisely is a default reason is left ambiguous in Reiter's logic. This problem is identified by Horty in the following passage about the *reification* of reasons (Horty, 2007, p. 6):

> Suppose, as in our example, that the agent's background theory contains the default $B \rightarrow F$, an instance for Tweety of the general default that birds fly, together with B, the proposition that Tweety is a bird, so that the default is triggered. In this case, it seems plain that the agent has a reason to conclude that Tweety flies. But how, exactly, should this reason be reified? Should it be identified with the default $B \rightarrow F$ itself, or with the proposition B?

Horty's conclusion is that this question "like many questions concerning reification, is somewhat artificial" and that "when it comes to reification, the reason relation could be projected in either direction, toward defaults or propositions, and the choice is largely arbitrary".

The goal of this section is to show that the question of reification is important and that giving an answer to it opens up new paths in formalizing defeasible reasoning. In particular, our focus is on showing benefits of formalizing default reasons with the language of justification logic, which is expressive enough to encode the structure of default inferences within its reason terms. To compare the two logics via the problem of reification, we illustrate the advantages of the expressiveness that our logic has in comparison to Reiter's default logic by means of an example with undercut.

Consider an agent reasoning about whether a KLM Boeing 737 aircraft has cleared the take-off protocol or not, given that a source of information says that "the crosswind component at the default runway is at the speed of 35 knots" (C). Knowing that, at this speed of the crosswind component, the Boeing 737 type of aircraft is usually not allowed to proceed with the take-off, the agent concludes that "the KLM Boeing 737 flight has been delayed" (K), according to the following default rule:

$$\delta_1 = \frac{r : C :: (s \cdot r) : K}{(s \cdot r) : K}.$$

The default can be read as follows: "If r is a reason justifying that the

crosswind component at the default runway is at 35 knots and it is consistent to assume that $(s \cdot r)$ is a reason justifying that the KLM Boeing 737 flight has been delayed, then $(s \cdot r)$ is a defeasible reason justifying that the KLM Boeing 737 flight has been delayed".

If the agent receives additional information that it is not the case that "the SAS Boeing 737 aircraft has been delayed" (S), then the agent has a reason to assume that "the aircraft can be allocated to an alternative runway" (R).

$$\delta_2 = \frac{t : \neg S :: (u \cdot t) : R}{(u \cdot t) : R}.$$

On a runway of a different orientation, the initial readings of the crosswind may even turn into a favorable headwind component. The information that there is an alternative runway undercuts the initial piece of reasoning codified by s, according to the following default rule:

$$\delta_3 = \frac{(u \cdot t) : R :: (v \cdot (u \cdot t)) : \neg[s : (C \rightarrow K)]}{(v \cdot (u \cdot t)) : \neg[s : (C \rightarrow K)]}.$$

The consequent reads as follows: "$(v \cdot (u \cdot t))$ is a defeasible reason denying that the reason s justifies that if the crosswind component for the default runway is at the speed of 35 knots, then the KLM Boeing 737 flight has been delayed". Additionally, the agent has a reason to conclude that the KLM flight has not been delayed, grounded on the reasoning about an alternative runway:

$$\delta_4 = \frac{(u \cdot t) : R :: (w \cdot (u \cdot t)) : \neg K}{(w \cdot (u \cdot t)) : \neg K}.$$

Were it the case that the course of the agent's reasoning follows the proposed order, the agent would have to revise the conclusion supported by the reason $(s \cdot r)$. For a default theory $T_1 = (W, D)$ with $W = \{r : C, t : \neg S\}$ and $D = \{\delta_1, \delta_2, \delta_3, \delta_4\}$, the process $(\delta_1, \delta_2, \delta_3)$ corresponds to such course of reasoning with a revised $\mathbf{JT_{CS}}$-admissible extension. Figure 3.3 shows all the possible processes of T.

Is there a way to model undercut in Reiter's default logic, without extending the logic with, say, default priorities, as done by Horty (2012) and Brewka (1994)? To answer this, we need to view default logic from the perspective of formal argumentation. The relation between formal argumentation and Reiter's default logic is known. Dung (1995) shows that Reiter's default logic extensions can be defined in terms of stable

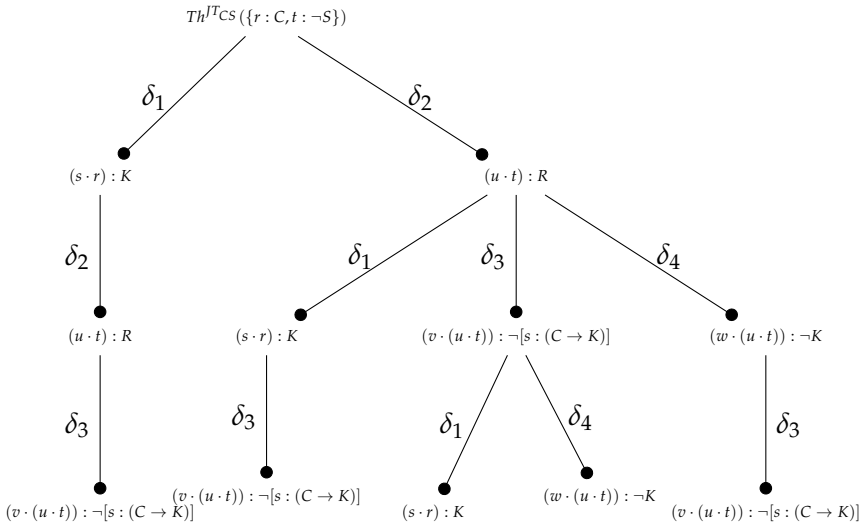

Figure 3.3: Each node of the process tree of T is labeled with an *In*-set after a default rule (edges) has been applied. Visually, we display only the formulas that are added to *In*-sets as a result of applying the available defaults.

extensions of abstract argumentation frameworks and this chapter shows that a large subclass of abstract argumentation frameworks is a special case of our logic. But besides finding formal correspondences between their extensions, it is interesting to look at the conceptual relation of Reiter's logic to our justification logic. This relation is not straightforward, because the two logics are based on different underlying languages. This difference does not cause divergence in the way the two logics model rebuttal. Rebuttal is based on the workings of multiple incompatible extensions: two formulas extending some knowledge base rebut each other if they cannot both be included in a same default extension. However, the comparison of the ways in which the two logics deal with the concept of undercut reveals some immediate benefits of reifying default reasons in justification logic.

We recall here some basic definitions of Reiter's default logic from Chapter 1. First, the general form of a default rule in Reiter's logic is

$$\frac{\varphi : \psi_1, \ldots, \psi_m}{\chi},$$

for some predicate logic formulas $\varphi, \psi_1, \ldots, \psi_m$ and χ. The operational semantics for default logic is similar to ours, but besides closure, Antoniou (1997) introduces an additional condition on the extension-producing processes of Reiter's default theories, namely *success*. This concept turns out to be fundamental for us to model exclusionary reasons. Recall that a process is successful if each of the justifications $\psi_1, \ldots, \psi_m$ is consistent with the consequents added to an *In*-set after all the other applicable defaults have been applied. To capture this formally, we use the set $Out(\Pi) = \{\neg\psi \mid \psi \in just(\delta) \text{ for some } \delta \in \Pi\}$, for some justification $just(\delta)$ of a rule δ in a process Π of a default theory $\Delta = (W, D)$. None of the formulas from an *Out*-set should become a part of an *In*-set for the same process. The notions of closure and success give a formal characterization of extensions: given a set of first-order formulas $E = In(\Pi)$, E is an extension in Reiter's logic if and only if Π is both closed and successful.

In our logic, it is possible to consider consequents of defaults as arguments based on their underlying warrants. This enables us to represent conflicts simply by opposing reasons. In Reiter's logic, reasons are not reified and their conflicts cannot be reflected in the logical language. It is, however, possible to take the perspective of formal argumentation on Reiter's logic. To take such perspective, we follow Prakken (2018, p. 52) in defining arguments in terms of finite processes of Reiter's theory and their mutual attacks through conflicts of *In*-sets with *Out*-sets. We start from defining attacks in terms of finite processes of a theory $\Delta = (W, D)$:

- Π attacks Π' if $\varphi \in In(\Pi)$ for some $\varphi \in Out(\Pi')$,

where Π and Π' are some finite processes of Δ. We can develop further on this definition to specify different kinds of attack:

- If all the default rules from Π and Π' could possibly form a finite process Π'' of Δ (in any possible order of the sequence), then the attack between Π and Π' is undercut. Otherwise, it is a rebuttal between Π and Π'.

The idea behind the refinement of the attack definition is that in Reiter's logic, non-normal default rules can be seen as a way to introduce "exclusionary reasons" (Horty, 2012) and undercut in Reiter's theory. Consider the following two Reiter's default rules:

$$\delta' = \frac{35Knots(crosswind) : \neg alternative(runway)}{delayed(KLM)},$$

saying that "if the crosswind component at the default runway is at 35 knots, and it is consistent to assume that the aircraft cannot be allocated to an alternative runway, then the KLM Boeing 737 flight has been delayed" and

$$\delta'' = \frac{35Knots(crosswind) \wedge \neg delayed(SAS) : alternative(runway)}{alternative(runway)},$$

saying that "if the crosswind component at the default runway is at 35 knots and the SAS Boeing 737 flight has not been delayed, and it is consistent to assume that the flight can be allocated to an alternative runway, then the flight can be allocated to an alternative runway". Moreover, the following default is available to the agent:

$$\delta''' = \frac{alternative(runway) : \neg delayed(KLM)}{\neg delayed(KLM)}.$$

Take $\Delta = (W, D)$ to be a Reiter's default theory with $W = \{35Knots(crosswind), \neg delayed(SAS)\}$ and $D = \{\delta', \delta'', \delta'''\}$. The process tree for the Reiter's theory Δ is found in Figure 3.4. The idea of undercut can be illustrated by the way in which $\Pi' = (\delta'')$ attacks $\Pi = (\delta')$ via affirming the circumstance in which the conclusion of Π would not be reasonable any more, but it does not affirm the negation of that conclusion. This corresponds to the idea of undercut. Notice that, in contrast to δ_1 above, Reiter's rule δ' needs to include the negation of the exclusionary circumstance $alternative(runway)$ in conveying the idea of undercut.

Notice that the mechanism whereby Π' attacks Π is also responsible for the problem of process "destruction" (Antoniou, 1997, p. 63). An example of process destruction occurs in the unsuccessful process $\Pi'' = (\delta', \delta'')$, where, after δ'' has been applied, Π'' becomes closed and unsuccessful. As it can be seen from the process tree of T, justification logic-based default theories are able to model interaction among default rules and undercut without having to resort to the use of process destruction. This is mainly due to the fact that the rule δ_3, which is "missing" in the Reiter's logic rendition of the example, brings forth the undercut of the default reason from δ_1 as a part of the example description.

Several remarks are at hand by comparing the structures of the process trees of T and Δ. Although the theories model the same phenomenon, Reiter's logic noticeably simplifies the example. While some ways of simplifying are desirable, there are several reasons to prefer

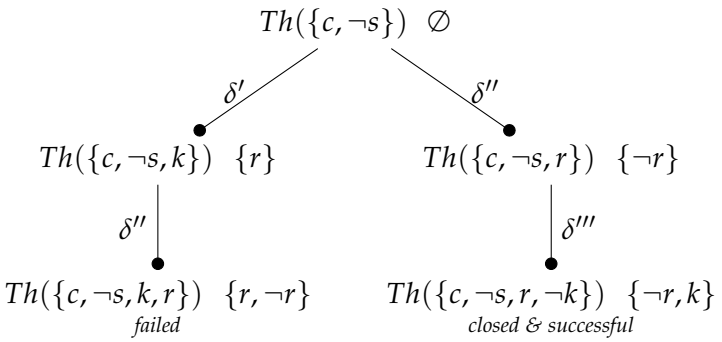

Figure 3.4: The process tree of Reiter's default theory Δ. Undercut is modelled by the failed process branch.

the representation of undercut in justification logic. One of them is that failed processes disable reinstatement of reasons that could, in principle, be reinstated by undercutting their undercutters. In our example above, even if the SAS flight has not been delayed, it might be the case that the current demand for the runway reassignment exceeds the operational capacities of the alternative runway. This, in turn, provides a reason to reinstate the initial reason in support of the claim that the KLM flight is delayed. Justification logic is able to represent such reason reinstatement and processes that are, in principle, infinitely extendable.

As can be seen from the definitions of $\mathbf{JT_{CS}}$ extensions, only $\mathbf{JT_{CS}}$-admissible extensions can be identified by looking at a single branch of the process tree of T. To determine the status of other $\mathbf{JT_{CS}}$ extensions, all reasons need to be taken into account.[6] In contrast, Reiter's default processes are self-contained with respect to the extensions status. Therefore, $\mathbf{JT_{CS}}$ extensions have more in common with the notion of Reiter's logic consequence relation (*credulous* and *skeptical*). For instance, $\mathbf{JT_{CS}}$-grounded extensions correspond to the skeptical notion of validity, which amounts to the intersection of all Reiter extensions. It is only in taking the argumentation perspective on Reiter processes that we need to look at the dependencies of different process tree branches.

Conceptually, the most important advantage of representing defeasible reasoning in justification logic is that neither for reaching a defeasi-

[6]This could be further amended by considering "anytime reasoning" methods as, for example, those proposed by Cadoli and Schaerf (1994) for Reiter's logic.

ble conclusion nor for undercutting that reason, agents do not need to anticipate exclusionary reasons. In an important sense, anticipating exclusionary reasons with $\neg alternative(runway)$ in the Reiter rule δ' above goes against the idea of default reasoning. Namely, in the sense that the number of these conditions may be infinite. The need to anticipate exclusionary reasons brings us back to the initial problem: we want to find out how to *avoid* anticipating numerous exceptions before an agent is able to reach a conclusion. It is an advantage of our theory to be able to model exceptions to rules via undercut, but without the drawback of guessing all the conditions of undercut within a default theory.

Finally, notice that by using Reiter's non-normal defaults, we are not only able to define undercut, but also to define theories such as $\Delta^* = (W^*, D^*)$, where $W^* = \emptyset$ and $D^* = \{\delta'''' = \frac{\top : \neg A}{A}\}$. The rule $\frac{\top : \neg A}{A}$ invalidates its own applicability. Using the above defined translation to argument frameworks, it is possible to build a single-argument attack cycle in terms of the argument $\Pi = (\delta'''')$. This kind of attack is well-known from Dung's (1995) abstract argumentation frameworks. In this chapter, we prove that defining single-argument attack cycles is not possible once the warrants of arguments have been included as underlying rules for each default. This shows that although δ'''' sanctions "jumping" to the conclusion A, this does not mean that the type of inference it instantiates counts a reasoned default step.

3.5 Conclusions

In this chapter, we explored relations that default justification logic has to formal argumentation frameworks and standard default logic. While we provided correspondence results between our logic and Dung's AFs, we did not show correspondence results on how the logic relates to structured argumentation frameworks or Reiter's default logic. Instead of focusing on formal relations to structured argumentation frameworks, we instead showed that our logic meets all of the standard postulates for structured argumentation. To show relations to Reiters's default logic, we chose to focus on their conceptual differences regarding the problem of reification. It is, however, reasonable to assume that there are correspondence-like results since the relation to Dung's argumentation frameworks has been established in this chapter and the relation between Reiter's default logic and Dung's argumentation frameworks are known

from (Dung, 1995, Section 4.1).

The results from this chapter show that the logic of default justification has a similar connection to abstract argumentation frameworks as standard justification logic systems have to their modal logic counterparts. Artemov (2001) provided a proof of the *Realization Theorem* that connects the logic of arithmetic proofs LP with the modal logic S4. The result has been followed up by similar theorems for many other modal logics with known "explicit" justification counterparts.[7] In our paper we show that our logic can be considered as an explicit justification logic counterpart to a substantial subclass of abstract argumentation frameworks called warranted frameworks.

Several interesting paths could be followed in further connecting the logic of default justifications with formal argumentation frameworks. Among frameworks with abstract arguments, the AFRA framework (Baroni et al., 2011) with recursive attacks offers a possibility of representing attacks to attacks. This conceptual advance is useful in connecting default reasons to abstract arguments. More obviously, our logic is closely related to the frameworks with structured arguments, which is why connections with systems such as ASPIC+ (Prakken, 2010), DeLP (García and Simari, 2004), SG (Hecham et al., 2018) and the logic-based argumentation framework by (Besnard and Hunter, 2001) are interesting to explore. Since each of these frameworks elaborates on the notion of defeat, a thorough comparison to our logic would shed light on their formal connections. Caminada and Gabbay (2009) and Grossi (2010) give a different logic-based perspective on argumentation frameworks. Both papers start from the idea of studying attack graphs and formalizing notions of extensions from abstract argumentation theory using modal logic, with the former approach being proof-theoretical and the latter model-theoretical. A further interesting research venue in the field of argumentation theory is Verheij's (2003) logical interpretation of *prima facie* justified assumptions. The DefLog system which is developed there is closely related to ours in motivation, but it develops from a perspective of a sentence-based theory of defeasible reasoning instead of a rule-based or argument-based approach.

Some existing extensions of default theories that can deal with the problem of exclusionary reasons come close to our intention of reifying default reasons. Most notably, approaches that are based on reasoning

[7] See (Fitting, 2016) for a good overview of realization theorems.

about default rule priorities such as (Brewka, 1994) and (Horty, 2007, 2012) include a variant of reasoning about other reasons, namely, by reasoning about the relative weights of defaults. Strictly speaking, reasoning about default priorities reifies default rules, not default reasons, by extending the underlying language of default logic with default names and a predicate symbol that represents priorities among defaults. In such default theories, agents may arrive to conclusions about which ordering of defaults is a preferred one and to, thereby, consider higher priority as a source of defeat. Priority weighing in the style of Brewka (1994) can be represented in process trees, as done in (Antoniou, 1997, pp. 97-98). It can be noticed that one of the difficulties with such reasoning is that before applying a default, an agent needs to consider all other applicable defaults. As in Reiter's logic, processes are possibly failed, where failure is now due to making application choices that are inconsistent with a valid ordering among defaults. Horty (2012, p. 124) defines "exclusionary default theories" where he explicitly includes undercutters, but his undercut is logically only a predicate saying that a rule has been excluded.

A more elaborate study of the different ways to defeat reasons is carried out in argumentation theory, from the classical account of undercut and rebuttal in (Pollock, 1987) to some later formal argumentation frameworks such as, e.g., (Prakken, 2010), (Besnard and Hunter, 2001) and (Verheij, 2016). These frameworks are not based on default logic nor do they base their formalism on a language with formulas that feature reason terms. Hence, as for our current discussion on the problem of reification, such systems do not provide explicit answers. Among justification logic systems, some of them Baltag et al. (2014), Renne (2012) combine belief revision and dynamic epistemic logic techniques to model defeat, which is closest in its kind to undermining. However, none of them is able to model undercut or to encode defeasibility in the structure of reason terms.

Finally, our answer to the problem of reification is that both a prerequisite of a default rule and the rule itself are involved in reifying default reasons. This is reflected in the way in which default application codifies default steps from warrants and prerequisites of defaults to their consequents. An immediate advantage of reification is that, by referring to such reason-producing steps within the object language, we can provide a fine-grained logical account of defeat.

Chapter 4

Argumentation dynamics: Undermining in default justification logic

4.1 Introduction

This chapter uses belief revision methods to study information changes in default justification logic with argumentation semantics. The default justification logic introduced in Chapter 2 models non-monotonic behavior that results from extending incomplete information, but it does not deal with the consequences of information changes. We want to make our system to be adaptive to such changes. To add this new component, we introduce dynamic operators that combine tools from belief revision and default logic to define both prioritized and non-prioritized operations of contraction, expansion and revision for justification logic-based default theories. This combination enriches both default logics and belief revision techniques. We argue that the argumentative attack called "undermining" amounts to those operations that contract a knowledge base by an attacked formula.

4.2 Dynamics in formal argumentation

In this part of the thesis, we investigate the dynamics of default theories with justification logic formulas. The basic format of our default theo-

ries has been presented in Chapter 2, where we interpret justification formulas of the form $t : F$ as arguments that can defeat other formulas by means of undercut or rebuttal. Technically, the workings of undercut and rebuttal rely on defining default theories with default rules based on justification logic formulas. In such rules, justification terms codify defeasible inferences in their structure. In this chapter, we will make a further step and add reasoning about changes to default theories.

The existing work about dynamics in formal argumentation (Booth et al., 2013, Coste-Marquis et al., 2014, Doutre et al., 2014, Diller et al., 2015, de Saint-Cyr et al., 2016) almost entirely focuses on abstract argumentation frameworks in the style of Dung (1995). The literature on the dynamics of structured argumentation is limited to the DeLP framework (Alfano et al., 2018), where the dynamics is understood as adding or removing strict and defeasible rules, and ASPIC+ (Modgil and Prakken, 2012), where the dynamic component is meant to resolve symmetric attacks by updating preferences. The current chapter, based on (Pandžić, 2020), advances this line of research by specifying a variety of dynamic operators for modeling changes of argument systems based on justification logic.

We will show that introducing dynamic operators for justification logic default theories enables us to model an additional kind of defeat: *undermining*. According to van Eemeren et al. (2014), an argument is undermined if its premises or assumptions are attacked. Defeating an argument by attacking its premise or its assumption is not new to structured argumentation. In assumption-based argumentation (ABA) (Dung et al., 2009), all attacks are reduced to this type and in ASPIC+ (Prakken, 2010), ordinary premises of an argument are susceptible to undermining. However, these systems do not provide an insight into the logical workings of undermining, because they specify neither a concrete logical language nor inference rules.[1]

In our default theories, undermining can be given a precise logical interpretation. While undercut and rebuttal rely on the uncertainty of default arguments, undermining changes the context from which agents make further inferences. For a specific default theory, this con-

[1] In fact, ABA does not distinguish between different kinds of attacks and models each attack as an attack on premises and thus reduces all attacks to the type of attack that we call here undermining. In ASPIC+, undermining is taken as a primitive notion of attack, which is different from rebuttal or undercut only by virtue of targeting "ordinary" premises of an argument.

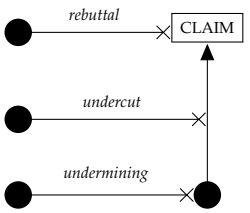

Figure 4.1: Three types of defeat between arguments

text is determined by the set of starting premises, W. Our idea is that, since undermining targets the given premises, it should be modeled as a result of non-inferential information inputs that require contracting the set of premises of a default theory. This means that we will define undermining by "climbing up" the definitions of more fundamental operations of default theory changes. To elicit the reasoning process behind undermining, we specify four different logical operations that model undermining: prioritized and non-prioritized contraction and prioritized and non-prioritized revision. Figure 4.1 illustrates the differences between undercutting, rebutting and undermining attacks.

The chapter is organized as follows. The logic introduced in Chapter 2 already models non-monotonicity with the use of undercut and rebuttal. We use this logic to exemplify Toulmin's (1958/2003) argumentation model in our default theories. The example is used to illustrate conceptual differences in the workings of undercutting, rebutting and undermining. After specifying the elements of arguments in default justification logic, we describe how we plan to connect dynamic operations for such default theories with undermining defeaters. Section 4.4 is the main technical contribution of this chapter, where we define dynamic operations for default theories with justification formulas. The operations we introduce combine Hansson's (1999a) base revision operations with a specific kind of standard Reiter default rules. Our approach to defining the dynamic operators for default theory revision has most in common with Antoniou's (2002) approach, which deals with the dynamics of Reiter's default theories. We show by the end of the chapter that undermining attacks on premises correspond to those dynamic operations that involve either contraction or a variant of non-prioritized contraction defined in Section 4.4.

4.3 Formalizing Toulmin's example

The above presented account of default reasons suffices to represent reasoning from an incomplete knowledge base, but it does not represent reasoning with information changes that alter the premises from which an agent starts to reason. Still, the basic account can already model one type of non-monotonic behavior induced by the definition of undercut. We will refer to this way of revising as "inferential" revision. The workings of undercut can be illustrated with Toulmin's famous example (Toulmin, 1958/2003, p. 92) of arguing for the claim that Harry is a British subject. This claim "can be defended by appeal to the information that he was born in Bermuda, for this datum lends support to our conclusion on account of the warrants implicit in the British Nationality Acts...". The example is translated into a justification logic default as follows. Given the fact that Harry was born in Bermuda (B), an agent can conclude that Harry is a British subject (S):

$$\delta_1 = \frac{r : B :: (s \cdot r) : S}{(s \cdot r) : S}.$$

The default can be read as follows: "If r is a reason justifying that Harry was born in Bermuda and it is consistent to assume that $(s \cdot r)$ is a reason justifying that Harry is a British subject, then $(s \cdot r)$ is a defeasible reason justifying that Harry is a British subject".

However, if the agent were in possession of the additional information saying that both Harry's parents are aliens (P), the "general authority of the warrant" $s : (B \to S)$ for the claim S would have to be set aside. This is modeled with the following rule that introduces an undercutting reason:

$$\delta_2 = \frac{t : P :: (u \cdot t) : \neg[s : (B \to S)]}{(u \cdot t) : \neg[s : (B \to S)]}.$$

The consequent can be read as follows: "$(u \cdot t)$ is a defeasible reason denying that the reason s justifies that if Harry was born in Bermuda, then Harry is a British subject". This is a classical argumentation theory example of a defeater that leads to the suspension of the conclusion supported by the reason $(s \cdot r)$. For a default theory $T_1 = (W, D)$ with $W = \{r : B, t : P\}$ and $D = \{\delta_1, \delta_2\}$, the process (δ_1, δ_2) corresponds to such course of reasoning with revised $\mathbf{JT_{CS}}$ extensions. Notice that the warrant underlying δ_2 can also be questioned in a further course of

reasoning. For example, one could find out that one of Harry's parents was settled in Bermuda at the time when he was born, thus reinstating the authority of the warrant used in δ_1. Formally, this would mean that there is an available default rule that gives you a reason denying that the warrant of δ_2 is true.

Notice that in the logic above, the argument $(s \cdot r) : S$ is susceptible to attack due to the fallibility of inference δ_1 that is characteristic for *default* reasoning. For the argument $(s \cdot r) : S$ to be undermined, we consider a wider Toulminian interpretation of the argument that includes the formula $r : B$ as the data for the argument. Since $r : B$ is in the set W, the only possibility to attack $r : B$ is to remove it from W and to thereby undermine $(s \cdot r) : S$. This kind of attack on arguments is studied under the paradigm of *plausible* reasoning.[2] In this paradigm, arguments are taken to be susceptible to attack due to the uncertainty of their premises. The aim of the current work is to unify the two paradigms in a single logical system.

4.4 Dynamic operations for default theories: Introducing undermining attack

As mentioned above, undermining can be interpreted as an attack on the formulas that are considered to be facts. In our view, undermining is essentially non-inferential because introducing conflicting information that undermines facts cannot be done with the use of warrants.[3] For a default theory, these facts are represented by justification logic formulas from the set of premises W and, in constructing a default argument, such formulas can be prerequisites of default rules. A plausible interpretation of undermining defeaters would be that they propose alternative states of facts which ground further reasoning steps. To be able to incorporate factual changes, we need methods from belief revision. Our selection of the belief-revision operations follows the way in which default theories are defined — since the set of premises W is typically finite, it is natural

[2] See (Prakken, 2017, pp. 59-61) for more details on this distinction.

[3] The non-inferential view of information change is also relevant for human interaction. As Hlobil (2018) argues, we can believe by accepting testimonies, but we cannot make inferences by merely accepting testimony. Two testimonies that contradict each other are to be, *ceteris paribus*, equally treated and the acceptance of new information is not the same process as inferentially extending the existing (incomplete) information.

to use operators for sets that do not require closure. Therefore, our choice is to make use of base revision operators (Hansson, 1999a) instead of the AGM operators (Alchourrón et al., 1985).

To model changes to default theories, we will use the capacity of default logic to represent two levels of information certainty. The top-level of information certainty is represented by formulas that are included in all extensions. Typical examples of such formulas are those contained in a set of premises W. The lower-level of information certainty is represented by formulas whose status is contingent on whether they become defeated by other available information. Such formulas are typically consequents of default rules. Our goal is to use the two levels and to define dynamic operators that can bring about the changes that fix whether a formula is included in or excluded from all extensions, but also to define *non-prioritizing* operations that leave the status of a formula undecided.

To be able to model the dynamics at the two levels of information certainty, we extend the above defined default theories with *unwarranted* defaults,[4] which correspond to Reiter's supernormal defaults, but (possibly) containing justification assertions:

$$\delta = \frac{\top :: F}{F}.$$

Standard default rules with justification assertions encode inferential steps supported by warrants. In contrast to inferential steps, supernormal defaults will be used to represent non-inferential, information-changing actions in which an agent accepts that a formula can be included in (at least) one extension. We will extend sets of defaults with supernormal defaults whenever we represent introducing uncertain information to a theory T or relegate information from W to the status of uncertain information.

Why would we want to make changes only to the lower-level of information certainty or alter a default theory at the level of some, instead of all extensions? Sometimes, an agent has doubts with respect to whether it is safe to include some information or not and, analogously, whether it is safe to remove some information or not. In the standard base revision approach to modeling information change, incoming information is always prioritized over existing information, which is ensured by the success postulate. Consider again the example of the agent reasoning

[4]By referring to defaults as "unwarranted", we mean only that such default rules introduce their consequents without using warrants as defined in Chapter 2.

about Harry's eligibility for British nationality. It is possible that, according to the census record, Harry was born in Bermuda and, according to the military record, he was born outside Bermuda. The fact that the agent first collected the census record data and then collected the military record data cannot justify the prioritization of the newly acquired information. If the agent does not know which information source is reliable, the order of data input is irrelevant. In these cases, default logic can avoid the "naive" priority ascription by the use of multiple extensions. The rest of this section gives a solution to the problem of non-prioritized change of default theories, along with the more standard prioritized change. In a case of non-prioritized change, the corresponding dynamic operator uses supernormal defaults with an aim to alter the lower-level of information certainty of a default theory. On our interpretation of undermining attacks, whether undermining fully or partially realizes its defeating potential depends on whether the new information is prioritized or not.

4.4.1 Default theory expansion

We consider three kinds of change: expansion, contraction and revision. The first kind of change corresponds to learning new information. For example, adding a formula F to a set of premises W can be based on the information provided by some information channel. The formal operation that naively adds new information without checking the joint consistency of the resulting set of beliefs is called "expansion":

Definition 4.1 (Expansion)**.** *For a default theory $T = (W, D)$ and a formula F, $T_F^+ = (W_F^+, D)$ is the expansion of the default theory T, where W_F^+ is the base expansion of the set W such that $W_F^+ = W \cup \{F\}$.*

If the added information results in an inconsistent set W_F^+, any definable $\mathbf{JT_{CS}}$ extension will be inconsistent. Notice that default theory expansion can already cause non-monotonic behaviour on the level of default theory extensions. For example, if the added formula is a prerequisite for a default rule with an undercutter for some other default consequent, the new information can result in removing elements from $\mathbf{JT_{CS}}$ extensions of T.

An agent can approach accepting incoming information more cautiously. If the agent accepts new information as a plausible premise, but hesitates to consider it a fact, the change is made to the set of default rules:

Definition 4.2 (Conservative Expansion)**.** *For a default theory $T = (W, D)$ and a formula F, $T_F^{\times} = (W, D \cup \{\delta_F\})$ is the conservative expansion of the default theory T with F, where $\delta_F = \frac{\top ::F}{F}$.*

Notice that the operation $\times$ opens up a possibility that the formula F is included in all extensions, but it can also be excluded from all extensions. For example, if $\neg F$ is contained in $Th^{JT_{CS}}(W)$, then δ_F is not applicable. The introduced operators have the following properties:

Proposition 4.3. *For a default theory $T = (W, D)$ with unwarranted default rules and a* $\mathbf{JT_{CS}}$ *formula F, it holds that*

a) If F is not a contradiction, then F is contained in each $\mathbf{JT_{CS}}$ *extension of the theory T_F^{+}.*

b) If F is not a contradiction and if $\neg F$ is not contained in any $\mathbf{JT_{CS}}$ *extension of T, then F is contained in each* $\mathbf{JT_{CS}}$ *extension of the theory $T_F^{\times}$.*

c) If W is not inconsistent and if $\neg F$ is contained in $Th^{JT_{CS}}(W)$, then F is not contained in any $\mathbf{JT_{CS}}$ *extension of the theory $T_F^{\times}$.*

Proving Proposition 4.3 is straightforward.

4.4.2 Default theory contraction

How does an agent discard some information that is no longer considered to be reliable? We will again differentiate between two strategies of discarding information or, more technically, of *contracting* default theories: one aims to remove information when an agent is confident that the information is unreliable and another aims to relegate the status of information reliability to a lower level. In our default theories, this will mean that the first operation removes a formula from all extensions while the second operation leaves the possibility that extensions still contain the formula. One problem we face in removing a formula from all theory extensions is that the base contraction of a set of premises W is necessary, but not sufficient to secure that the formula will not be reintroduced by the application of a default rule. To illustrate the need for such operation, consider that changes in information may cause that a certain source of justification t is denied its reliability as a reason for some formula F.

To deal with this problem, we propose to put restrictions on the application of default rules. The aim of restrictions on application is

to prevent an unwanted $\mathbf{JT_{CS}}$ formula to become a part of any default theory extension. For a $\mathbf{JT_{CS}}$ formula R, the application of a default rule $\delta = \frac{F::G}{G}$ to a $\mathbf{JT_{CS}}$-closed set of formulas Γ is *restricted* by R if δ is applicable to Γ iff:

- $F \in \Gamma$ and
- $\neg G \notin {}_R\Gamma$,

where ${}_R\Gamma = Th^{JT_{CS}}(\Gamma \cup \{R\})$. We say that δ is *restrictedly-applicable* to Γ according to the restriction R. The role of the restriction R is to prevent the formula $\neg R$ to extend the $\mathbf{JT_{CS}}$ theory Γ by blocking the applicability of each default whose consequent formula would introduce $\neg R$ to the extended theory, if the consequent is added to Γ.

Using the restricted variant of default applicability, we will define restricted variants of default theories and their processes. Default theories in which we restrict the application of defaults by a $\mathbf{JT_{CS}}$ formula R can be defined for those sets of premises W that do not entail $\neg R$:

Definition 4.4 (Application-Restricted Default Theory)**.** *For a* $\mathbf{JT_{CS}}$ *formula R, an application-restricted default theory ${}_{[R]}T$ is defined as a pair (W, D), where the set W is a finite set of* $\mathbf{JT_{CS}}$ *formulas such that $\neg R \notin Th^{JT_{CS}}(W)$ and D is a countable set of default rules such that the application of each $\delta \in D$ is restricted by R.*

Application-restricted default theories differ from default theories defined in Section 2.4 only in the view of restrictions that might eliminate some possible ways to build default processes that would otherwise be possible without restrictions.

To define processes for an application-restricted theory ${}_{[R]}T$, recall that the *In*-set from Definition 2.9 is a closed set of $\mathbf{JT_{CS}}$ formulas that represents current evidence base.

Definition 4.5 (Application-Restricted Process)**.** *A sequence of default rules Π is a* process *of an application-restricted default theory ${}_{[R]}T = (W, D)$ iff every k such that $\delta_k \in \Pi$ is* restrictedly-applicable *to the set $In(\Pi[k])$ according to the restriction R, where $\Pi[k] = (\delta_0, \ldots \delta_{k-1})$.*

When the application of δ is restricted restricted by R, we need to check whether the negation of $req(\delta)$ is contained in the following set of $\mathbf{JT_{CS}}$ formulas:

Definition 4.6. ${}_R In(\Pi) = Th^{JT_{CS}}(In(\Pi) \cup \{R\})$.

For a theory ${}_{[R]}T$, an application-restricted process Π is said to be closed iff every $\delta \in D$ that is restrictedly-applicable to $In(\Pi)$ according to the restriction R is already in Π.

Notice that the definitions of warrant, undercut, acceptability, as well as potential, $\mathbf{JT_{CS}}$-admissible, $\mathbf{JT_{CS}}$-preferred, $\mathbf{JT_{CS}}$-stable, $\mathbf{JT_{CS}}$-complete and$\mathbf{JT_{CS}}$-grounded extensions from Section 2.5 all depend on the definitions of *In*-sets and default processes. Therefore, the definitions of these concepts have equivalent formulations for application-restricted default theory pairs (W, D). In fact, application-restricted default theories are a generalization of default theories from Definition 2.4. Each default theory $T = (W, D)$ can be defined as an application-restricted default theory ${}_{[\top]}T = (W, D)$, where the restriction formula is a tautology. For the theory ${}_{[\top]}T$, the sets $In(\Pi)$ and ${}_{\top}In(\Pi)$ coincide.

For any application-restricted default theory, the expansion operation $({}_{[F]}T)^{+}_{G}$ and the conservative expansion operation $({}_{[F]}T)^{\times}_{G}$ are both defined analogously to the corresponding default theory operations. Notice the following exception: expanding an application restricted theory ${}_{[F]}T$ with the formula $\neg F$, that is, $({}_{[F]}T)^{+}_{\neg F}$. According to Definition 4.4, an application-restricted theory can only be defined for a set of premises W that does not entail the negation of a restriction formula, so the theory $({}_{[F]}T)^{+}_{\neg F}$ is not defined. In general, the following holds in such cases of expansion:

- For an application-restricted theory ${}_{[F]}T$ and a $\mathbf{JT_{CS}}$ formula F, if $\neg F \in Th^{JT_{CS}}(W \cup \{G\})$, then the expansion of an application-constrained theory ${}_{[F]}T$ with a formula G is defined as the the default theory T^{+}_{G}.

Therefore, the expansion of ${}_{[F]}T$ with $\neg F$ is the default theory with unwarranted rules $T^{+}_{\neg F}$. After expanding ${}_{[F]}T$ with $\neg F$, $W^{+}_{\neg F}$ contains the formula $\neg F$, which means that the resulting theory cannot be application-restricted by the formula F.

We can now define a contraction operation that aims at removing a formula at the level of a whole default theory. The operation corresponds to the action of removing information when an agent is confident that the information is not reliable. To achieve this in a default theory, a formula has to be removed from the set of premises by a base contraction and its reintroduction should be prevented. In the definition of contraction,

remainder sets will be used: for any set of $\mathbf{JT_{CS}}$ formulas Γ and a formula F, the remainder set $\Gamma \perp F$ is defined as the set of maximal subsets of Γ that do not entail F.

Definition 4.7 (Contraction)**.** *For a default theory $T = (W, D)$ and a formula F, the application-restricted theory ${}_{[\neg F]}T_F^- = (W_F^-, D \cup D_{!F}^-)$ is the contraction of the default theory T by F, where*

1. *W_F^- is the (full) meet contraction of the set W such that $W_F^- = \bigcap(W \perp F)$ and*
2. *$D_{!F}^- = \{\delta_G \mid \delta_G = \frac{\top ::G}{G}$ for every $G \in W \setminus (\bigcap(W \perp F) \cup \{F\})\}$.*

Notice that an application-restricted default theory ${}_{[\neg F]}T_F^-$ is definable for any theory T since, due to condition 1, the formula F cannot be an element of the set $Th^{JT_{CS}}(W_F^-)$.

The combination of the restriction $\neg F$ and the set of default rules $D_{!F}^-$ provides a balanced solution for avoiding extremely cautious and extremely incautious behavior. Since the set of formulas $W \perp F$ usually contains many elements, theory contraction operations need to include a procedure of selecting the formulas that can be kept after contracting by F, excluding F itself. It is difficult to define such procedures in a principled and intuitively plausible way. In default theory contraction, we do not need to force selection by a function. Instead, the choice of formulas selected upon contraction depends on the type of extension that is being computed. For example, a $\mathbf{JT_{CS}}$ preferred extension corresponds to the idea of maxichoice contraction, while $\mathbf{JT_{CS}}$ extension corresponds to the idea of full meet contraction (Hansson, 1999a, pp. 12-13).

Using again the two-leveled perspective on changing default theories, we can define a more conservative way of giving up a belief. In conservative contraction, agents are reluctant to entirely give up on some information, but the information is no longer considered to be a fact. To relegate the status of a formula in such a way within a default theory, the formula is removed from the set of premises W and then reintroduced through application of a supernormal default rule.

Definition 4.8 (Conservative Contraction)**.** *For a default theory $T = (W, D)$ and a formula F, $T_F^{\div} = (W_F^-, D \cup D_{!F})$ is the conservative contraction of the default theory T by F, where*

1. *W_F^- is the (full) meet contraction of the set W such that $W_F^- = \bigcap(W \perp F)$ and*

2. $D_{!F} = \{\delta_G \mid \delta_G = \frac{\top ::G}{G}$ *for every* $G \in W \setminus \bigcap(W \perp F)\}$.

Analogously to $T_F^{\dot{-}}$, we define the conservative contraction ${}_{[\neg F]}T_F^{\dot{-}}$ that realizes the same set of $\mathbf{JT_{CS}}$ extensions as the theory ${}_{[\neg F]}T_F^{-}$.

Conservative contraction is an open-ended operation in the sense that it does not preclude the possibility of reintroducing a formula F in an extension through a default rule application. Furthermore, it leaves open the possibility that F occurs in all extensions of the resulting default theory.[5] Antoniou (2002, p. 1149) takes a different approach in defining a more conservative contraction operation for Reiter's default theories. His idea is to secure that there is at least one extension added that does not contain the formula removed from the set of premises. In our view, it is unnecessary to have such an operation. If some formula is not regarded to be a fact, but it is still plausible that the formula is true, accepting it as the only available information might be the only reasonable action. Instead of "forcing" an extension without the formula, conservative contraction enables the possibility of an extension without the formula. If there is no support for the contrary statement whatsoever, an agent might still need to hold on to the only available information. The following statement follows from Definition 4.7:

Proposition 4.9. *For a default theory* $T = (W, D)$ *with unwarranted default rules and a non-tautological* $\mathbf{JT_{CS}}$ *formula* F*, it holds that* F *is not contained in any* $\mathbf{JT_{CS}}$ *extension of the theory* ${}_{[\neg F]}T_F^{-}$.

Proof. From the condition 1 of the definition, we know that the full meet contraction of W removes the formula F from the set of premises of ${}_{[\neg F]}T_F^{-}$. Moreover, from condition 2 and the fact that the application of each default rule from $D_{!F}^{-}$ in ${}_{[\neg F]}T_F^{-}$ is restricted by $\neg F$, we know that F cannot be reintroduced into an evidence base $In(\Pi)$ by applying defaults from a process Π for any process Π of ${}_{[\neg F]}T_F^{-}$. Therefore, F cannot be contained in any $\mathbf{JT_{CS}}$ extension of ${}_{[\neg F]}T_F^{-}$. □

4.4.3 Default theory revision

The task of adding new information to the set of premises by the expansion operation (Definition 4.1) can lead to an inconsistent set of premises.

[5] Analogously, conservative expansion might not guarantee that there will be any extension containing a formula F, after a default theory has been conservatively expanded with F.

A more realistic dynamic operator for adding information needs to specify a process by which an agent adds information inconsistent with W without being committed to an inconsistent set of premises. One possible way is to only add information via the conservative expansion operation (Definition 4.2), but this comes with an obvious flaw: an agent is not able to confidently replace an old, unreliable piece of information with a new, reliable one. This is one of the motivations to define a default theory revision operator that not only adds a formula, but also removes inconsistent formulas at one of the two levels of the default theory.

A revision operation can be defined from a combination of the expansion and contraction operations.[6] In our approach, we will follow the traditional arrangement of the operations as proposed by Hansson (1999a, p. 203), namely, removing formulas will precede adding a formula. Those revision operations in which contraction is followed by expansion are called "internal revision" operators. The choice of a revision operation used for a particular revision example depends on both whether old information is to be removed confidently or conservatively and whether new information is to be added confidently or conservatively. We define the following four internal revision operators for each combination of the operations.

Definition 4.10 (Revision Operators)**.** *For a default theory $T = (W, D)$ and a formula F, (internal) revision operators for T are defined as follows:*

1. $T_F^{\mp} = ({}_{[F]}T^{-}_{\neg F})^{+}_{F}$
2. $T_F^{\not\times} = ({}_{[F]}T^{-}_{\neg F})^{\times}_{F}$
3. $T_F^{\dot{+}} = (T^{\dot{\div}}_{\neg F})^{+}_{F}$
4. $T_F^{\divideontimes} = (T^{\dot{\div}}_{\neg F})^{\times}_{F}$

The variety of possible revision operators raises the question about what kinds of revision strategies they represent.

We can show that the four operations amount to two strategies. Again, as in the cases of expansion and contraction, one strategy is meant to

[6] If we were to exhaust all possible combinations, eight revision operators could be defined. Note that the revision operation symbols used here reflect the composition of the introduced revision operations that are defined in terms of contraction and expansion variants. The symbols are not intended to be in continuity with the standard usage of revision operation symbols.

revise confidently and the other strategy to revise more conservatively. The key to show this is to prove that the operations $\mp$, $\curlywedge$ and $\dot{+}$ give equivalent extensions in revising a default theory with some formula F. This is the result stated in the a) clause of Proposition 4.11:

Proposition 4.11. *For a default theory $T = (W, D)$ with unwarranted default rules and a* $\mathbf{JT_{CS}}$ *formula F, it holds that*

a) If F is not a contradiction, then for all $\mathbf{JT_{CS}}$ *extensions Γ of the theories $T_F^{\mp}$, $T_F^{\curlywedge}$ and $T_F^{\dot{+}}$, it holds that $F \in \Gamma$.*

b) If F is not a contradiction, then there is a $\mathbf{JT_{CS}}$ *extension Γ of the theory T_F^{*} such that $F \in \Gamma$.*

c) If F is not a contradiction, then there is a $\mathbf{JT_{CS}}$ *extension Γ of the theory T_F^{*} such that $\neg F \notin \Gamma$.*

Proof. To prove that a) holds, consider the three revision operators $\mp$, $\curlywedge$ and $\dot{+}$ and the resulting theories $T_F^{\mp}$, $T_F^{\curlywedge}$ and $T_F^{\dot{+}}$. For the case of the default theory $T_F^{\mp}$, it follows from Proposition 4.9 that $\neg F$ is not contained in any $\mathbf{JT_{CS}}$ extension of ${}_{[F]}T^{-}_{\neg F}$. By Proposition 4.3 a), F is contained in each $\mathbf{JT_{CS}}$ extension of $({}_{[F]}T^{-}_{\neg F})^{+}_{F}$.

For the case of the default theory $T_F^{\curlywedge}$, it follows from Proposition 4.9 that $\neg F$ is not contained in any $\mathbf{JT_{CS}}$ extension of ${}_{[F]}T^{-}_{\neg F}$. Moreover, the set ${}_{F}In(\Pi)$ contains the formula F and it is not $\mathbf{JT_{CS}}$ inconsistent, which means that the default rule $\frac{\top :: F}{F}$ is restrictedly-applicable to any $\mathbf{JT_{CS}}$ extension of the conservative expansion $({}_{[F]}T^{-}_{\neg F})^{\times}_{F}$ of the theory ${}_{[F]}T^{-}_{\neg F}$. Therefore, F is contained in each $\mathbf{JT_{CS}}$ extension of $({}_{[F]}T^{-}_{\neg F})^{\times}_{F}$.

For the case of the default theory $T_F^{\dot{+}}$, consider that the base contraction of W ensures that $\neg F$ cannot be contained in the set of premises $W^{-}_{\neg F}$ of the default theory $T^{\dot{-}}_{\neg F}$, but $\neg F$ can still be reintroduced by applying the defaults from $D_{!\neg F}$. However, after expanding the theory $T^{\dot{-}}_{\neg F}$ by F, the inclusion of the formula $\neg F$ into any $\mathbf{JT_{CS}}$ extension of the theory $T_F^{\dot{+}}$ is blocked and, by Proposition 4.3 a), F is in contained each $\mathbf{JT_{CS}}$ extension of $T_F^{\dot{+}}$.

To prove that b) holds, consider that the base contraction of W ensures that $\neg F$ cannot be contained in the set of premises $W^{-}_{\neg F}$ for the conservative contraction $T^{\dot{-}}_{\neg F}$. This means that, for the conservative expansion $(T^{\dot{-}}_{\neg F})^{\times}_{F}$, it holds that the default rule $\frac{\top :: F}{F}$ is applicable to $Th^{JT_{CS}}(W)$ and, therefore, contained in at least one $\mathbf{JT_{CS}}$ extension of T_F^{*}.

To prove c), consider that after the base contraction of W by $\neg F$, $\mathbf{JT_{CS}}$ extensions of $T^{\dot{-}}_{\neg F}$ are $\mathbf{JT_{CS}}$ consistent. Since we also know that after the conservative expansion $(T^{\dot{-}}_{\neg F})^{\curlywedge}_{F}$, b) holds, then c) holds. □

To show the equivalence of the operators $\mp$, $\curlywedge$ and $\dot{+}$, we first say that for any σ-extension, where

$$\sigma \in \{\mathbf{JT_{CS}}\text{-admissible}, \mathbf{JT_{CS}}\text{-complete}, \mathbf{JT_{CS}}\text{-grounded}, \mathbf{JT_{CS}}\text{-preferred}, \mathbf{JT_{CS}}\text{-stable}\},$$

$\sigma(T)$ is the set of all σ-extensions for a theory T. Then we prove that for any default theory T, the default theories $T^{\mp}_{F}$, $T^{\curlywedge}_{F}$ and $T^{\dot{+}}_{F}$ realize the same set of extensions under any $\mathbf{JT_{CS}}$ extension-based semantics for default theories. The following result is obtainable from Proposition 4.11 a) together with the fact that none of the three operators $\mp$, $\curlywedge$ and $\dot{+}$ change the status of formulas that do not take part in F-implying sets:

Theorem 4.12. *For any default theory $T = (W, D)$, a $\mathbf{JT_{CS}}$ formula F and the (internal) revision operators $\mp$, $\curlywedge$ and $\dot{+}$, it holds that $\sigma(T^{\mp}_{F}) = \sigma(T^{\curlywedge}_{F}) = \sigma(T^{\dot{+}}_{F})$.*

Intuitively, the three operations represent a type of revision in which an agent confidently includes new and possibly inconsistent information into all $\mathbf{JT_{CS}}$ extensions. Another option specified by the operator $\divideontimes$ is to accept the new information in some extensions while maintaining the old information in other extensions. The revision operators comply to the two-leveled view of default semantics: the first three revision operators of Definition 4.10 fix the status of a revision at the level of a default theory as a whole, while the last revision operator targets at modifying only some extensions. Any of the three operations $T^{\mp}_{F}$, $T^{\curlywedge}_{F}$ and $T^{\dot{+}}_{F}$ will be referred to as the *Revision* of T with F and the operation $T^{\divideontimes}_{F}$ will be referred to as the *Conservative Revision* of T with F.

4.4.4 The notion of undermining

Finally, we are now able to say in what way the dynamic operations connect to the notion of undermining defeat. It was mentioned in the Introduction that by undermining we understand the attack whereby argument premises are being questioned. This intuition can now be cashed out by using those dynamic operators for default theories that involve contracting a default theory.

Definition 4.13 (Undermining). *For a default theory $T = (W, D)$ and a* $\mathbf{JT_{CS}}$ *formula F such that $F \in W$ and $F = pre(\delta)$ for some $\delta \in D$, F is undermined iff W is contracted by F by applying any of the following operations to T:*

1. ${}_{[\neg F]}T^{-}_{F}$ *(Contraction)*
2. $T^{\div}_{F}$ *(Conservative Contraction)*
3. $T^{\mp}_{G}$, $T^{\not\asymp}_{G}$ *or* $T^{\dagger}_{G}$ *for* $\mathbf{JT_{CS}}$ *inconsistent formulas F and G (Revision)*
4. T^{*}_{G} *for* $\mathbf{JT_{CS}}$ *inconsistent formulas F and G (Conservative Revision).*

Notice that there is no requirement on the structure of F. However, each meaningful undermining targets justification assertions because W cannot be successfully contracted by a tautology and justification assertions are the only other type of formula occurring as a default prerequisite. Not every attack on the premises results in confidently revising the set W. It is possible that undermining leaves an agent undecided as to whether newly acquired information or older information should be prioritized.

Starting from the theory T_1 defined in Section 4.3, we can give a formalized undermining example to show the difference between inferential and non-inferential ways of information acquisition. Recall that the agent started to reason from the information that Harry was born in Bermuda. This piece of information is represented in the set of premises W with the formula $r : B$, where r can now be taken to reflect the source of information as, e.g., data from census records. However, if the information based on military records says that Harry was born outside Bermuda, and having no means to resolve this conflict of information, the theory T_1 needs to be conservatively revised. The theory $T^{*}_{1_{v:\neg B}}$ is the revision of T_1 with the formula $v : \neg B$, where v reflects the new source of information for the claim that Harry was not born in Bermuda.

To see this revision in more detail, recall that the theory $T_1 = (W, D)$ consists of the set of premises $W = \{r : B, t : P\}$ and the set of defaults $D = \{\delta_1, \delta_2\}$ with

$$\delta_1 = \frac{r : B :: (s \cdot r) : S}{(s \cdot r) : S} \text{ and } \delta_2 = \frac{t : P :: (u \cdot t) : \neg[s : (B \to S)]}{(u \cdot t) : \neg[s : (B \to S)]}.$$

The first dynamic operation in revising with $v : \neg B$ is contracting the theory by $\neg v : \neg B$. The resulting theory $T^{\div}_{1_{\neg v:\neg B}} = (W^{-}_{\neg v:\neg B}, D \cup D_{!\neg v:\neg B})$

consists of the set of premises $W^{-}_{\neg v:\neg B} = \{t : P\}$ and the set of defaults D extended with the default $\delta_{r:B} = \frac{\top::r:B}{r:B}$.

Finally, the agent conservatively expands the theory $T^{\div}_{1_{\neg v:\neg B}}$ with the information that Harry was not born in Bermuda. The new default theory is defined as $T^{*}_{1_{v:\neg B}} = (T^{\div}_{1_{\neg v:\neg B}})^{\times}_{v:\neg B}$. The change of the theory after conservative expansion with $v : \neg B$ amounts to adding the new default rule $\delta_{v:\neg B} = \frac{\top::v:\neg B}{v:\neg B}$, which means that the new set of defaults is $D \cup D_{!\neg v:\neg B} \cup \{\delta_{v:\neg B}\}$. The revised theory $T^{*}_{1_{v:\neg B}}$ changes the default processes in which the agent reasons about Harry's nationality and, eventually, changes the structure of acceptable reasons by changing the way in which $\mathbf{JT_{CS}}$ extensions are computed.

4.5 Conclusions

As mentioned in the introduction to this chapter, our approach to structured argumentation dynamics builds on similar ideas as Antoniou's (2002) approach to the dynamics of standard default theories. Antoniou's approach significantly differs from ours in the way he treats those changes that add or remove a formula at the level of some, but not necessarily all extensions. Unlike our conservative expansion and conservative contraction, none of Antoniou's operations leaves the inclusion status of a formula undecided. For instance, to secure that a formula is not contained in at least one extension, Antoniou (2002, p. 1149) adds a new extension where introducing the formula is blocked by adding either a new atom or its negation to any default, dependent on whether they are allowed to be in a same extension or not.

In this chapter, we defined *local* change operations that do not, in general, give a recipe on how to further change their output default theories. But to represent actual dynamic contexts of argumentation, we need to make our operators *global*, rather than local, and enable iterated revision. Note that, for example, the second output theory $({}_{[F]}T^{-}_{\neg F})^{\times}_{F}$ of Definition 4.10 is an application-restricted default theory. If we want to allow for iterated contraction and generalize the contraction operation to application-restricted theories such as $({}_{[F]}T^{-}_{\neg F})^{\times}_{F}$, we need to deal with multiple restrictions to the output theory, and possibly mutually inconsistent restrictions. This could be done if we allow that an application-restricted theory ${}_{[F]}T$ can be further restricted by a formula G in such a way that, for any default δ, we need to check if $req(\delta)$ is consistent with

both ${}_F In(\Pi)$ and ${}_G In(\Pi)$, thus defining the application-restricted default theory ${}_{[G,F]}T$.

Some approaches to default reasoning such as (van Linder et al., 1997) and (Meyer and van der Hoek, 1990) represent the idea of defaults in dynamic epistemic logic. The main focus of van Linder et al. (1997) is to embed supernormal defaults in a multi-agent modal logic with knowledge, belief and update modalities. The authors show that Reiter's extensions can be represented as a result of consecutive jump actions to default conclusions, but they do not focus on how such extensions are revised due to information changes. Meyer and van der Hoek (1990) introduce a preference modality to distinguish between known and (provisionally) preferred information. A non-monotonic belief revision component consists in changing preferences as a result of obtaining knowledge.

Baltag et al. (2012, 2014) and Renne (2012) define extensions of justification logic in which agents may acquire new information that defeats the reasons they accepted. The logics combine belief revision and dynamic epistemic logic techniques to model a kind of defeat that seems to correspond to undermining. However, each of the logics assumes prioritizing new information and none of these logics is able to model undercut and rebuttal. Even so, approaches based on dynamic epistemic logic are attractive because they open up a possibility of developing a multi-agent justification logic with defeaters.

We indicated in the introduction to this chapter that the work in the area of the dynamics of argumentation frameworks without argument structures is already well-developed. Among the approaches, it is worth mentioning those that follow the belief revision methods applied to Dung's argumentation frameworks, such as that of Booth et al. (2013) and Diller et al. (2015). Booth et al. (2013) start from a labelling approach to Dung's argumentation frameworks and constraints on a framework's output. Their focus is on finding the best way to recover a rational output given a framework and a constraint on its output. For this, they use ordering of conflict-free labellings in a way that the most rational conflict-free labelling is chosen when none of complete labellings respects the constraint. In the work by Diller et al. (2015), we find two kinds of revision operators. One of them revises an abstract argumentation framework by taking a propositional formula as a means to represent the new information, while the other operation revises an input framework by information in the form of another framework. Both operations give a

single output framework respecting a particular type of rankings on extensions.

Finally, our work contributes to the study of non-prioritized belief revision operations, that is, such operations for which the new information has no special priority due to its novelty (Hansson, 1997, 1999b). The way in which our operators are defined meaningfully combines resources from both belief revision and default logic. The relation between belief revision and non-monotonic reasoning has long been a matter of discussion (Gärdenfors, 1990, Makinson and Gärdenfors, 1991) among AI researchers. Although it was not our aim to discuss the relation between modeling reasoning with *incomplete information* in default theories and modeling reasoning with *changing information* in belief revision, we showed that our justification logic creates a useful junction for the two approaches.

As a result of connecting the two reasoning paradigms, the logic presented here, we can model both plausible and default reasoning. According to Prakken (2017, p. 2198), argumentation models of plausible reasoning locate all fallibility of an argument in its premises, while argumentation models of default reasoning locate all fallibility in its defeasible inferences. To the best of our knowledge, the system presented here is the first *logic* to combine the two types of argumentation models by capturing all the standard notions of defeat in AI: rebuttal, undercut and undermining.

Chapter 5

A default logic framework for normative rules in human reasoning

5.1 Introduction

In this chapter, we deal with commonsense reasoners in the context of the philosophical debate on the normative role of logic in human reasoning. We argue for weak psychologism — the claim that logical rules are normative for human reasoning. We particularly stress the importance of making right choices when deciding on a logical format for norms in human reasoning. The problem we find in the traditional criticism of weak psychologism is that it starts from the idea that there are exceptionless normative principles that require human agents to follow the rules of classical logic — the so called "bridge principles" between logic and human reasoning. We want to show that the facts of ordinary reasoning, such as uncertainty of information and filtering out irrelevant information, vouch for relaxing the assumptions on normative rules in ordinary reasoning.

To this end, we offer a new, default logic perspective on the normativity of logic. First we discuss Harman's (1986) proposed counterexamples to the normativity of classical logic. Harman argues that classical logic has neither a normative role nor an explanatory role in human reasoning — a position known as "anti-psychologism". Harman's argument hinges on one crucial claim, that there is no adequate bridge principle between

logic and human reasoning. This is right, but, contrary to what Harman claims, we argue that this does not suffice to refute weak psychologism. Instead, we argue that Harmanian bridge principles presuppose two requirements that a normative principle cannot meet, namely the non-defeasibility requirement and the relevance requirement. We show that both requirements are unnecessary. Moreover, we define a new variant of default logic for ordinary reasoning as an alternative framework for normative rules. Using this default logic, we model the kind of reasoning that Harmanian bridge principles describe. Finally, we present a picture of how logic is normative for human reasoning.

5.2 Motivation

Is logic normative for human reasoning? Gottlob Frege set the stage for the twentieth-century debate on the normativity of logic in his *Grundgesetze der Arithmetik* (1893). He stated (1893/1964, p. 12) that the laws of logic "prescribe universally the way in which one ought to think if one is to think at all". Frege's position on the relation between formal logic and human reasoning is a common one and here it will be called "weak psychologism".[1] According to weak psychologism, logic describes what follows from what and thereby provides normative criteria for human reasoning. The other notable and more extreme historical position on the role of logic is the view that logic provides an explanation of the psychological facts of reasoning. This position will be referred to as "strong psychologism". This explanatory role of logic was famously rejected by Frege (1893/1964, p. 12), who claims that laws of logic "do not make explicit the nature of human thinking and change as it changes". Although this claim was widely considered as a rejection of psychologism in general, Frege did maintain that formal logic is normative for human reasoning.

The difference between weak and strong psychologism can be presented by way of an example taken from Levesque (1986). Let it be given that:

> *Jack is looking at Ann but Ann is looking at George. Jack is married but George is not.*

[1] We will follow Haack's (1978, p. 238) classification of types of psychologism. Although the terms "weak" and "strong" suggest otherwise, weak and strong psychologism differ in kind rather than in degree.

The question is whether the following claim is true or not according to the available information:

A married person is looking at an unmarried one.

Toplak and Stanovich (2002) argue that the logical structure of the example, that is, the disjunctive syllogism, hardly affects the way in which persons form their beliefs when presented with the example above.[2] Their survey confirms Frege's thesis about strong psychologism — the logical form of disjunctive syllogism does not *explain* the actual reasoning that underlies belief formation in human subjects. In contrast, the disjunctive syllogism *prescribes* how to correctly form our beliefs after being asked to decide on the question above.

In evaluating whether subjects formed their beliefs correctly, we have to find out whether the actual reasoning that underlies their belief formation conforms to what the formal logical structure prescribes. This is a simple way of saying that the disjunctive syllogism is normative for the reasoning of the participants. The example points to the role that classical logic has in adding, retaining or giving up some of our beliefs: logic serves as a "measuring rod" for correct reasoning, as weak psychologism suggests.

So far, we might be convinced that weak psychologism is an uncontroversial position. However, this view is widely contested in the philosophy of logic in what is known as the "bridge principle" debate. In this chapter, our aims are: (1) to present some key objections to the normativity of logic initiated by Gilbert Harman in the 1980s; (2) to argue that the format of Harmanian bridge principles comes with unnecessary requirements that shape our understanding of the role of logic in ordinary reasoning; (3) to define a new default logic as an alternative framework for normative rules in reasoning; and finally, (4) to use this default logic to show that Harman's counterexamples to weak psychologism are not successful.

[2] According to (Toplak and Stanovich, 2002), only a small number of participants (16 out of 125 or 13% of them) gave a correct answer to the "disjunctive syllogism insight" problem. The answers authors offered were a) Yes, b) No, and c) Cannot be determined. The last answer was chosen by 107 participants or 86% of the total number, while the correct answer is "Yes". Notice that the claim is true if Ann is married and also if Ann is not married.

5.3 Outlining the "bridge principle" debate

In this section we present some of the most prominent counterarguments to weak psychologism. We emphasize the role of the format of normative rules in these counterarguments and motivate an alternative system for normative rules in human reasoning.

5.3.1 Harman's criticism of the relevance of logic for reasoning

Starting in the 1980s, Gilbert Harman has argued that logic is not "specially relevant" (1986, p. 20) for human reasoning. In a series of papers, he holds that logical theory is neither a normative nor a descriptive theory of human reasoning — a position conveniently labelled as "anti-psychologism".[3] Harman's argument hinges on the claim that there is no exceptionless principle that captures the relevant role of logic for reasoning. His misgivings triggered a debate on "bridge principles", principles that connect formal logic with human reasoning. John MacFarlane (2004) first used the term "bridge principle" as a metaphor for the disputed connection between formal logic and human reasoning.[4] Ever since Harman raised the challenge, the possibility and the exact formulation of bridge principles have been recognized as a genuine problem in the philosophy of logic.[5]

What is a Harmanian bridge principle supposed to look like? Bridge principles bring together facts about logical entailment and normative aspects of ordinary reasoning. MacFarlane (2004, p. 6) proposes a general form for a bridge principle:

If $P_1 \dots P_n \vDash Q$, *then (normative claim about believing* $P_1 \dots P_n$ *and* Q*).*

Harman's anti-psychologism results from a failure to find a bridge principle that admits of no exceptions. In fact, the cornerstone of the bridge

[3] In his more recent work, Harman (2002, p. 171) restates his anti-psychologist position: "Principles of implication are not normative (outside of deontic logic) and do not have a psychological subject matter (outside of the logic of belief)". Likewise in (Harman, 2009, p. 333) where he claims that deductive logic "is not a particular psychological subject and is not a particularly normative subject".

[4] In 1966, Hempel (1966) discussed bridge principles from a philosophy of science standpoint where a bridge principle connects theoretical concepts with empirical phenomena. A similar term "bridge-law" appears in (Nagel, 1961).

[5] Recent contributions to the debate include (Dutilh Novaes, 2015), (Fitelson, 2008), (Field, 2009), (Steinberger, 2016), (Steinberger, 2019a) and (Streumer, 2007). Not all of them refer to the resulting formulations by "bridge principle".

principle debate is Harman's conclusion (1986, p. 11) that every principle he considers only holds *other things being equal*. Elsewhere (1986, p. 5), Harman suggests that even if there was a principle of reasoning corresponding to a logical principle, it would have to be a "different principle" because "the logical principle holds without exception, whereas there would be exceptions to the corresponding principle of belief revision".[6] As an obvious candidate principle, *modus ponens* states that if P and $P \rightarrow Q$ are true, then necessarily Q is also true. We might try to formulate a normatively exceptionless bridge principle that exemplifies *modus ponens*. Could the following principle bridge logic and human reasoning in the desired way?

> *Given that* $P, P \rightarrow Q \vDash Q$*, if you believe that* P *and* $P \rightarrow Q$ *you ought to believe* Q*.*

Harman and most of the authors involved in the debate on his "skeptical challenge"[7] answer in the negative.

In his *Change in View* (1986), Harman supports his anti-psychologist claims by offering counterexamples to bridge-principle candidates. He (1986, p. Ch. 2) reasons by cases to show that bridge principles are impossible while at the same time maintaining that such principles are necessary to relate logic to ordinary reasoning. First, Harman (1986, p. 11) presents two bridge principle candidates, namely the "Logical Implication Principle" (LIM) and the "Logical Inconsistency Principle" (LIN):

> *The fact that one's view logically implies X can be a reason to accept X.* (LIM)

> *Logical inconsistency is to be avoided.* (LIN)

Harman argues that both principles are defeasible and that they therefore do not meet the requirement that proper bridge principles must meet. He offers the following counterexample to the LIM (Harman, 1986, pp. 11-12):

> Remember Mary who came to believe three inconsistent things: If she looks in the closet she will see a box of Cheerios,

[6] Harman is using the term "belief revision" as interchangeable with the terms "ordinary reasoning", "human reasoning" and "change in view". He is not referring here to the logic of belief revision (Alchourrón et al., 1985).

[7] The phrase is taken from Steinberger (2019a).

> she is looking in the closet, but she does not see a box of Cheerios. Mary should not at this point infer that she does see a box of Cheerios from her first two beliefs.

It shows that although Mary's set of beliefs entails that "she will see a box of Cheerios", she is bound to revise her initial set of beliefs by giving up her belief about the entailment, rather than adding the entailed one. Thus, logical principles hold without exceptions, but they do not have their exceptionless counterparts in principles of reasoning. Harman concludes that logical principles are not about the regulation of beliefs and that there are no principles of ordinary belief revision that correspond to logical principles like *modus ponens* (Harman, 1984, p. 107; Harman, 1986, p. 5).

To do justice to logical entailment, Harman (1986, p. 12) also considers a modified version of LIM, the "Logical Closure Principle" (LCP). Unlike LIM, LCP prescribes that one's beliefs should be closed under logical entailment:

> *If there is a proposition logically implied by one's set of beliefs which one does not already believe, in that case one should either add the implied proposition to one's beliefs or give up one of the implying beliefs.* (LCP)

According to Harman, LCP is not the required principle either. Since our current set of beliefs entails an infinite set of other propositions, it is unreasonable and "worse than pointless" to add entailed trivialities (Harman, 1986, p. 12). Thus in reasoning we are faced with a demand to avoid cluttering our mind with trivialities. It seems that LCP contradicts this practical demand on how we should go about revising our beliefs.

On what grounds does Harman reject LIN, his second natural bridge principle candidate? Here, Harman invokes the possibility of having to deal with conflicting information. Contrary to what LIN advises, we might find ourselves in situations where we are required to retain inconsistent beliefs. Harman (1986, p. 15) argues that "sometimes one discovers one's views are inconsistent and does not know how to revise them in order to avoid inconsistency without great cost. In that case the best response may be to keep the inconsistency and try to avoid inferences that exploit it." Harman's argument seem to leave no place for the normativity of classical logic or any other "explosive" logic while discussing the possibility of holding inconsistent beliefs.[8]

[8]In addition, Harman (1986, p. Ch. 2) considers the "Liar paradox" and a version of

The objections above are Harman's stepping stones to anti-psychologism, although they do not exhaust his arguments against the normativity of logic. Besides defending an anti-psychologist position, Harman's skeptical challenge laid the groundwork for the further debate over bridge principle candidates.[9] Some authors claim that it is possible to come up with a satisfying Harmanian bridge principle, while others deny its possibility. However, most of them accept Harman's criticism of the role that logic has in reasoning. It is especially worth noting that even the authors who argue that logic is somehow normative for human reasoning accept the format of Harmanian bridge principles as well as Harman's key assumptions on what the requirements are that a correct bridge principle needs to fulfill.[10]

MacFarlane (2004) was the first author to bring to attention the form of Harmanian bridge principles. He proposed a systematic overview of possibilities on how to interpret the normative requirement by way of bridge principles. MacFarlane (2004, p. 7) takes into account the following three key parameters to differentiate between bridge principle candidates: type of the deontic operator ("ought to", "may" and "have a reason to" impose different normative constraints), polarity (normative requirement to believe or not to disbelieve), and the scope of the deontic operator (embedded in the consequent, embedded in both antecedent and consequent and ranging over the entire conditional).[11] We argue in 5.5 that "ought to" is the only deontic operator relevant to weak psychologism. However, it is not possible to come up with a bridge principle based on it. In what follows, we first show that the problem of bridge principles rests upon two unnecessary requirements on normative principles contained in (Harman, 1984) and (Harman, 1986).

---

the "Preface paradox" (discussed in Chapter 6 of this thesis) as objections to LIN. In Subsection 5.5.3, we discuss whether paradoxes and "rational inconsistencies" provide arguments against LIN.

[9] According to Steinberger (2017), Harman's challenge has been particularly influential and Steinberger dedicates an entire section to it in his *Stanford Encyclopedia of Philosophy* entry on the normativity of logic.

[10] Among them are, e.g., (Steinberger, 2019a), (Dutilh Novaes, 2015), (Field, 2009) and (MacFarlane, 2004).

[11] Dutilh Novaes (2015) includes a dialogical, multi-agent perspective on bridge principles and Field (2009) proposes a probabilistic type of constraint for bridge principles relying on the degree of beliefs. Steinberger (2019a) offers a bridge principle that features "have a reason to" operator, which we discuss in Section 5.5.

5.3.2 Defeasibility of normative rules and the "frame problem" of Harmanian bridge principles

Harman's skeptical challenge ultimately rests on two unnecessary requirements, as we will argue henceforth. Harman's *first requirement* is that if logical rules are normative for human reasoning, they are to impose *non-defeasible* norms only. Because logical rules are non-defeasible, Harman assumes that bridge principles are to be non-defeasible as well and rejects his bridge principle candidates by showing that they are defeasible. He (1984, p. 108) argues that while logical principles hold universally, without exceptions:

> the corresponding principles of belief revision would be at best *prima facie* principles, which do not always hold. It is not always true that, if one believes p and believes if p then q, one may infer q. The proposition q may be absurd or otherwise unacceptable in the light of one's other beliefs, so that one should give up either one's belief in p or one's belief in if p then q rather than believe q.

In his arguments against principles LIN and LIM, Harman (1986, p. 11) explicitly mentions their defeasibility: "Neither principle is exceptionless as it stands. Each holds, as it were, other things being equal. Each is defeasible". Later on, he rejects (1986, p. 16) both of them because "we take logic to require precise principles with precise boundaries, not principles that hold merely *normally* or *other things being equal*".

We disagree with the non-defeasibility requirement. In ordinary reasoning, humans acquire beliefs in various ways that do not guarantee their truth. This, in turn, causes the kinds of situations where an entailed proposition "may be absurd or otherwise unacceptable in the light of one's other beliefs". Typically, one could face a normative conflict by following *modus ponens* in reasoning from a set of inconsistent beliefs. But the fact that one has to reason from defeasible and, sometimes, inconsistent information does not bear on the question of whether *modus ponens* is normative for human reasoning or not.[12] The best we can hope for

[12] In Chapter 4, we discuss two ways of defeasibility. First, information that is represented by *premises* is plausible, but it can be defeated by new information. Secondly, *default inferences* can be defeated in the course of reasoning. In Section 5.4, both types of defeasibility play role in shaping a system in which we represent Harman's counterexamples. By assuming that premises are defeasible, nothing has been said about

in this kind of normative conflict is to either assign a higher priority to one of the norms or, when resolving the conflict is not possible, continue reasoning. In the first scenario, a lower-priority norm gets defeated in the course of reasoning and the undefeated norm results in an obligation for a reasoning task at hand. As for the second scenario, we are still normatively bounded by logical rules, despite being in a normative conflict.[13]

Despite its importance, Harman does not offer an argument that supports the non-defeasibility requirement on normative principles. However, normative rules typically *are* defeasible. Consider the following example of conflicting norms in ethics taken from Horty (2003). You ought to meet your friend, given that you have promised to do so. You also ought to save a drowning child, given the obligation to save lives. We can imagine the circumstances where the two norms apply, but it is impossible to fulfill both of them. An intuitive response in such a case is to conclude that you only need to perform the second action. Such examples show that normativity and defeasibility do not exclude each other. Following Harman, we take logical rules to be defeasible norms in reasoning, but we do not accept that defeasibility refutes their normative status. What Harman's arguments unsurprisingly show is that given a set of norms triggered for some human reasoning task, it is possible that some of them could get defeated.

Harman's *second requirement* is that if logical rules are prescriptive for human reasoning, we must be able to specify *a priori* which entailments are normative given a set of initial beliefs. The "relevance requirement", as we will call it, is only made implicitly by Harman and his followers in their suggestions of "ought-based" and prescriptive bridge principles. Recall that LCP suggested that "one's beliefs should be closed under logical implication" (Harman, 1986, p. 12). Although obviously mistaken, the principle comes closest to the paradigmatic case of normativity we discussed in the "Married-unmarried" puzzle: it *prescribes* how to form our beliefs in ordinary reasoning. What is the mistake behind the normative

---

the normativity question. With that assumption, however, we are able to circumvent the quandary of the so called "(non)attitudinal" Harmanian bridge principles and what Steinberger (2017) calls the "bootstrapping objection".

[13] Authors from different fields acknowledge that normative rules are often defeasible rules. For a deontic logic account of defeasible rules see (Horty, 1994), (Horty, 2003) and (Horty, 2012); for a study of exceptions to rules in ethics see (Miller, 1956); for AI and law literature see (Boonin, 1966), (Prakken and Sartor, 2004) and (Verheij, 1996).

demand imposed by LCP?

The answer is that LCP fails, not because logic is not normative for human reasoning, but because it is impossible to *a priori* separate the entailed propositions into what is normatively required and what is best considered a triviality. This can be shown on MacFarlane's format of prescriptive bridge principles. Consider this simplified version of LCP:

If $P_1 \dots P_n \vDash Q$*, then if you believe that* $P_1 \dots P_n$ *you ought to believe* Q.

The propositions $P_1 \dots P_n$ and Q may be any arbitrary propositions satisfying the entailment relation $P_1 \dots P_n \vDash Q$ and Q is a placeholder for infinitely many propositions entailed by the propositions $P_1 \dots P_n$ that you believe. But you reason *about* a specific proposition Q and it is impossible to know which instance is relevant for your reasoning task by only looking into the entailment relation and your initial beliefs, as the principle recommends. Given surrounding context for a reasoning task, e.g., an appropriate set of background beliefs, any belief is potentially relevant to any other. This is the property of cognitive systems that Fodor (1983, p. 105) calls "isotropy". The normative commitments that LCP recommends disregard an infinite number of things that a reasoning task might potentially be about. Therefore, by imposing the bridge principle format, Harman unjustly assumes that on the basis of entailment alone, and given a set of initially believed propositions, a principle should answer which information counts as relevant.

It is impossible to live up to this demand. This is one of the lessons learned from the "frame problem" of artificial intelligence.[14] In particular, bridge principles are undermined by the "relevance problem". According to it, since any piece of information is *potentially* relevant to any other piece of information, it is impossible to determine *a priori* what information bears on an actual reasoning task. Therefore, we can easily side

[14]See (McCarthy and Hayes, 1969). The original frame problem is the problem of describing the effects of an action without having to attend to an infinite number of the non-effects. The upshot of the problem is that for a monotonic first-order representation of actions, there are no obviously trivial non-effects. Thus, logic-based artificial intelligence had to find a way to mimic anthropomorphic intelligence of taking actions as affecting only a limited number of properties while avoiding to consider an "explosion" of propositions about what is not affected by the action at hand. See (Chow, 2013) for a discussion on different implications of the original frame problem. Generalizations and epistemic guises of the frame problem were first considered by Fodor (1983) and Dennett (1984).

with those authors who claim that a principle like LCP is by no means the right bridge principle candidate. However, the failure of LCP is a failure of solving the relevance problem. This has nothing to do with the question of whether logical rules are normative for human reasoning.

Up to now, as far as we know, the frame problem has not been recognized as the problem of the format of bridge principles.[15] The fact that the authors take the "Clutter avoidance" issue to be one of the major objections to the normativity of prescriptive rules confirms this claim. As Harman (1984, p. 108) puts it:

> Many trivial things are implied by one's view which it would be worse than pointless to add to what one believes. For example, if one believes P, one's view trivially implies "either P or Q," "either P or P," "P and either P or R," and so on. There is no point in cluttering one's mind with all these propositions.

While believing trivialities is unreasonable, it is wrong to think that we are required to do so because logic is normative for human reasoning. The cluttering issue results from a stronger claim than the claim of weak psychologism, namely, that if logical rules are prescriptively normative, then there would necessarily be a unique bridge principle that holds for any possible reasoning task.

Harman's anti-psychologism and the bridge principle debate are premised on the requirements of non-defeasibility and relevance. In order to give a fair chance to the weak psychologism thesis, we need to separate the thesis itself from the ramifications of conceding the two requirements. This is our task throughout the rest of the chapter. We propose to deal with Harman's objections from the perspective of a formal system. The purpose of introducing such a system is threefold. First, Harman's counterexamples suffer from ambiguities that can be disambiguated with the help of a formal system. Secondly, the system enables us to differentiate between normative rules and resulting obligations. This is important because not every triggered rule necessarily

[15] Pollock (1987, p. 505-506) briefly discusses the need to appeal to reasoning interests in the context of Harmanian principles. The problem of considering only those implications that are "in question" is recognized by Field (2009, p. 259) and the problem of "reasons to consider" implications by Steinberger (2019a, p. 315), but they do not question the format of bridge principles as normative rules.

results in an obligation.[16] Thirdly, the system enables us to directly put our criticism of the two requirements above into effect. We want to assess the weak psychologism thesis once we set aside the non-defeasibility requirement and the relevance requirement.

Instead of the Harmanian framework, we offer a formal system that enables us to take a case-by-case perspective on obligations that result from those normative rules that apply to a particular case. The first step is introducing Slow Default Logic (SDL) for ordinary reasoning to reconstruct logical principles as defeasible normative rules. The logic we develop is a generalization of standard default logic. After defining the system, we use it to stake out the weak psychologism position amid Harmanian bridge principle candidates.

5.4 Slow default logic for ordinary reasoning

Let us now focus on a new logical system that enables us to model the kind of reasoning that Harmanian bridge principles aim to describe. The system reflects our criticism of the two requirements of bridge principles. First, we model logical norms in reasoning as being defeasible rules only. Secondly, the choice of rules is made a part of a system designer's input. Having defined the system accordingly, we investigate whether logical principles still exert normative force in the cases that Harman considers counterexamples to weak psychologism.

We first need to offer an appropriate logical language of default rules. The logic developed here is a generalization of Reiter's (1980) standard default logic. Contrary to Reiter's approach, we do not rely on logical closure and logical consistency in defining our logic. This decision enables us to represent *ordinary* reasoning rules in a logical system. Before defining our logic, we first look at some common features of default logics. We will use a standard way to represent defaults (Reiter, 1980):

$$\frac{bird(Tweety) : flies(Tweety)}{flies(Tweety)}.$$

We read the default as "If Tweety is a bird and if it is consistent to assume

[16] It seems as if the distinction between a norm and an obligation is not appreciated in Harman's skeptical challenge. While a reasoning task might trigger many coexisting norms, not each of them necessarily gives rise to an obligation. Makinson (1999) and Makinson and van der Torre (2000) introduce this important distinction in deontic logic.

that Tweety flies, then we conclude that Tweety flies". The reasoning behind this default tells us that, *usually*, if we know that something is a bird and it is consistent with what we already know that it flies, then we defeasibly infer that it indeed flies. The conclusion that Tweety flies is a plausible extension of what we already knew about it. However, if in the course of reasoning we also learn that Tweety is in fact a penguin, then, given the background knowledge that penguins do not fly, we cannot infer anymore that it flies. The new piece of information on Tweety undermines the applicability of a default that we initially allowed, because it is not consistent with our knowledge anymore to assume that Tweety flies. We will focus on defeasible reasoning in more detail and adapt it for the purposes of representing the type of reasoning we considered in Harman's objections to the bridge principle candidates.

5.4.1 Syntax of slow default logic

The general form of a default δ is:

$$\frac{\varphi_1, \ldots, \varphi_n : \psi_1, \ldots, \psi_m}{\chi}$$

where φ, ψ and χ are propositional logic formulas. Usually, formulas $\varphi_1, \ldots, \varphi_n$ are called the prerequisites, formulas $\psi_1, \ldots, \psi_m$ are called the justifications, and formula χ is called the consequent. The sets of a default δ's prerequisites and justifications and the consequent are also referred to with $pre(\delta)$, $just(\delta)$, $cons(\delta)$, respectively. Accordingly, $pre(\delta) = \{\varphi_1, \ldots, \varphi_n\}$, $just(\delta) = \{\psi_1, \ldots, \psi_m\}$, and $cons(\delta) = \{\chi\}$. Interpreted as a norm in human reasoning, we read the whole default as "If $\varphi_1, \ldots, \varphi_n$ are believed, and for every $\psi_1, \ldots, \psi_m$ it holds that they are not disbelieved, then conclude χ".

Particularly, we are able to express classical entailment with the default rules. For example, we can take the simple case of conjunction as the default rule

$$\frac{\varphi, \psi : \varphi \wedge \psi}{\varphi \wedge \psi}$$

where φ and ψ are some specific propositional formulas. Both standard defaults and "classical" defaults are given as rule instantiations only, not as rule schemes. This decision reflects our intention to interpret all classical logical rules as defeasible normative rules, not simply to express classical logical rules by means of default rules. Although classical logical

rules are valid inference schemes, the reasoning *norms* they give rise to are only defeasible and need to be considered as being closer to material rules of inference. For example, the reasoning norm $\frac{\varphi,\psi:\varphi\wedge\psi}{\varphi\wedge\psi}$ does not necessarily hold after replacing φ or ψ by some propositional formula χ. The logical properties of reasoning norms are not to be confused with the properties of classical logical *rules*. In particular, representing classical entailment with defaults makes it possible to interpret logical rules as defeasible norms in human reasoning that hold all other things being equal: "If φ and ψ are believed, and $\varphi \wedge \psi$ is not disbelieved, then conclude $\varphi \wedge \psi$".

Note that here as well as in the general form we mention only believed and disbelieved facts as prerequisites of a default δ and not known facts. The reason to interpret them doxastically rather than epistemically is that we do not want to assume that the initial beliefs are veridical. In other words, we are only here implementing the assumption that in ordinary reasoning we reason from defeasible information. One of the direct advantages of dealing with beliefs, rather than with *knowledge* or *facts*, is to be able to represent inconsistent beliefs. It is possible to reason with inconsistent beliefs because SDL does not rely on logical closure and logical consistency. With this, however, we are already anticipating the semantics of our default logic.

5.4.2 Operational semantics of slow default logic

The semantics of standard default logic is centered around the notion of theory *extensions*.[17] As suggested by Reiter (1980), the intuition behind extensions is that an "extension specifies one coherent view of an incompletely specified world", while "many such coherent views are possible, one for each extension of the default theory". Default reasoning is remarkable for representing conclusions that are best assumed to follow from the existing information, but this relation is stronger than logical entailment. We also represent classical logical rules as default reasoning rules to capture the defeasibility of normative rules in human reasoning. Technically, including such defaults will lead to several changes of a default theory. As a result, the intuitive idea behind the definition of extensions will be changed from a specification of one *coherent* view of

[17] For a fixed-point definition of an extension see (Reiter, 1980, pp. 88-94) and for a method to compute extensions see (Antoniou, 1997, pp. 27-37).

an *incompletely* specified world to a specification of an *actual* view of an *incompletely* and *uncertainly* specified world. Let us now first define our basic default theory and see how to build its extensions.

We will start from a simple default theory $\Delta = (B, D)$ where B is a finite set of propositional formulas and D is an enumerable set of default rules. The set of beliefs B corresponds to the set of initial facts in standard default theories, but with some important differences between the two. Crucially, we do not assume that the set of beliefs B is consistent.[18] Before getting to the definition of default theory extensions, we will define some standard technical notions following Antoniou (1997, pp. 31-32). First we take Π to be a sequence of the defaults $\delta_1, \delta_2, \ldots \in D$ without repetitions. Intuitively, Π is a possible order of applying the list of defaults from D. The initial segment of Π is denoted as $\Pi[k]$ where k stands for the number of elements contained in the segment. In particular, $k = 0$ is the empty sequence. Every segment $\Pi[k]$ is itself also a sequence. With any sequence of defaults Π we associate two sets $In(\Pi)$ and $Out(\Pi)$:

- $In(\Pi) = B \cup \{cons(\delta) \mid \delta \in \Pi\}$;
- $Out(\Pi) = \{\neg\psi \mid \psi \in just(\delta)$ for some $\delta \in \Pi\} \cup \{\xi \mid \psi \in just(\delta)$ for some $\delta \in \Pi$ and ψ is of the form $\neg\xi\}$.

We say that a default δ is *applicable* to the set of formulas $In(\Pi[k])$ iff:

$$\varphi \in In(\Pi[k]) \text{ for all } \varphi \in pre(\delta) \text{ and } \neg\psi \notin In(\Pi[k]) \text{ for all } \psi \in just(\delta) \text{ and } \xi \notin In(\Pi[k]) \text{ for all } \psi \in just(\delta) \text{ of the form } \neg\xi.$$

The set of formulas $Out(\Pi)$ is assumed not to become a part of the belief set B throughout the application of all defaults from Π. On this point, our theory does not depart from standard default theories. However, the new definition of the set $In(\Pi)$ will make a difference to the definition of a default theory extension.

The usual idea behind the set $In(\Pi)$ is to apply the available defaults and keep on classical reasoning on the basis of default conclusions *as long as possible*. That is why the set $In(\Pi)$ is normally closed under classical entailment. Here, we adopt a different idea of reasoning *as long as*

[18] Unlike so called "axioms" or "facts" (see in (Antoniou, 1997, p. 19)) of W in the usual definition of default theories (W, D), beliefs can be inconsistent. We do not want to assume that having a set of beliefs necessarily means having a consistent set of beliefs, because this is often not the case with the non-ideal reasoning that we are modeling here.

required by the application of available defaults. In this way, we are able to look at the normativity of rules for each modeled case of reasoning, which was our central desideratum for the system. Moreover, we avoid reduplicating the effects of classical reasoning in a slow default logic theory Δ. Depending on a definition of D, any formula $\varphi \in Th(In(\Pi[k]))$ could be a formula $cons(\delta_k)$ for some default rule δ_k with classical entailment that has not been applied yet to $In(\Pi[k])$. However, whether a specific classical logic rule is to be represented by defaults in D fully depends on the relevance considerations. As mentioned earlier, while classical logic inference rules are schematic, they give rise to norms that have properties of material inferences. For example, classical norms are context-dependent as a result of relevance considerations. Therefore, if a default theory already represented classical entailment as norms within default rules from D, then we would need to justify further why we assume that deductive closure prompts an infinite sequence (or sequences) of classically entailed formulas to be added to $In(\Pi)$. The two processes would be working at cross-purposes, resulting in a different treatment of classical consequence throughout the same default theory.

To see how this decision affects the logical consequences of a default theory, we first define the notion of a *process* of Δ. A process is a stepwise procedure in which we apply default rules while respecting the beliefs that have been collected thus far. For some sequence of defaults Π, we say that Π is a *process* if and only if δ_k is applicable to $In(\Pi[k])$ for every k such that δ_k is in Π. There are two main properties of default processes we are interested in:

- A process Π is *closed* iff every $\delta \in D$ applicable to $In(\Pi)$ is included in Π;
- A process Π is *successful* iff $In(\Pi) \cap Out(\Pi) = \emptyset$, otherwise it is a *failed* process.

Finally, we are now able to define the extension of a theory Δ.

Definition 5.1 (SDL Extension)**.** *A set of formulas E is an extension of the default theory $\Delta = (B, D)$ iff there is a closed and successful process Π of Δ such that $E = In(\Pi)$.*

The new definition of an extension of Δ is dependent upon the modification of a set $In(\Pi)$. In particular, it does not meet the condition of

(Antoniou, 1997, p. 41) that $Th(E) = E$ or that an extension has to be deductively closed.

To illustrate such default theory, take the following simple example where default rules with classical entailment combine with standard defaults:
$\Delta_0 = (B_0, D_0)$, with $B_0 = \{bird(Tweety), bird(Tweety) \rightarrow \neg fish(Tweety)\}$ and $D = \{\delta_a, \delta_b\}$ with

$$\delta_a = \frac{bird(Tweety) : flies(Tweety)}{flies(Tweety)} \text{ and}$$

$$\delta_b = \frac{bird(Tweety), bird(Tweety) \rightarrow \neg fish(Tweety) : \neg fish(Tweety)}{\neg fish(Tweety)}.$$

The theory Δ_0 has two successful and closed processes, $\Pi_1 = (\delta_a, \delta_b)$ and $\Pi_2 = (\delta_b, \delta_a)$, and a single extension

$$E = \{bird(Tweety), bird(Tweety) \rightarrow \neg fish(Tweety), flies(Tweety), \neg fish(Tweety)\}.$$

What type of reasoning does Δ_0 represent? Consequent $cons(\delta_a)$ is an instance of a plausible conclusion that is typical of situations where we must draw a conclusion despite restricted access to the relevant information. On the other hand, $cons(\delta_b)$ is a default representation of *modus ponens*. In particular, what warrants adding $cons(\delta_b)$ to B_0 are the relevant beliefs from the set $pre(\delta_b)$. The conclusion $cons(\delta_b)$ is thus warranted provided that the set of required beliefs $pre(\delta_b)$ has not been given up throughout the course of reasoning, thereby changing the default theory for which δ_b is an applicable rule. We define the notion of validity for the default theory in the following way:

Definition 5.2 (SDL validity)**.** *Let $\Delta = (B, D)$ be a default theory and φ a propositional formula. Then $\Delta \mid\sim_s \varphi$ iff φ is in all extensions of Δ.*

With Definition 5.2, we define what is called "skeptical consequence" (Antoniou, 1997, 172): something is believed under the condition that it is supported by each line of reasoning available. We will later give examples where the importance of this definition will become more obvious.

5.5 Bridge principles in slow default logic

Slow default logic is a system that is neutral toward weak psychologism. By "neutral" we mean that logical norms are neither favoured, say, by

relying on logical closure, nor limited by the two requirements that limit Harmanian bridge principles. Since classical reasoning is not assumed outside the rules with classical entailment, the system is neutral in the first sense. The second sense of "neutral" is realized through removing both the non-defeasibility and the relevance requirement on normative rules. In this section we present advantages of this neutral stance. We start our analysis of bridge principles by showing that, although non-defeasible, LIM is not a *normative* principle.

5.5.1 A non-defeasible principle LIM

We first show in some details why Harman's "Mary example" does not involve a violation of *modus ponens* and why LIM is in fact a non-defeasible bridge principle. Recall Harman's first objection against the principle LIM. According to him, Mary came to believe that if she opens the closet, she will see a box of Cheerios. She did open the closet, but there was no box of Cheerios to be found there. We are now able to reconstruct Harman's description of LIM within a default theory. Take $\Delta_1 = (B_1, D_1)$ to be a default theory with $B_1 = \{open(Closet) \rightarrow see(Cheerios), open(Closet), \neg see(Cheerios)\}$ and a single default rule

$$\delta_1 = \frac{open(Closet) \rightarrow see(Cheerios), open(Closet) : see(Cheerios)}{see(Cheerios)}.$$

As is obvious, the default rule δ_1 cannot be applied: $just(\delta_1) = see(Cheerios)$ and $\neg see(Cheerios) \in In(( ))$ where "$( )$" stands for the empty sequence of defaults. Therefore, Δ_1 has only one extension, namely $E = B_1$. Apparently, the default theory Δ_1 supports Harman's conclusion that *modus ponens* does not exert any normative force on this particular fragment of Mary's reasoning. Despite her seeming commitment to follow *modus ponens* imposed by the rule δ_1, $cons(\delta_1)$ is not an element of E. The normative role of that logical principle has been compromised by the counterexample Δ_1.

It seems that there is still more to Harman's example than this simple theory shows. The first thing that has to be challenged is that this example consists of a single default theory. For Harman (1986, p. 5), Mary's example gives sufficient support for the claim that if it were the case that Mary followed *modus ponens* expressed with the rule δ_1, then she would be obliged to accept two inconsistent beliefs:

> Mary believes that if she looks in the cupboard, she will see a box of Cheerios. She comes to believe that she is looking in the cupboard and that she does not see a box of Cheerios. At this point, Mary's beliefs are jointly inconsistent and therefore imply any proposition whatsoever.

Harman's conclusion that Mary's beliefs are inconsistent does not follow because his counterexample is ambiguous. In fact, Harman's counterexample can only be modeled with *two* default theories, Δ_2 and Δ_3 (see below), neither of which equals Δ_1 and in neither of which δ_1 is applicable.

To be able to apply the default δ_1 from Δ_1, there are two necessary prerequisites, namely $open(Closet) \rightarrow see(Cheerios)$ and $open(Closet)$. Harman's example, however, does not specify a theory with the two beliefs occurring together within a set of beliefs. What it does specify, leaves us with two possibilities of how Mary could have acquired beliefs about the box of Cheerios, neither of which involves *modus ponens*. The first possibility is just defeasible reasoning from an incomplete set of information before Mary took any action. The second is a change of a default theory resulting from Mary's actions. Harman makes the change of a theory implicit through the example's temporal dimension. Initially, Mary falsely believed in an implication. Subsequent to her looking at the cupboard, she held beliefs in the antecedent and the negation of the consequent of that conditional and, thereafter, ceased to believe the conditional itself. Thus, Harman's conclusion hinges on his false assumption that Mary was actually holding two relevant beliefs together throughout the course of change of the default theories. Accordingly, Harman disregards the effects of information-defeasibility: an incoming piece of information causes Mary to give up some of her previously held beliefs.

We now provide a detailed formal layout of the relevant possibilities. The first observed theory is $\Delta_2 = (B_2, D_2)$ with $B_2 = \{open(Closet) \rightarrow see(Cheerios)\}$. The set of beliefs B_2 does not include the formulas $open(Closet)$ and $\neg see(Cheerios)$ until Mary performs the actions that make them true. For the obvious reasons, the default rule δ_1 is not applicable to the described theory, that is, to the set $In(( ))$. Still, Mary could have formed a belief about Cheerios being in the closet before she was able to check. Thus we still need to explain where that belief comes from since, contrary to Harman, we claim that Mary could not rely on *modus ponens* to obtain the belief that there are Cheerios in the closet.

Relying on her initial belief state, Mary got engaged in defeasible reasoning about Cheerios. Her reasonable expectation to see Cheerios in the closet is represented as the following default rule:

$$\delta_2 = \frac{open(Closet) \rightarrow see(Cheerios) : willsee(Cheerios)}{willsee(Cheerios)}.$$

The rule δ_2 extends Mary's initial belief state B_2 in a way that is typical of practical reasoning tasks. In short, Mary's false belief in the conditional $open(Closet) \rightarrow see(Cheerios)$ grounded her reasonable expectation that she would see the box. However, without actually seeing the box, Mary couldn't have formed a belief that she does see the box: the two propositions, $see(Cheerios)$ and $willsee(Cheerios)$, give rise to different belief states. The default theory Δ_2 with the rule δ_2 contained in D_2 represents the change in her doxastic state prior to opening the closet. The theory now has a successful and closed process in $\Pi = (\delta_2)$ and the corresponding extension $E = B_2 \cup cons(\delta_2)$. The theory Δ_2, however, does not represent an instance of the *modus ponens* reasoning that Harman's criticism targets.[19]

Now consider the set of beliefs B_3 that captures Mary's belief state after she actually opens the closet. Then, Mary's background beliefs have changed because of her observation. To model the change in her view, we need a revised set of background beliefs B_3 without the conditional $open(Closet) \rightarrow see(Cheerios)$. Take $\Delta_3 = (B_3, D_3)$ with $B_3 = \{open(Closet), \neg see(Cheerios)\}$ and consider again adding the controversial rule δ_1 to the set D_3. We immediately see that the default theory Δ_3 would not result in any process because δ_1 is inapplicable to $In(( ))$. The application of that rule has been blocked by the belief in $\neg see(Cheerios)$. Accordingly, Δ_3 does not comply to the scenario of the counterexample to *modus ponens* described by the default theory Δ_1. Therefore, the alleged violation of *modus ponens* results from disregarding the change in Mary's background beliefs and conflating two different default theories Δ_2 and Δ_3 into a single one, namely Δ_1.

[19]Standard deafult rules in our theory may be plausibly held as a projection of practical interests in reasoning. According to Harman (1986, p. 2), in reasoned revision "one should make minimal changes in one's view that increase its coherence as much as possible while promising suitable satisfaction of one's ends". A suitable satisfaction of one's ends sometimes might be impossible without non-deductive defeasible reasoning towards the needed conclusion. The very fact that we are often directed towards forms of reasoning that differ from deductive reasoning is just another argument against strong psychologism.

What was eventually shown by rejecting Harman's counterexample to LIM? Recall that LIM modestly claims that logical entailment *can* be a reason to accept an entailed proposition. Even if Harman was right about his counterexample to the applicability of *modus ponens*, this would not prove the principle LIM to be false. As Knorpp (1997, p. 87) aptly argues, to show that LIM is false, Harman would need to show that when one's view logically entails X, then this *cannot* be a reason to accept X. But this is not the claim that he argues for. In fact, Harman accepts logical rules as *prima facie* rules in reasoning and, thereby, the claim that our beliefs can be founded on logical entailment. Does that mean that Harman himself suggests LIM as a valid, non-defeasible principle?

5.5.2 The relevance problem of LCP

The principle LIM meets Harman's non-defeasibility requirement, but this comes at the cost of not being connected to the weak psychologism thesis at all. To see why, recall from the Introduction that weak psychologism is a view according to which there is a prescriptive side to logical norms. As a consequence of violating a prescriptive norm, an agent is rightly deemed as rationally culpable and we simply say that the agent is committing a reasoning mistake. Now, consider again LIM and the claim that logical entailment *can be* a reason to accept a proposition. The principle LIM does not impose any obligation on an agent's doxastic perspective and, therefore, whatever the outcome of a reasoning process, an agent cannot be mistaken even in systematically dodging to follow logical norms. The LIM principle claims only that logical rules can legitimately ground our beliefs. This is not controversial, but only because it is not *normative* either.

The problem with the *defeasible* operators of the type "can be a reason", "have reasons to" or "has (defeasible) reason for" is that they do not directly bear on the standards of rightdoing and wrongdoing. Bridge principles based on any such operator do not prescribe "what one ought to do" neither they provide "a standard for the evaluation of one's conduct as good or bad", both of which are defining features of normative laws (MacFarlane, 2002, 35). Having a reason to perform an action is compatible with it not being the case that you ought to perform that action and even with the case where you are not allowed to. One cannot use the mentioned operators as deontic ones without a further qualification that a reason is in fact a non-defeated or a perfect one, thereby imposing a

normative requirement.[20] Most importantly for the normativity debate, by devising a principle with this kind of operator, we meet Harman's challenge merely by avoiding the central problem of the debate, which is whether logical rules commit human agents to follow logical rules in their belief formation or not.[21]

A non-defeasible principle with a defeasible operator such as "can be a reason", even if true, is irrelevant for the weak psychologism thesis. Weak psychologism requires prescriptive normative principles. Accordingly, a bridge principle which articulates the normative role of logic in human reasoning needs to feature an operator that qualifies as typically deontic. Note here that this eliminates some existing bridge principles, e.g. Steinberger's (2019a) and MacFarlane's (2004), based on the operator "have a (defeasible) reason to" for which Steinberger (2019b) argues that they are *defeasible norms*. According to our analysis, these principles are neither defeasible nor do they express norms related to weak psychologism.[22]

This directs us toward a stronger, prescriptive, normative requirement proposed in LCP. Unlike LIM, however, any prescriptive principle with a deontic operator that is strong enough to express the weak psychologism thesis will only be a defeasible rule. As we argue in Section 5.3, information-defeasibility causes normative principles of the type *you ought to add* X *to your set of beliefs* to be defeasible requirements. Above that, LCP is simply wrong because it fails to meet the relevance requirement. With the help of slow default logic, we now show in detail how LCP fails to meet the relevance requirement.

[20]This corresponds to the standard vocabulary of deontic logics where permission, restriction and prescription are considered (Åqvist, 1984, von Wright, 1951). Note that we are not denying here the importance of *pro tanto* reasons, that is, those reasons that support only to a certain extent, for normativity. Reasons may even form a major part of normative considerations, but the bridge principle formula is a way to set up the normative standard (if any) which constrains human reasoning, not only to acknowledge that logical entailment can be taken into account.

[21]This is our way of interpreting why the "Strictness test" Steinberger (2017) goes against "have a reason to" bridge principles.

[22]Steinberger (2019b) argues for the difference between three normative roles that logic has in human reasoning. Unlike "objective" *evaluations*, *directives* and *appraisals* take into account an agent's perspective and its ability to live up to logical standards. His take on what a bridge principle is supposed to look like articulates a different role for logic from the traditional prescriptive role, which is central to weak psychologism. In particular, he is interested in directives or "first-personal norms that offer *advice of a sort a person can take*" (Steinberger, 2019a).

Harman (1986, p. 12) suggested LCP as a modified version of the LIM principle that avoids its shortcomings. A closer analysis reveals how LIM and LCP advise two utterly different positions on what is the role of logic in human reasoning. Recall that according to LCP, if a proposition is entailed by our beliefs, we should add the proposition or revise the entailing beliefs when faced with their falsity. Let us translate this demand into an adequate default theory. We define a theory $\Delta_4 = (B_4, D_4)$, where B_4 is a finite set of beliefs and $D_4 = \{\delta_m, \delta_n, \ldots\}$ has countably many defaults. Since in SDL the *In*-set is not deductively closed, D_4 contains all the possible rules with whatever is entailed by the deductive closure of B_4. The theory Δ_4 specifies infinitely many closed and successful processes and, provided that B_4 is consistent, a single extension containing the deductive closure of a set B_4. How to interpret the normative requirements specified by the theory Δ_4?

The theory Δ_4 is an example of extreme normative requirements. From the perspective of weak psychologism, Δ_4 describes the reasoning one has to carry out if asked to find out *whatever* is true according to one's beliefs. It is then, according to the "reasoning task" Δ_4, normatively required that whatever is entailed by that agent's belief set also needs to be included in that set and if the agent discovers that B needs to be revised, then the agent would face equally demanding logical commitments arising from a different default theory. It is highly improbable that any agent ever faces such unreasonable requirements. Together with Harman, we reject the plausibility of LCP, but with a different conclusion drawn from it.

For Harman, the rejection of LCP paves the way to advance his anti-psychologism regarding logic. In contrast, we deny that Δ_4 articulates weak psychologism. Assume, for the sake of argument, that Δ_4 correctly articulates the weak psychologism thesis. As a consequence, if we started from the belief set B_4, then the same subset of rules would be normative for us, regardless of what the actual reasoning task is that the theory describes. But this is impossible. The example of the theory Δ_4 shows that LCP requirements overlap only with the "pathological" reasoning task where one is asked to find out whatever is entailed by what he believes to be true. There are no reasons, however, to think that Δ_4 uniquely specifies normative requirements for any other possible reasoning task as LCP implies.[23]

[23] As argued in Section 5.3.2, to claim that relevant entailments are relative to a given reasoning task is not bringing in the same issue as Harman's (1986, p. 12) advice to avoid

Any prescriptive Harmanian bridge principle candidate would prove itself to be inadequate for the sensitive task of specifying only the relevant entailments of a set of beliefs. That being so, each SDL theory resulting from a prescriptive bridge principle is equally naive as Δ_4. This is the way in which the relevance problem hinders the format of bridge principles. In order to avoid the relevance problem, weak psychologism deserves to be judged without Harmanian principles. In our system the question of relevance is relegated to the level of the designers's input of a set D. This decision enables us to look at a variety of different reasoning tasks, each with their own normative rules.

5.5.3 Harman's objection to the principle LIN

Recall the principle LIN discussed in Section 5.3. This principle says that in ordinary reasoning, one is to avoid logical inconsistency. Harman warns us that the normative status of this principle is threatened because the principle cannot be applied to cases of reasoning with defeasible premises. Imagine you find out that your beliefs are inconsistent and you have no available updates to resolve the inconsistency. Harman advises us that in "that case the best response may be to keep the inconsistency and try to avoid inferences that exploit it". But we still might be interested in how we reason from inconsistent premises and why to think that logical norms stop being normative in doing so. We follow Harman's line of argument to see which default theory formalizes a typical case of dealing with inconsistency.

Let us take a default theory $\Delta_5 = (B_5, D_5)$ with $B_5 = \{p, q, p \to r, q \to \neg r\}$ and $D_5 = \{\delta_3, \delta_4\}$. The defaults δ_3 and δ_4 are defined as follows:

$$\delta_3 = \frac{p \to r, p : r}{r} \text{ and } \delta_4 = \frac{q \to \neg r, q : \neg r}{\neg r}.$$

Δ_5 has two processes, $\Pi_1 = (\delta_3)$ and $\Pi_2 = (\delta_4)$. It is easy to check why neither $\Pi_3 = (\delta_3, \delta_4)$ nor $\Pi_4 = (\delta_4, \delta_3)$ are processes of Δ_5: once δ_3 has been applied, the conclusion r is added in $In(\delta_3)$, then δ_4 cannot be applied because r directly contradicts the justification $\neg r$. The theory Δ_5 has two extensions. The first extension is $E_1 = \{p, q, p \to r, q \to \neg r, r\}$

cluttering our mind with trivialities. The cluttering avoidance issue is a practical advice that becomes redundant on a correct interpretation of the weak psychologism position. Any specific default rule representing a logical norm for reasoning is included in D only if it is related to the reasoning task at hand.

and the second extension is $E_2 = \{p, q, p \to r, q \to \neg r, \neg r\}$. The two extensions show that we can still go on reasoning even upon collecting the contradicting information r and $\neg r$ from the applied defaults δ_3 and δ_4. Yet this example does not support Harman's conclusion that at this point logical rules cease to be normative. On the contrary, reasoning with inconsistent beliefs still needs to be done in accordance with logical rules.

Facing irresolvable inconsistencies inevitably results from applying logical rules to defeasible premises. Eventually most of the conflicts between inconsistent sources get resolved by an update that gives precedence to one of them. Thus by removing a culprit, say p from the set B_5 of Δ_5, δ_3 becomes defeated. As we argued in Section 5.3, defeasibility of premises does not bear on the question of whether logical rules are normative or not. For Harman, however, a conflict of logical norms and defeated logical norms are both incompatible with the normativity thesis. This is a result of his non-defeasibility requirement.

In assuming non-defeasibility, Harman overlooks that logical rules retain their normative status regardless of how (un)certain or (in)consistent information may be. What Harman (1984, p. 108) suggests is that on discovering irresolvable inconsistency, "one should (at least sometimes) simply acquiesce in the contradiction while trying to keep it fairly isolated". Again, his argument is imprecise: it is not clear what exactly does "acquiescing in the contradiction" mean. We suggest that the act of acquiescing in the contradiction still makes it possible to continue reasoning on the basis of contradictory information but without holding contradictory beliefs. Harman's suggestion leaves the possibility of a different interpretation of "acquiescing in the contradiction": to hold contradictory beliefs and not to reason further with the contradiction. This is why we need to give more precise examples to see what role logic has in dealing with inconsistency and contradiction.

The theory Δ_5 exemplifies only a specific aspect of the normativity thesis, namely the *restrictive* role of logic in reasoning. According to it, we are not supposed to hold contradictory beliefs. Formally, this aspect was captured by the inability to apply default rules that would result in extensions containing contradictions. The restriction to avoid believing in contradictions is the reason why Π_3 and Π_4 are not processes of Δ_5. However, this does not entail that we are not able to reason further with contradicting information, as shown by the two extensions of Δ_5. We concede Harman's point that we are sometimes bound to reason with inconsistency, and even contradictions that need to be kept isolated. But

in keeping the contradiction isolated, we are again bound to follow logical principles.[24] The restrictive role of logic would therefore be more aptly expressed in the following way: "Avoid directly contradicting beliefs", instead of "Avoid inconsistency". There are, however, further reasons to be cautious in formulating an exceptionless restrictive bridge principle since even avoiding contradicting beliefs could turn out to be a defeasible imperative as well.

5.5.4 Rational inconsistencies

Recently, the normativity debate has been strongly influenced by the possibility of rationally holding inconsistent beliefs. Authors who take Harman's skeptical challenge as a starting point, including MacFarlane (2004) and Steinberger (2017), consider the possibility of true contradictions and the view called "dialetheism". It is hard to say what are the merits of dialetheism for the normativity debate. It was mainly motivated by the paradoxes of self-reference involving some abstract concepts, such as the notion of set or the notion of semantic definition of truth (Priest and Berto, 2017). These concepts at best belong to the peripheral cases of human reasoning and stretch the scope of what we are interested in here. For one thing, there is a striking difference between the cases of reasoning when one is to reason about an upcoming breakfast and the type of reasoning that we find in, say, "Russell's paradox". It would be surprising, to say the least, if Mary acted on the grounds of entertaining both a belief that she sees Cheerios and a belief that she does not see Cheerios. In any case, it is not clear that the semantical and logical paradoxes present us with reasons for believing contradictory everyday statements.

A similar line of reasoning extends to a group of arguments that attracted the attention of philosophers in the normativity debate. We will call them "arguments from rational inconsistencies" since they start from a premise that holding inconsistent or even contradictory beliefs is not only tolerable, but also normatively required in certain circumstances. These arguments are based on a discovery of a paradox, as it is the case with the arguments for dialetheism above. Our intention here is not to argue that their main premise is wrong. On the contrary, given the complexity of all the possible tasks subsumed under "human reasoning",

[24]Note that any fear of the "Principle of explosion" effects on normative rules becomes unfounded since SDL rules avoid problems of the relevance requirement.

it should be expected there are such circumstances in which one cannot avoid self-referentiality, arbitrariness or, otherwise, any sort of vagueness that is built in a proposition. The problem, however, is to argue that the possibility of creating a context in which one rationally accepts inconsistencies overthrows the normative role of logical rules.

A good number of these arguments rely on a version of the "Preface paradox" (Makinson, 1965).[25] A brief version of the paradox is given by Harman (1986, p. 16):

> For example, there is the sort of inconsistency that arises when one believes that not all one's beliefs could be true. One might well be justified in continuing to believe that and each of one's other beliefs as well.

Together with the objection we mentioned in Section 5.3, Harman takes this example to show that the principle LIN should be abandoned. For Steinberger (2019a, p. 16), the preface paradox is one of the cardinal objections to the normative role of logic, "which spells so much trouble for qualitative bridge principles". These arguments are examples of the *ignoratio elenchi* fallacy. In what follows, we explain the origins of the fallacy in more detail.

The discovery of an extreme example where the rule is not applicable does not support the conclusion that LIN is not normative. What it does support is that there are normal cases of application and non-normal cases where a normative rule might be overridden. In the preface paradox case, one is faced with an extreme situation where LIN conflicts with an obligation to adopt a belief about the totality of one's beliefs.[26] What follows from this case is that the complexity of belief systems is such that it allows for the quandaries of the preface scenario and that human reasoning lacks the expected uniformity. However, nothing more than this follows. To think that the preface paradox or dialetheias overthrow the normative role of LIN is a part of what Foley (1992, p. 120) calls "the unfounded worry that if consistencies are allowed anywhere they will have to follow everywhere". This applies to all the mentioned arguments from rational inconsistencies. Therefore, Harman's and Steinberger's skepticism is unfounded and their conclusions do not follow from any specific extreme example of a human reasoning

[25] Chapter 6 discusses the preface paradox in the context of a more general problem of how rational, but fallible, agents should state their doxastic modesty.

[26] Evnine (1999), however, argues against the rationality of the preface sentence.

task. However, the examples show that LIN cannot be considered as a non-defeasible restriction for *any* possible instance of a human reasoning task.

Such conclusion about rational inconsistencies is anticipated in the argument from Section 5.3 where we advocated a system of normative rules based on default logic. Again, an analogy with deontic rules illustrates the point we make about the case of rational inconsistencies. Take a paradigmatic moral restriction that says "Thou shalt not kill!" and a scenario where one is faced with a fierce murderer in the midst of a killing spree. Without any doubts, many would be compelled to advise that one should violate the imperative of not killing. In contrast, it is irrational to question the normative role of the restriction not to kill in normal cases only on account of the killing spree scenario. Instead, we should accept that there are such peripheral cases where the rule application is not clearly prescribed or may be otherwise judged as inadequate. If anything, this is the lesson about human reasoning that we learn from the preface scenario and similar cases.

5.6 A positive account of weak psychologism

In this section, we support the claim that classical logic is normative for human reasoning. Instead of altering the weak psychologism claim so that classical entailment provides only probabilistic constrains on degrees of belief (Field, 2009, p. 258) or first-personal norms that offer advice of a sort a person can take (Steinberger, 2019a, p. 318), we present a view that logic is normative in the traditional prescriptive sense of providing norms that one *ought to* follow.

5.6.1 Alternative notions of logical entailment

Up to now, we have argued against Harman's skeptical challenge to the normative role of logic in human reasoning. Although presenting Harman's counterexamples in SDL theories mitigates the skeptical attack on logic's role, the SDL system is in principle neutral about what principles can be plugged into it. To model Harman's counterexamples to the normativity of classical logic, SDL represents both classical reasoning and defeasible reasoning in the format of default rules. However, the critique of Harman's arguments does not on itself give a positive account

for the claim that *classical* logic is normative for human reasoning.

The decision on which notion of entailment should be plugged into the normative principles we represent in SDL theories seems to be a necessary reflection of the most convincing arguments in favor of some particular notion of logical entailment. In particular, the claim that some specific logical principles are normative should be tied up with the claim that there is a specific notion of entailment that corresponds best to regularities in the world. This issue belongs to the debate on logical pluralism where the traditional role of classical entailment has been reevaluated against many non-classical notions of entailment. The main criticism that proponents of classical logic face is that there are other notions of reasoning, captured by different notions of entailment, such that they outperform classical reasoning for some aspects of reality that classical entailment misrepresents.

Answering the criticism raised by logical pluralists would require a thorough discussion of the alternative notions of logical entailment that goes beyond the purpose of the present chapter. The purpose of this chapter is to show that Harman's skeptical challenge fails in the basic way of proposing the inadequate format of non-defeasible bridge principles between logic and reasoning. That being so, we propose two arguments that relate the choice of classical entailment for SDL rules with the notion of correct entailment. However, this holds only for the specific context of the normativity debate we are discussing here.

First, in the normativity debate we are concerned with the practice of human reasoning which is constrained by normalcy. As we saw on the example of rational inconsistencies, not every aspect of human reasoning has to be clearly regimented by classical entailment to claim that classical logic is normative for human reasoning. This question is again closely related to the defeasible nature of norms and their application. While most reasoning tasks as the married-unmarried puzzle follow the standard of correctness posited by classical entailment, there are cases where the standard is not clearly prescribed. However, the crucial point here is that it is impossible to make sense out of the idea that most reasoning cases do not comply with a standard and that clear-cut cases cannot be assessed against any standard. Here we follow the tradition of differentiating between normal and abnormal cases of rule application that stems from Wittgenstein (1953/1958) and Geach (1956).[27] According

[27]This line of reasoning has been recently adopted by Milne (2009, p. 295): "Public rule-

to it, if one is not able to correctly claim that only 13% of participants gave the correct answer to the married-unmarried puzzle, the practice of reasoning would become meaningless and would collapse. Therefore, by the normalcy of rule application in human reasoning, arguments for non-classical notions of entailment are not central to the normativity debate, given the centrality of classical entailment for assessment of human reasoning.

Another argument comes from surveying the motivation for adopting a different notion of entailment. Our claim is that once we set apart the two requirements for normative rules that impede the format of bridge principles, we remove most of the motivation for adopting an alternative logic. For example, since the relevance requirement does not affect SDL theories, the initial problems that motivate the adoption of paraconsistent logics are alleviated. The most straightforward example is that of motivating relevance logics. The problem because of which relevance assumptions are introduced are directly eliminated with the treatment of rules in SDL. Furthermore, since the property of explosion does not occur in a system where the relevance problem is taken into account, most of the motivation for many-valued logics has already been successfully dealt with.[28] Aside from paraconsistent logics, some logics that are grouped with non-classical logics already built on a motivation to represent obligation to believe a statement (modal logic) or to model justification or provability of a statement and, thereby, obligation to believe a statement (intuitionistic logic). These examples of non-classical logics are therefore better not considered alongside classical entailment.

The arguments offered above are by no means conclusive of what is the correct notion of entailment in general, if there is such logic at all. To answer that question we would also have to consider the arguments for empirical adequacy of classical logic, to mention only one traditional problem connected to it. For now, we leave these questions aside.

---

or convention-governed practices do not need rules to cover every possibility, nor even conventions on the setting of precedents. And semantic paradox does not stalk the land laying waste our every discourse". In §141 and §142, Wittgenstein (1953/1958) focuses on normalcy criteria for a rule-governed character of language: "And if things were quite different from what they actually are — (...) if rule became exception and exception rule; or if both became phenomena of roughly equal frequency – this would make our normal language-games lose their point".

[28] Steinberger (2016) discusses whether there are bridge principles that could be "normative" arguments for paraconsistent logics.

5.6.2 Weak psychologism without bridge principles

An opponent of weak psychologism might still be unsatisfied by our choice of the SDL theories up to now. After all, we claimed at the beginning that there is a prescriptive role that logic has in belief formation. Yet, the only example of prescriptivity we gave is the theory Δ_4, which is a highly non-representative reasoning task. Fortunately, there is an abundance of reasoning tasks that can be modeled in SDL to clearly present the *prescriptive* role of logic in human reasoning. To round off the chapter, our decision falls on the married-unmarried example from Section 5.2 since it exemplifies both the prescriptive and restrictive normative role of logic in a single reasoning task.

Take the theory $\Delta_6 = (B_6, D_6)$ with $B_6 = \{M(j), \neg M(g), L(j,a), L(a,g)\}$ and $D_6 = \{\delta_5, \delta_6, \delta_7, \delta_8\}$. The constants a, g and j stand for *Ann*, *George* and *Jack* respectively, M is a unary predicate standing for the property *being married* and L stands for the binary relation *looking at*. The default rules are defined as follows:

$$\delta_5 = \frac{\mathit{empty} : M(a)}{M(a)}, \delta_6 = \frac{\mathit{empty} : \neg M(a)}{\neg M(a)},$$

$$\delta_7 = \frac{M(j), \neg M(a), L(j,a) : (M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))}{(M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))}$$

and

$$\delta_8 = \frac{M(a), \neg M(g), L(a,g) : (M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))}{(M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))}.$$

Here is an informal description of reasoning steps described by Δ_6. We start with an incomplete set of information as to whether Anne is married or not. However, to find out whether there is a married person looking at an unmarried one, we are determined to reason with the propositions describing Anne's current status. From the theory Δ_6 we learn that, according to D_6, we are not able to decide whether $M(a)$ or $\neg M(a)$ holds according to B_6. Instead of that, we are only able to do an ordinary default step δ_5 or δ_6, still leaving B_6 underdetermined as to whether the relevant proposition holds according to B_6. Despite that, we are now able to reason further on two branches of default processes.

More formally, Δ_6 has two closed and successful processes, namely $\Pi_1 = (\delta_5, \delta_8)$ and $\Pi_2 = (\delta_6, \delta_7)$. First, we consider the only two applicable

defaults to $In(\,)$: δ_5 and δ_6. After we apply one of them, say the default δ_5, δ_6 becomes inapplicable to $In(\delta_5)$ since $just(\delta_6) = \neg M(a)$ and $M(a) \in In(\delta_5)$. Now the only applicable default that is left to be applied to $In(\delta_5)$ is δ_8. Similar reasoning applies to the default process Π_2. Now we know that we have exactly two extensions of Δ_6:

$$E_1 = B \cup \{M(a), (M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))\}$$

$$E_2 = B \cup \{\neg M(a), (M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))\}$$

According to Definition 3.2, the only formula that needs to be added is the disjunction $(M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))$, corresponding to the correct solution of the informal example description from Section 5.2, namely that a married person is looking at an unmarried one. While we still remain skeptical about which of $M(a)$ or $\neg M(a)$ holds, we take the following conclusion as a valid one:

$$\Delta_6 \mid\sim_s (M(j) \wedge \neg M(a) \wedge L(j,a)) \vee (M(a) \wedge \neg M(g) \wedge L(a,g))$$

With the theory Δ_6 we show that logical rules do not only restrict, but also prescribe reasoning steps. Were we to solve the Married-unmarried puzzle, we would be advised to reason from contradictory assumptions. All the while our reasoning needs to follow logical rules in finding out the right answer. Moreover, we are now able to see why the set D_6 does not contain an infinite number of rules and what the benefits are of denying the relevance requirement. In particular, why not to include a default rule such as

$$\delta_9 = \frac{\neg M(g) : \neg M(g) \vee L(g,a)}{\neg M(g) \vee L(g,a)}$$

to the set of rules D_6? The rule δ_9 would extend our beliefs by a valid conclusion to a proposition containing the disjunct $L(g,a)$ we read as "George is looking at Ann". The answer is that the theory Δ_6 represents the reasoning task that one needs to perform to solve the Married-unmarried puzzle. The puzzle is not asking about who is George looking at and we are not required to add this conclusion to our set of beliefs. This entailment surely *can* be a reason to add the $cons(\delta_9)$ as the principle LIM suggests. But the rule δ_9 is best considered a part of a different default theory where reasoning is in fact about $cons(\delta_9)$. Similar argument applies to an infinite number of other defaults that could be generated by logical entailment closure over B_6.

5.7 Conclusions

Our goal in this chapter is to defend weak psychologism as saying that classical logical principles are prescriptive and, thereby, normative for human reasoning. Harman's skeptical challenge to the normativity of logic set the standard of Harmanian bridge principles for the further debate. On a closer look, we uncover two requirements that make bridge principles impossible, namely the non-defeasibility requirement and the relevance requirement. We develop slow default logic, which is capable of representing every rule, including classical entailment rules, within the form of a default rule. In this way we amend the negative effects of the non-defeasibility requirement on normative principles. To deal with the relevance problem, the choice of rules that an agent has to follow is made a part of a system designer's specification of a set D.

In Section 5.5, we present models that challenge Harmanian bridge principles LIM, LCP and LIN. We show that the shortcomings of these principles do not challenge the weak psychologism thesis. Among the suggested principles, LCP is the closest articulation of how logical principles could be prescriptive for belief formation. However, this principle is undercut by the assumption that one needs to solve the relevance problem to answer the normativity question. Our conclusion is that a single overarching Harmanian bridge principle is impossible and, above that, irrelevant to weak psychologism.

Once we avoid ramifications of the Harmanian bridge principle format, we are able to vindicate Frege's claim that classical logic is prescriptive for human reasoning. According to our analysis, the best approximation of the normativity thesis based on a simplistic principle is the following one:

> *For some specific propositional formulas $P_1, \ldots, P_n$ and Q, if $P_1, \ldots, P_n \models Q$ and Q is relevant to an actual reasoning task, then if one believes that $P_1, \ldots, P_n$ in this context, one should, other things being equal, add Q to one's set of beliefs*.

But in formulating this approximation, we do much less than if we systematically look for the normative role of logic in actual human cognition tasks. Only in taking a case-by-case perspective can we gain insight into how logical rules have the obligatory character that is elusive to any Harmanian bridge principle. The major advantage of such perspective is

that instead of a gap between logic and reasoning, we see a woven web of defeasible connections between the two.

Chapter 6

On modest reasoners who believe that they believe falsely

One can mistrust one's own senses, but not one's own belief. If there were a verb meaning 'to believe falsely', it would not have any significant first person present indicative.

—Wittgenstein (1953/1958, p. 190,§ X)

6.1 Introduction

In this chapter, we focus on an important feature of agents who engage in commonsense reasoning, namely, their fallibility. More specifically, we start from the premise that rational agents who are aware of their own mistakes and mistakes of their fallible peers have good reasons to be doxastically modest. One of the most interesting questions that a modest, fallible agent is faced with is "What beliefs about my own (fallible) beliefs should I hold, given doxastic modesty?". Some agents may interpret doxastic modesty as a requirement to believe that they hold some false beliefs. In this chapter, we argue that an agent cannot in principle form a belief in the statement "At least one of my beliefs is false". Once an agent has learned this statement, the agent is not committed to the same belief any more. Agents encounter a problem of the same kind

when learning Moorean statements. We explain this analogy in detail and argue that learning the statement "At least one of my beliefs is false" is not *successful*, that is, once it has been learned, it should not believed any more. The same diagnosis of the problem is accepted for Mooorean statements. We examine two additional versions of doxastic modesty statements that could avoid the problem of unsuccessful learning. In these two versions, agents can refer to their totality of beliefs slightly differently and, thereby, still be committed to believe that they hold false beliefs. We argue that each of the two *ad hoc* solutions to the unsuccessful learning problem that we discuss is controversial. We instead suggest that doxastic modesty justifies suspension of the belief in the conjunction of one's beliefs as well as believing more modest statements that do not commit one to believe falsely. Finally, we propose that the connections between the general doxastic modesty statements and Makinson's (1965) "Paradox of the Preface" are not as straightforward as it has been usually assumed in the debate.

6.2 Doxastic modesty statements

As a rational human being, you are aware of your own past beliefs that turned out to be false. Moreover, you know that other rational humans also change their beliefs as a result of discovering that they have believed falsely. Therefore, you have good reasons to be modest about your present beliefs, because these could turn out to be false as well. But how exactly to take your fallibility into account in order to state that you are doxastically modest? One of the possible ways is to uncontroversially state that "Some of my beliefs may turn out to be false". You might also be convinced that doxastic modesty requires you to go further in your Doxastic Modesty Statement (DMS) and claim that:

(1) "At least one of my beliefs is false".

The DMS (1) is inconsistent with your other beliefs, which makes the belief in (1) controversial. It is usually held that, after you learn (1), you are justified in holding jointly incompatible beliefs: doxastic modesty justifies your belief in (1), while it does not disqualify your reasons to believe each of your other beliefs.[1]

[1]By "learning" we mean nothing more than acquiring beliefs and we do not assume that learning is factive. As an obvious example, an agent can learn a false statement and,

Contrary to the accepted view, we will argue that after learning (1), your beliefs are not jointly inconsistent because what you believe is not the proposition (1).[2] What we show instead is that (1) contradicts the following reasonable expectation on learning a proposition:

- After learning a proposition P, P is believed.

This is the so called "Success axiom" that is known from belief revision (Alchourrón et al., 1985).

Then we will show that it is possible to circumvent the problem of learning (1) by referring to the totality of your beliefs in a slightly different manner. Namely, you could state in DMS that DMS is itself fallible as one of your beliefs or you could simply state that it is the other beliefs that are fallible. These considerations open up two additional DMS candidates, which claim that at least one of your beliefs is false:

(2) "At least one of my beliefs other than this one is false" or

(3) "At least one of my beliefs including this one is false".

We will argue that these two solutions to the problem of learning (1) result in different problems that undermine the rationality of believing (2) and (3). The statement (2) introduces an *ad hoc* distinction among your beliefs which cannot be adequately accounted for. On the other hand, believing the statement (3) commits you to believe the denial of (3) as well.

It is usual in the literature on doxastic modesty to assume that the following *conjunction* principle is the strongest argument against the above statements:[3]

- If you believe that A and you believe that B, you should believe that $A \wedge B$.

The main purpose of this chapter is to show that believing any one of the statements (1)-(3) comes with problems that are specific to the statements

thereby, acquire a false belief. Therefore, learning does not entail knowledge.

[2]For the sake of argument, we assume that your beliefs were not jointly inconsistent already before learning (1).

[3]Most authors argue against this principle. Among them are Kyburg (1970), Foley (1979, p. 249), Williams (1987, p. 121) and Christensen (2004, p. 36). Kyburg (1970) introduced the term "conjunctivitis" as a pejorative term for the principle. Evnine (1999) is one of the defenders of the disputed principle.

themselves, independently of whether we accept that our beliefs are closed under conjunction or not. We do not, however, argue that one should conceitedly believe that all of one's beliefs are true, but only that one cannot believe (1) and that one should abstain from believing (2)-(3) as well. The focus in the literature on the conjunction principle is, in our view, misleading and the rationality of believing the statements above can be questioned on different grounds.

The chapter has the following structure. First, we introduce the notation used throughout the chapter. Then we investigate the three possibilities to formulate DMS and explain what are the problems of believing them. After showing that one cannot in principle believe (1), we turn to the suggestions that avoid the problems of (1). The statement (2) avoids the problem at the price of not being adequately supported by evidence that grounds the requirement to be doxastically modest. While the statement (3) is commonly discarded for its connection with the semantic paradoxes, we analyze the problem of believing (3) in terms of paradoxical commitments of a belief set that contains (3). In Section 6.4, we argue that doxastic modesty has to be seen as a result of a certain kind of higher-order evidence that is best understood as a requirement to suspend a belief in the conjunction of one's beliefs. Then, in Section 6.5, we connect general doxastic modesty statements with prefatorial book statements, first proposed by Makinson (1965). Prefatorial statements are doxastic modesty statements particular to book prefaces. We argue that the rationality of prefatorial statements might not be entirely assessable in the light of the DMS problems. This is mainly because book statements and preface statements may come with assumptions that do not hold for ordinary beliefs. Finally, we follow Evnine (2001) to suggest an alternative DMS formulation that does not commit you to hold false beliefs.

6.3 Three problems of doxastic modesty statements

To disambiguate the propositional content of the statements (1), (2) and (3), we use elementary set notation. Say that you hold some number n of beliefs. Your current set of explicit beliefs *Bel* is then defined as follows:

$$\{P_1, P_2, \ldots, P_n\} = Bel,$$

where for any k such that $1 \leq k \leq n$, P_k is one of the propositions you explicitly believe. Whatever version of DMS is discussed below, we

assume that you reason from the belief set *Bel*.

An immediate question about the belief set is whether we should talk about deductively closed sets and, therefore, infinite sets, instead of talking about finite sets of explicit beliefs. We propose that it is reasonable to focus on explicit beliefs in the context of the doxastic modesty debate. Taking credit or blame for your right and mistaken beliefs relates only to your explicitly held beliefs and these form a finite set. Consider that although there are infinitely many mathematical truths that are entailed by your current beliefs, some of them take a lot of effort until you can reach an explicit belief in them. It makes no sense to take credit for truthfully believing any of these propositions just because they are entailed by your beliefs. On the other hand, mistaken beliefs could be detected on the grounds of showing that they entail a contradiction, but the blame is attached to your explicit beliefs that need to be changed, not to the entailed contradiction.[4] We can say that in judging that a belief is mistaken we adhere to the variant of the legal principle *Nullum crimen sine actu*, where the required act is interpreted as explicitly holding the belief.

6.3.1 Case 1: Unsuccessful learning

We first argue that adding a belief in the DMS (1) (see p. 152) prompts its own revision. By adding (1) to your set of beliefs, the totality of your beliefs changes. Since (1) claims fallibility of your beliefs with reference to the totality of your beliefs, learning (1) results in believing that (1) is fallible as well. That is, instead of believing that at least one of your beliefs is false, you believe that:

(1′) "At least one of my beliefs is false or all my beliefs are true".

Learning the statement (1) is thus self-undermining: the statement rules out the possibility of all your beliefs being true, but (1) is itself one of those beliefs.

To see why learning (1) fails, one needs to consider the change of the scope of your beliefs. Your initial state of beliefs is *Bel*, as defined above.

[4]Williams (1981, p. 601) makes a similar point: "It would be ludicrous to credit someone with beliefs of things of which he had never heard or of which he had no understanding". Notice that, in the logic of belief revision, the term "belief set" is taken to imply that a set is deductively closed, while "belief bases" are not necessarily deductively closed.

The DMS candidate (1) corresponds to the proposition:

$$P_{n+1} = \sim P_1 \vee \sim P_2 \vee \cdots \vee \sim P_n.$$

If P_{n+1} is added to *Bel*, your new set of beliefs is now described by

$$\{P_1, P_2, \ldots, P_n, P_{n+1}\} = Bel^1.$$

To capture the revision of (1) with (1′) precisely, we look at the change of the scope of your beliefs after adding P_{n+1}. This change causes the revision of the belief in (1), since (1) quantifies over your entire set of beliefs. Before learning (1), the proposition P_{n+1} refers to your entire set of beliefs. However, after P_{n+1} itself is added to your set of beliefs, the proposition corresponding to DMS has to be revised accordingly. What you now believe is that P_{n+1}, as one of your beliefs, is also among the propositions of which you claim that at least one is false. Instead of believing P_{n+1}, learning (1) results in believing

$$P_{n+2} = \sim P_1 \vee \sim P_2 \vee \cdots \vee \sim P_n \vee \sim(\sim P_1 \vee \sim P_2 \vee \cdots \vee \sim P_n),$$

a tautology which corresponds to (1′). Therefore, you eventually learn nothing more than that the DMS statement (1) is itself fallible, as any other belief. This argument shows that the DMS statement (1) in principle cannot be learned.

Notice that our notation loses some important information about the quantification over beliefs both in (1) and (1′). Not only does the proposition P_{n+2} state that the beliefs $P_1, P_2, \ldots, P_n$ could be true, but also generally that it is possible that *all* beliefs are true and, therefore, that it is possibly false that *there exists* at least one false belief among your beliefs. Otherwise, one could think that now after P_{n+2} is added, a new DMS is needed to state its fallibility as well. That would mean that the DMS should be revised *ad infinitum*. This is not the case because, upon recognizing that "At least one of my beliefs is false" is itself fallible, you recognize that it is possible that all your beliefs are true. This means that, once you learn (1′), you are not any more believing that at least one of your beliefs is false. In stating (1′), you are not stating that something holds about *some* of your beliefs as the statement (1) does, but that it might hold of *some* or of *none* of your beliefs.

One could still doubt if the learned proposition is the same as the believed one. Is it possible that the only actual change is adding a new

domain member over which (1) quantifies? We can settle the doubts about the difference between the learned and the believed proposition by showing that their basic semantic properties differ. First, assume that your beliefs were consistent before learning (1). The statement (1) is inconsistent with your beliefs, since were it the case that all your other beliefs are true, (1) would be false. But if (1) becomes one of your beliefs, then it refers to itself and makes it the case that what you now believe is that if all your other beliefs are true then at least *this one* is false, where "this one" stands for "At least one of my beliefs is false". The resulting belief is infallible, regardless of whether some other beliefs are false or none of them is. But that means that your beliefs are not inconsistent. Believing a statement which is always true cannot cause inconsistency among your beliefs, no matter what other beliefs you might hold. Therefore, since a set of propositions is either consistent or not and since none of the other believed propositions changed, the proposition you believe cannot be the same as (1).

While it was noticed in the literature (Sorensen, 1988, pp. 23-24) that the proposition (1) is possibly false, but that the believed proposition is never false, the two propositions were not considered to be different. But if the proposition (1) and the believed proposition were the same proposition, the belief in (1) would also be contingent, rather than infallible as (1′). It is known, however, that this is not the case.

According to the argument above, we can state two general results that outline our argument against believing (1). First, it follows from the semantic properties of the learned and the believed proposition that the DMS (1) is a counterexample to this generalization:

- If a proposition P is inconsistent with one's beliefs, after learning P, one's beliefs will be consistent only if some other beliefs are removed.

In the case of learning (1), none of the other beliefs change, but your beliefs are consistent. This leaves (1) as the only possible belief that has been revised. Therefore, the second result is that learning (1) contradicts the success axiom: you do not believe (1), after it has been learned.

The reason why learning (1) fails is that the learning action induces changes to the state of higher-order facts. These changes may affect higher-order beliefs so as to cause the learning action itself to trigger the requirement to revise the statement after learning the statement takes

place. Notorious counterexamples to the success axiom are Moorean statements. We find it absurd that one believes the Moorean statement "It is raining, but I don't believe it", although the statement may be true before one learns that it is. In fact, once an agent has learned the statement, it becomes false and the agent is committed to believe its negation, since after learning that it rains one has to believe it rains — contrary to the statement. Thus, the change to the state of higher-order facts makes it absurd for a rational agent to believe the statement after it has been learned.[5]

Moorean statements are, however, more straightforward counterexamples to the success axiom than the sentence (1) is. The statement "At least one of my beliefs is false" is not self-refuting in the sense that it is not believed any more and that *its negation becomes believed*, but its being learned fails only in the sense that *it is not believed* any more after it has been learned. The statement (1) is not self-refuting because, unlike learning Moorean statements, learning (1) does not result from an information update with *hard facts*. In learning the Moorean statement above, one knows that it is raining and this fact is "hard-coded" throughout the description of the example. In the scenario of learning (1), the only kind of factual change is the change of higher-order facts caused by the action of learning (1). The status of the other (higher-order or hard) facts that the propositions $P_1, P_2, \ldots, P_n$ refer to remains unknown and unchanged throughout the example. Still, the action of learning (1) changes the scope of one's beliefs in such a way to make it true that (1) is not believed any more and that ($1'$) is believed instead. The way we show that (1) contradicts the success axiom cannot be direct as in the case of Moorean

[5]Note that we do not inquire here into whether believing the statement (1) and Moorean statements is psychologically possible or not. Instead, our goal is to argue a rational agent should not believe the statement (1) because, given its truth-value conditions, such belief would be paradoxical. However, this does not influence the psychological fact that many agents do not realize the paradoxical nature of (1). It is even psychologically plausible for an agent to entertain a belief in a Moorean statement. De Almeida (2001, p. 56) gives example of a person who forms a belief in the statement "My father loves me, but I do not believe so" after the person has been told so by a reliable psychoanalyst.

It is shown by Rieger (2015) that avoiding beliefs in Moorean statements in doxastic logic requires the "negative infallibility" principle, which says that if you believe you don't believe something, you really don't believe it. However, it is unclear if this principle is also sufficient for the absence of the belief in the statement (1). The statement (1) cannot be easily formalized in doxastic logic and the question of minimal rationality requirements for avoiding (1) cannot be answered here.

statements, but we can trace the change from the learned to the believed proposition by focusing on their semantic properties.[6]

Learning the statement (1) does not result in rationally holding inconsistent beliefs and it does not preclude "the possibility of all one's beliefs being true" (Sorensen, 1988, p. 23), as usually held. The DMS (1) provides a complex counterexample to the success axiom. The analysis of learning (1) reflects the difficulty of the attempt to turn yourself against your own beliefs. The unsuccessful learning of (1) is, in a broad sense, a corollary of Wittgenstein's claim (see the quote on p. 151) that a verb meaning "to believe falsely" would not have any significant first-person present indicative. Wittgenstein first noticed the problems of believing falsely in connection to Moorean statements and the absurdity of their first-person versions. Through the connection of Moorean statements and the statement (1) via the success axiom, we can now extend Wittgenstein's argument to hold for the statement (1). It is absurd for you to state that it is raining, but you don't believe it, if stating this implies that you believe that it is raining. By the same principle, it is absurd for you to state that at least one of your beliefs is false if this implies that you believe that.[7]

We can further develop the parallel with Moorean problems to show a possible way out of the problems of learning the DMS candidate (1). Notice that in both Moorean statements and the statement (1), the first-person perspective is essential for the inability to learn the statements. It is absurd for you to believe that "It is raining, but I don't believe it" and, by our extension of the argument, it is absurd for you to believe that "At least one of my beliefs is false". By contrast, it is not absurd for

[6]For a more in-depth technical analysis of Moorean statements in dynamic doxastic logic, see (Baltag and Smets, 2016). For a formal study of unsuccessful updates in dynamic epistemic logic, see (Holliday and Icard III, 2010, van Ditmarsch and Kooi, 2006). From a belief-dynamics perspective, it is not exceptional that once you have learned something, your learning action itself changes the state of the world that your belief happens to be about. Consider the statement: "All my beliefs ultimately derive from experience". Once you have learned that this is the case, this statement becomes one of your beliefs and you will believe that this belief also ultimately derives from experience. Furthermore, the tenability of your claim will depend, among other beliefs, on this particular belief that you hold.

[7]Connections between the statement (1) and Moorean statements are often discussed in the literature (see (Evnine, 2001, p. 160) or see (Sorensen, 2018, Section 1) for modesty statements related to book prefaces), but they have never been fully explained to show the impossibility of learning (1), which resounds with similar results for Moorean statements. Our use of the success axiom enables us to be explicit about the similarities between the statement (1) and Moorean statements.

someone else to say about you that "It is raining, but you don't believe it" or that "At least one of your beliefs is false". But notice that if someone announces the latter statement to you, the scope of your beliefs to which the statement refers could, under some assumptions, be taken as limited to your beliefs up to the present point. For example, if someone knows that there is at least one specific belief that you now hold, which is false, then the reference to your beliefs does not include any future beliefs you may hold. Therefore, you can successfully learn from such announcement that "At least one of my beliefs other than this one is false", which is equivalent to the DMS (2) (see p. 153).

Why not, then, to simply state that "At least one my beliefs other than this one is false" in order to state that you are doxastically modest? Learning the statement (2) seems to avoid the problems of believing falsely. In the next section we investigate whether doxastic modesty justifies you in believing that (2).[8]

6.3.2 Case 2: Underdetermined beliefs

Adding the proviso to exempt DMS from its own scope results in believing that "At least one of my beliefs other than this one is false". Since (2) explicitly refers to the fallibility of your *other* beliefs, the subsequent change of the scope of your beliefs is irrelevant for the belief in (2). The important difference to the statement (1) is that, after learning (2), your beliefs are also inconsistent. This strategy is promising for a proponent of the claim that doxastic modesty obliges you to believe that you hold some false beliefs. If you have good grounds to believe that "At least one of my beliefs is false", then it seems safe to exempt the belief in DMS from fallible beliefs. The number of your other explicit beliefs, your own past false beliefs and false beliefs of others provide you with sound reasons to suspect that there will be false beliefs among your other beliefs.

Despite its intuitive appeal, the strategy of exempting DMS from its own scope comes at a cost of irrationally believing (2). Our argument against believing (2) builds on Evnine's (1999, p. 205) remark that the statement (2) "makes invidious distinctions among our beliefs and gives a special status to some that it does not give to others, namely, exemption

[8]The statement (2) is endorsed by many philosophers to be a proof that a rational and reasonably modest person should always be inconsistent. Some of the proponents of the DMS (2) are Harman (1973), Foley (1979) and, more recently, Christensen (2004). Harman (1973, p. 119) attributed this version to Nozick, but without mentioning the source.

from possible error" and, later on (1999, p. 224), that "it is a case of special pleading". The *ad hoc* nature of "special pleading" in (2) is in our view best explained as a case of a belief *underdetermined* by the available evidence.

For a simple example of underdetermined beliefs, suppose that you claim that "Most crows are black, but Snowflake isn't" and that you have evidence to support this claim. What would your evidence necessarily include? First it would necessarily include a support in favour of the claim that most crows are black — typically, some sort of inductive support. But the latter part of the claim needs to be independently supported by additional evidence — typically, a convincing piece of evidence that makes it clear that Snowflake is a crow that is not black. Most importantly, this evidence is not the same as the evidence that some crows, in general, are white. Your evidence underdetermines the belief in your claim, which invites for additional evidence.

A more striking example would include high stakes. Imagine, for example, that you have compelling evidence to assert that "There is a murderer among us" (say that you are a part of a group of people present at a boat at the time when a murder took place). Your evidence should undoubtedly be taken as a valid support for the claim that "The murderer is among us". But if you say instead that "The murderer is one of us, but it couldn't be me", then the implicit assumption is that, if ever needed, you have additional evidence in support of the exemption of yourself from the range of suspects. Indeed, a court would need more than just evidence for the claim that the murderer is among us to establish the veracity of the claim that it is someone else than yourself. While DMS (2) does not imply practical high-stakes of this type, the role of the statement (2) in the alleged paradox of inconsistent beliefs requires us to at least adhere to unambiguous rationality standards for assessing when is a statement supported by the available evidence.

We can now fully develop the analogy between such examples and the belief in the statement (2) by focusing on the evidence to believe (2). The reason for you to form a belief that you hold a false belief is your doxastic modesty. The requirement of doxastic modesty itself is grounded in your past experience. Namely, your own past mistakes and mistakes of others constitute a sound *inductive evidence* base for you to believe that, given a sufficiently large number of beliefs, there are some false beliefs among your beliefs. But the statement (2) does not only existentially quantify over your beliefs. In asserting (2), you additionally

propose a difference between your beliefs. Namely, that your belief in DMS should be exempted from the scope of fallible beliefs. We argue here that the exemption of DMS from its own scope cannot be supported by the available evidence, even if the exemption is needed to avoid the problems of believing falsely. Moreover, we can show that taking the exemption in (2) to be justified leads to absurd consequences for your doxastic state.

The problem we find in believing that "At least one of my beliefs other than this one, is false" is that the available inductive evidence underdetermines what belief(s) you should believe to be false as a response to your evidence. Your belief in (2) cannot be supported by the evidence you possess any more than a belief in the proposition "One of my beliefs other than P_k is false", for any other proposition P_k that you believe. If all that you know is that you have previously believed falsely and that you have witnessed others believing falsely, you do not know enough to eliminate the possibility of the belief in DMS being false and your other beliefs being true. In other words, your evidence could at best support existential generalization that *there is* a false belief, but not that P_k is not false, for any P_k from *Bel*.

Occasionally, proponents of the DMS (2) admit the problem of underdetermination by conflating the statement (2) with the tautological statement (1′), which is not equivalent to any of the DMS (1)-(3). Consider, for example, Harman's (1973, p. 119) version: "It has occasionally been suggested that a rational man believes that he has at least some (other) false beliefs. If so, it follows logically that at least one thing he believes is false (if nothing else, then his belief that he has other false beliefs); a rational man will know that.". While the latter sentence with the parenthetic proviso just rephrases (1′), which claims that *either all* your beliefs are true *or some* beliefs are false, the first sentence *excludes itself* from the range of fallible beliefs.[9] This suggests that, according to Harman, a rational person knows that the belief in (2) is also fallible, but the person can, nevertheless, choose to believe that it is one of the other beliefs that is false. But if you do not know which belief is false,

[9]Foley, one of the proponents of the version (2), argues (1979, p. 252) that "although we have evidence sufficient to justify each of these beliefs, we also have equally good evidence for believing that at least one of these other beliefs is false". But the same paragraph also reads that "it seems likely that at least one of" the beliefs is false, "even if we have no idea which one it is". The latter claim shows that the belief in the DMS (2) is underdetermined.

the claim that it is the other beliefs that are false is only an *ad hoc* hypothesis that serves your purposes of avoiding the unsuccessful learning of (1). Harman's argument for the rationality of believing (2) is, therefore, self-undermining.

At this point, a proponent of the DMS (2) could defend the belief in (2) by arguing that the only additional support needed to support (2), above our fallibility, is the fact that "the number of present beliefs in question here is very, very large" (Foley, 1979, p. 252). Even at a probability of 0.99 for each explicit belief, the probability of a thousand of beliefs being true is insignificant. It seems likely that there are some false beliefs among those other beliefs.

Although the intuition about the number of beliefs is appealing, proponents of the DMS (2) cannot justify the belief in (2) by resorting to the large number of beliefs. If the number of your beliefs justifies the statement (2), then it also justifies n additional statements of the type "One of my beliefs other than P_k is false", for any P_k from *Bel*. Since the number of beliefs is supposedly very large, you have a strong probabilistic justification to believe each of $n + 1$ statements. This means that the same evidence that needs to justify you in believing that you hold at least some (other) false beliefs, justifies you to believe that each of n other beliefs you hold is infallible.

Hereby you run into several additional problems concerning the evidential support for the belief in (2). First, it is questionable if you should ever accept an evidence base as reliable if it justifies inconsistent beliefs. All the more reason not to accept the number of beliefs as evidence for (2) is that you cannot resolve the inconsistency by favoring some proposition P_k over any other equally supported proposition. Secondly, if you are faced with $n + 1$ propositions that are justified in the same way as (2) and you know that they are jointly inconsistent, as a rational person, you would want your beliefs to be consistent. A clear solution is that, since the number of beliefs underdetermines $n + 1$ propositions of the type "One of my beliefs other than P_k is false", you should consider the fact that each of your other n beliefs is supported by some independent evidence. A rational response is to continue holding those beliefs and disregard (2).

6.3.3 Case 3: Truth-commitment glut

This leaves us with one more option to state that at least some of your beliefs, taken in their totality, are false. It is possible to anticipate the problems of learning (1) so as to include the DMS into its own scope. This amounts to the doxastic modesty statement (3): "At least one of my beliefs including this one is false". The statement (3) can be learned and it does not claim its own infallibility as (2) does.[10] Assuming again that you hold n beliefs, the DMS (3) corresponds to the following statement:

$$P^*_{n+1} = \sim P_1 \vee \sim P_2 \vee \cdots \vee \sim P_n \vee \sim P^*_{n+1}.$$

In this formulation of DMS, including a DMS into its own scope precludes the change of the scope of your beliefs, which is responsible for the failure of learning (1). Your belief set after learning P^*_{n+1} is defined as:

$$\{P_1, P_2, \ldots, P_n, P^*_{n+1}\} = Bel^2.$$

The "bootstrapping" proposal in (3) to include DMS into its own scope is notorious in the literature for its problem of self-referentiality. As a result, many authors associate the statement (3) with semantic paradoxes of the liar-type, e.g. Evnine (2001, p. 160) and New (1978, p. 344). In our argument, we do not use the connection to semantic paradoxes, but we do identify a doxastic paradox that results from a belief in (3).

The doxastic paradox of believing P^*_{n+1} comes from your truth commitments. By the definition of belief, you cannot believe some proposition P_k without claiming its truthfulness. Beliefs satisfy the following condition:

$$Bel \subseteq CommitTrue,$$

for a set of the propositions *CommitTrue* that you claim to be true. Notice that propositions from the set *CommitTrue* need not actually be true. We do not know their actual truth value, but only the provision that you

[10] A version of this DMS is discussed in relation to book prefaces: "At least one of the claims in this book including this one is false". That version was named "sophistical" by New (1978, p. 341) and it was first formulated by Prior (1971, pp. 84-87). New (1978, p. 344) spells out the high demands involved in formulating such doxastic modesty statements: "It takes a certain amount of perverse ingenuity to generate the sophistical paradox, an amount which —— fortunately —— most authors lack". He claims that this statement is one of the variants of the Liar paradox.

claim that they are true.[11] Therefore, the set of beliefs Bel^2 incurs the following set of commitments:

$$\{P_1, P_2, \ldots, P_n, P^*_{n+1}\} \subseteq CommitTrue.$$

But the problem is that you are also committed to the falsity of P^*_{n+1}:

$$\{P^*_{n+1}\} \subseteq CommitFalse.$$

This is so because if you are committed to the truth of $P_1, P_2, \ldots, P_n$ and P^*_{n+1}, then, by the definition of P^*_{n+1} and assuming that all of the propositions $P_1, P_2, \ldots, P_n$ and P^*_{n+1} are true, P^*_{n+1} is also false. Assuming that neither the truth nor the falsity commitment above can be disregarded, in believing P^*_{n+1}, you are committed to disbelieve it. We will refer to such commitments as truth-commitment gluts.[12]

It follows from the truth-commitment glut that your belief set has to include $\sim P^*_{n+1}$ as well:

$$\{P_1, P_2, \ldots, P_n, P^*_{n+1}, \sim P^*_{n+1}\} = Bel^3.$$

The set Bel^3 shows the doxastic paradox of believing the DMS (3). The paradox is unavoidable since the obligation to believe a proposition contradicting (3) follows from the definition of (3) and the fact that (3) is believed. Since the set Bel^3 contains an outright contradiction, it is difficult to defend its rationality without a strong assumption of true contradictions or "dialetheias". While claiming that some statement is dialetheia does not on itself make an argument against believing it, it does make an argument against believing that the statement is true without believing its negation to be true as well. To make the latter option plausible, a proponent of (3) needs to provide an argument that justifies disregarding the truth-commitment glut and explain why to opt for the commitment to truth only. It is, however, unlikely that there is any such argument available, because the truth-commitment glut of (3)

[11] That is to say, beliefs are not necessarily factive. In the logic of beliefs **KD45**, this means that we do not accept the formula $Bp \rightarrow p$ to be an axiom, where B is a belief operator and p a believed proposition.

[12] In some many-valued logics, a sentence can be both true and false, in which case it exemplifies a truth-value glut. For example, the sentence "This sentence is false" can be evaluated as both true and false. See (Priest, 2008, pp. 127-133) for more details on semantic truth-value gluts.

ensues from the same principle as semantic truth-value gluts. These are, by their definition, unresolvable by resorting to a single value.

While the statements (2) and (3) can be successfully learned, they are not entirely devoid, each in its own way, of the problems we ascribe to the inability to believe of yourself that you believe falsely. Our conclusion is that each of the two problems eventually make the attempts to avoid the learning issue of (1) unacceptable for a rational modest reasoner.

6.4 Doxastic modesty and higher-order evidence

The problems we discussed in relation to the DMS candidates (1)-(3) do not eliminate the very requirement to be doxastically modest. We left open the question of how you are supposed to reasonably fulfil this requirement. In this section we propose that, besides avoiding the candidates above, there are further general restrictions on how to state doxastic modesty. Most importantly, evidence that consists in your past mistakes and mistakes of other people, which motivates your doxastic modesty, can only justify suspension of your belief in the conjunction of all of your beliefs and does not justify believing its denial.

The requirement of doxastic modesty is grounded in evidence of your fallibility, that is, your past mistakes and your awareness of mistakes that other reasoners commit. When facing such evidence, you need to decide which beliefs it supports and which beliefs it eliminates. In making this decision, it is helpful to look at the kind of evidence that your fallibility provides to you. This evidence does not point out a specific false proposition you happen to believe. It is also clear that it does not bear directly on the facts being believed. On the contrary, the evidence of fallibility is supposed to justify your belief that you hold some false beliefs, though regardless of what you actually believe. Such evidence, therefore, cannot be considered as *first-order* evidence.

The evidence of fallibility can only be what epistemologists call "higher-order" evidence: evidence about the quality of evidence or evidence about one's ability to adequately respond to certain first-order evidence. Typically, higher-order evidence results in a change of beliefs "because it indicates that my former beliefs were rationally sub-par" (Christensen, 2010, p. 185) and, thereby, generates a kind of self-doubt. Let us focus on such higher-order evidence against some state of beliefs. For example, learning that you were administered with a hallucinogen

last evening calls your beliefs about the events that took place thereafter into question. Such higher-order evidence does not refer directly to the facts. It rather works as an *undercutting* reason to one's initial reason to form some belief — it attacks the connection between the evidence and the conclusion, rather than attacking the conclusion itself. After learning about your hallucinogen-driven belief formation, you are justified in retracting your beliefs about last evening because your initial reasons to believe are defeated. However, if you were to see some footage of the actual course of events that makes it obvious to you that many of your beliefs about last evening are in fact mistaken, you would not only be provided with a reason not to believe your initial beliefs, but also to believe their denial. The latter kind of reasons are called *rebutting* reasons.[13]

Epistemologists agree that higher-order evidence indicating that a belief is formed in a sub-par way primarily functions as an undercutting reason for the belief. When facing such evidence "I must in some sense, and to at least some extent, put aside or bracket my original reasons for my answer. In a sense, I am barred from giving a certain part of my evidence its due" (Christensen, 2010, p. 195). Now consider again doxastic modesty and evidence of your fallibility.

If you believe that each of the propositions $P_1, P_2, \ldots, P_n$ is true, deductive closure gives you at least a *prima facie* reason to believe the conjunction of these propositions. However, the fact that you have made mistakes before becomes relevant for the beliefs about whether your totality of beliefs is true. How does such evidence defeat your reason to believe the conjunction of your beliefs? Evidence of your fallibility can only undercut the reason to believe the conjunction of your beliefs. Higher-order evidence of fallibility does not differ from other kinds of higher-order evidence that we usually take to be a reason to doubt one's capacities to form correct beliefs, such as the hallucinogen example above. The fact that you are fallible is irrelevant to the truth of propositions $P_1, P_2, \ldots, P_n$. As an undercutting defeater, it does justify you to suspend the belief in the conjunction of your beliefs.[14]

[13]For a more extensive discussion on this distinction, see Chapter 2.

[14]Possession of higher-order evidence is usually taken not to provide one with a rebutting defeater. Arguably, a counterexample to this qualification could be peer disagreement cases where someone who is of comparable doxastic dispositions faces the same evidence, but comes to a contradicting conclusion to the one you obtained. These cases are, however, far from what we discuss in doxastic modesty statements. Lasonen-

Admittedly, there is something counter-intuitive about not being able to believe statements that appear at least probabilistically well-justified, as the DMS candidates appear to be justified. The probability of multiple statements being true together is always lower or equal than the probability of the least probable statement among them. Proponents of believing either of the DMS (1)-(3) could object that in avoiding to take a position on (1)-(3), you seem to indirectly go against the so called "Principal principle" that recommends you to conform your credences to your estimation of objective probabilities (Lewis, 1980).[15] To this, we can only answer that believing against the conjunction principle and the possibility of the first-person *believing falsely* statements (1)-(3) both come across as being at least as counter-intuitive. This seems to be an outcome of taking any kind of doxastic attitude toward the totality of your beliefs — perhaps many intuitions about your *ordinary* beliefs do not apply in this case. This should not be surprising because the statements (1)-(3) combine higher-order beliefs, which are notoriously delicate with respect to the success of learning outcomes, with the reference to the totality of beliefs, which already indicates a possibility of a (doxastic) paradox.

The difficulty to take an objective view of our own totality of beliefs is well-captured by Evnine (2001, p. 171):

> The higher-order activity of taking a view of our beliefs is thus, at least in part, of the same nature as its first-order subject matter. If we adopt certain objective beliefs about our own beliefs, those higher-order beliefs will be as much subject to that special commitment we owe to our own beliefs as are the first-order beliefs of which we are trying to obtain an objective view. While we may be able to fragment our beliefs and use some to obtain an objective view of others, we cannot obtain a comprehensive objective view of our beliefs *in toto*.

Evnine's analysis of the doxastic modesty is close to ours, especially in his appreciation of the importance of the first-person perspective (2001, p. 165):

> For there are significant ways in which the fact that some beliefs are our current beliefs precludes us from taking certain

Aarnio (2014, p. 317, footnote 9) hints at a different possibility: "Perhaps, for instance, higher-order defeaters often or even always have some rebutting force: evidence that I came to believe p as a result of a flawed process may be weak evidence that p is false."

[15] Of course, we are here primarily talking about full beliefs, not credences.

> attitudes to them that would be quite unproblematic in the case that they were someone else's beliefs. It is, of course, not the content of our beliefs that makes them special to us, but merely the fact that they are ours.

With some important differences from Evnine's account, the main contribution of our account can be found in the detailed elaboration of the first-person perspective problems of the DMS candidates (1)-(3).[16]

6.5 The preface paradox and doxastic modesty

According to Makinson's (1965) "Paradox of the Preface", doxastic modesty obliges book authors to believe that there is at least one false statement in their book. Imagine that you have written a book in a field of your expertise. While you do believe each of the claims from the book taken on itself, you also have reliable inductive evidence to believe that the book is likely to contain at least one undetected error. Evidence of your fallibility indicates that your book most likely contains at least one error. Therefore it is rational for you to state in the book preface that there are errors in your book. But then, Makinson argues, you cannot avoid holding inconsistent beliefs since you hold a number of individual beliefs from the body of the book, together with the prefatorial belief in the negation of the conjunction of those beliefs taken together.

Most commentators agree that Makinson's prefatorial statement is a special case of doxastic modesty statements.[17] This is plausible to assume because such prefatorial acknowledgements result from the nature of beliefs, that is, "from belief in fallibility of one's beliefs, of which the

[16]While some of Evnine's general conclusions are in the same spirit as ours, especially his emphasis on the first-person perspective, Evnine's arguments differ in some important respects. First, Evnine (2001, p. 158) does not distinguish between the statement (1) and (3) and he does not identify the problem of unsuccessful learning. Secondly, he thinks (1999, pp. 202-204) that the conjunction of one's beliefs is not identical to all of one's beliefs and he argues that believing the conjunction of one's beliefs is justified. Thirdly, in his arguments against the DMS (2) and (3) Evnine (2001, p. 159) uses "the idea of a creature with only one belief" and the idea of "an ideally rational belief set" to argue that they "cannot be part of an ideally rational set". He does not argue that they are "irrational *simpliciter*", although his starting premises of commitments that one has to one's own beliefs point out in this direction. These are only some of the most important differences.

[17]Among the authors who discuss the version of the paradox with the prefatorial statement is Pollock (1986). Unlike our logical analysis, Pollock's analysis uses probabilities.

prefatorial form is merely a particular example" (New, 1978, p. 342).[18] In our view, however, there are good reasons to avoid the original book version in the discussion about doxastic modesty.

One of them is that only those statements from the body of a book are naturally seen as the main contribution of that book to the field in which it has been published in. Its preface and acknowledgements belong to a customary part of a book-writing process and they are usually written in a more relaxed fashion. For example, they are neither reviewed nor do they necessarily contribute to the field in which the book has been written. It is not clear whether the body of the book and its preface are comparable in the sense that they both represent doxastic attitudes of a comparable status. In any case, they are naturally intended to serve for different purposes and this is reflected in their mutual independence both in writing and, subsequently, in reading books. Such natural distinctions between preface statements and the body of the book statements obscure the fact that it is the statements *qua* believed statements that cause the alleged paradox.

The unclarity as to what exactly is the doxastic attitude toward book prefaces justifies us to suspect willingness of their authors to revise their prefatorial statement in the light of considerations about general doxastic modesty statements. The risk involved, for example, in neglecting the fact that exempting the prefatorial statement from its own scope is underdetermined by evidence is insignificant for the prospects of the entire book. It is hard to imagine that anyone would demand from a book author to account for recklessly excluding the prefatorial statement from the scope of the fallible statements. Therefore, it is difficult to apply the same rationality standards on beliefs about doxastic modesty and prefatorial statements.

Further ambiguities arise from the status of the statements in the body of the book, regardless of the preface statement. Are the statements from the body of the book all believed in the same sense? According to Leitgeb (2014, p. 15), "by uttering or publishing a great many declarative sentences in assertoric mode, one does not actually assert that their conjunction is true — one rather asserts that the vast majority of these sentences are true". Ryan (1991, p. 300) shares similar views and argues that hard work and intellectual responsibility usually do not give one

[18] One of the exceptions is Christensen (2004, p. 37), who thinks that the totality of beliefs version "makes less transparent the relations between the first- and second-order beliefs".

sufficiently good reason to believe every single statement in the book. If Leitgeb and Ryan are right about this, there are workarounds to the book preface problems that do not hold for the general doxastic modesty statements (1)-(3). The discussion about doxastic modesty should not be obfuscated by problems (or solutions) of the preface paradox related only to book prefaces. Instead, the discussion should focus on the nature of beliefs, which is the alleged source of the preface paradox and the reason why the paradox is taken to be an argument that supports the rationality of inconsistent beliefs.

6.6 Conclusions

Given that the problems of believing falsely undermine believing the candidate doxastic modesty statement (1), and given that the ways out of the problem are not satisfactory, our conclusion is that doxastic modesty statements should not oblige you to believe in the falsity of your beliefs. However, we do not imply by this that one should believe the negation of, say, the statement (1). This is sometimes represented as the only available alternative. But believing the negation of (1) is known to be problematic as well. If you believe that none of your beliefs is false, you are being doxastically conceited.[19]

You can, therefore, subscribe to Kyburg's (1970, p. 59) analysis of conjunctivitis and state that: "of everything that I believe, it is correct to say that I believe it to be true; but it is not correct to say that I believe everything I believe to be true". But, according to the problems discussed, you should not take a step further by believing that something you believe is false. This leaves you with a simple recommendation with respect to the doxastic modesty statement problem: to abstain from believing the DMS candidates (1)-(3) and their denials as well. There are many uncontroversial doxastic modesty statements that could be believed instead. One of them is submitted by Evnine (2001, p. 173)

[19]Smullyan (1986, p. 344) shows that doxastic conceitedness, i.e. believing in one's belief accuracy (inerrancy or factivity of beliefs), leads to inaccuracy, even for ideal reasoners. That is, assuming that reasoners have a complete knowledge of propositional logic and have beliefs that are closed under *modus ponens*, there will be a proposition that disproves their accuracy. On the other hand, modest reasoners in doxastic logic do not believe that something they believe is accurate unless they believe it (Smullyan, 1986, p. 351). This captures the commitment to the truth of beliefs, but it also ensures that reasoners do not unconditionally believe into their accuracy.

as "the right way for a rational creature to satisfy" the requirement of doxastic modesty: "Some of my beliefs may be false". After all, a modest reasoner would recognize that doxastic modesty should not be claimed so boldly and would use the appropriately modest modality "may".

Conclusion

Results

In the introduction to this thesis, we submitted that argumentation theory needs to rejoin formal logic as a part of a more encompassing idea of building bridges between ordinary reasoning and logic. We pointed out that the dominant philosophical thought of the twentieth century saw logic as insufficiently rich to give a proper account of the complex phenomenon of argumentation. Despite their close historical connections, logic and argumentation theory have grown apart.

In Chapters 2, 3, and 4, we set out to define a new logical system that is capable of meeting the challenge of being a proper *logic* of arguments. Instead of discarding philosophical criticism of formal approaches, we drew inspiration from philosophical findings on the complexity of arguments and the logical limits of determining their tenability. The resulting system is a unique logical answer to modeling arguments with internal structure: **default justification logic**. We will mention several features that make our default justification logic stand out from the existing formal systems for argumentation.

We show that our novel combination of justification logic and default logic produces a logical theory of default reasons. As mentioned in the Introduction, one of the expectations for a logical theory of default reasons is to be able to model the now-standard Pollock-type conflicts among reasons, namely undercut and rebuttal. To define a logic with conflicting reasons, we build on the strengths of justification logic and default logic.

The strength of justification logic in this context lies in the expressivity of justification assertions of the type $t : F$. Such formulas give us an immediate object-level representation of premise-conclusion pairs. With the use of default rules, we also gain a method to represent situations in

which adding new information compromises the existing default reasons.

One succinct characterization of the idea behind our logic is the following: while Reiter's defaults generalize *modus ponens* inferences for those cases where conditional claims hold only *ceteris paribus*, our defaults generalize *modus ponens* inferences evidenced by formal operations on **reasons** for those cases where reasons for conditional claims hold only *ceteris paribus*. Justification assertions produced with the new type of rules will help in establishing justification logic as a mathematical theory of reasons in general, not only as a theory of ideal reasons such as mathematical proofs. This should be read in the sense that the logic we defined might also support deontic (*prima facie* reasons), linguistic (grammatical evidentials), causal and possibly other interpretations of reasons.

There are several ways in which default justification logic can be seen as a junction of research areas. We already described how this holds for default reasoning and justification logic. Moreover, our logic integrates the study of arguments within AI with formal logic. The goal of defining a logical system that represents structures of arguments has been set by the AI community since the 1980s. Default justification logic lives up to this long-anticipated goal and perhaps unsurprisingly so, given the connections between the principles of defeasible reasons and argumentation theory. The connections between Dung's argumentation frameworks on one side and non-monotonic logics, logic programming and modal logic on the other side have already been investigated before. This thesis relates Dung's abstract argumentation frameworks and justification logic for the first time.

In certain ways, justification logic and abstract argumentation frameworks had to become ingredients for a general theory of reasoning with reasons. Consider that abstract argumentation frameworks study oppositions of arguments out of the logical context. Standard justification logic studies pairs of reasons and conclusions, but without a way to represent oppositions among reasons. This is where the two approaches immediately appear as complementary theories. In this thesis, we provide formal correspondence between default justification logic and abstract argumentation frameworks to make the intuitive connections between the two systems concrete. Correspondence results follow from the fact that each standard abstract argumentation semantics can be defined in default justification logic. Since justification logic naturally represents arguments as formulas, the connection to argumentation frameworks is

not limited to simply expressing one system by means of another system. Instead, we show that Dung's attack graphs capture only a single aspect of our logic, namely, the direction of undercutting and rebutting attacks. One the other hand, Dung's graphs can be realized into multiple default theories with justification formulas in such a way that we can find actual justification assertions that are counterparts to Dung's atomic arguments.

In addition, we also went beyond our initial expectations. For example, default justification logic turned out to be a logical system capable of modeling complex components of the Toulmin model of arguments, such as warrants and backings. The Toulmin model is typically taken to be an informal model of reasoning. This is one of the reasons why we claim that default justification logic fulfils the goal of bringing general argumentation theory back within the scope of formal logic. In our view, default justification logic has a potential of becoming a foundational system for the rich interdisciplinary study of models for reasoning and argumentation.

Besides the mentioned formal connections between abstract argumentation, default logic and justification logic, we also explored several ways to enrich our logic with belief revision operations. This was done in Chapter 4. There, we took a step beyond the default reasoning paradigm and we investigated the inclusion of plausible reasoning patterns. In plausible reasoning, an agent needs to deal with uncertainty of premises and inconsistent premises. In contrast, the basic default justification logic assumed certainty of premises and uncertainty of inferences. The purpose of considering plausible reasoning was to provide a unified logical system for all three standard types of argumentative attack, namely undercutting, rebutting and undermining. Again, justification assertions turned out to be an appropriate logical format for representing undermining attacks. In the plausible-reasoning extension of our logic, an assertion $t : F$ from a set of premises W is interpreted in such a way that t represents a source that supports the formula F. On our account, undermining is then interpreted as a removal of the unreliable source of information t. What we needed is to define operations that are able to alter even the set of premises W. To this end, we used techniques from belief revision to model information dynamics of justification logic default theories.

The relation between default reasoning and belief revision is interesting and has been discussed in the literature.[20] Our system with justifica-

[20]Gärdenfors (1990) sees them as the two sides of the same coin. His views are based

tion assertions provides a fresh outlook on this relation. We were able to flesh out the intuition that default reasoning deals with inconsistencies that arise from extending the starting premises, while belief revision deals with inconsistencies that result from receiving information incompatible with the starting premises. When this assumption is applied to justification assertions, it becomes clear that belief revision has more to do with the plausible reasoning paradigm than with default reasoning paradigm. Once again, justification logic acted as an intermediary system, this time by throwing light on the relation between default and plausible reasoning that was translated into relations between default logic and belief-revision techniques. With their fine-grained representation of reasons, justification assertions gave us a better perspective on the ways in which incoming information causes different types of non-monotonic behavior in belief revision and default logic. Using those insights, we succeeded in defining a logical theory of all three standard types of argumentative attack in AI, namely rebuttal, undercut and undermining.

In Chapters 5 and 6, we kept on dealing with the issues related to ordinary reasoning, but with an emphasis on the role of classical logic in it. Chapters 2, 3 and 4 argued that ordinary reasoning should be, first and foremost, studied through such phenomena as defeasible inferences, belief revision and reasoning errors. Given such state of facts, do we still have a place for classical logic and valid inference patterns in ordinary reasoning? Our answer is "Yes". We did not argue that deductive reasoning could provide an appropriate account of how humans reason. But the fact that most of our reasoning tasks do not fall under the category of deductive inference does not imply that the rules of classical deductive logic are not normative, once they in fact become *applicable* as norms. To show this, we define a logical system for defeasible norms that makes clear the role that classical logic has in regimenting reasoning tasks that are markedly riddled with errors and limitations of commonsense reasoning.

The role of classical logic rules in human reasoning may be overshadowed by our need to deal with inconsistent or incomplete information, and all this with limited cognitive resources. However, this says much more about the conditions in which ordinary reasoning proceeds and

---

on the possibilities to translate belief revision postulates into different non-monotonic inference variants. Makinson and Gärdenfors (1991) also follow the direction of translating between belief-revision postulates and conditions on non-monotonic inference conditions.

the capacities of reasoners, than it does about the question whether a logical norm is *prescriptive*, once it becomes relevant to a reasoning task. It seems that the critiques of the normative status of logic have taken the imperfections of ordinary reasoning as an indication that classical logic is too limited to be of any actual relevance. In Chapter 5, we defended the normative status of classical logic in the face of such criticism. However, Chapter 5 is at the same time an appreciation of non-monotonic techniques and their efficiency in *describing* commonsense reasoning. In this sense, Chapter 5 underscores the basic assumption of the thesis that there are both normative logical systems and descriptive logical systems for ordinary reasoning. By making the right design choices, we can develop such systems that combine the limitations of ordinary reasoning with the rigour of formal logics.

Chapter 6 investigates one renowned counterexample to the classical logic restriction against holding inconsistent beliefs: doxastic modesty statements.[21] It has been claimed that a modest and fallible agent is destined to have inconsistent beliefs, because the agent has to believe that at least one of the beliefs that the agent holds is false. We argued that this is not as easy as presented. Learning the statement "At least one of my beliefs is false" has all the needed ingredients for causing a paradox, such as self-referentiality and change of higher-order beliefs, but it does not cause inconsistency of beliefs as argued. We concluded Chapter 6 with a more positive outlook on classical norms that recommend retaining the consistency of beliefs. Critiques of classical logical norms are right to draw our attention to circumstances in which inconsistency of beliefs is irresolvable and, thus, best retained in the interest of the economy of belief formation. The problem is that such situations cannot result from the nature of beliefs themselves, as it has been claimed in the debate on doxastic modesty statements.

Future research

It still remains to be seen how our logical theory of defeasible arguments could add to the study of computational aspects of argumentation. The

[21] Doxastic modesty statements are most often connected to the "Paradox of the preface" by Makinson (1965). In the preface paradox scenario, modest book authors to be justified in believing that at least one of the statements from the book is false. Since they also believe each statement taken on itself, they seem to entertain justified inconsistent beliefs.

history of AI has been closely tied to the development of knowledge-based systems that mimic commonsense reasoning and defeasible argumentation is an important aspect of commonsense reasoning. Default justification logic offers some foundational solutions to the problems of understanding defeasible argumentation, but the solutions are yet to be tested in practice. One possible research avenue could be carried out with the help of cognitive psychologists to find out if there could be plausible computational models of human-like argumentation based on our logic. Another collaboration could be possible with linguists to find out whether default justification could explain phenomena like grammatical evidentials that indicate sources of evidence for statements in natural languages. Moreover, a systematic study of the properties of default justification logic and the study of its relation to the existing structured argumentation frameworks are still needed. On the implementation side, one of the most helpful insights could be provided by a study of the computational complexity of default justification logic. In general, default reasoning does not fare well in this aspect as compared to classical propositional logic.

An interesting feature of default justification logic is that it can model recovery from reasoning errors and exclusion and reinstatement of default hypotheses (here called "warrants"). Detecting the winning hypothesis bears some conceptual similarity to the use of data in machine learning to learn a function that best maps inputs to outputs. One far-fetched goal would be to explore whether there are ways to use the logic to extract rules and increase transparency of machine learning models. Some work has already been done in combining classification models and concept learning with abstract argumentation (Amgoud and Serrurier, 2008). Another worthwhile idea is to explore possibilities to combine the logic with arguments with formal learning theory. Epistemic logic and learning theory have already been connected by, for example, Gierasimczuk (2009), who explores bridging dynamic epistemic logic and learning theory.

We will also mention some possible technical developments of default justification logic that we did not consider in this thesis. First, there is a first-order variant of justification logic that could be a promising extension of the underlying logic of propositional justification assertions, from which we started defining its default variant. This first-order justification logic is closer to Reiter's original default logic, which has first-order logic as its underlying logic. With the added quantification over individual terms, default rules might also be expressed as schemes. Secondly, it will

be interesting to look at a multi-agent variant of default justification logic. This is especially interesting with respect to its applications to argument analysis, since argumentation is typically considered to be a multi-agent practice. One challenge in this direction is to find out how to make sense of an argumentative discourse in which participants do not disclose the structure of reasons in such a way that they make it possible to obtain warrant or backing of an argument.

In Chapter 5, we put emphasis on the defeasibility of normative rules for human reasoning and the unsolvability of the problem of relevance.[22] Both of these issues, in their own way, gave rise to Harman's skeptical challenge regarding the normative role of logic. There are good reasons to think that the problem of normativity and the problem of logical omniscience in epistemic logic are closely related. We assume that the relevance problem in particular could shed light on why is it notoriously difficult in epistemic logic to tell relevant from irrelevant logical consequences of an agent's knowledge.

Chapter 6 was an attempt to show how even learning from oneself does not need to always be successful, as doxastic modesty statements exemplify. We would like to know what is the class of statements that such doxastic modesty statements belong to. For this question to be answered, we need to be able to formalize such paradoxical expressions as the doxastic modesty statements that were considered here.

[22] In a nutshell, the problem of relevance consists in the impossibility to tell *a priori* what information could potentially become relevant to any other piece of information for any actual reasoning task.

Summary

This thesis inquires into logical and philosophical underpinnings for ordinary reasoning and argumentation. In Chapters 2, 3, and 4 we develop default justification logic that models structured argumentation and we contextualize this logic within the study of formal argumentation and defeasible reasoning. In Chapter 5, we defend the normative role of logic for human reasoning by using slow default logic. Finally, in Chapter 6, we analyze some problems with learning doxastic modesty statements that have been proposed as a way in which modest reasoners acknowledge their fallibility.

We define default justification logic that models defeasible argumentation. In default justification logic, we represent undercut and rebuttal as the two basic types of conflicts among defeasible reasons. We show that each standard abstract argumentation semantics, namely grounded, complete, preferred and stable, can be defined in this logic. We argue that having a logical system with an object level representation of structured arguments has some advantages over using structured argumentation frameworks. In particular, our justification logic does not need to use any meta-level rules or build on a language of another logic to represent arguments. Instead, object level formulas of the type $t : F$ called "justification assertions", and inference rules of default justification logic are the only requirements to represent arguments. In this sense, it provides a unique logical solution to the question of modelling defeasible arguments, which has been widely discussed both in AI and in philosophy. One of the main results we show is the relation of our logic to Dung's abstract argumentation frameworks. In particular, we show that our logic provides "realizations" of Dung's frameworks, that is, replaces Dung's abstract and implicit arguments with structured and explicit justification assertions. It is also possible to, conversely, obtain Dung's frameworks from justification logic default theories.

This logic is further extended with dynamic operators that model changes to the basic justification logic default theories. We present the expansion, contraction and revision operations, each in both prioritized and non-prioritized version, which differ as to whether they always give priority to the incoming information over the existing information. We argue that the kind of attack called "undermining" in argumentation theory amounts to those operations that contract a set of premises for some default theory. When a set of premises is contracted by a formula, this formula is considered attacked. Since for a default theory $T = (W, D)$, the set of facts W represents premises for default reasoning, removing a formula from W is naturally interpreted as an undermining attack.

By defining rebuttal, undercut and undermining, we succeed in giving a logical theory of all three standard types of argumentative attack in AI. Furthermore, the logic of default justifications, together with the default theory revision operations, provides a junction for the default and plausible reasoning paradigms in AI. In the basic default justification logic we deal with the classical default reasoning examples where exclusionary reasons play an important role. Here, the importance of the justification logic formalization of what Toulmin calls warrants improves the-state-of-the-art solutions to modeling exclusionary reasons. With the addition of undermining, we also enable handling inconsistent information inputs and elimination of unreliable sources of justification, thus rounding off our justification logic approach to argumentation.

We also used systems for non-monotonic reasoning to show that Harman's objections to the normative role of classical logic in human reasoning are not successful. We especially argue against the idea of exceptionless prescriptive rules and against the idea of a single "bridge principle" that succinctly articulates the normative role of logic. In our slow default logic, we take normative rules as being defeasible rules only and we respect the fact that, given a set of initial beliefs, it is, in principle, impossible to know what is a reasoning task *about*. The latter fact is one of the facets of the frame problem known as "the relevance problem". We detect this problem as one of the underlying issues in defining prescriptive bridge principle candidates.

Finally, we look into the problem of doxastic modesty statements that result from an agent's awareness of its own fallibility. According to some philosophers, a doxastically modest agent needs to recognize its fallibility and believe that it holds at least one false belief. Therefore, doxastically modest agents seem to be justified in holding inconsistent beliefs. We

argue that believing the direct expression of doxastic modesty with the statement "At least one of my beliefs is false" does not in fact lead to inconsistency.

Authors who argue that doxastically modest reasoners do have to be inconsistent noticed that they should reformulate the mentioned doxastic modesty statement to exempt the statement itself from the scope of believed statements, thus settling on the statement "At least one of my beliefs other than this one is false". We argue against this *ad hoc* solution on the ground that the statement unjustifiably gives special status to the doxastic modesty statement itself.

Samenvatting

Dit proefschrift onderzoekt de logische en filosofische onderbouwingen van alledaagse redeneer- en argumentatiepatronen. In de hoofdstukken 2, 3 en 4 ontwikkelen we een *default justification logic* (een defaultlogica van rechtvaardigingen) die gestructureerde argumentatie modelleert, en plaatsen wij deze logica binnen de studie naar formele argumentatie en herroepbare redeneringen. In hoofdstuk 5 verdedigen we de normatieve rol van logica voor menselijk redeneren met behulp van zogeheten langzame defaultlogica. Ten slotte analyseren we in hoofdstuk 6 enkele problemen die voorkomen bij het leren van doxastisch bescheiden beweringen, beweringen die de zelfverklaarde feilbaarheid van redeneerders uitdrukken.

We definiëren *default justification logic* als een logica die herroepbare redeneringen modelleert. In de *default justification logic* presenteren wij *undercutting* (ondergraven) en *rebuttal* (weerleggen) als de twee basistypes van conflict in herroepbare redeneringen. We tonen aan dat elke standaard abstracte argumentatiesemantiek, namelijk de gegronde, de volledige, de stabiele en de *preferred* (waaraan de voorkeur wordt gegeven) semantiek, gedefinieerd kan worden binnen deze logica. We beargumenteren dat het hebben van een logisch systeem dat kan representeren structurele argumenten op object-niveau bepaalde voordelen heeft boven het gebruik van gestructureerde argumentatieraamwerken. In het bijzonder maakt onze *justification logic* geen gebruik van regels op meta-niveau en is deze niet gestoeld op een ander soort logica om argumenten te representeren. In plaats daarvan zijn de enige benodigdheden voor het representeren van argumenten formules op object-niveau van het type t:F, genaamd "rechtvaardigingsbeweringen", en afleidingsregels van de *default justification logic*. In deze zin biedt de *default justification logic* dus een unieke logische oplossing voor het vraagstuk omtrent het modelleren van herroepbare argumenten, een onderwerp waarover breed

gediscussieerd wordt binnen zowel de kunstmatige intelligentie als de filosofie. Een van de belangrijkste resultaten is dat we de relatie aantonen tussen onze logica en Dungs abstracte argumentatieraamwerken. We laten zien dat onze logica "realisaties" biedt van Dungs raamwerken; dat wil zeggen, onze logica vervangt Dungs abstracte en impliciete argumenten door expliciete rechtvaardigingsbeweringen. Het is echter ook mogelijk om Dungs raamwerken te verkrijgen uit default-theorieën van de *justification logic*.

Deze logica wordt verder uitgebreid met dynamische operatoren die wijzigingen in de default-theorieën van de *default justification logic* uitdrukken. We presenteren de uitbreiding-, samentrekking- en herzieningsoperaties, zowel in geprioriteerde als non-geprioriteerde versies, die verschillen in of ze altijd prioriteit geven aan de binnenkomende informatie ten opzichte van de al bestaande informatie. We beargumenteren dat het soort aanval genaamd "*undermining*" (ondermijnen) in de argumentatietheorie neerkomt op operaties die een verzameling feiten samentrekt voor sommige defaulttheorieën. Wanneer een formule een verzameling feiten samentrekt, beschouwen we deze formule als "aangevallen". Aangezien voor een default-theorie T = (W, D) geldt dat de verzameling feiten W de premissen voor een default-redenering representeert, wordt het verwijderen van een formule uit W geïnterpreteerd als een *undermining* (ondermijnende) aanval.

Door het definiëren van *rebuttal*, *undercut* en *undermining* zijn we erin geslaagd om een logische theorie te verschaffen voor alle drie de standaardtypen van argumentatieve aanvallen binnen de kunstmatige intelligentie. Daarnaast bieden de logica van default-rechtvaardiging, samen met de herzieningsoperaties op defaulttheorieën, een knooppunt tussen de default- en plausibele redeneerparadigma's binnen de kunstmatige intelligentie. In de *default justification logic* van hoofdstuk 2 hebben we te maken met de klassieke default redeervoorbeelden waarin uitsluitingsredenen een belangrijke rol spelen. In dit geval is het belang van het formaliseren binnen de *justification logic* van wat Toulmin "*warrants*" (rechtvaardigingen) noemt, dat het de state-of-the-art oplossingen voor het modelleren van uitsluitende redeneringen verbetert. Met de toevoeging van ondermijning maken we de omgang mogelijk met inconsistente informatieaanlevering en eliminatie van onbetrouwbare bronnen van rechtvaardiging, waarmee we onze aanpak van argumentatie met *justification logic* afronden.

We gebruiken ook systemen voor niet-monotoon redeneren om aan

te tonen dat Harmans argumenten tegen de normatieve rol van klassieke logica voor het menselijk redeneren niet succesvol zijn. We argumenteren vooral tegen het idee van uitzonderingsloze prescriptieve regels en tegen het idee van een enkel "*bridge principle*" (brugprincipe) dat de normatieve rol van de logica beknopt uitdrukt. In onze langzame defaultlogica beschouwen we normatieve regels enkel als herroepbare regels en we respecteren het feit dat het, uitgaande van een verzameling aanvankelijke overtuigingen, in principe onmogelijk is om te weten waar een redeneertaak precies over gaat. Dit laatste feit is een van de facetten van het frame-probleem dat bekend staat als "het relevantieprobleem", en we ontdekten dit probleem als een van de onderliggende problemen bij het definiëren van kandidaten voor een prescriptief *bridge principle*.

Ten slotte kijken we naar het probleem van doxastische bescheidenheidsbeweringen die het gevolg zijn van het erkennen door een actor van zijn eigen feilbaarheid. Volgens sommige filosofen moet een doxastisch bescheiden agent zijn feilbaarheid erkennen en geloven dat hij ten minste één onware overtuiging heeft. Daarom lijken doxastisch bescheiden actoren gerechtvaardigd om inconsistente overtuigingen te hebben. We stellen dat het overtuigd zijn van de directe uitspraak van doxastische bescheidenheid met de bewering "Ten minste één van mijn overtuigingen is onwaar" in feite niet leidt tot inconsistentie.

Auteurs die argumenteren dat doxastisch bescheiden redeneerders inconsistent moeten zijn, merken op dat ze de genoemde doxastische bescheidenheidsbewering moeten herformuleren om de bewering zelf van de reikwijdte van de geloofde beweringen vrij te stellen, en dus genoegen nemen met de bewering: "Ten minste één van mijn andere overtuigingen dan deze is onwaar". We argumenteren tegen deze ad-hoc-oplossing omdat deze ongerechtvaardigd een speciale status geeft aan de doxastische bescheidenheidsbewering zelf.

Bibliography

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

Gianvincenzo Alfano, Sergio Greco, Francesco Parisi, Gerardo Ignacio Simari, and Guillermo Ricardo Simari. An incremental approach to structured argumentation over dynamic knowledge bases. In Michael Thielscher, Francesca Toni, and Frank Wolter, editors, *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning, KR 2018*, pages 78–87, 2018.

Leila Amgoud. Postulates for logic-based argumentation systems. *International Journal of Approximate Reasoning*, 55(9):2028–2048, 2014.

Leila Amgoud and Philippe Besnard. A formal characterization of the outcomes of rule-based argumentation systems. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *International Conference on Scalable Uncertainty Management, SUM 2013*, volume 8078 of LNCS, pages 78–91. Springer, 2013.

Leila Amgoud and Mathieu Serrurier. Agents that argue and explain classifications. *Autonomous Agents and Multi-Agent Systems*, 16(2):187–209, 2008.

Grigoris Antoniou. *Nonmonotonic Reasoning*. Cambridge, MA: MIT Press, 1997.

Grigoris Antoniou. On the dynamics of default reasoning. *International Journal of Intelligent Systems*, 17(12):1143–1155, 2002.

Lennart Åqvist. Deontic logic. In Dov Gabbay and Frank H. Guenthner, editors, *Handbook of Philosophical Logic*, pages 605–714. Springer, 1984.

Sergei N. Artemov. Explicit provability and constructive semantics. *Bulletin of Symbolic Logic*, pages 1–36, 2001.

Sergei N. Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4):477–513, 2008.

Sergei N. Artemov. Justification awareness models. In Sergei N. Artemov and Anil Nerode, editors, *International Symposium on Logical Foundations of Computer Science*, volume 10703 of LNCS, pages 22–36. Springer, 2018.

Sergei N. Artemov and Melvin Fitting. Justification logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

Sergei N. Artemov and Melvin Fitting. *Justification Logic: Reasoning with Reasons*, volume 216 of *Cambridge tracts in mathematics*. Cambridge University Press, 2019.

Sergei N. Artemov and Elena Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, 2005.

Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Horacio Arló-Costa, Vincent F. Hendricks, Johan van Benthem, Henrik Boensvang, and Rasmus K. Rendsvig, editors, *Readings in Formal Epistemology*, pages 813–858. Springer, 2016.

Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In Luke Ong and Ruy de Queiroz, editors, *International Workshop on Logic, Language, Information, and Computation*, pages 168–190. Springer, 2012.

Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief, explicit knowledge, and conclusive evidence. *Annals of Pure and Applied Logic*, 165(1):49–81, 2014.

Henk Barendregt, Wil Dekkers, and Richard Statman. *Lambda Calculus with Types*. Cambridge University Press, 2013.

Pietro Baroni and Massimiliano Giacomin. Solving semantic problems with odd-length cycles in argumentation. In Thomas Dyhre Nielsen and Nevin Lianwen Zhang, editors, *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, volume 2711 of LNCS, pages 440–451. Springer-Verlag, 2003.

Pietro Baroni, Federico Cerutti, Massimiliano Giacomin, and Giovanni Guida. AFRA: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37, 2011.

Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.

Philippe Besnard and Anthony Hunter. A logic-based theory of deductive arguments. *Artificial Intelligence*, 128(1-2):203–235, 2001.

Philippe Besnard and Anthony Hunter. Practical first-order argumentation. In *Proceedings of the National Conference on Artificial Intelligence, AAAI'05*, volume 20(2), page 590. AAAI Press, 2005.

Philippe Besnard and Anthony Hunter. Constructing argument graphs with deductive arguments: A tutorial. *Argument & Computation*, 5(1): 5–30, 2014.

Andrei Bondarenko, Phan Minh Dung, Robert A Kowalski, and Francesca Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1-2):63–101, 1997.

Leonard G. Boonin. Concerning the defeasibility of legal rules. *Philosophy and Phenomenological Research*, 26(3):371–378, 1966.

Richard Booth, Souhila Kaci, Tjitze Rienstra, and Leendert van der Torre. A logical theory about dynamics in abstract argumentation. In Weiru Liu, V. S. Subrahmanian, and Jef Wijsen, editors, *International Conference on Scalable Uncertainty Management, SUM 2013*, volume 8078 of LNCS, pages 148–161. Springer, 2013.

Gerhard Brewka. Reasoning about priorities in default logic. In *Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI'94*, volume 2, pages 940–945. AAAI Press/The MIT Press, 1994.

Vladimir Brezhnev. On the logic of proofs. In Kristina Striegnitz, editor, *Proceedings of the Sixth ESSLLI Student Session, Helsinki*, pages 35–46, 2001.

Marco Cadoli and Marco Schaerf. Approximate inference in default logic and circumscription. *Fundamenta Informaticae*, 21(1, 2):103–112, 1994.

Martin W. A. Caminada. Contamination in formal argumentation systems. In Katja Verbeeck, Karl Tuyls, Ann Nowé, Bernard Manderick, and Bart Kuijpers, editors, *Proceedings of the 17th Belgium-Netherlands Conference on Artificial Intelligence, BNAIC 2005*. Koninklijke Vlaamse Academie van Belie voor Wetenschappen en Kunsten, 2005.

Martin W. A. Caminada. A gentle introduction to argumentation semantics. *Unpublished manuscript*, Summer 2008. Lecture material.

Martin W. A. Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

Martin W. A. Caminada and Dov M. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109, 2009.

Roderick M. Chisholm. *Theory of Knowledge*. Englewood Cliffs, NJ: Prentice-Hall, 1966.

Sheldon J. Chow. What's the problem with the frame problem? *Review of Philosophy and Psychology*, 4(2):309–331, 2013.

David Christensen. *Putting Logic in its Place: Formal Constraints on Rational Belief*. Oxford University Press, 2004.

David Christensen. Higher-order evidence. *Philosophy and Phenomenological Research*, 81(1):185–215, 2010.

Sylvie Coste-Marquis, Sébastien Konieczny, Jean-Guy Mailly, and Pierre Marquis. On the revision of argumentation systems: Minimal change of arguments statuses. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning, KR 2014*, 2014.

Claudio De Almeida. What Moore's paradox is about. *Philosophy and Phenomenological Research*, 62(1):33–58, 2001.

Florence Dupin de Saint-Cyr, Pierre Bisquert, Claudette Cayrol, and Marie-Christine Lagasquie-Schiex. Argumentation update in YALLA (yet another logic language for argumentation). *International Journal of Approximate Reasoning*, 75:57–92, 2016.

James P. Delgrande and Torsten Schaub. Expressing preferences in default logic. *Artificial Intelligence*, 123(1-2):41–87, 2000.

Daniel C. Dennett. Cognitive wheels: The frame problem of AI. In Christopher Hookway, editor, *Minds, Machines, and Evolution*, pages 129–152. Cambridge University Press, 1984.

Martin Diller, Adrian Haret, Thomas Linsbichler, Stefan Rümmele, and Stefan Woltran. An extension-based approach to belief revision in abstract argumentation. In Qiang Yang and Michael Wooldridge, editors, *Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, 2015.

Sylvie Doutre, Andreas Herzig, and Laurent Perrussel. A dynamic logic framework for abstract argumentation. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning, KR 2014*, 2014.

Fred Dretske. Is knowledge closed under known entailment? In Matthias Steup, John Turri, and Ernest Sosa, editors, *Contemporary Debates in Epistemology*, pages 13–26. Malden, MA: Wiley Blackwell, 2nd edition, 2005.

Phan M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

Phan Minh Dung, Robert A. Kowalski, and Francesca Toni. Assumption-based argumentation. In Iyad Rahwan and Guillermo R. Simari, editors, *Argumentation in artificial intelligence*, pages 199–218. Springer, 2009.

Catarina Dutilh Novaes. A dialogical, multi-agent account of the normativity of logic. *Dialectica*, 69(4):587–609, 2015.

Morten Elvang-Gøransson, Paul Krause, and John Fox. Dialectic reasoning with inconsistent information. In David Heckerman and Abe Mamdani, editors, *Proceedings of the Ninth International Conference on Uncertainty in Artificial Intelligence*, pages 114–121. Morgan Kaufmann Publishers Inc., 1993.

Simon J. Evnine. Believing conjunctions. *Synthese*, 118(2):201–227, 1999.

Simon J. Evnine. Learning from one's mistakes: Epistemic modesty and the nature of belief. *Pacific Philosophical Quarterly*, 82(2):157–177, 2001.

Tuan-Fang Fan and Churn-Jung Liau. A logic for reasoning about justified uncertain beliefs. In Qiang Yang and Michael Wooldridge, editors, *Proceedings of the IJCAI 2015*, pages 2948–2954. AAAI Press, 2015.

Hartry Field. What is the normative role of logic? *Aristotelian Society Supplementary Volume*, 83(1):251–268, 2009.

Branden Fitelson. Goodman's "new riddle". *Journal of Philosophical Logic*, 37(6):613–643, 2008.

Melvin Fitting. A logic of explicit knowledge. In Libor Běhounek and Marta Bílková, editors, *Logica Yearbook 2004*, pages 11–22. Prague: Filosofia, 2005a.

Melvin Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005b.

Melvin Fitting. Justification logics, logics of knowledge, and conservativity. *Annals of Mathematics and Artificial Intelligence*, 53(1-4):153–167, 2008.

Melvin Fitting. Reasoning with justifications. In *Towards Mathematical Philosophy*, pages 107–123. Springer, 2009.

Melvin Fitting. Possible world semantics for first-order logic of proofs. *Annals of Pure and Applied Logic*, 165(1):225–240, 2014.

Melvin Fitting. Modal logics, justification logics, and realization. *Annals of Pure and Applied Logic*, 167(8):615–648, 2016.

Melvin Fitting. Paraconsistent logic, evidence, and justification. *Studia Logica*, 105(6):1149–1166, 2017.

Jerry A. Fodor. *The Modularity of Mind: An Essay on Faculty Psychology*. Cmabridge, MA: MIT press, 1983.

Richard Foley. Justified inconsistent beliefs. *American Philosophical Quarterly*, 16(4):247–257, 1979.

Richard Foley. The epistemology of belief and the epistemology of degrees of belief. *American Philosophical Quarterly*, 29(2):111–124, 1992.

John Fox, David Glasspool, and Jonathan Bury. Quantitative and qualitative approaches to reasoning under uncertainty in medical decision making. In Silvana Quaglini, Pedro Barahona, and Steen Andreassen, editors, *Conference on Artificial Intelligence in Medicine in Europe, AIME 2001*, pages 272–282. Springer, 2001.

Gottlob Frege. *The Basic Laws of Arithmetic: Exposition of the System*. University of California Press, 1893/1964. [Translation of the 1893 original *Grundgesetze der Arithmetik*].

Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(1+ 2):95–138, 2004.

Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: Delp-servers, contextual queries, and explanations for answers. *Argument & Computation*, 5(1):63–88, 2014.

Peter Gärdenfors. Belief revision and nonmonotonic logic: Two sides of the same coin? In Jan van Eijck, editor, *European Workshop on Logics in Artificial Intelligence, JELIA 1990*, pages 52–54. Springer, 1990.

Peter T. Geach. Good and evil. *Analysis*, 17(2):33–42, 1956.

Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6): 121–123, 1963.

Nina Gierasimczuk. Bridging learning theory and dynamic epistemic logic. *Synthese*, 169(2):371–384, 2009.

Kurt Gödel. Vortrag bei Zilsel/Lecture at Zilsel's (1938a). In *Kurt Gödel: Collected Works: Volume III: Unpublished Essays and Lectures*, volume 3, pages 87–114. Oxford University Press, 1995.

Davide Grossi. Argumentation in the view of modal logic. In Peter McBurney, Iyad Rahwan, and Simon Parsons, editors, *7th International Workshop on Argumentation in Multi-Agent Systems, ArgMAS 2010*, volume 6614 of LNCS, pages 190–208. Springer, 2010.

Susan Haack. *Philosophy of Logics*. Cambridge University Press, 1978.

Dale Hample. The arguers. *Informal Logic*, 27(2):163–178, 2007.

Sven O. Hansson, editor. *Special issue on non-prioritized belief revision*, volume 63(1-2) of *Theoria*. Chichester: Wiley, 1997.

Sven O. Hansson. *A Textbook of Belief Dynamics: Theory Change and Database Updating*. Dordrecht: Kluwer, 1999a.

Sven O. Hansson. A survey of non-prioritized belief revision. *Erkenntnis*, 50(2-3):413–427, 1999b.

Gilbert Harman. *Thought*. Princeton, NJ: Princeton University Press, 1973.

Gilbert Harman. Logic and reasoning. *Synthese*, 60(1):107–127, 1984.

Gilbert Harman. *Change in View*. Cambridge, MA: MIT Press, 1986.

Gilbert Harman. Internal critique: A logic is not a theory of reasoning and a theory of reasoning is not a logic. *Studies in Logic and Practical Reasoning*, 1:171–186, 2002.

Gilbert Harman. Field on the normative role of logic. *Proceedings of the Aristotelian Society*, 109(1 pt 3):333–335, 2009.

Abdelraouf Hecham, Pierre Bisquert, and Madalina Croitoru. On a flexible representation for defeasible reasoning variants. In Mehdi Dastani, Gita Sukthankar, Elisabeth André, and Sven Koenig, editors, *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018*, pages 1123–1131. International Foundation for Autonomous Agents and Multiagent Systems, 2018.

Carl G. Hempel. *Philosophy of Natural Science*. Englewood Cliffs: Prentice-Hall, 1966.

Ulf Hlobil. We cannot infer by accepting testimony. *Philosophical Studies*, pages 1–10, 2018. doi: 10.1007/s11098-018-1142-3.

Wesley H. Holliday and Thomas F. Icard III. Moorean phenomena in epistemic logic. In Lev Beklemishev, Valentin Goranko, and Valentin Shehtman, editors, *Advances in Modal Logic 8*, pages 178–199. London: College Publications, 2010.

John F. Horty. Moral dilemmas and nonmonotonic logic. *Journal of Philosophical Logic*, 23(1):35–65, 1994.

John F. Horty. Reasoning with moral conflicts. *Noûs*, 37(4):557–605, 2003.

John F. Horty. Reasons as defaults. *Philosopher's Imprint*, 7(3):1–28, 2007.

John F. Horty. *Reasons as Defaults*. Oxford University Press, 2012.

Immanuel Kant. *Critique of Pure Reason*. Cambridge University Press, 1781/1998.

William M. Knorpp. The relevance of logic to reasoning and belief revision: Harman on 'Change in View'. *Pacific Philosophical Quarterly*, 78(1):78–92, 1997.

Ioannis Kokkinis, Petar Maksimović, Zoran Ognjanović, and Thomas Studer. First steps towards probabilistic justification logic. *Logic Journal of the IGPL*, 23(4):662–687, 2015.

Ioannis Kokkinis, Zoran Ognjanović, and Thomas Studer. Probabilistic justification logic. In Sergei N. Artemov and Anil Nerode, editors, *International Symposium on Logical Foundations of Computer Science*, volume 9537 of LNCS, pages 174–186. Springer, 2016.

Robert Koons. Defeasible reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition, 2017.

Paul Krause, Simon Ambler, Morten Elvang-Gøransson, and John Fox. A logic of argumentation for reasoning under uncertainty. *Computational Intelligence*, 11(1):113–131, 1995.

Roman Kuznets. On the complexity of explicit modal logics. In Peter G. Clote and Helmut Schwichtenberg, editors, *Computer Science Logic: 14th International Workshop, CSL 2000*, volume 1862 of LNCS, pages 371–383. Springer-Verlag, 2000.

Roman Kuznets and Thomas Studer. *Logics of Proofs and Justifications*. College Publications, 2019.

Henry E. Kyburg. Conjunctivitis. In Marshall Swain, editor, *Induction, Acceptance and Rational Belief*, pages 55–82. Springer, 1970.

Maria Lasonen-Aarnio. Higher-order evidence and the limits of defeat. *Philosophy and Phenomenological Research*, 88(2):314–345, 2014.

Hannes Leitgeb. A way out of the preface paradox? *Analysis*, 74(1):11–15, 2014.

Hector J. Levesque. Making believers out of computers. *Artificial Intelligence*, 30(1):81–108, 1986.

David Lewis. A subjectivist's guide to objective chance. In Richard C. Jeffrey, editor, *Studies in Inductive Logic and Probability*, volume 2, page 263. Berkeley: University of California Press, 1980.

John MacFarlane. Frege, Kant, and the logic in logicism. *The Philosophical Review*, 111(1):25–65, 2002.

John MacFarlane. In what sense (if any) is logic normative for thought? *Unpublished manuscript*, 2004. Delivered at the American Philosophical Association Central Division meeting.

David C. Makinson. The paradox of the preface. *Analysis*, 25(6):205–207, 1965.

David C. Makinson. On a fundamental problem of deontic logic. In Paul McNamara and Henry Prakken, editors, *Norms, Logics and Information Systems. New Studies on Deontic Logic and Computer Science*, volume 49 of *Frontiers in Artificial Intelligence and Applications*, pages 29–54. Amsterdam: IOS Press, 1999.

David C. Makinson and Peter Gärdenfors. Relations between the logic of theory change and nonmonotonic logic. In André Fuhrmann and Michael Morreau, editors, *The Logic of Theory Change: Workshop, Konstanz, FRG, October 13-15, 1989, Proceedings*, pages 183–205. Springer, 1991.

David C. Makinson and Leendert van der Torre. Input/output logics. *Journal of Philosophical Logic*, 29(4):383–408, 2000.

John McCarthy. Circumscription — a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1-2):27–39, 1980.

John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence*, volume 4, pages 463–502. Edinburgh University Press, 1969.

John-Jules Ch. Meyer and Wiebe van der Hoek. Non-monotonic reasoning by monotonic means. In Jan van Eijck, editor, *European Workshop on Logics in Artificial Intelligence, JELIA 1990*, pages 399–411. Springer, 1990.

Leonard G. Miller. Rules and exceptions. *Ethics*, 66(4):262–270, 1956.

Peter Milne. What is the normative role of logic? *Aristotelian Society Supplementary Volume*, 83(1):269–298, 2009.

Robert S. Milnikel. Derivability in certain subsystems of the logic of proofs is Π_2^p-complete. *Annals of Pure and Applied Logic*, 145(3):223–239, 2007.

Robert S. Milnikel. The logic of uncertain justifications. *Annals of Pure and Applied Logic*, 165(1):305–315, 2014.

Alexey Mkrtychev. Models for the logic of proofs. In Sergei Adian and Anil Nerode, editors, *Logical Foundations of Computer Science, 4th International Symposium, LFCS '97*, volume 1234 of LNCS, pages 266–275. Springer-Verlag, 1997.

Sanjay Modgil and Henry Prakken. Resolutions in structured argumentation. In Bart H. Verheij, Stefan Szeider, and Stefan Woltran, editors, *Computational Models of Argument: Proceedings of COMMA 2012*, pages 310–321. IOS Press, 2012.

Sanjay Modgil and Henry Prakken. The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation*, 5(1):31–62, 2014.

Robert C. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25(1):75–94, 1985.

Ernest Nagel. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace and World, 1961.

Christopher New. A note on the paradox of the preface. *The Philosophical Quarterly (1950-)*, 28(113):341–344, 1978.

Søren H. Nielsen and Simon Parsons. A generalization of Dung's abstract framework for argumentation: Arguing with sets of attacking arguments. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 54–73. Springer, 2006.

Friedrich W. Nietzsche. *The Gay Science: With a Prelude in German Rhymes and an Appendix of Songs*. Cambridge: Cambridge University Press, 1882/2001. [Translation of the 1882 original *Die fröhliche Wissenschaft*].

Zoran Ognjanović, Nenad Savić, and Thomas Studer. Justification logic with approximate conditional probabilities. In Alexandru Baltag, Jeremy Seligman, and Tomoyuki Yamada, editors, *Logic, Rationality and Interaction, 6th International Workshop, LORI 2017*, volume 10455 of LNCS, pages 681–686. Springer, 2017.

Stipe Pandžić. A logic of default justifications. In Eduardo Fermé and Serena Villata, editors, *Nonmonotonic Reasoning, 17th International Workshop, NMR 2018*, pages 126–135, 2018.

Stipe Pandžić. Reifying default reasons in justification logic. In Christoph Beierle, Marco Ragni, Frieder Stolzenburg, and Matthias Thimm, editors, *Proceedings of the KI 2019 Workshop on Formal and Cognitive Reasoning, DKB-KIK 2019*, volume 2445, pages 59–70. CEUR Workshop Proceedings, 2019.

Stipe Pandžić. On the dynamics of structured argumentation: Modeling changes in default justification logic. In Andreas Herzig and Juha Kontinen, editors, *Foundations of Information and Knowledge Systems, 11th International Symposium, FoIKS 2020*, volume 12012 of LNCS, pages 222–241. Springer, 2020.

John L. Pollock. The paradox of the preface. *Philosophy of Science*, 53(2): 246–258, 1986.

John L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

John L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57(1): 1–42, 1992.

John L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person*. Cambridge, MA: MIT Press, 1995.

John L. Pollock. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(1-2):233–282, 2001.

John L Pollock. A recursive semantics for defeasible reasoning. In Iyad Rahwan and Guillermo R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 173–197. Springer, 2009.

Henry Prakken. An argumentation framework in default logic. *Annals of Mathematics and Artificial Intelligence*, 9(1-2):93–132, 1993.

Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

Henry Prakken. Historical overview of formal argumentation. *IfCoLog Journal of Logics and their Applications*, 4(8):2183–2262, 2017.

Henry Prakken. *Commonsense Reasoning and Argumentation*. Utrecht University, 2018. Course reader.

Henry Prakken and John F. Horty. An appreciation of John Pollock's work on the computational study of argument. *Argument & Computation*, 3 (1):1–19, 2012.

Henry Prakken and Giovanni Sartor. The three faces of defeasibility in the law. *Ratio Juris*, 17(1):118–139, 2004.

Graham Priest. Intensional paradoxes. *Notre Dame Journal of Formal Logic*, 32(2):193–211, 1991.

Graham Priest. *An Introduction to Non-Classical Logic: From If to Is*. Cambridge University Press, 2008.

Graham Priest and Francesco Berto. Dialetheism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, spring 2017 edition, 2017.

Arthur N. Prior. On a family of paradoxes. *Notre Dame Journal of Formal Logic*, 2(1):16–32, 1961.

Arthur N. Prior. *Objects of Thought*. Oxford: Clarendon Press, 1971.

Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13 (1-2):81–132, 1980.

Bryan Renne. Multi-agent justification logic: Communication and evidence elimination. *Synthese*, 185(1):43–82, 2012.

Nicholas Rescher. *Plausible Reasoning: An Introduction to the Theory and Practice of Plausibilistic Inference*. Assen: Van Gorcum, 1976.

Nicholas Rescher. *Dialectics: A Controversy-Oriented Approach to the Theory of Knowledge*. Albany, NY: SUNY Press, 1977.

Adam Rieger. Moore's paradox, introspection and doxastic logic. *Thought: A Journal of Philosophy*, 4(4):215–227, 2015.

Bertrand Russell. *The Problems of Philosophy*. Oxford: Oxford University Press, 1912.

Gillian Russell. Logic isn't normative. *Inquiry*, pages 1–18, 2017. doi: 10.1080/0020174X.2017.1372305.

Sharon Ryan. The preface paradox. *Philosophical Studies*, 64(3):293–307, 1991.

Raymond M. Smullyan. Logicians who reason about themselves. In Joseph Y. Halpern, editor, *Proceedings of the 1986 conference on Theoretical Aspects of Reasoning about Knowledge*, pages 341–352. Morgan Kaufmann Publishers Inc., 1986.

Roy A. Sorensen. *Blindspots*. Oxford: Clarendon Press, 1988.

Roy A. Sorensen. Epistemic paradoxes. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, summer 2018 edition, 2018.

Florian Steinberger. Explosion and the normativity of logic. *Mind*, 125 (498):385–419, 2016.

Florian Steinberger. The normative status of logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, spring 2017 edition, 2017.

Florian Steinberger. Consequence and normative guidance. *Philosophy and Phenomenological Research*, 98(2):306–328, 2019a.

Florian Steinberger. Three ways in which logic might be normative. *Journal of Philosophy*, 116(1):5–31, 2019b.

Bart Streumer. Reasons and entailment. *Erkenntnis*, 66(3):353–374, 2007.

Che-Ping Su, Tuan-Fang Fan, and Churn-Jung Liau. Possibilistic justification logic: Reasoning about justified uncertain beliefs. *ACM Transactions on Computational Logic (TOCL)*, 18(2):15, 2017.

Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5(2):285–309, 1955.

Francesca Toni. A tutorial on assumption-based argumentation. *Argument & Computation*, 5(1):89–117, 2014.

Maggie E. Toplak and Keith E. Stanovich. The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 94(1):197, 2002.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958/2003.

Gabriel Uzquiano. Quantifiers and quantification. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2020 edition, 2020.

Hans van Ditmarsch and Barteld Kooi. The secret of my success. *Synthese*, 151(2):201–232, 2006.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, H. Bart Verheij, and Jean H. M. Wagemans. Argumentation and artificial intelligence. In *Handbook of Argumentation Theory*, pages 615–675. Springer, 2014.

Bert van Linder, Wiebe van der Hoek, and John-Jules Ch. Meyer. The dynamics of default reasoning. *Data & Knowledge Engineering*, 3(21): 317–346, 1997.

Bart Verheij. *Rules, Reasons, Arguments: Formal Studies of Argumentation and Defeat*. PhD thesis, Maastricht: Universiteit Maastricht, 1996.

Bart Verheij. DefLog: On the logical interpretation of prima facie justified assumptions. *Journal of Logic and Computation*, 13(3):319–346, 2003.

Bart Verheij. The Toulmin argument model in artificial intelligence. In Iyad Rahwan and Guillermo R. Simari, editors, *Argumentation in Artificial Intelligence*, pages 219–238. Springer, 2009.

Bart Verheij. Correct grounded reasoning with presumptive arguments. In Michael Loizos and Antonis Kakas, editors, *European Conference on Logics in Artificial Intelligence, JELIA 2016*, pages 481–496. Springer, 2016.

Georg H. von Wright. Deontic logic. *Mind*, 60(237):1–15, 1951.

Gerard A. W. Vreeswijk. *Studies in defeasible argumentation*. PhD thesis, Free University of Amsterdam, 1993.

John N. Williams. Inconsistency and contradiction. *Mind*, 90(360):600–602, 1981.

John N. Williams. The preface paradox dissolved. *Theoria*, 53(2-3):121–140, 1987.

Ludwig Wittgenstein. *Philosophical Investigations*. Basil Blackwell, 2nd edition, 1953/1958. [Translation of the original manuscript *Philosophische Untersuchungen*].

Max Zorn. A remark on method in transfinite algebra. *Bulletin of the American Mathematical Society*, 41(10):667–670, 1935.

Appendix A

Proof of Theorem 2.5. The claim from left to right is obvious. For the other direction, take $\mathcal{CS}$ to be some specific axiomatically appropriate and injective constant specification. We first show that if a set Γ is $\mathbf{JT_{CS}}$-finitely satisfiable, then for all formulas $F \in Fm$, it holds that $\Gamma \cup \{F\}$ or $\Gamma \cup \{\neg F\}$ is $\mathbf{JT_{CS}}$-finitely satisfiable. Suppose that Γ is $\mathbf{JT_{CS}}$-finitely satisfiable and that $\Gamma \cup \{F\}$ and $\Gamma \cup \{\neg F\}$ are both not $\mathbf{JT_{CS}}$-finitely satisfiable. Then there would be finite subsets Γ' and Γ'' of Γ such that $\Gamma' \cup \{F\}$ and $\Gamma'' \cup \{\neg F\}$ are not $\mathbf{JT_{CS}}$ satisfiable. Since for no interpretation $\mathcal{I}$ it holds that $\mathcal{I} \models \{F, \neg F\}$, $\Gamma' \cup \{F, \neg F\}$ is never $\mathbf{JT_{CS}}$ satisfiable. Note that for any possible interpretation $\mathcal{I}$ one of the formulas F or $\neg F$ holds. This means that $\mathcal{I} \models \Gamma' \subseteq \mathcal{A}'$ for a class of interpretations $\mathcal{A}'$ such that for each $\mathcal{I}' \in \mathcal{A}'$, it holds that $\mathcal{I}' \models \neg F$. In a similar way we get that $\mathcal{I} \models \Gamma'' \subseteq \mathcal{A}''$ for a class $\mathcal{A}''$ consisting of the interpretations $\mathcal{I}''$ such that $\mathcal{I}'' \models F$. Therefore, we have that $\mathcal{I} \models \Gamma' \cap \mathcal{I} \models \Gamma'' = \emptyset$ and, thus, $\Gamma' \cup \Gamma''$ is not $\mathbf{JT_{CS}}$-satisfiable. But $\Gamma' \cup \Gamma''$ is a finite subset of Γ and this contradicts the assumption that Γ is $\mathbf{JT_{CS}}$-finitely satisfiable.

The next step is proving a $\mathbf{JT_{CS}}$ variant of the Lindenbaum lemma. Using the above-proven statement that for any $\mathbf{JT_{CS}}$-finitely satisfiable set of formulas Γ and any formula F, $\Gamma \cup \{F\}$ or $\Gamma \cup \{\neg F\}$ is $\mathbf{JT_{CS}}$-finitely satisfiable together with the fact that $\Gamma \cup \{F, \neg F\}$ is never $\mathbf{JT_{CS}}$-finitely satisfiable, we can construct maximally $\mathbf{JT_{CS}}$-finitely satisfiable sets. Let $F_1, F_2, F_3, \ldots$ be an enumeration of $F \in Fm$. For a $\mathbf{JT}_{CS}$-finitely satisfiable set Γ and for all $i \in \mathbb{N}$ define an increasing sequence of sets of formulas as follows:

$\Gamma_0 = \Gamma$

$\Gamma_{i+1} = \Gamma_i \cup \{F_i\}$ if $\Gamma_i \cup \{F_1\}$ is $\mathbf{JT}_{CS}$-finitely satisfiable, otherwise $\Gamma_{i+1} = \Gamma_i \cup \{\neg F_i\}$

$$\Gamma' = \bigcup_{i=0}^{\infty} \Gamma_i$$

We can prove that Γ' is $\mathbf{JT}_{CS}$-finitely satisfiable by induction. The base case $\Gamma_0 = \Gamma$ holds by assumption. Then we claim that for all $i \in \mathbb{N}$, Γ_i is $\mathbf{JT_{CS}}$-finitely satisfiable. For some $n \in \mathbb{N}$, take Γ_n to be $\mathbf{JT_{CS}}$-finitely satisfiable. Then either $\Gamma \cup \{F_n\}$ or $\Gamma \cup \{\neg F_n\}$ is $\mathbf{JT_{CS}}$-finitely satisfiable and, therefore, Γ_{n+1} is also $\mathbf{JT_{CS}}$-finitely satisfiable.

From the construction of the increasing sequence, we have that for any finite set $\Gamma_k \subseteq \Gamma'$ there is a $\mathbf{JT_{CS}}$-finitely satisfiable finite set $\Gamma_{k+1} \subseteq \Gamma'$ such that $\Gamma_k \subseteq \Gamma_{k+1}$ and, therefore, Γ_k is $\mathbf{JT_{CS}}$-satisfiable. Since any finite subset of Γ' is $\mathbf{JT_{CS}}$ satisfiable, Γ' is $\mathbf{JT_{CS}}$-finitely satisfiable. The set Γ' is maximal according to the enumeration of the set of formulas Fm and contains exactly one of F_i or $\neg F_i$ for all $i \in \mathbb{N}$.

Now we define a valuation v such that $v(P) = \mathit{True}$ iff $P \in \Gamma'$ and the reason assignment $*(t) = \{F \mid t : F \in \Gamma'\}$. We only need to check the conditions on the reason assignment function. First, we show that $*(\cdot)$ satisfies the application condition. Since the formula $t : (F \to G) \to (u : F \to (t \cdot u) : G)$ is $\mathbf{JT_{CS}}$ valid, it is contained in Γ'. If $F \to G \in *(t)$ and $F \in *(u)$, then $\{t : (F \to G), u : F\} \in \Gamma'$. Since Γ is closed under *Modus ponens*, we have that $(t \cdot u) : G \in \Gamma'$ and, therefore, $G \in *(t \cdot u)$. Similarly, since the formulas $t : F \to (t + u) : F$ and $u : F \to (t + u) : F$ are both in Γ' we can easily check that the sum condition holds for $*(\cdot)$.

Finally, we have defined an interpretation $\mathcal{I} = (*, v)$ that meets $\mathcal{CS}$ and we need to prove that truth in this interpretation is equivalent to inclusion in Γ':

$$\mathcal{I} \models F \text{ iff } F \in \Gamma'$$

The proof is by induction on the structure of F. For the base case, suppose F is an atomic formula P: $\mathcal{I} \models P$ iff $v(P) = \mathit{True}$ iff $P \in \Gamma'$.

For the inductive step, suppose that if the result holds for F and G, then it also holds for $\neg F$, $F \wedge G$, $F \vee G$, $F \to G$ and $t : F$. For the negation case: $\mathcal{I} \models \neg F$ iff $\mathcal{I} \not\models F$. By the inductive hypothesis, $\mathcal{I} \not\models F$ iff $F \notin \Gamma'$. By the maximality of Γ', we have that $F \notin \Gamma'$ iff $\neg F \in \Gamma'$.

For the conjunction case: $\mathcal{I} \models F \wedge G$ iff $\mathcal{I} \models F$ and $\mathcal{I} \models G$. By the inductive hypothesis, $\mathcal{I} \models F$ and $\mathcal{I} \models G$ iff $F \in \Gamma'$ and $G \in \Gamma'$ iff $F \wedge G \in \Gamma'$. Since other connectives are definable in terms of $\neg$ and $\wedge$, we skip the remaining cases.

Finally for the justified formula case: $\mathcal{I} \models t : F$ iff $F \in *(t)$. By the definition of $*(\cdot)$, it holds that $F \in *(t)$ iff $t : F \in \Gamma'$.

Therefore, for any $\mathbf{JT_{CS}}$-finitely satisfiable set Γ there is an interpretation $\mathcal{I}$ based on a maximal $\mathbf{JT_{CS}}$-finitely satisfiable extension Γ' of Γ such that $\mathcal{I} \models \Gamma$. $\square$

Acknowledgements

Reflecting on my PhD project, I feel that I have progressed in leaps and bounds compared to where I was four years ago. All other things aside, none of this would be possible without a high-quality relation that I had with my thesis supervisors. It is already a good sign if a PhD student has one supervisor to rely on, but in my case, I had a privilege to rely on a team of three excellent professionals: Barteld Kooi, Rineke Verbrugge and Allard Tamminga.

I am grateful to Barteld for being open to all sorts of different approaches that did not always fit into our initial plans. I fully enjoyed the freedom to choose my own research path and the breadth of imagination with which you followed me along that path. I will especially remember many occasions on which I bounced some of my thoughts off you and you would effortlessly engage with those thoughts in your unique, almost playful ways. I am now confident to say that after my detour into non-monotonic logics, which you had kindly supported, I now fully understand the original problem that you wanted me to solve at the start of this project — and all this only four years too late.

I find it very difficult to put into words how important Rineke's role was in my project, and this is ranging from its very small details to its large scale prospects. It is not uncommon to see that a person with your calibre of abstract thinking capacities has no other significant traits, but you certainly are much more than just a brilliant scholar. I am sure that many colleagues would feel intimidated by your talent, if it weren't combined with your warmhearted and uncomplicated personality. I wish to thank you for having patience to answer to almost any question that bothered me, be it on a philosophical, mathematical or organizational level. I wish to also thank you for your selfless guidance of my project and your flawless management skills, without which things might have gotten out of order on so many occasions. I am proud to have been one

of your PhD students and, if I get a chance to have students of my own, I will at least try my best to guide them so as to follow your example. In fact, I believe that there are many professors out there who could use a couple of lessons from you.

Since the start of my project, I have had a couple of decent ideas, but I can hardly imagine how would my thesis ever pan out without numerous invaluable comments that I received from Allard. I think that your didactic approach based on honesty and astute criticism was the crucial factor in learning about the process of getting from raw intuitions to clearer ideas. I am aware that my drafts occasionally took you on an unexpected "roller coaster in the dark" that you had to endure. But with each new draft, you were ready to have a fresh outlook on my ideas, even when I lost my patience to work on them. I have always admired your dedication to supervision and your impeccable sense of the role that each small detail has in an overall idea. I feel that it is almost impossible that any flimsy idea would go unnoticed under your radar. The breadth of your knowledge has been a great resource for me to learn about academic crafts and beyond. I will also always remember how you made me feel welcome in Groningen when I moved here four years ago. Thank you for everything you have done for me professionally and personally.

I am grateful to the members of the assessment committee of my thesis, namely to Jan Broersen, Thomas Studer and Bart Verheij, for their time and their valuable comments given on this thesis. I am especially grateful to Bart from whom I learned most of the things I know about formal argumentation, be it on different occasions when we had interesting conversations or by reading his inspiring work.

At the beginning of this year, I had a chance to visit Professor Sergei Artemov and his group at the Graduate Center of the City University of New York with a support of the Evert Willem Beth Foundation. I wish to thank Sergei for giving me a chance to present my work at the panel discussion during the symposium on "Logical Foundations of Computer Science". Professor Artemov and his former and current PhD students made me feel at home during my visit, not least because of their generous support of my work.

During the last four years, I have been lucky to work with many intelligent, friendly and supportive colleagues. First of all, I wish to thank my paranymphs and colleagues Kritika and Yuri. I am especially indebted to Yuri for many times that he commented on my work as well as for many occasions on which we laughed at the ways in which

academics spoil the beauty of things that they study. It really helps to know that there is someone who will always understand you. To both of you, thanks for everything that you have done for me.

I spent most of my PhD project working at the Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, and I truly enjoyed being there. Thanks to all the members of the Multi-Agent Systems group for many interesting meetings and to all the PhD students for all the nice moments outside the work. Thanks to Xingchi and Hang for being wonderful friends and great office mates. Special thanks goes to Ana who always figures out how to galvanize and "Balkanize" everyone on unremarkable dull days.

I am grateful to have been a part of the vibrant Faculty of Philosophy community. The PhD office overlooking the Oude Boteringestraat is a source of fond memories to me. Thanks to Li-Chih, Sanne, Job, Crystel, Merel, Yanjun and (as a cherry on top) to my long-time friend Petar for all the nice moments we got to share. I am especially grateful to Lieuwe for his sincere friendship; for all the serious conversations, silly jokes and all the strange mixtures of the two. I regret to inform you, however, that naming any of my future children after you is not a serious option — it's just too many vowels that don't make much sense to us.

Among many friends in Croatia whom I like to thank are Ana, Toni, Damir, Dado, Višeslav and Vatroslav. Special thanks goes to my always positive in-laws, cheerful family Vukasović: Vesna, Zoran, Sonja, Kruno, Eva and their honorary members *Baka* Katica and Boni. Moreover, I am grateful to professor Srećko Kovač, who sparked my interest in formal philosophy and who first introduced me to justification logic.

I wish to also kindly thank my family in Herzegovina, to my parents Iva and Milan and my sister Danijela, for giving me all the love and support that I could ever ask for. I dedicate this thesis to my brothers Filip and Vladimir. It was due to your selfless support that I had been able to pursue any education whatsoever in the first place. I have always admired your persistence in high moral standards and the values that you stand for, despite having to fight your ways through a surrounding of often corrupted morals and inverted values.

(*Zahvaljujem od srca svojoj obitelji u Hercegovini, mojim roditeljima Ivi i Milanu te sestri Danijeli, za svu ljubav i podršku koju sam ikada mogao tražiti. Ovu disertaciju posvećujem svojoj braći Filipu i Vladimiru. Pristup obrazovanju imao sam samo zahvaljujući vašoj nesebičnoj podršci. Oduvijek cijenim vašu posvećenost visokim moralnim standardima i vrijednosti koje zastupate, usprkos*

okruženju narušenog morala i izvrnutih vrijednosti u kojem ste se osamostalili.)

The last, and the happiest, word of thanks and the last part of this small thesis goes to where things that matter most start — to my wife Karmen. Thank you for sharing a wonderful life with me. Let's be honest here, I would have ended up being quite a weird guy without your exuberant personality that made me realize that it is fine to spend some time outside my own head. I can't wait to see all the things that are still ahead of us!