The Carneades Model of Argument and Burden of Proof

Thomas F. Gordon

Fraunhofer FOKUS
Berlin, Germany
email: thomas.gordon@fokus.fraunhofer.de

Henry Prakken

Department of Information and Computing Sciences, Utrecht University
and Faculty of Law, University of Groningen
Utrecht and Groningen, The Netherlands
email: henry@cs.uu.nl

Douglas Walton

Department of Philosophy, University of Winnipeg
Winnipeg, Manitoba, Canada
email: d.walton@uwinnipeg.ca

Abstract

We present a formal, mathematical model of argument structure and evaluation, taking seriously the procedural and dialogical aspects of argumentation. The model applies proof standards to determine the acceptability of statements on an issue-by-issue basis. The model uses different types of premises (ordinary premises, assumptions and exceptions) and information about the dialectical status of statements (stated, questioned, accepted or rejected) to allow the burden of proof to be allocated to the proponent or the respondent, as appropriate, for each premise separately. Our approach allows the burden of proof for a premise to be assigned to a different party than the one who has the burden of proving the conclusion of the argument, and also to change the burden of proof or applicable proof standard as the dialogue progresses from stage to stage. Useful for modeling legal dialogues, the *burden of production* and *burden of persuasion* can be handled separately, with a different responsible party and applicable proof standard for each. Carneades enables critical questions of argumentation schemes to be modeled as additional premises, using premise types to capture the varying effect on the burden of proof of different kinds of questions.

Key words: Argument Structure, Argument Graphs, Argument Evaluation, Argumentation Schemes, Burden of Proof, Proof Standards, Legal Argument

1 Introduction

This article presents a functional model of the evaluation of commonsense arguments, taking seriously the procedural and dialogical aspects of argumentation. The model, called Carneades in honor of the Greek skeptic philosopher who emphasized the importance of plausible reasoning [9, vol. 1, p. 33–34], applies proof standards [10] to determine the acceptability of statements on an issue-by-issue basis. The model has been implemented using a functional programming language. This system, also called Carneades, supports a range of argumentation tasks, including argument reconstruction, evaluation and visualization.

Carneades is meant to overcome several limitations of current 'mainstream' AI work on argumentation. The mainstream approach essentially regards the problem of argument evaluation to be a question of defining the appropriate (nonmonotonic) relation between sets of statements: a statement 'follows' from a set of statements if an argument for it can be constructed that survives all attacks by counterarguments that can be constructed from the same set of statements. (This idea can be refined in various ways but that is not the issue here). This 'relational' approach ignores that arguments are embedded in a procedural context, in that they can be seen as having been put forward on one side or the other of an issue during a dialogue between (human and/or artificial) agents. This dialogical context must be taken into account when evaluating arguments. The information provided by a dialogue for constructing and evaluating arguments is richer than just a set of sentences. The context can tell us whether some party has questioned or conceded a statement, or whether a decision has been taken, perhaps by a third party, to accept or reject a claim, taking into consideration the arguments which have been made. Such decisions may be taken at intermediate states of the dialogue, for example when some alternatives are eliminated after brainstorming during a deliberation. Moreover, the dialogue context may provide information about the allocation of the burden of proof for each statement. Legal applications may require the *burden of production* and *burden of persuasion* to be handled separately, with a different responsible party and applicable proof standard for each [35]. This allocation may well change over the course of the dialogue. In the law of civil procedure, for example, the burden of proof may be allocated to the party who has the better access to the evidence. Finally, the proof standard may depend on the phase of the dialogue. For example, a weak proof standard may be appropriate during the brainstorming phase of a deliberation or during the pleading phase of a legal conflict.

The Carneades model has been designed to be applied in such rich dialogical contexts. The evaluation of arguments in Carneades depends on whether statements have been questioned or decided; the allocation of the burden of proof; and the proof standard applicable to questioned statements. All these elements depend on the stage and context of the dialogue in which the arguments have been put forward.

An influential classification of dialogue types is that of Walton and Krabbe [51]. For present purposes their distinction between persuasion and deliberation dialogue is especially relevant. The goal of a deliberation dialogue is to solve a problem while the goal of a persuasion dialogue is to test whether a claim is acceptable. The present version of Carneades is meant to support persuasion dialogues. In such dialogues, two or more participants try to resolve a difference of opinion by arguing about the tenability of a claim, each trying to persuade the other participants to adopt their point of view. Dialogue systems regulate such things as the preconditions and effects of speech acts, including their effects on the commitments of the participants, as well as criteria for terminating the dialogue and determining its outcome. Good dialogue systems regulate all this in such a way that conflicting viewpoints can be resolved in a way that is both fair and effective [24].

It is important to note that although Carneades has been designed to be embedded in a procedural context, it does not itself define a dialogue protocol. No roles, speech acts, termination criteria, or procedural rules are defined. Instead Carneades is intended to be a reusable component providing services generally needed when specifying such argumentation protocols.

In line with prior AI research, arguments in Carneades are defeasible, i.e, arguments can be defeated by counterarguments. The Carneades model of defeasible argument is founded on Walton's theory of argumentation schemes [48]. Argumentation schemes express reasoning policies, i.e. conventional patterns of reasoning, and thus are dependent on the norms of the community. Arguments in Carneades are designed to model instantiations of argumentation schemes. Besides being defeasible, argumentation schemes have a dialogical aspect in that they come with a set of critical questions [20], which enumerate ways of challenging arguments created using the scheme. Critical questions differ with regard to their impact on the burden of proof [3,43]. For some critical questions, merely asking the question is enough to shift the burden of proof back to the party who put forward the argument to answer the question. For other critical questions, the party who raised the question also has the burden of answering it. Carneades models critical questions as additional premises of an argument, with a different type of premise, called assumptions and exceptions, for each kind of question. [1]

The Carneades model is the latest result of a research effort that started with the Pleadings Game [14], a computational model of civil pleading in Anglo-American Law. Besides a dialogue protocol, the Pleadings Game also included a component for evaluating arguments in states of a dialogue. Like Carneades, the status of state-

[1] In an earlier version of Carneades, as reported in [17], assumptions were called 'presumptions'. This usage however conflicted with the meaning of 'presumption' in our main intended application field, the legal domain. In the law, merely questioning a legal presumption is not enough to shift the burden of proof to the other party; the burden of proof is on the party interested in rebutting the presumption. For example, the presumption of innocence in criminal cases places the burden on the prosecution to prove guilt.

ments in the dialogue state was taken into account by this evaluation. (Burden of proof and proof standards were not modeled in the Pleadings Game.) Whereas statements and arguments were modeled in the Pleadings Game using a specific logical calculus, Geffner's logic of Conditional Entailment [12], Carneades is designed to be an open integration framework for various kinds of argumentation schemes, using whatever kind of knowledge representation is appropriate for each kind of scheme. Thus, like Dung's abstract model of argument [8], Carneades does not depend on a particular logical language for expressing statements, inference rules or argumentation schemes.

Another ancestor of Carneades is Zeno [15], an argumentation model based on Horst Rittel's idea of an Issue-Based Information System (IBIS) [38]. In IBIS, issues can be raised, ideas to resolve them can be proposed and arguments pro and con the various ideas can be put forward. Zeno was intended to be simple enough for use in web-based mediation systems and targeted to support practical decisions in deliberation dialogues, about what action to take. Later versions of Zeno were used in several e-democracy pilot projects [16]. Like the present work, Zeno included a model of proof standards and was intended as an integration framework for arguments from "heterogeneous information sources and models". Whereas Zeno was designed for deliberation dialogues and based on the Rittel's IBIS model, Carneades has been designed primarily for persuasion dialogues and is based on Walton's philosophy of argumentation. It remains for future work to see whether Carneades will also be suitable for deliberation dialogues. [2]

The rest of the article is structured as follows. Section 2 presents some features of (natural) argumentation that motivate the design of Carneades, in particular argumentation schemes, dialogical attitudes towards statements, and the notion of burden of proof. Section 3 defines the structure of argument graphs and illustrates them with examples from related work by Toulmin and Pollock. Section 4 formally defines the acceptability of statements in argument graphs in terms of proof standards, premise types, and the dialectical status of statements in dialogues. Section 5 returns to the subject of critical questions and illustrates how they can be modeled using Carneades. Section 6 discusses how to model the distribution of the burdens of production and persuasion using Carneades. Section 7 presents related work on computational models of argument. Finally, we close in Section 8 with a recapitulation of conclusions and ideas for future work.

[2] Persuasion dialogues, like deliberation, can be about a course of action. The difference between persuasion and deliberation is the starting point and goal of the dialogue, not its subject matter. A persuasion dialogue begins with the making of a claim and has the goal of persuading the opponent to accept this claim. Such claims can be about what course of action would be best, or any other topic. Deliberation dialogues, on the other hand, begin with a problem and have the goal of finding a solution to this problem. Typically, the first phase of a deliberation has the goal of identifying stakeholders and their interests. Only later, perhaps after brainstorming, are proposals for actions collected. When such a proposal is defended, a deliberation dialogue can temporarily shift to a persuasion dialogue.

2 Theoretical Background

In this section we look in some more detail at some features of natural argumentation that motivate the design of Carneades. In particular, we discuss argumentation schemes, dialogical attitudes towards statements, and the notion of burden of proof.

Argumentation schemes are stereotypical patterns of reasoning used in everyday conversational argumentation, and in other contexts as well, like legal and scientific argumentation. They represent patterns of non-deductive reasoning that have long been studied in argumentation theory. Historically, they are the descendants of the so-called topics of Aristotle, thought to be useful for inventing arguments as well as for evaluating them. In recent times, many schemes have been identified and studied [11]. [3] Despite this body of work, little has been done in argumentation theory to formalize the logical structure of the schemes. By contrast, artificial intelligence has become increasingly interested in argumentation schemes, due to their potential for making significant improvements in the reasoning capabilities of artificial agents [37,43,3].

In argumentation theory, the device used for evaluating schemes [20] is that of a specific set of critical questions matching each scheme. The questions represent attacks, challenges or criticisms that, if not answered adequately, make the argument fitting the scheme default. The critical questions evidently need to be formalized along with the scheme, in order to capture the logic of each scheme as a defeasible argument. The biggest problem is how to carry out this task, since the relationship of a set of critical questions to a scheme brings in the notion of a dialogue sequence between a questioner and an answerer. [4]

Let us illustrate the concept of argumentation schemes with the scheme for arguments from expert opinion, as formulated in [49, p. 210], with some minor notational changes:

Major Premise. Source E is an expert in the subject domain S containing proposition A.
Minor Premise. E asserts that proposition A in domain S is true.
Conclusion. A may plausibly be taken as true.

The six basic critical questions matching the appeal to expert opinion [49, p. 223] are the following.

[3] We do not address in this article how argumentation schemes develop or evolve, or how to design good argumentation schemes.
[4] Dialogues do not require multiple agents. A single agent can be viewed as arguing with himself, switching back and forth between pro and con viewpoints. The proponent and opponent of a dialogue are roles, not agents. Any number of agents can occupy these roles.

(1) How credible is E as an expert source?
(2) Is E an expert in the field that A is in?
(3) Does E's testimony imply A?
(4) Is E reliable?
(5) Is A consistent with the testimony of other experts?
(6) Is A supported by evidence?

The method of evaluating an argument like one from expert opinion is by a shifting of burden of proof in a dialogue [49]. When the respondent asks one of the six critical questions a burden of proof shifts back to the proponent's side, defeating or undercutting the argument temporarily until the critical question has been answered successfully. However, as will be suggested below, there are differences between the critical questions on how strongly or weakly they produce such a shift.

The notion of a shift in the burden of proof from one side to another is a problematic topic in the special kind of case in which an argument has been put forward and a critical question matching the scheme for that argument has been asked. Does merely asking the question make the argument default, or is the burden on the questioner to provide evidence? Providing a dialogue mechanism to handle a representation of how the burden of proof should shift in such an exchange goes beyond what is currently available. Such observations have led to two theories about the allocation of the burden of proof when critical questions are asked [50].

According to one theory, when the respondent asks a critical question, the burden of proof automatically shifts back to the proponent's side to provide an answer, and if she fails to do so, the argument defaults (is defeated). On this theory, only if the proponent provides an appropriate answer is the plausibility of the original argument restored. This could be called the shifting burden theory, or SB theory.

According to the other theory, asking a critical question should not be enough to make the original argument default. The question, if asked, needs to be backed up with some evidence before it can shift any burden back to the proponent. This could be called the backup evidence theory, or BE theory.

In AI it has recently been observed that argumentation schemes are related to defeasible inference rules in nonmonotonic logic. Bex et al. [3] argue that argumentation schemes can be formalized as Pollock's [29] defeasible reasons and that critical questions that should be supported by backup evidence can be regarded as undercutters to such defeasible reasons. (Undercutters [29] providing reasons for breaking the connection between premises and conclusion of another reason in the given circumstances. They can be used to express exceptions to defeasible reasons.) Verheij [43] proposes a similar account, suggesting that it may be useful to handle different kinds of questions differently. He begins by showing that critical questions can have four different kinds of roles.

(1) They can be used to question whether a premise of a scheme holds.

(2) They can point to exceptional situations in which a scheme should not be used.
(3) They can set conditions for the proper use of a scheme.
(4) They can point to other arguments that might be used to attack the scheme.

Verheij then applies these ideas within his Deflog logic [42]. Critical questions of type (2) undercut an argument, those of type (3) refute specific implicit assumptions on which the argument rests, while those of type (4) point to 'rebutting' counter-arguments. (Rebuttals are in Pollock (1995) counterarguments with a conclusion that is contradictory to that of its target.) Finally, Verheij [43] argues that critical questions that address the premises of a scheme are redundant, because they merely ask whether the premise is true.

Such accounts in terms of ordinary premises and exceptions of a defeasible inference rule provide a natural explanation as to why some critical questions shift the burden of answering back to the proponent of the argument while other questions must be backed up with evidence. However, they leave another issue unaddressed, viz. the dialogical status of (ordinary) premises of an argumentation scheme. Verheij's claim that critical questions about the premises of a scheme are redundant makes sense in a logical setting, with a static information state, in which the only issue is whether a premise is given or derivable from what is given. In a dialogical setting, however, with a possibly changing information state, this is different. Suppose, for example, that an argument instantiating the scheme from Expert Opinion has been put forward and the other side has not yet said anything in response. The question then is whether in the current dialogical context the conclusion of the argument is acceptable. This boils down to the question whether silence implies consent to the major and minor premise or not. Or to give a legal example, suppose the plaintiff in a legal dispute claims that he and the defendant have a contract, arguing that he made an offer accepted by the defendant and citing the legal rule that an offer and acceptance constitute a contract. Although the argument is logically perfectly acceptable, according to Dutch civil law, e.g., it is not sufficient, since claims about offer and acceptance must be backed by evidence. Suppose the plaintiff provides such evidence, for example in the form of witness testimonies. If the defendant continues not to reply, the plantiff's argument will now be acceptable, since the Dutch law of civil procedure states that factual claims not questioned by the other party must be accepted by the court.

The upshot of all this is that in a dialogical context different kinds of premises must be distinguished: those that must always be supported with further grounds; those that can be assumed until they have been questioned; and those that do not hold in the absence of evidence to the contrary. In this article we will call the first kind 'ordinary premises', the second kind 'assumptions' and third kind 'exceptions'. Note that assumptions are not the same as the negation of exceptions. [5] We claim that

[5] Thus our use of the term 'assumption' deviates, e.g., from Bondarenko et al. [4], where it is used for negations of what we call exceptions.

'static', relational AI models of argumentation cannot handle the distinctions between these kinds of premises, since these distinctions only make sense in a procedural context. Carneades, by contrast, has been designed for use in such procedural contexts.

The final feature of argumentation motivating the design of Carneades concerns the allocation of the burden of proof, especially in legal domain. The general intuitions here are very much related to the distinction between the types of premises just discussed. When a party provides an argument instantiating some applicable scheme, the burden is on that party to prove its ordinary premises and (once challenged) its assumptions, after which the burden shifts to the other party to defeat the argument by rebutting it or pointing out exceptional circumstances. Legal systems make fine-grained distinctions between different kinds of burdens of proof and we want Carneades to account for these distinctions.

In civil law suits, the plaintiff generally has the burden of proof. This seems simple enough, but the notion of burden of proof can refer to two different things [52, p. 270], often called the risk of non-persuasion, or burden of persuasion, and the burden of production: "The risk of non-persuasion operates when the case has come into the hands of the jury, while the duty of producing evidence implies a liability to a ruling by the judge disposing of the issue without leaving the question open to the jury's deliberations." Wigmore wrote that the risk of non-persuasion never shifts, but the duty of producing evidence to satisfy the judge does have this shifting characteristic [52, pp. 285-286]. McCormick on Evidence [40, p. 425] defines the burden of producing evidence in these terms: "The burden of producing evidence on an issue means the liability to an adverse ruling (generally a finding or directed verdict) if evidence on the issue has not been produced.". In contrast, the burden of persuasion [40, p. 426] means that if the party having that burden has failed to satisfy it, the issue is to be decided against that party. Park, Leonard and Goldberg [26, p. 88] wrote that the burden of proof involves two things – the amount of evidence required to establish the ultimate question of fact, and the allocation of the risk of non-persuasion to that degree. The burden of persuasion, in their terms [26, p. 89], defines the degree to which the fact-finder must be persuaded in order for the ultimate claim to be proved, and which party must meet that burden. There are various degrees, like "more likely than not", "beyond reasonable doubt" and so forth. This leads us in turn to the question of proof standards.

Both burdens can be distributed over the parties. For example, the plaintiff usually has the burden of production for the 'legal-operative' facts of his main claim, while the defendant has the burden of production for exceptions. The same can happen with the burden of persuasion: while the plaintiff has the burden of persuasion for his main claim, as observed by Prakken and Sartor [35] the burden of persuasion for other statements can de distributed over the parties. In the case of exceptions, sometimes only the burden of production rests on the defendant; once this burden is satisfied, the plaintiff has the burden of persuading the court that the exception

does not hold. This is the case, for instance, with all exceptions in criminal law. Carneades, we claim, is able to handle these different ways of distributing the burdens of production and persuasion.

We now see that, when discussing above how critical questions affect the burden of proof, we meant the burden of production in particular. In formal accounts, distributions of the burden of production can be simply expressed by formulating exceptions. However, as shown by Prakken [30], distributions of the burden of persuasion cannot be expressed by current 'mainstream' AI formalisms for argumentation. In particular, they cannot express the distinction between cases where the exception has to be proven (i.e., a decision maker has to be convinced that the exception is true) and cases where the exception merely has to be made plausible (i.e., a decision maker would have reason to doubt it is not true). Current nonmonotonic logics can only model the latter kind of situation. Accordingly, the system of Prakken and Sartor [34], based on Dung's grounded semantics, was adjusted in [30] to cope with this.

The approach taken in Carneades to the problem of determining how the burden of proof should be distributed is as follows. The burden of production is distributed by dividing premises into different types: evidence for ordinary premises and (once challenged) assumptions must be produced by the proponent of the argument with these premises, while evidence for exceptions must be produced by the respondent. In addition some initially low proof standard needs to be assigned to the statement of each premise. After the burden of production has been met, the burden of persuasion can be distributed by raising the proof standard assigned to the statement. Changing the proof standard can also cause the burden to shift from one party to another.

Summarizing, Carneades allows the burdens of production and persuasion to be allocated, separately, to either the proponent or the respondent and modified during the course of the dialogue. The initial allocation of the burden of production is regulated by the premise types of the argumentation scheme applied. The burden of persuasion is allocated by assigning the appropriate proof standard. As the dialogue progresses, subject to the argumentation protocol, the burdens may be reallocated by changing the assignment of premise types and proof standards via speech acts designed for this purpose.

3 Argument Graphs

We begin by defining the structure of argument graphs. Our conception of argument graphs is similar to Pollock's concept of an *inference graph* [29], in that there are nodes in the graph representing statements (propositions) and links which indicate inference relations between statements. Unlike Dung's model [8], in which

the internal structure of arguments is irrelevant, our model of the acceptability of statements makes use of and depends on the usual conception of argument in the argumentation theory literature, where arguments are instantiations of argumentation schemes linking a set of premises to a conclusion. The premises and the conclusion of arguments are statements about the world, whether empirical or institutional, which may be accepted or rejected. For the purpose of evaluating arguments, the syntax of statements is not important. We only require the ability to determine whether two statements are syntactically equal and some way to denote the logical complement of a statement.

Definition 1 (Statements) *Let* $\langle \text{statement}, =, \text{complement} \rangle$ *be a structure, where* statement *denotes the set of declarative sentences in some language,* $=$ *is an equality relation, modeled as a function of type* $\text{statement} \times \text{statement} \rightarrow \text{boolean}$*, and* complement *is a function of type* $\text{statement} \rightarrow \text{statement}$ *mapping a statement to its logical complement. If* s *is a statement, the complement of* s *is denoted* $\overline{s}$*.*

Next, to support defeasible argumentation and allow the burden of proof to be distributed, we distinguish several types of premises.

Definition 2 (Premises) *Let* premise *denote the set of premises. There are the following types of premises:*

(1) If s *is a* statement*, then* $\diamond s$*, called an* ordinary premise*, is a premise.*
(2) If s *is a* statement*, then* $\bullet s$*, called an* assumption*, is a premise.*
(3) If s *is a* statement*, then* $\circ s$*, called an* exception*, is a premise.*
(4) Nothing else is a premise.

Now we are ready to define the structure of arguments.

Definition 3 (Arguments) *An* argument *is a tuple* $\langle c, d, p \rangle$*, where* c *is a* statement*,* $d \in \{\text{pro}, \text{con}\}$ *and* $p \in 2^{\text{premise}}$*. If* a *is an argument* $\langle c, d, p \rangle$*, then* $\text{conclusion}(a) = c$*,* $\text{direction}(a) = d$ *and* $\text{premises}(a) = p$*.*

Note that an argument may have an empty set of premises. One difference between arguments and inference rules is evident in the distinction between pro and con arguments. Semantically, con arguments are instances of presumptive inference rules for the negation of the conclusion; if the premises of a con argument hold, this provides reasons to reject the conclusion or, equivalently, to accept its logical complement. Our approach abstracts completely from the syntax of the language for statements. Also, the use of both pro and con arguments is in accordance with our view of argumentation as a dialectical process. The arguments pro and con some statement need to be ordered or otherwise aggregated to resolve the conflict.

We assume arguments are put forward by the participants of a dialogue, where a dialogue is a sequence of speech acts such as asserting claims, conceding or

questioning claims and making arguments. Speech acts can be modeled as functions which map a state of the dialogue to another state. Argument graphs, defined next, have been designed to be a part of these dialogue states, as a structure for keeping track of the arguments which have been made and their relations.

An argument graph is a kind of proof tree in that it provides a basis for explanations and justifications. The *acceptability* relation between argument graphs and statements is intended to model the sufficiency of the proof: intuitively, a statement is acceptable given the arguments if and only if the argument graph is a proof of the statement. This distinguishes the acceptability relation from the defeasible consequence relation of nonmonotonic logics. In a calculus for such logics, assuming it is correct and complete, a statement is a defeasible consequence of a set of statements if and only if the statement is *derivable* in the calculus, whether or not such a proof has in fact been *derived*.

Argument graphs have two kinds of nodes, statement nodes and argument nodes. The edges of the graph link up the premises and conclusions of the arguments. [6] In Carneades argument graphs, at most one statement node is allowed for every statement s and its complement, $\overline{s}$. A statement and its complement are not just two unrelated statements. A dispute about s is also a dispute about $\overline{s}$. An argument pro one is an argument con the other. If one of these statements is accepted, the other must be rejected. There are no restrictions on the use of statements in premises. Both s and $\overline{s}$ may be used in premises, no matter which statement is represented by a statement node in the argument graph. Restricting statement nodes to at most one node for each s and $\overline{s}$ pair avoids the duplication of all arguments pro s as arguments con $\overline{s}$ and helps to reduce the complexity of diagrams of the graphs. If s is represented by a statement node in an argumentation graph, then a premise using $\overline{s}$ is called a *negative premise*. Ordinary premises, assumptions and exceptions may all be negative, using $\diamond\overline{s}$, $\bullet\overline{s}$, and $\circ\overline{s}$.

In the diagrams of argument graphs which follow, statements are displayed as boxes and arguments as circles or rounded boxes. Different arrowhead shapes are used to distinguish pro and con arguments as well as the different kinds of premises. Pro arguments are indicated using ordinary arrowheads; con arguments with open arrowheads. Ordinary premises are represented as edges with no arrowheads, assumptions with closed-dot arrowheads and exceptions with open-dot arrowheads. Negative premises have an additional tee mark (a short perpendicular line). No-

[6] Notice that argument graphs in Carneades are different than the *dialectical graphs* of the first author's Pleadings Game [13,15]. These dialectical graphs had only one kind of node, for arguments represented as sets of statements, and their edges modeled support, defeat and rebuttal relations between arguments. Argument graphs in Carneades have a bit more in common with the dialectical graphs of the same author's Zeno system [15], whose *position* nodes are like statement nodes, but arguments in Zeno are modeled as a binary relation between two positions and thus do not support arguments with multiple premises, let alone different types of premises.

tice that the premise type cannot be adequately represented using statement labels: since argument graphs are not restricted to trees, a statement may be used in several premises with different types.

Figure 1 illustrates this method of diagramming argument graphs with a toy legal argument. The issue is whether there is a contract. One con argument, a2, states there is not a contract because the agreement is not in writing although it is for the sale of real estate. The other con argument, a3, states the agreement is not in writing since is was by email. Argument a1 states there is a contract, because agreements are contracts unless one of the parties is a minor. Argument a5 states the agreement is for the sale of real estate, assuming there is a deed, i.e. a written instrument conveying the property. Argument a4 uses the deed as evidence of there having been an agreement. These last two arguments, especially, may not be realistic, but were contrived only to illustrate how argument graphs are not restricted to trees and how a statement can be used in different types of premises in several arguments.

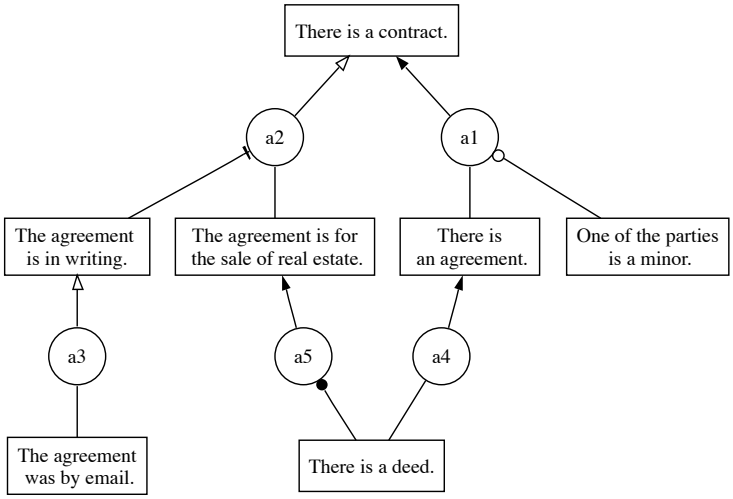

Fig. 1. Argument Graph

Although argument graphs are not restricted to trees, they are not completely general; we do not allow cycles. This restriction is intended to assure the decidability of the acceptability property of statements. At first sight, the condition that argument graphs be acyclic would seem to be a severe limitation. However, things are not that serious. Firstly, in systems using Dung's approach most cycles in realistic examples are two-cycles between arguments with incompatible conclusions. In Carneades, these can be represented as a pair of arguments pro and con the same statement, which does not introduce cycles into the argument graph. Next, cycles caused by indirectly using a statement in support of itself are also excluded in many other systems. This leaves cycles through exceptions, as shown in Figure 2.

Such examples have always been somewhat problematic in nonmonotonic logic. For instance, in Dung's [8] abstract framework defeat graphs with odd cycles may have no stable extensions. Also, intuitions differ on the proper treatment of cycles

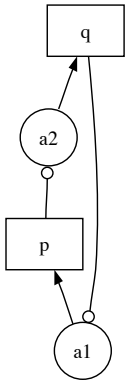

Fig. 2. Exception Cycle

[1]. Prior 'relational' approaches to this problem, such as Dung's preferred and grounded semantics as alternatives for stable semantics, are not directly transferable to Carneades, with its dialogical and procedural elements. We therefore leave an extension to graphs that allow for cycles through exceptions for future work.

Definition 4 (Argument Graphs) *An* argument-graph *is a labeled, finite, directed, acyclic, bipartite graph, consisting of* argument *nodes and* statement *nodes. The edges link the argument nodes to the statements in the premises and conclusion of each argument. At most one statement node is allowed for each statement* s *and its complement,* $\overline{s}$*.*

This completes the formal definition of the structure of arguments and argument graphs. Let us now discuss briefly the expressiveness of this model, beginning by comparing our approach with Toulmin's argumentation scheme [41]. Toulmin's scheme can be instantiated to produce arguments in Carneades with four premises: a minor premise called the *datum*; a defeasible major premise called the *warrant*, an additional piece of data supporting the warrant, called *backing*, and an exception, somewhat confusingly called a *rebuttal*. Toulmin's model also contains a *qualifier* stating the probative value of the inference (e.g. presumably, or necessarily). The strength of arguments is modeled by a partial order in Carneades, as explained in Section 4 .

To illustrate, Figure 3 reconstructs Toulmin's standard example about British citizenship in Carneades. Notice that the 'rebuttal' is modeled as an exception and the backing as an assumption. Both the datum and warrant are ordinary premises. Alternatively, backing could be modeled as the premise of an additional argument pro the warrant, by generalizing the concept of an argumentation scheme to cover patterns with multiple arguments.

In this diagram, the premise links have been labeled with the role each premise plays in Toulmin's scheme. This also illustrates the distinction between premise types in Carneades and the roles of premises in argumentation schemes. These are orthogonal properties of premises.

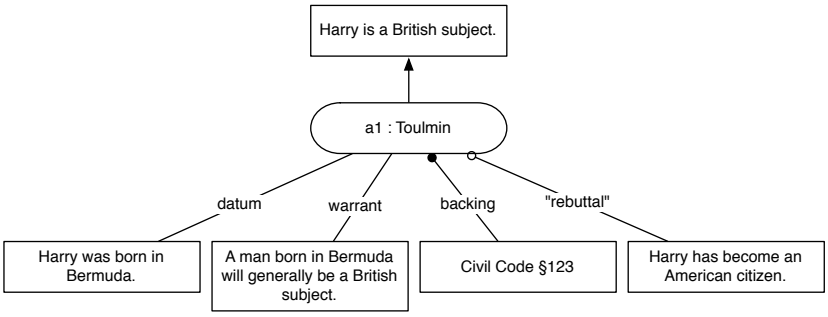

Fig. 3. Reconstruction of Toulmin Diagrams

Our model of argument is rich enough to handle Pollock's rebuttals and undercutting defeaters [29] as well as premise defeat. Rebuttals can be modeled as arguments in the opposite direction for the same conclusion. [7] (If an argument a_1 is pro some statement s, then some argument a_2 con s is a rebuttal of a_1, and vice versa.) Premises can be defeated by further arguments pro or con the statement of the premise, depending on the type of premise.

As for undercutting defeaters, according to Pollock these are exceptions to defeasible 'reasons'. Accordingly, undercutters can be modeled directly using Carneades exceptions. Consider Pollock's example of something looking red being a reason for believing it is red, unless it turns out to be illuminated by a red light. Figure 4 shows a reconstruction in Carneades of this argument.

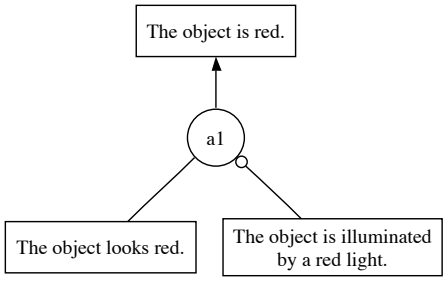

Fig. 4. Undercutting Defeater Example

Exceptions to defeasible generalizations, such as legal rules, can be understood as critical questions of argumentation schemes developed for reasoning with these generalizations. Consider, for example, a legal rule prohibiting vehicles in parks [19]. An argumentation scheme for reasoning with legal rules might include critical questions for challenging the validity of the rule or asking whether the rule is excluded by some other rule with higher priority in a particular case [44]. For example, one could use the second critical question to undercut an argument for some ambulance not being allowed in a park, by citing another rule excluding ambulances on duty. Such critical questions can also be modeled using exceptions in

[7] One must be careful not to confuse Pollock's rebuttals with Toulmin's, which would be represented as undercutters in Pollock's model.

Carneades.

Two kinds of arguments, *convergent* and *linked*, are distinguished in the literature [47]. Convergent arguments provide multiple reasons for a conclusion, each of which alone can be sufficient to accept the conclusion. Convergent arguments are handled in our approach by multiple arguments for the same conclusion. Linked arguments, on the other hand, consist of two or more premises which all must hold for the argument to provide significant support for its conclusion. They are handled in our approach by defining arguments to consist of a set of premises, rather than a single premise, and defining arguments to be defensible only if all of their premises hold. (The concept of argument defensibility is defined below.)

Assumptions and exceptions can be used to model Walton's concept of critical questions [48]. Critical questions enumerate specific ways to defeat arguments matching some argumentation scheme. But so long as an issue has not been raised by actually asking some critical question, we would like to be able to express which answer to assume. The distinction between assumptions and exceptions here provides this ability. Section 5 illustrates one way to reconstruct the scheme for arguments from expert opinion using Carneades.

4 Argument Evaluation

By argument evaluation we mean determining whether a statement is *acceptable* in an argument graph. [8] Intuitively, a statement is acceptable if a decision to accept the statement as true can be justified or explained given the arguments which have been put forward in the dialogue. The definition of the acceptability of statements is recursive. The acceptability of a statement depends on its *proof standard*. Whether or not a statement's proof standard is *satisfied* depends on the *defensibility* of the arguments pro and con this statement. The defensibility of an argument depends on whether or not its premises *hold*. Finally, we end up where we began: whether or not a premise holds can depend on whether or not the premise's statement is acceptable. Since the definitions are recursive, we cannot avoid making forward references to functions which will be defined later.

To evaluate a set of arguments in an argument graph, we require some additional information. Firstly, we need to know the current dialectical *status* of each statement in the dialogue, i.e. whether it is stated, questioned, accepted, or rejected. This status information is pragmatic; the status of statements is set by speech acts in the dialogue, such as asking a question, putting forward an argument or making

[8] In argumentation theory, argument evaluation has a somewhat broader meaning, including such tasks as matching arguments to argumentation schemes and revealing missing premises.

a decision. Secondly, we assume that a proof standard has been assigned to each statement. In the following, let proof-standard be an enumeration of some proof standards. Finally, we assume a strict partial ordering on arguments, which we will denote with $>$. Let $a1$ and $a2$ be arguments. If $a1 > a2$ we say that $a1$ *has priority over* $a2$. Let us formalize these requirements by postulating an *argument context* as follows.

Definition 5 (Argument Context) *Let $\mathcal{C}$, the argument context, be a tuple $\langle \text{status}, \text{ps}, > \rangle$, where* status *is a function of type* $\text{statement} \rightarrow \{\text{stated}, \text{questioned}, \text{accepted}, \text{rejected}\}$, ps *is a function of type* $\text{statement} \rightarrow \text{proof-standard}$ *and $>$ is a strict partial ordering on arguments.* [9] *For every statement s and its complement $\overline{s}$, the proof standard assigned to $\overline{s}$ is the complement of the proof standard assigned to s and*

- *if* $\text{status}(s) = \text{stated}$ *then* $\text{status}(\overline{s}) = \text{stated}$,
- *if* $\text{status}(s) = \text{questioned}$ *then* $\text{status}(\overline{s}) = \text{questioned}$,
- *if* $\text{status}(s) = \text{accepted}$ *then* $\text{status}(\overline{s}) = \text{rejected}$, *and*
- *if* $\text{status}(s) = \text{rejected}$ *then* $\text{status}(\overline{s}) = \text{accepted}$.

The constraints on the status of statements in this definition serve two purposes: First, they assure that whenever a statement is stated or questioned its complement is also, implicitly, stated or questioned. Stating or questioning a statement does imply the assertion of some position or viewpoint pro or con the statement. Stating a statement merely introduces it into the dialogue. Questioning a statement merely makes an issue out of the statement. Second, the constraints assure that whenever a statement is accepted its complement is rejected and vice versa. A decision to accept one is simultaneously a decision to reject the alternative.

In the definitions which follow, an argument context is presumed.

Definition 6 (Acceptability of Statements) *Let* acceptable *be a function of type* $\text{statement} \times \text{argument-graph} \rightarrow \text{boolean}$. *A statement s is* acceptable *in an argument graph G if and only if it satisfies its proof standard:*

$$\text{acceptable}(s, G) = \text{satisfies}(s, \text{ps}(s), G).$$

Definition 7 (Satisfaction of Proof Standards) *A proof standard is a function of type* $\text{statement} \times \text{argument-graph} \rightarrow \text{boolean}$. *A statement s is* satisfied *by a proof standard f in an argument graph G if and only if $f(s, G)$ is true.*

The following proof standards are defined in this article. We do not claim these particular proof standards are exhaustive. Others can be defined similarly.

[9] Should the need arise to make the proof standard of a statement depend on the argument in which the statement is used as a premise, the ps function of the context could be extended to be of type $\text{statement} \times \text{argument} \rightarrow \text{proof-standard}$.

SE (Scintilla of Evidence). A statement meets this standard iff it is supported by at least one defensible pro argument.

BA (Best Argument). A statement meets this standard iff it is supported by some defensible pro argument with priority over all defensible con arguments.

DV (Dialectical Validity). A statement meets this standard iff it is supported by at least one defensible pro argument and none of its con arguments are defensible.

In addition, a proof standard can be derived from another by switching the roles of pro and con arguments in the standard:

Definition 8 (Complement of a Proof Standard) *The* complement *of a proof standard σ, denoted $\overline{\sigma}$, is the standard which results by switching the roles of pro and con arguments in the definition of σ.*

For example, the complement of the BA proof standard, $\overline{\mathrm{BA}}$, is satisfied iff the statement is supported by some defensible *con* argument with priority over its strongest defensible *pro* argument. In principal a statement can satisfy both a proof standard and its complement, highlighting the difference between acceptability and truth. The arguments can be sufficient to justify a decision either way without being inconsistent. In practice, however, this will be the case only with very weak proof standards, such as scintilla of evidence. If there exists a defensible pro argument and a defensible con argument, then both SE and its complement, $\overline{\mathrm{SE}}$, are satisfied.

Further research is required to define and validate models of proof standards for various applications. In particular we do not claim that any of the standards defined here adequately model legal proof standards.

The standards defined here can be ordered by their relative strength:

$$\mathrm{DV} > \mathrm{BA} > \mathrm{SE}$$

If a statement satisfies a proof standard, it will also satisfy the weaker proof standards. The best argument (BA) standard uses a strict partial ordering on arguments to resolve conflicts. Recall that we assume such an ordering as part of the argument context.

All of the proof standards defined above depend on a determination of the *defensiblity* of arguments. Defensibility is defined next.

Definition 9 (Defensibility of Arguments) *Let* $\mathrm{defensible}$ *be a function of type* $\mathrm{argument} \times \mathrm{argument\text{-}graph} \rightarrow \mathrm{boolean}$*. An argument α is defensible in an argument graph G if and only if all of its premises* hold *in the argument graph:* $\mathrm{defensible}(\alpha, G) = \mathrm{all}(\lambda p.\, \mathrm{holds}(p, G))(\mathrm{premises}\, \alpha)$. [10]

[10] Here 'all' is a higher-order function, not a quantifier, applied to an anonymous function, represented with λ, as in lambda calculus.

Now we come to the last definition required for evaluating arguments, of the holds predicate. This is where the dialectical status of a statement and the type of premises come into play.

Definition 10 (Holding of Premises) *Let* holds *be a function of type* premise $\times$ argument-graph $\rightarrow$ boolean. *Whether or not a premise holds depends on its type. Thus, there are the following cases:*

If p is an ordinary premise, $\diamond s$, then

$$\text{holds}(p, G) = \begin{cases} acceptable(s, G) & \textit{if } \text{status}(s) = \text{stated} \\ acceptable(s, G) & \textit{if } \text{status}(s) = \text{questioned} \\ true & \textit{if } \text{status}(s) = \text{accepted} \\ false & \textit{if } \text{status}(s) = \text{rejected} \end{cases}$$

If p is an assumption, $\bullet s$, then

$$\text{holds}(p, G) = \begin{cases} true & \textit{if } \text{status}(s) = \text{stated} \\ acceptable(s, G) & \textit{if } \text{status}(s) = \text{questioned} \\ true & \textit{if } \text{status}(s) = \text{accepted} \\ false & \textit{if } \text{status}(s) = \text{rejected} \end{cases}$$

Finally, if p is an exception, $\circ s$, then

$$\text{holds}(p, G) = \begin{cases} \neg acceptable(s, G) & \textit{if } \text{status}(s) = \text{stated} \\ \neg acceptable(s, G) & \textit{if } \text{status}(s) = \text{questioned} \\ false & \textit{if } \text{status}(s) = \text{accepted} \\ true & \textit{if } \text{status}(s) = \text{rejected} \end{cases}$$

It can be proven that there is always a unique and complete assignment of 'acceptable' or 'not acceptable' to statements, and of 'holds' or 'not holds' to premises.

Theorem 1 *For every argument context C with proof standards* SE, BA *or* DV *and every argument graph G:* acceptable *and* holds *are total functions.*

Moreover, if the BA or DV proof standard is assigned to a statement in the context, it can never be the case that both the statement and its complement are acceptable.

Theorem 2 *For every argument context C with proof standards* BA *or* DV *and every argument graph G, no statement s exists such that s and $\overline{s}$ are both acceptable.*

The proofs are in the appendix. The important thing to notice is that whether or not a premise holds depends in this model not only on the arguments which have been put forward, but also on the type of premise and the status of the premise's statement in the context. We assume that the status of a statement progresses in the course of the dialogue:

(1) Statements are introduced, i.e. 'stated', by using them in arguments. Whether or not a premise which uses this statement holds at this stage of the dialogue depends on the kind of premise. Ordinary premises hold only if the statement is acceptable given whatever arguments have been put forward; assumptions hold unconditionally. [11] This is the only difference between ordinary premises and assumptions in the model. Exceptions holds at this stage only if the statement is not acceptable given the arguments put forward.
(2) At some point a participant may explicitly make an issue out of a statement, by questioning it. Now both ordinary premises and assumptions which use this statement hold only if they are acceptable. Exceptions continue to hold only if the statement is not acceptable. We assume that arguments will be exchanged in a dialogue for some period of time, and that during this phase the acceptability of statements will be in flux.
(3) Finally, at some point a decision will be made to either accept or reject some statement. Accepting a statement implicitly rejects its complement and vice versa. The model constrains the decisions which can be justified or explained. Any interested person can check whether the decisions are justified given the arguments made and the applicable proof standards. After a decision has been made, it is respected by the model. For example, ordinary premises with accepted statements hold and ordinary premises with rejected statements do not hold.

5 Modeling Critical Questions in Carneades

When the scheme for arguments from expert opinion is instantiated to create a specific argument, the critical questions can be represented in Carneades as assumptions and exceptions. Whether an assumption or exception is appropriate depends on who should have the burden of production, which is a policy issue dependent on the domain. If the respondent, the person who poses the critical question, should have the burden of production, then the critical question should be modeled as an exception. If, on the other hand, the proponent, the party who used the scheme

[11] Both a statement and its complement may be assumed at the same time, but only in the context of premises of different arguments.

to construct the argument, should have the burden of production, then the critical question should be modeled as an assumption.

The distinction between assumptions and exceptions in Carneades allows answers to be assumed for critical questions which have yet to be asked. Which type of premise is appropriate depends on the policy of the argumentation scheme. If a new argumentation scheme is being modeled, the premise type can be used to clearly express the desired policy. Modeling an existing argumentation scheme requires the scheme to be interpreted to determine the policy intended by its drafters, similar to the way legislation must be interpreted. In our interpretation of the scheme for arguments from expert opinion, the *shifting burden* theory seems appropriate for the expertise and backup evidence questions, since experts are generally credible sources of knowledge in their field, who typically base their testimony on evidence. Thus these questions have been modeled as assumptions. But the *backing evidence theory* seems more appropriate for the trustworthiness and consistency questions, since it is easier to produce evidence of instances of unreliable or inconsistent behavior than to try to prove that no instances of such behavior have occurred. Applying the general principal of allocating the burden of production to the party with better access to the evidence, we have modeled these questions as exceptions.

The scheme for argument from expert opinion can be recast to fit the Carneades model as follows, where 'premise' here means *ordinary premise*:

Premise. E is an expert in the subject domain S containing the proposition A.
Premise. E asserts A.
Assumption. E is a credible expert.
Exception. E is not reliable.
Exception. A is not consistent with the testimony of other experts.
Assumption. A is based on evidence.
Conclusion. A.

6 Modeling Distributions of Proof Burdens in Carneades

In this section we illustrate how premise types and proof standards in Carneades can be used to allocate various burdens of proof, and the importance of being able to do so, reusing Prakken and Sartor's example about a murder trial [35].

According to Section §187 of the California Penal Code, murder is (unlawful) killing with 'malice aforethought'. Section §197 states an exception for self-defense. Let's suppose the criminal proceedings have been initiated by the prosecution filing a complaint in which it makes its first arguments, by applying a scheme for arguments from legal rules to Section §187 and providing enough evidence of killing and malice aforethought to meet the burden of production required to bring the case

to trial. Let us assume the defense has accepted the killing and malice elements of the crime, so these are not at issue. Using a simplified scheme for arguments from legal rules derived from Reason-Based Logic [44], the argument can be modeled in Carneades as shown in Figure 5.

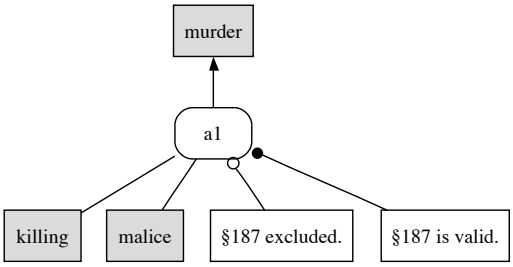

Fig. 5. The Murder Complaint

Figure 5 illustrates how premise types are assigned by applying argumentation schemes. The appropriate type of premise is a policy issue which must be addressed when developing argumentation schemes and legal rules. In the scheme for arguments from legal rules used here, the validity of a legal rule is assumed, and thus must be supported with evidence only if questioned; being excluded by some other rule is an exception, and the elements of the crime, killing and malice here, are ordinary premises which must always be supported with evidence. In the example we assume sufficient evidence, not shown, was provided to justify the acceptance of the killing and malice ordinary premises.

This figure, as well as the others in this section, were generated and labeled by the implementation of Carneades. The statements shown in gray have been accepted (killing, malice) or are acceptable (murder). In the initial context, all statements are assigned the lowest proof standard, scintilla of the evidence (SE), and no argument has priority over any other. The murder charge is acceptable in this initial context, given the argument and the claims accepted by the defense. The prosecution has met its burden of production.

Next, the defense puts forward its self-defense argument, citing Section §197 of the Penal Code, and calls a witness who testifies that the defendant was attacked with a knife by the victim. After these moves, the argument graph is as shown in Figure 6. The argument concluding that Section §187 is excluded, a2, is another instance of the scheme for arguments from legal rules. The third argument applies a simplified scheme for arguments from testimony. Since the witness testified before the court, we assume the fact this testimony was made is accepted. (This does not imply accepting that the defendant was attacked.) This is enough to meet the defense's burden of production for the self-defense claim, again applying the SE proof standard. Thus, as can be seen in the Figure 6, the counterargument succeeds in making the murder charge no longer acceptable.

But the testimony can still be challenged by the prosecution in various ways. One

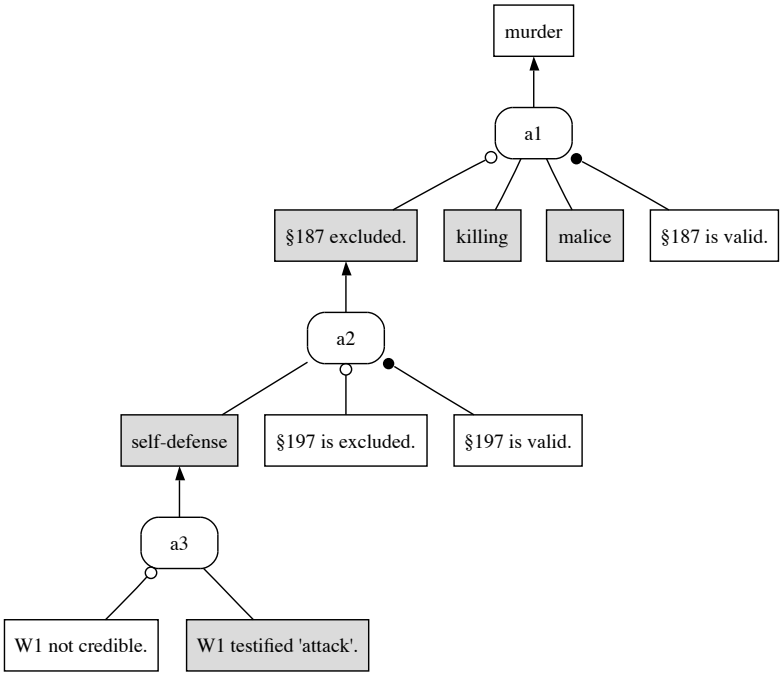

Fig. 6. The Argument Graph after W1's Testimony

would be to question the credibility of the witness, which is a critical question of the scheme for arguments from testimony, modeled here as an exception. The prosecution in our example, however, chooses another move, by calling another witness to the stand to testify that the defendant had enough time to run away. The state of the argument graph after this move is shown in Figure 7. This second application of the scheme for arguments from testimony also illustrates how rebuttals are modeled as con arguments in Carneades. Notice, however, that this rebuttal does not (yet) succeed. The self-defense claim is still acceptable. This is because the prosecution has the burden of persuasion in criminal cases, also for exceptions such as the self-defense claim here. After the defendant met his burden of production for self-defense, the proof standard of the self-defense statement was changed in the model to a standard which reflects the prosecution's burden of persuasion. This proof standard is satisfied if it is not the case that the complement of the best argument (BA) standard is satisfied. Recall that the complement of the BA standard is satisfied only if the best con argument has priority over the best pro argument. But since the arguments have not been ordered, this is not yet the case. Thus the self-defense claim is still acceptable.

The source of the priority ordering on arguments depends on the application scenario. In a criminal proceeding, determining the relative strength of evidence, such as witness testimony, is a task for the jury. How best to model the 'beyond a reasonable doubt' standard applicable in criminal proceedings is a question for future research. One approach, suggested by Prakken and Sartor [35], is to apply the BA standard, but to consider one argument to have priority over another only if it is sufficiently much stronger.

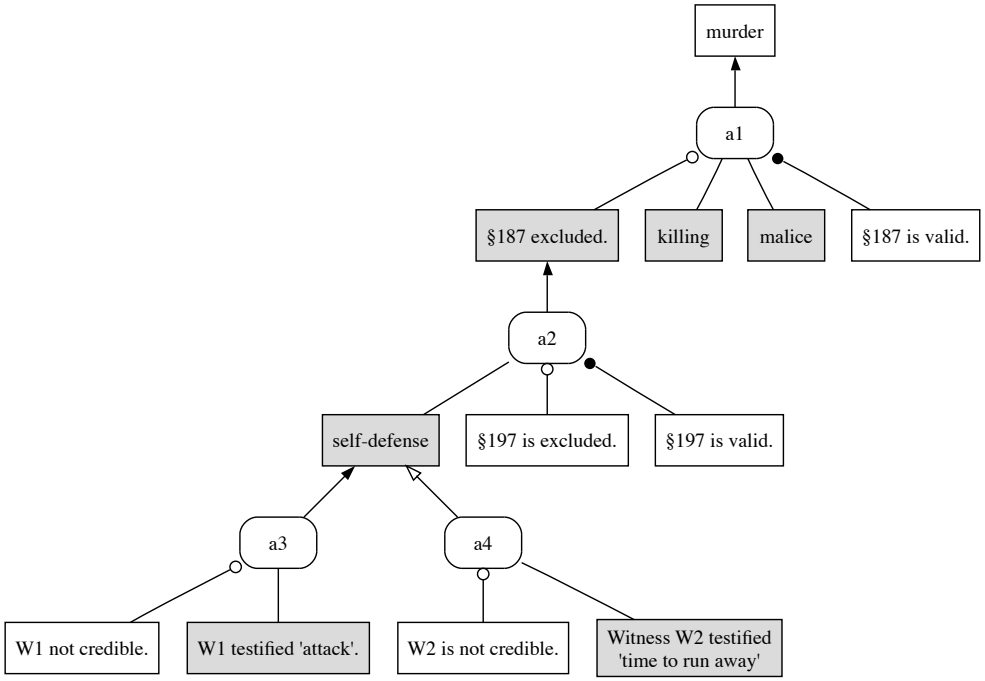

Fig. 7. The Argument Graph After W2's Testimony

The burden of proof, whether it be the burden of production or persuasion, is assigned to some party in the dialogue, either the proponent or respondent. These roles are relative to some statement, such as a claim or thesis. One party may be the proponent with regard to one statement and the respondent with respect to another. The example illustrates how this is modeled in Carneades. Whereas the prosecution is the proponent of the main claim, i.e. the murder charge, the defense is the proponent of the legal claim that Section §187 is excluded by the self-defense rule of Section §197. The defense is also the proponent of the factual claim that the defendant did, in fact, act in self defense. However, due to the prosecution's burden of persuasion in criminal cases, it has the burden of persuading the jury that that defendant did not, in fact, act in self defense, even though it has the respondent role and the defense had the burden of production, with respect to the self-defense issue. Had this been a civil case, the defense would have had the burden of persuasion for the self-defense issue and the proof standard for the burden of persuasion would have been the lower 'preponderance of the evidence' standard. To appreciate what a difference this can make, consider the O.J. Simpson case, where he was found not guilty of murder in the criminal proceeding but lost the later civil case against him about the same killing. Given the arguments shown in Figure 7, assuming the arguments pro and con self-defense have equal priority in the context, 'self-defence' would be acceptable in a criminal case but not in a civil case. This illustrates how burden of proof assignments depend on the procedural context.

In addition to selecting the appropriate premise type when developing argumentation schemes, to handle burden of proof one must also design the argumentation protocol, at the procedural level, so that the appropriate proof standard is assigned

to statements at each stage of the proceedings. One way to do this, as we have done in this example, is select some standard to apply by default to all statements, and then assign some other standard to particular statements as speech acts are performed, as required by the protocol.

The burden of persuasion typically requires a more stringent proof standard to be satisfied than the burden of production. Moreover, the party who has the burden of persuasion need not be the same as the party who had the burden of production. If the same party has both the burden of production and the burden of persuasion, then all that is required is to change the proof standard to one requiring the arguments pro the statement to be stronger than the arguments con the statement, where the proof standard defines just how much stronger the pro arguments must be. Given the example proof standards defined in this article, any proof standard except SE would be suitable.

As illustrated in the example, however, modeling the situation where the party who has the burden of persuasion is not the same as the party who had the burden of production is a more subtle problem, which can require some thought when changing the proof standard after the burden of production has been met. In the example, the appropriate standard was the negation of the complement of the best argument standard.

7 Related Work

The idea of developing a computer model for managing support and justification relationships between propositions goes back to research on truth and reason maintenance systems in Artificial Intelligence, beginning with Jon Doyle's Truth Maintenance System [7]. Probably the most famous system of this kind is Johan de Kleer's Assumption-Based Truth Maintenance System [6]. The system presented here is more like Doyle's original system; the way it supports defeasible reasoning using arguments with different kinds of premises is reminiscent of Doyle's use of *in* and *out* lists in justifications.

The idea of using different kinds of premises to help structure dialogues was also made by [2], which applied Toulmin's argumentation scheme [41] to annotate the conditions in the body of Horn clauses with their role in the scheme (data, warrant, backing, rebuttal) and used these annotations to structure explanation dialogues with users. The role of a premise in an argumentation scheme such as Toulmin's, however, is orthogonal to the premise types defined here. They serve different purposes and can be used together.

The first logical models of argumentation in AI were proposed by Loui [23] and Pollock [28], followed by, among others, Simari and Loui [39], Vreeswijk [45], Pollock [29], Dung, [8], Bondarenko et al. [4] and much subsequent work. For an overview, see Prakken & Vreeswijk [36]. As noted in the introduction, all this work essentially takes a relational approach, ignoring or abstracting from the dialogical context of argumentation and hardwiring proof burdens and proof standards into the formalism. Below we briefly compare the 'logical' features of Carneades with this work.

As for *the structure of arguments*, in AI two alternative approaches are taken, depending on whether defeasibility arises from the use of a special kind of statement or from the use of a special kind of inference rule. In the assumption-based approach [4], the information state is divided into a theory, containing certain statements, and a set of default assumptions, comparable to negations of our exceptions. Arguments are sets of assumptions that can be consistently added to the theory. Conclusions can be drawn by applying some deductive consequence notion to the combination of theory and argument. Arguments can be attacked by constructing arguments that conclude the *contrary* of an assumption in the argument to be attacked. In the inference-rule approach [45,29], the inference rules are divided into strict and defeasible inference rules and arguments are conceived of as (possibly composite) premise-conclusion structures instantiating some inference rule. Arguments can be attacked on their defeasible inferences, for instance, with an argument for a contradictory conclusion (Pollock's rebuttals) or with an argument denying that a defeasible inference rule applies in the case at hand (Pollock's undercutters).

Carneades combines elements of both approaches. Its use of exceptions is similar to the use of assumptions, while its idea that arguments instantiate argument schemes allows for defeasible inference rules. Like Pollock and Vreeswijk, Carneades abstracts from particular inference rules and the logical language. [12] Carneades shares with Pollock [29] the feature that elementary arguments are combined in an argument graph. Pollock calls these *inference graphs*. In such a graph, each statement occurs only once and arguments are recorded via links between the various statements, passing through argument nodes. These nodes are similar to *and* links in *and/or* graphs. While Pollock's graphs contain special links between statements for attack relations, in Carneades such attack relations are expressed either as a pair of arguments pro and con the same statement or as an argument pro an exception. A statement s and its complement $\overline{s}$ are not unrelated in Carneades. An argument pro a s is equivalent to an argument con $\overline{s}$. And the conclusion of an argument pro s logically contradicts the conclusion of an argument con s. Also, Carneades dis-

[12] Pollock and Vreeswijk, however, do rely on a negation operator to define argument conflict.

tinguishes the types of premises in the graphs, to allocate the burden of proof for statements among the parties.

As for *argument attack*, as explained above in Section 3, Pollock's undercutters can be represented using exceptions. Carneades' distinction between pro and con arguments is essentially the same as Pollock's notion of rebutting arguments. Con arguments should not be confused with Verheij's [42] arguments with dialectical negation. When in Verheij's Deflog system arguments for a statement and its dialectical negation can be constructed, the latter automatically prevails; in Carneades, by contrast, arguments pro and con the same statement may need to be compared to see which one has priority, depending on the applicable proof standard. Note finally that in Carneades, unlike in systems like those of Pollock and Vreeswijk, argument premises also can be attacked. The reason is that Carneades assumes a dialogical context in which any premise can be questioned, while in 'relational' approaches the premises are given and only default assumptions can be attacked.

We next briefly compare Carneades' *argument evaluation* mechanism with this prior AI work. Because of its embedding in a dialogical context and its variable proof burdens and proof standards, such a comparison is not straightforward. Nevertheless, it is worthwhile disregarding these additional features of Carneades and consider its 'logical' or 'relational' core.

Firstly, Theorem 1 shows that Carneades adopts the 'unique assignment' instead of the 'multiple-status assignment' approach [36]. If, for instance, a statement has one defensible pro argument and one defensible con argument and neither argument has priority, then Carneades does not induce multiple status assignments (one in which the statement is acceptable and one in which it is not). Instead it assigns 'acceptable' to the statement if it satisfies its proof standard and it assigns 'not acceptable' to it otherwise.

Secondly, Carneades' relational core sometimes exhibits so-called 'ambiguity-blocking' behavior. Informally, a nonmonotonic logic is called ambiguity-blocking if, when a conflict between two lines of reasoning with contradictory conclusions cannot be resolved, both lines of reasoning are cut-off and neither of the conclusions can be used for further reasoning. In such logics it may happen that other lines of reasoning remain undefeated even though one of the cut-off lines of reasoning interferes with it and is not weaker. Consider the Carneades argument graph shown in Figure 8.

Suppose that the proof standard for all statements has been set to BA, that t, s and r are accepted and that neither of the two arguments pro and con q has priority. Then q is not acceptable. However p is acceptable, since the argument con p is not defensible. For the same reason, p would remain acceptable even if the con argument had priority over the pro argument.

It has been argued that such an outcome is counterintuitive. For instance, Makinson & Schlechta [25] have argued using similar examples that since the status of q is

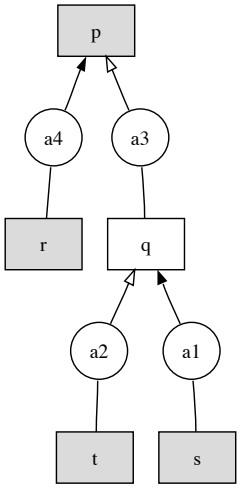

Fig. 8. Ambiguity Blocking Example

ambiguous, the argument con p, although not alive, is not fully dead either, so it should retain its interfering force as a 'zombie' argument. Logics in which 'zombie paths' retaining their interfering force are sometimes called 'ambiguity propagating'.

Others, e.g. Horty [21] and Governatori et. al [18] have instead argued that ambiguity propagating is counterintuitive. Carneades' procedural perspective offers another justification for this position. In full Carneades, not reduced to its relational core, the status of q is not ambiguous, because the proof standard assigned to q clearly regulates the acceptability of q, also when no argument has priority. By assigning the BA proof standard to q, a policy decision was made to assign the burden of proof to the proponent of q. This burden of proof has not been met; thus q is certainly unacceptable. The proponent of q has lost this part of the argument. On this procedural account, there is nothing unintuitive about the argument con p failing. Since its only premise unequivocally does not hold, one should not expect it to be taken into consideration when evaluating the acceptability of p.

7.2 Other Dialogical Work on Argumentation

As stated in the introduction, Carneades is meant to be an evaluative component of systems supporting persuasion dialogues. In such dialogues, two or more participants try to resolve a difference of opinion by arguing about the tenability of a claim, each trying to persuade the other participants to adopt its point of view. Dialogue systems for persuasion regulate what utterances the participants can make and under which conditions they can make them, what the effects of their utterances are on their propositional commitments, when a dialogue terminates and what the outcome of a dialogue is. Dialogue systems for persuasion have been defined mainly in the areas of argumentation theory, AI (& Law) and multi-agent systems.

See Prakken [32] for a recent overview.

As for the outcome of dialogues, some dialogue systems only focus on whether a dialogue participant has changed his internal, subjective beliefs in the course of a dialogue, which may not be observable for other agents. Systems of this kind can especially be found in research on multi-agent systems, e.g. [27]. Other systems focus on inter-subjective consensus between the dialogue participants, by looking at whether they have (publicly) committed themselves to or rejected certain statements. Basically, consensus has been achieved at the end of a dialogue if both parties are committed or not committed to the main claim. This was the original approach in argumentation theory, e.g. in [46]. Finally, there are dialogue systems that besides the participants' publicly declared standpoints also take standards imposed from the outside into account, so that a statement can be declared acceptable or not in a certain dialogue state even if no consensus has been reached. This is the class of systems Carneades is meant to support.

In some AI work these standards come from a (nonmonotonic) logic, which is applied to the *current* explicit consensus in a dialogue state. This was the approach taken in the Pleading Game ([14], using conditional entailment [12], and also by, for instance, Loui [24], who used a variant of Simari & Loui's system [39], and Brewka [5], who used prioritized default logic. In Prakken [31] this idea is refined. First, a purely dialogical notion of acceptability is defined, in terms of attacking and surrendering reply relations between dialogue utterances: On the basis of these relations it is defined whether the initial move of a dialogue is *in* or *out*. Then it is proven for a class of inference-based argumentation systems satisfying grounded semantics that the initial move is *in* if and only if the main claim is logically implied by the current consensus, assuming the participants make logically optimal moves in the dialogue. Carneades takes Prakken's approach to the extreme by replacing his use of a relational consequence notion with the purely procedural notions of proof burdens and proof standards. Also, Carneades replaces Prakken's reply relations between dialogue utterances with its dialectical status of statements, thus fully abstracting from the speech acts of dialogues.

7.3 Other Work on Modeling Burden of Proof

As noted in the introduction, Carneades' modeling of proof standards is inspired by Freeman & Farley's DART system [10], an implemented model of dialectical argumentation based on a semi-formal specification. DART models various forms of defeasible reasoning with rules, including standard logic principles such as modus ponens and modus tollens but also nonstandard ones, such as for abductive and *a contrario* reasoning. The status of arguments is defined in terms of an argument game parameterized by a proof level for the main claim. DART does not allow for different levels of proof for premises of the main claim or for shifts of the burden

of proof to the other party and it does not take the dialogical status of statements into account. Also, the distinction between burden of production and burden of persuasion is not explicitly addressed.

As discussed above in Section 2, Prakken [30] adapts Prakken & Sartor's [34] argumentation system based on grounded semantics to let it allow for shifts of the burden of proof. The modified system assumes as input not just an (ordered) set of rules but also an allocation of proof burdens for literals to plaintiff and defendant. The argument game is modified to allow the two players (plaintiff and defendant) to have different dialectical roles (proponent or opponent) for different propositions. Although Prakken [30] does not explicitly distinguish the burdens of production and persuasion, Prakken & Sartor [35] argue that this work in fact models distribution of the burden of persuasion. They also argue that different proof levels can be modeled by adopting different definitions of the binary defeat relation between arguments assumed by a Dung-style approach. Our account in Section 6 of how proof burdens can be represented in Carneades was meant to respect Prakken & Sartor's observations while retaining the special features of Carneades.

Finally, Prakken et al. [33] modified a persuasion dialogue game [31] to allow for debates on who has the burden of proof. Allocations of proof burdens are expressed with a special speech act, which can be moved in reply to a challenge of a claim. Allocations can also be challenged, which gives rise to metadialogues about who has the burden of proof. The resulting dialogue system provides a possible dialogical context for Carneades: the burden allocations agreed on in a dialogue could be translated into the assignment of premise types and proof standards.

8 Conclusion

Carneades is a formal, mathematical model of argument structure and evaluation which applies proof standards to determine the acceptability of statements on an issue-by-issue basis. The main original contribution of Carneades is its method for allocating the burden of proof to the proponent or respondent, for each premise separately, using premises type, proof standards and the dialectical status of statements. Useful for modeling legal dialogues, the *burden of production* and *burden of persuasion* can be handled separately, with a different responsible party and applicable proof standard for each [35]. Following [30], our approach allows the burdens of proof to be distributed over the parties and the applicable proof standard to be changed as the dialogue progresses from stage to stage. Finally, Carneades' allows critical questions of argumentation schemes to be modeled with additional premises, using assumptions and exceptions.

Carneades is a semantic model of argumentation, not a calculus. However, since this semantic model is formulated in terms of computable functions, the lambda

calculus may be used as a formal system. As a semantic model, the question of the soundness or completeness of Carneades does not arise. Rather, the relevant question concerns the validity of the semantic model. Are these semantics sufficient for providing the kind of argumentation support required by our intended application scenarios? This question cannot be answered by purely formal means, but rather requires experiments with realistic test cases.

Functional programming languages can be used to easily implement the model in software. Indeed, Carneades has been fully implemented in this way, using the Scheme programming language [22]. The implementation computes the acceptability of statements, given an argument graph and context as inputs. It can also generate argument diagrams. The diagrams in Section 6 of this article were generated by the program.

Further work is required on modeling legal proof standards, such as 'preponderance of the evidence' and 'beyond a reasonable doubt'. Although our validation efforts thus far have been mainly in the legal domain, Carneades is intended to be a general model of argumentation, not restricted to some application domain. Outside the legal context, we plan to evaluate the suitability of Carneades for practical reasoning in deliberation dialogues. When completed, Carneades will support a range of argumentation use cases, including argument construction, reconstruction, evaluation and visualization.

Acknowledgments

Tom Gordon was supported by the European ESTRELLA project (IST-2004-027655). Doug Walton was supported by a grant (410-2005-0398) from the Social Sciences and Humanities Research Council of Canada for a project on Dialogue Systems for Argumentation in Artificial Intelligence and Law. We thank Giovanni Sartor for helping us to understand the burdens of production and persuasion. Finally, we would like to thank the anonymous reviewers for their constructive and detailed comments.

A Proofs

Theorem 1 *For every argument context C with proof standard* SE, BA *or* DV *and every argument graph G:* acceptable *and* holds *are total functions.*

Proof: The proof is with induction on the structure of argument graphs. Recall that G is finite and acyclic and every statement occurs in it only once. We first consider

the acceptability of any statement node s with no parents. A statement is acceptable iff it satisfies its proof standard. Since G does not contain any argument pro s, with all three proof standards s trivially is not acceptable. We next examine the holds function for any premise p containing s. If $p = \diamond s$ then if s is accepted, p clearly holds, if s is rejected, p clearly does not hold, and if s is stated or questioned, p does not hold since s is not acceptable. If $p = \bullet s$ then if s is stated or accepted, p holds, if s is rejected, p does not hold and if s is questioned, p does not hold since s is not acceptable. Finally, if $p = \circ s$ then if s is accepted, p clearly does not hold, if s is rejected, p clearly holds and if s is at stated or questioned, p holds since s is not acceptable.

Consider next any statement node s in G that has a parent node. Again s is acceptable iff it satisfies its proof standard. All three proof standards depend only on the arguments pro or con s that are in G. These arguments can be unambiguously identified since by the induction hypothesis the theorem holds for all parent statements of s and all premises using them. Then clearly with all three proof standards s is unambiguously either acceptable or not acceptable. We finally examine the holds function for any premise using s. Since this function depends only on the acceptability and dialectical status of s, clearly all premises using s unambiguously either hold or do not hold. □

Theorem 2 *For every argument context $\mathcal{C}$ with proof standards* BA *or* DV *and every argument graph G, no statement s exists such that s and $\overline{s}$ are both acceptable.*

Proof: Suppose s is acceptable. Then if $\mathrm{ps}(s) = \mathrm{BA}$, there exists a defensible argument A pro s such that for all defensible arguments B con s it holds that $A > B$. If $\overline{s}$ satisfies $\overline{\mathrm{BA}}$ then there exists a defensible argument B con s such that for all defensible arguments A pro s it holds that $B > A$. But then $>$ is not a strict partial order. If $\mathrm{ps}(s) = \mathrm{DV}$, then there exists a defensible argument pro s and there is no defensible argument con s. The latter immediately implies that $\overline{s}$ does not satisfy $\overline{\mathrm{DV}}$. □

References

[1] B. Baroni, M. Giacom, and G. Guida. SCC-recursiveness: a general schema for argumentation semantics. *Artificial Intelligence*, 168:162–210, 2005.

[2] T. Bench-Capon, D. Lowes, and M. McEnery. Using Toulmin's argument schema to explain logic programs. *Knowledge-Based Systems*, 4(3):177–183, September 1991.

[3] F. Bex, H. Prakken, C. Reed, and D. Walton. Towards a formal account of reasoning with evidence: Argumentation schemes and generalizations. *Artificial Intelligence and Law*, 11(2-3):125–165, 2003.

[4] A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93(1-2):63–101, 1997.

[5] G. Brewka. Dynamic argument systems: A formal model of argumentation processes based on situation calculus. *Journal of Logic and Computation*, 11(2):257–282, 2001.

[6] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[7] J. Doyle. A truth maintenance system. *Artificial Intelligence*, 12:231–272, 1979.

[8] P. 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.

[9] P. Edwards. *The Encyclopedia of Philosophy*, volume 1. Macmillian and Free Press, 1972.

[10] K. Freeman and A. M. Farley. A model of argumentation and its application to legal reasoning. *Artificial Intelligence and Law*, 4(3-4):163–197, 1996.

[11] B. Garssen. Argument schemes. In F. H. van Eemeren, editor, *Critical Concepts in Argumentation Theory*, pages 81–100. Amsterdam University Press, 2001.

[12] H. Geffner and J. Pearl. Conditional entailment: Bridging two approaches to default reasoning. *Artificial Intelligence*, 53(2-3):209–244, 1992.

[13] T. F. Gordon. The Pleadings Game — an exercise in computational dialectics. *Artificial Intelligence and Law*, 2(4):239–292, 1994.

[14] T. F. Gordon. *The Pleadings Game; An Artificial Intelligence Model of Procedural Justice*. Springer, New York, 1995. Book version of 1993 Ph.D. Thesis; University of Darmstadt.

[15] T. F. Gordon and N. Karacapilidis. The Zeno argumentation framework. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, pages 10–18, Melbourne, Australia, 1997. ACM Press.

[16] T. F. Gordon and G. Richter. Discourse support systems for deliberative democracy. In R. Traunmüller and K. Lenk, editors, *eGovernment: State of the Art and Perspectives (EGOV02)*, pages 248–255, Aix-en-Provence, 2002. Springer Verlag.

[17] T. F. Gordon and D. Walton. The Carneades argumentation framework — using presumptions and exceptions to model critical questions. In P. E. Dunne and T. J. Bench-Capon, editors, *Computational Models of Argument. Proceedings of COMMA 2006*, pages 195–207, Amsterdam, September 2006. IOS Press.

[18] G. Governatori, M. Maher, G. Antoniou, and D. Billington. Argumentation semantics for defeasible logic. *Journal of Logic and Computation*, 14:675–702, 2004.

[19] H. L. A. Hart. *The Concept of Law*. Clarendon Press, Oxford, 1961.

[20] A. C. Hastings. *A Reformulation of the Modes of Reasoning in Argumentation*. PhD thesis, Northwestern University, Evanston, Illinois, 1963.

[21] J. Horty. Argument construction and reinstatement in logics for defeasible reasoning. *Artificial Intelligence and Law*, 9:1–28, 2001.

[22] R. Kelsey, W. Clinger, and J. Rees. Revised (5) report on the algorithmic language Scheme. *Higher-Order and Symbolic Computation*, 11(1):7–105, August 1998.

[23] R. Loui. Defeat among arguments. *Computational Intelligence*, 3:100–106, 1987.

[24] R. P. Loui. Process and policy: resource-bounded non-demonstrative reasoning. *Computational Intelligence*, 14:1–38, 1998.

[25] D. Makinson and K. Schlechta. Floating conclusions and zombie paths: two deep difficulties in the "directly sceptical" approach to defeasible inheritance nets. *Artificial Intelligence*, 48:199–209, 1991.

[26] R. C. Park, D. P. Leonard, and S. H. Goldberg. *Evidence Law*. West Group, St. Paul, Minnesota, 1998.

[27] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of some formal inter-agent dialogues. *Journal of Logic and Computation*, 13, 2003. 347-376.

[28] J. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

[29] J. Pollock. *Cognitive Carpentry*. MIT Press, Cambridge, Mass., 1995.

[30] H. Prakken. Modeling defeasibility in law: Logic or procedure? *Fundamenta Informaticae*, 48:253–271, 2001.

[31] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15:1009–1040, 2005.

[32] H. Prakken. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21:163–188, 2006.

[33] H. Prakken, C. Reed, and D. Walton. Dialogues about the burden of proof. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pages 85–94, Bologna, 2005. ACM Press.

[34] H. Prakken and G. Sartor. Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics*, 7:25–75, 1997.

[35] H. Prakken and G. Sartor. Presumptions and burden of proof. In T. van Engers, editor, *Legal Knowledge and Information Systems. JURIX 2006: The Nineteenth Annual Conference*, pages 21–30, Amsterdam, 2006. IOS Press.

[36] H. Prakken and G. Vreeswijk. Logics for defeasible argumentation. In D. M. Gabby and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 4, pages 219–318. Kluwer Academic Publishers, 2nd edition, 2002.

[37] C. Reed and T. J. Norman, editors. *Argumentation Machines — New Frontiers in Argument and Computation*. Kluwer Academic Publishers, Boston/Dordrecht/London, 2003.

[38] H. W. Rittel and M. M. Webber. Dilemmas in a general theory of planning. *Policy Science*, 4:155–169, 1973.

[39] G. R. Simari and R. P. Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artificial Intelligence*, 53(2-3):125–157, 1992.

[40] J. W. Strong. *McCormick on Evidence*. West Publishing Co., 4th edition, 1992.

[41] S. E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

[42] B. Verheij. Deflog: on the logical interpretation of prima facie justified assumptions. *Journal of Logic and Computation*, 13(3):1–28, 2003.

[43] B. Verheij. Dialectical argumentation with argumentation schemes: An approach to legal logic. *Artificial Intelligence and Law*, 11(2-3):167–195, 2003.

[44] B. Verheij, J. Hage, and J. v. d. Herik. An integrated view of rules and principles. *Artificial Intelligence and Law*, 6(1):3–26, 1998.

[45] G. Vreeswijk. Defeasible dialectics: A controversy-oriented approach towards defeasible argumentation. *Journal of Logic and Computation*, 3(3):317–334, 1993.

[46] D. Walton. *Logical Dialogue-Games and Fallacies*. University Press of America, Lanham, MD, 1984.

[47] D. Walton. *Argument Structure: a Pragmatic Theory*. Toronto studies in philosophy. University of Toronto Press, Toronto, 1996.

[48] D. Walton. *Argumentation Schemes for Presumptive Reasoning*. Erlbaum, 1996.

[49] D. Walton. *Appeal to Expert Opinion*. Penn State Press, University Park, 1997.

[50] D. Walton and D. Godden. The nature of critical questions in argumentation schemes. In D. Hitchcock, editor, *The Uses of Argument*, pages 476–484, Hamilton, Ontario, Canada, 2005. McMaster University.

[51] D. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY series in logic and language. State University of New York Press, Albany, 1995.

[52] J. H. Wigmore. *A Treatise on the Anglo-American System of Evidence*, volume 1. Little, Brown and Company, 1940.