Dialectical proof theory for defeasible argumentation with defeasible priorities (preliminary report)

Henry Prakken*

Department of Computer Science, Free University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
email: henry@cs.vu.nl

Abstract. In this paper a dialectical proof theory is proposed for logical systems for defeasible argumentation that fit a certain format. This format is the abstract theory developed by Dung, Kowalski and others. A main feature of the proof theory is that it also applies to systems in which reasoning about the standards for comparing arguments is possible. The proof theory could serve as the 'logical core' of protocols for dispute in multi-agent decision making processes.

1 Introduction

Recent nonmonotonic logics often have the form of a system for defeasible argumentation (e.g. [12, 21, 7, 15, 17, 24]). In such systems nonmonotonic reasoning is analyzed in terms of the interactions between arguments for alternative conclusions. Nonmonotonicity arises since arguments can be defeated by stronger counterarguments. In this paper a dialectical proof theory is proposed for systems of this kind that fit a certain abstract format, viz. the one defined by [7]. The use of dialectical proof theories for defeasible reasoning was earlier studied by Dung [6] and, inspired by Rescher [19], by Loui [10], Vreeswijk [22] and Brewka [4], while [20] also contains ideas that can be regarded as a dialectical proof theory. The general idea is based on game-theoretic notions of logical consequence developed in dialogue logic (for an overview see [1]). Here a proof of a formula takes the form of a dialogue game between a proponent and an opponent of the formula. Both players have certain ways available of attacking and defending a statement. A formula is provable iff it can be successfully defended against every possible attack.

In this paper first the general framework of Dung [7] will be described (Section 2), after which in section 3 the dialectical proof theory is presented. Then in Section 4 Dung's framework and the proof theory will be adapted in such a way that the standards used for comparing conflicting arguments are themselves (defeasible) consequences of the premises.

* The research reported in this paper was made possible by a research fellowship of the Royal Netherlands Academy of Arts and Sciences, and by Esprit WG 8319 'Modelage'.

The ideas of this paper were originally developed in [16], for a system of extended logic programming presented in [15], which in turn extended and revised [5]'s application of his semantics to extended logic programming. In [16] and [14] the system is applied to legal reasoning. The main purpose of the present paper is to show that the proof-theoretical ideas apply to any system of the format defined by [7]. For this reason the present paper does not express arguments in a formal language; it just assumes that this can be done.

2 An abstract framework for defeasible argumentation

Inspired by earlier work of Bondarenko, Kakas, Kowalski and Toni, [7] has proposed a very abstract and general argument-based framework. An up-to-date technical survey of this approach is [2]. The two basic notions of the framework are a set of arguments, and a binary relation of defeat among arguments. In terms of these notions, various notions of argument extensions are defined, which aim to capture various types of defeasible consequence. Then it is shown that many existing nonmonotonic logics can be reformulated as instances of the abstract framework.

The following version of this framework is kept in the abstract style of [7], with some adjustments proposed in [15, 17]. Important differences will be indicated when relevant.

Definition 1. An argument-based theory (AT) is a pair $(Args_{AT}, defeat_{AT})$,[2] where $Args_{AT}$ is a set of arguments, and $defeat_{AT}$ a binary relation on $Args_{AT}$.

- An AT is *finitary* iff each argument in $Args_{AT}$ is defeated by at most a finite number of arguments in $Args_{AT}$.
- An argument A *strictly defeats* an argument B iff A defeats B and B does not defeat A.
- A set of arguments is *conflict-free* iff no argument in the set is defeated by another argument in the set.

This definition abstracts from both the internal structure of an argument and the origin of the set of arguments. The idea is that an AT is defined by some nonmonotonic logic or system for defeasible argumentation. Usually the set $Args$ will be all arguments that can be constructed in these logics from a given set of premises, but this set might also just contain all arguments that a reasoner has actually constructed. In this paper I will (almost) completely abstract from the source of an AT. Moreover, unless stated otherwise, I will below implicitly assume an arbitrary but fixed AT.

The relation of $defeat$ is intended to be a weak notion: intuitively 'A defeats B' means that A and B are in conflict and that A is not worse than B. This means that two arguments can defeat each other. A typical example is the Nixon Diamond, with two arguments 'Nixon is a pacifist because he is a Quaker' and

[2] Below the subscripts will usually be left implicit.

'Nixon is not a pacifist because he is a Republican'. If there are no grounds for preferring one argument over the other, they defeat each other.

A stronger notion is captured by strict defeat (not used in Dung's work), which by definition is asymmetric. A standard example is the Tweety Triangle, where (if arguments are compared with specificity) the argument that Tweety flies because it is a bird is strictly defeated by the argument that Tweety doesn't fly since it is a penguin.

A central notion of Dung's framework is acceptability. Intuitively, it defines how an argument that cannot defend itself, can be protected from attacks by a set of arguments. Since [15, 17], on which this paper's proof theory is based, use a slightly different notion of acceptability, I will tag Dung's version with a *d*.

Definition 2. An argument A is *d-acceptable* with respect to a set S of arguments iff each argument defeating A is defeated by some argument in S.

The variant of Prakken & Sartor will just be called 'acceptability'.

Definition 3. An argument A is *acceptable* with respect to a set S of arguments iff each argument defeating A is strictly defeated by some argument in S.

So the only difference is that Dung uses 'defeat' where Prakken & Sartor use 'strict defeat'. In Section 4.1 I will comment on the significance of this difference.

To illustrate acceptability, consider the Tweety Triangle with A = 'Tweety is a bird, so Tweety flies', B = 'Tweety is a penguin, so Tweety does not fly' and C = 'Tweety is not a penguin', and assume that B strictly defeats A and C strictly defeats B. Then A is acceptable with respect to $\{C\}$, $\{A, C\}$, $\{B, C\}$ and $\{A, B, C\}$, but not with respect to $\emptyset$ and $\{B\}$.

Another central notion of Dung's framework is that of an admissible set.

Definition 4. A conflict-free set of arguments S is *admissible* iff each argument in S is d-acceptable with respect to S.

In the Tweety Triangle the sets $\emptyset$, $\{C\}$ and $\{A, C\}$ are admissible but all other subsets of $\{A, B, C\}$ are not admissible.

On the basis of these definitions several notions of 'argument extensions' can be defined. These notions are purely declarative, in that they just declare a set of arguments to be 'OK', without defining how such a set can be constructed. For instance, Dung defines the following credulous notions.

Definition 5. A conflict-free set S is a *stable extension* iff every argument that is not in S, is defeated by some argument in S.

Consider an AT called TT (the Tweety Triangle) where $Args_{TT} = \{A, B, C\}$ and $defeats_{TT} = \{(B, A), (C, B)\}$. TT has only one stable extension, viz. $\{A, C\}$. Consider next an AT called ND (the Nixon Diamond), with $Args_{ND} = \{A, B\}$, where A = 'Nixon is a quaker, so he is a pacifist', B = 'Nixon is a republican, so he is not a pacifist', and $defeats_{ND} = \{(A, B), (B, A)\}$. ND has two stable extensions, $\{A\}$ and $\{B\}$.

Since a stable extension is conflict-free, it reflects in some sense a coherent point of view. Moreover, it is a maximal point of view, in the sense that every possible argument is either accepted or rejected. The maximality requirement makes that not all AT's have stable extensions. Consider, for example, an AT with three arguments A, B and C, and such that A defeats B, B defeats C and C defeats A (such circular defeat relations can occur, for instance, in logic programming because of negation as failure, and in default logic because of the justification part of defaults.) To give also such AT's a credulous semantics, Dung defines the notion of a preferred extension.

Definition 6. A conflict-free set is a *preferred extension* iff it is a maximal (with respect to set inclusion) admissible set.

Clearly all stable extensions are preferred extensions, so in the Nixon Diamond and the Tweety Triangle the two semantics coincide. However, not all preferred extensions are stable: in the above example with circular defeat relations the empty set is a (unique) preferred extension, which is not stable.

Preferred and stable semantics clearly capture a credulous notion of defeasible consequence: in cases of an irresolvable conflict as in the Nixon diamond, two, mutually conflicting extensions are obtained. Dung also defines a notion of sceptical consequence, and this is for which I will define the dialectical proof theory. Application of the proof theory to the credulous semantics will be briefly discussed in Section 5. Dung defines the sceptical semantics with a monotonic operator, which for each set S of arguments returns the set of all arguments d-acceptable to S. Its least fixpoint captures the smallest set which contains every argument that is acceptable to it. I will use the variant with plain acceptability.

Definition 7. Let $AT = (Args, defeat)$ be an argument-based theory and S any subset of $Args$. The *characteristic function* of AT is:

- $F_{AT} : Pow(Args) \longrightarrow Pow(Args)$
- $F_{AT}(S) = \{A \in Args | A \text{ is acceptable with respect to } S\}$

I now give the, perhaps more intuitive, definition of [15], which by a result of [7] for finitary AT's is equivalent to the fixpoint version (which is also used in [17]). The formal results on the proof theory hold for both formulations, although for the fixpoint formulation completeness holds under the condition that the AT is finitary; cf. [7, 17].

Definition 8. For any AT $= (Args, defeat)$ we define the following sequence of subsets of $Args$.

- $F^0_{AT} = \emptyset$
- $F^{i+1}_{AT} = \{A \in Args \mid A \text{ is acceptable with respect to } F^i_{AT}\}$.

Then the set $JustArgs_{AT}$ of arguments that are justified on the basis of AT is $\cup^{\infty}_{i=0}(F^i_{AT})$.

In this definition the notion of acceptability captures reinstatement of arguments: if all arguments that defeat A are themselves defeated by an argument in F^i, then A is in F^{i+1}. To illustrate this with the Tweety Triangle: $F^1_{TT} = \{C\}$, $F^2_{TT} = \{A, C\}$, $F^3_{TT} = F^2_{TT}$, so A is reinstated at F^2 by C.

That this semantics is sceptical is illustrated by the Nixon Diamond: $F^1_{ND} = F^0_{ND} = \emptyset$.

3 A dialectical proof theory

3.1 General idea and illustrations

In this section a dialectical proof theory will be defined for the just-presented sceptical semantics. Essentially it is a notational variant of [6]'s dialogue game version of his sceptical semantics of extended logic programs. A proof of a formula takes the form of a dialogue tree, where each branch is a dialogue, and the root of the tree is an argument for the formula. The idea is that every move in a dialogue consists of an argument based on an implicitly assumed AT, and that each move attacks the last move of the opponent in a way that meets the player's burden of proof. That a move consists of a complete argument means that the search for an individual argument is conducted in a 'monological' fashion, determined by the nature of the underlying logic; only the process of considering counterarguments is modelled dialectically. The required force of a move depends on who states it, and is motivated by the definition of acceptability. Since the proponent wants a conclusion to be justified, a proponent's move has to be strictly defeating, while since the opponent only wants to prevent the conclusion from being justified, an opponent's move may be just defeating.

Let us illustrate this with an informal example of a dialogue (recall that it implicitly assumes a given AT). Let us denote the arguments stated by the proponent by P_i and those of the opponent by O_i. The proponent starts the dispute by asserting that P_1 is a justified argument.

P_1: Assuming the evidence concerning the glove was not forged,
it proves guilt of OJ.

(Many nonmonotonic logics allow the formalization of assumptions, e.g. logic programming with negation as failure and default logic with the justification part of a default.)

The opponent must defeat this argument. Suppose O can do so in only one way.

O_1: I know that the evidence concerning the glove was forged,
so your assumption is not warranted.

The proponent now has to counterattack with an argument that strictly defeats O_1. Consider the following argument

P_2: The evidence concerning the glove was not forged, since it was found by a police officer, and police officers don't forge evidence.

and suppose (for the sake of illustration) that defeat is determined by specificity considerations. Then P_2 strictly defeats O_1, so P_2 is a possible move. If the opponent has no new moves available from $Args_{AT}$, s/he loses, and the conclusion that OJ is guilty has been proved.

In dialectical proof systems a 'loop checker' can be implemented in a very natural way: no two moves of the proponent in the same branch of the dialogue may have the same content. It is easy to see that this rule will not harm P; if O had a move the first time P stated the argument, it will also have a move the second time, so no repetition by P can make P win a dialogue.

Assume for illustration that the arguments in $Args$ are those that can be made by chaining one or more of the following premises:

(1) Mr. F forged the glove-evidence
(2) Someone who forges evidence is not honest
(3) Mr. F is a police officer
(4) Police officers are honest
(5) Someone who is honest, does not forge evidence.

Assume again that defeat is determined by specificity, in the obvious way. Now the proponent argues that Mr. F did not forge the glove-evidence.

P_1: Mr. F is a police officer, so he is honest and therefore does not forge evidence.

O attacks this argument on its 'subconclusion' that Mr. F is honest; and since the counterargument is more specific, this is a defeating argument.

O_1: I know that F forged evidence, and this shows that he is not honest.

P now wants to attack O's argument in the same way as O attacked P's argument: by launching a more specific attack on O's 'subconclusion' that F forged the glove-evidence. However, P has already stated that argument at the beginning of the dispute, so the move is not allowed. And no other strictly defeating argument is available, so it is not provable that Mr. F did not forge the glove-evidence, not even that he is honest. However, by a completely symmetric line of reasoning we obtain that also the contrary conclusions are not provable. So no conclusion about whether Mr. F is honest or not, and forged evidence or not, is provably justified.

3.2 The proof theory

Now the dialectical proof theory will be formally defined. Again the definitions assume an arbitrary but fixed AT.

Definition 9. A *dialogue* is a finite nonempty sequence of moves $move_i = (Player_i, Arg_i)$ $(i > 0)$, such that

1. $Player_i = P$ iff i is odd; and $Player_i = O$ iff i is even;
2. If $Player_i = Player_j = P$ and $i \neq j$, then $Arg_i \neq Arg_j$;
3. If $Player_i = P$, then Arg_i strictly defeats Arg_{i-1};
4. If $Player_i = O$, then Arg_i defeats Arg_{i-1}.

The first condition says that the proponent begins and then the players take turns, while the second condition prevents the proponent from repeating its attacks. The last two conditions form the heart of the definition: they state the burdens of proof for P and O.

Definition 10. A *dialogue tree* is a tree of dialogues such that if $Player_i = P$ then $move_1$'s children of are all arguments that defeat Arg_i.

It is this definition that makes dialogue trees candidates for being proofs: it says that the tree should consider all possible ways in which O can attack an argument of P.

Definition 11. A player *wins a dialogue* iff the other player cannot move. And a player *wins a dialogue tree* iff it wins all branches of the tree.

The idea of this definition is that if P's last argument is undefeated, it reinstates all previous arguments of P that occur in the same branch of a tree, in particular the root of the tree.

Definition 12. An argument A is *provably justified* iff there is a dialogue tree with A as its root, and won by the proponent.

In [17] it is shown that this proof theory is sound and for finitary AT's also complete with respect to the sceptical fixpoint semantics. This is not surprising, since what the proof theory does is, basically, traversing the sequence defined by Definition 8 in the reverse direction. Note that the latter definition implies that an argument A is justified iff there is a sequence $F^1, \ldots, F^n$ such that A occurs for the first time in F^n (in the explicit fixpoint definition of [7, 17] this only holds for finitary AT's; in the general case only the 'if' part holds). We start with A, and then for any argument B defeating A we find an argument C in F^{n-1} that strictly defeats B and so indirectly supports A. Then any argument defeating C is met with a strict defeater from F^{n-2}, and so on. Since the sequence is finite, we end with an argument indirectly supporting A that cannot be defeated.

It should be noted that completeness here does not imply semi-decidability: if the logic for constructing individual arguments is not decidable, then the search for counterarguments is in general not even semi-decidable, since this search is essentially a consistency check.

4 Defeasible priorities

In several argumentation frameworks, as in many other nonmonotonic logics, the defeat relation is partly defined with the help of priority relations, usually defined on the premises, but sometimes directly on arguments. In most systems these priorities are undisputable and assumed consistent. However, as discussed in e.g. [8, 16, 9], these features are often unrealistic. In several domains of practical reasoning, such as legal reasoning, the priorities are themselves subject to debate, and therefore a full theory of defeasible argumentation should also be able to formalise arguments about priorities, and to adjudicate between such arguments.

This section presents a formalisation of this feature, which forms the main technical addition to [5, 6]. As the previous section, this section is also based on [15], in which the semantics of [5] is revised, and on [17], in which the same is done with the proof theory of [6]. The present section generalises these revisions to any system fitting the format of [7].

However the generalisation is only well-defined if the logic generating an AT satisfies some additional assumptions. Firstly, it must be assumed that for each AT a set O is defined of objects to be ordered. For most AT's the set O will contain the premises from which the arguments of the AT can be constructed; however, since some AT's instead define the priorities between sets of premises or even directly between arguments (as [24]), the content of O must be left undefined.

Next I assume that the defeat relation of an AT is determined by a strict partial ordering of O. In fact, this assumption transforms the defeat relation of an AT into a *set* of defeat relations $<$-$defeat$, where $<$ is any strict partial ordering of O.

On the basis of these assumptions the notion of a prioritised argument-based theory can be defined.

Definition 13. A prioritised argument-based theory (PAT for short) is a triple $(Args_{PAT}, O_{PAT}, defeat_{PAT})$,[3] where $Args_{PAT}$ is a set of arguments, and where $defeat_{PAT}$ is a set of binary relations $<$-$defeat$ on $Args_{PAT}$, $<$ being any strict partial order on O_{PAT}.

- A PAT is *finitary* iff for all $<$ each argument in $Args_{PAT}$ is $<$-defeated by at most a finite number of arguments in $Args_{PAT}$.
- An argument A *strictly* $<$-*defeats* an argument B iff A $<$-defeats B and B does not $<$-defeat A.
- A set of arguments is $<$-*conflict-free* iff no argument in the set is $<$-defeated by another argument in the set.

Finally, it must be assumed that the argument language of a PAT is sufficiently expressive to express partial orderings on O; i.e., that this language contains a distinguished twoplace predicate symbol $\prec$, intended to denote the relation $<$, and that there is a naming function $N : O \longrightarrow Names$, where $Names$ is a set of

[3] Below the subscripts will usually be left implicit.

terms. N is not assumed to be a bijection, since it might be convenient to assign the same name to more than one object.

4.1 Changing the semantics

Now how can we make the priorities that are needed to determine defeat, defeasible consequences of the AT, according to Definition 8? The idea is that in determining whether an argument is acceptable with respect to F^i_{PAT}, we look at those priority statements that are conclusions of arguments in F^i_{PAT}. To this end first the notion of an ordering expressed by a set of arguments must be defined.

Definition 14. For any set S of arguments

$$<_S = \{o < o' \mid N(o) \prec N(o') \text{ is a conclusion of some } A \in S\}$$

Below '$<_S$-defeat' will be abbreviated as 'S-defeat'; and for singleton sets $\{C\}$, '$\{C\}$-defeat' will be written as 'C-defeat'.

For arbitrary sets S it is not guaranteed that $<_S$ is a strict partial order. However, it is sufficient that the properties hold for each $<_{F^i_{AT}}$. In virtually any nonmonotonic logic this can be assured by including the axioms of a strict partial order for $\prec$ in the undebatable part of the premises (see [17] for an illustration in argument-based extended logic programming).

Next the notion of acceptability is redefined as follows (d-acceptability can be changed in the same way).

Definition 15. An argument A is *acceptable* with respect to a set S of arguments iff all arguments S-defeating A are strictly S-defeated by some argument in S.

Note that with this definition Dung's original definition is not only changed (by using strict defeat), but also refined: this is since Dung does not consider defeasible priorities and therefore does not make defeat relative to sets of arguments.

Definition 8 can now be applied with Definition 15. However, to make this application well-behaved, the notion of S-defeat should have the following two properties, which are crucial in proving that each F^i is contained in F^{i+1}; this in turn guarantees that each set of justified arguments is conflict-free. The properties are also crucial in proving that the explicit-fixpoint definition of [17] is monotonic. Note that they do not follow from the above definitions but must instead be enforced by a proper definition of the notion of defeat.

Property 4.1 *For any two conflict-free sets of arguments S and S' such that $S \subseteq S'$, and any two arguments A and B we have that*

1. *If A S'-defeats B, then A S-defeats B.*
2. *If A strictly S-defeats B, then A strictly S'-defeats B.*

Given our weak interpretation of the defeat notion, this property can easily be enforced: the idea is to define 'A S-defeats B' in terms of the absence of priorities in $<_S$ that would make A worse than B; then adding more priorities cannot create new defeat relations, while the only defeat relations that go away are one side of a mutual defeat relation.

Property 4.2 *For any conflict-free set of arguments S and arguments $A \in S$ and B: if A strictly S-defeats B, then some $C \in S$ strictly C-defeats B.*

This property also seems very natural. The intuition behind it is that C is the combination of A with the priority arguments in S that make A strictly S-defeat B; and C can then be used in a dialectical proof as a reply to B.

It is now appropriate to comment on the use of strict defeat in Definitions 3 and 15: Property 4.1(2) will usually not hold for defeat, while yet it is essential to make Definition 8 well-behaved when combined with Definition 15.

4.2 Changing the proof theory

Let us now see how the proof theory must be changed. The main problem here is on the basis of which priorities the defeating force of the moves should be determined. What is to be avoided is that we have to generate all priority arguments before we can determine the defeating force of a move. The pleasant surprise is that, to achieve this, a few very simple conditions suffice. For O it is sufficient that its move $\emptyset$-defeats P's previous move. This is so since Property 4.1 implies that if A is for some S an S-defeater of P's previous move, it is also an $\emptyset$-defeater of that move. So O does not have to take priorities into account. Let us illustrate this by modifying our informal glove dialogue as follows (we again leave it to the readers to formalise the arguments in their favourite formalism). Again the proponent starts with

P_1: Assuming the evidence concerning the glove was not forged, it proves guilt of OJ.

Suppose the opponent now replies with

O_1: I know that the evidence concerning the glove was forged, since I was told so, so your assumption is not warranted.

In agreement with most nonmonotonic logics, I assume that an attack on an assumption succeeds if no priority relations hold: i.e. O_1 $\emptyset$-defeats P_1.

P, on the other hand, should take some priorities into account, since strict defeat usually requires 'better than' relations between rules. However, it suffices to apply only those priorities that are stated by P's move; more priorities are not needed, since Property 4.1 also implies that if P's argument Arg_i strictly $Args_i$-defeats O's previous move, it also does so whatever more priorities are derived. So P can reply to O_1 with

P_2: The evidence concerning the glove was not forged, since it was found by a police officer, and as a general rule police officers don't forge evidence. This rule is more reliable than your rule that what you are told is true.

Because of the priority statement at the end, P_2 strictly P_2-defeats O_1.

However, this is not the only type of move that the proponent should be allowed to make. To see this, note that O can respond with repeating O_1 as O_2, at least assuming that O_1 $\emptyset$-defeats P_2, which in many systems it will do (e.g. in [15]). And because of the nonrepetition rule P cannot respond to O_2 with $P_3 = P_2$. Therefore P must be allowed to state a priority argument that neutralises the defeating force of O_2, i.e. to state an argument P_3 such that O_2 does not P_3-defeat P_1. If P is allowed to make such a move, it can in our example repeat the priority part of P_2:

P_3: The rule that police officers don't forge evidence is more reliable than your rule that what you are told is true.

O might now, depending on the content of $Args_{PAT}$, challenge P's priority argument, for instance, by saying that instead the 'what I am told is true' rule is more reliable since O only listens to very reliable people.

Let us see which changes in the proof theory are required to capture such priority debates. All that has to be changed is the burdens of proof in Definition 9:

(3) If $Player_i = P$ *then*

- Arg_i *strictly* Arg_i*-defeats* Arg_{i-1}*; or*
- Arg_{i-1} *does not* Arg_i*-defeat* A_{i-2}.

(4) If $Player_i = O$ *then* Arg_i $\emptyset$*-defeats* Arg_{i-1}.

The other definitions stay the same.

In [17] it is shown that the proof theory is, with respect to the fixpoint semantics, sound in the general case and complete for finitary AT's. The corresponding results for the system with fixed priorities are proven as a special case. Although these results are proven for a particular system, the proofs are based on only the definitions and properties presented in this paper.

4.3 A clash of intuitions

In some cases the semantics of this section gives results that seem debatable. Consider an AT with $Args_{AT} = \{A, B, C, D\}$ where A = 'John is an adult, so John is employed', B = 'John is a student, so John is unemployed', C = 'John is imprisoned, so John is unemployed' and D is a priority argument with conclusion $A \prec B \vee A \prec C$. Assume that this induces an ordering $<_{JustArgs_{AT}} = \emptyset$, so that none of the arguments is justified. Assume now that if this ordering were instead

$\{A < B\}$ or $\{A < C\}$, then B and C would be justified and A overruled. It might be argued that then this should also be the outcome in the original case. However, intuitions seem to differ here: from a constructive point of view the outcome of the present definitions seems acceptable.

Yet it is worthwhile investigating how the alternative, non-constructive intuition can be formalized. Probably techniques from [3] and [13] can be used, which formalize the non-constructive intuition for extension-based systems, but this has to be left for future research, as well as the corresponding proof theory. Alternatively, syntactic restrictions will do; practically this seems a feasible option, since in practical applications disjunctive priority information seems very rare.

5 Proof theory for credulous semantics

In this section I sketch how a dialectical proof theory can be developed for the credulous semantics discussed in Section 2. I will first focus on the case with fixed priorities. Defining a proof theory for stable semantics will not be easy, since it must always be proven that a stable extension exists. Let us therefore concentrate on preferred semantics. This is also relevant for stable semantics, since [7] identifies conditions under which preferred and stable semantics coincide.

Note first that the existence of a proof means that the argument is in *some* preferred extension. Now the idea is to reverse the burden of proof of P and O. P now only has to defeat O's arguments, while O now must strictly defeat P's moves. Moreover, the non-repetition rule now holds for O instead of for P, while the children of P's moves are now all its *strict* defeaters. Finally, since preferred extensions are conflict-free, we must require that in each dialogue the set of all moves of the proponent is conflict-free.

With respect to soundness and completeness, it is relevant that by definition every admissible set is contained in some preferred extension. Then soundness follows since it is easy to see that the union of all P's arguments in a dialogue tree is an admissible set. Completeness can be proven for the finite case, by showing that each finite admissible set corresponds to a proof for each of its members. For the infinite case there are obvious counterexamples. Consider e.g. an infinite set of arguments $\{A_1, \ldots, A_n, \ldots\}$, where each A_i $(i > 1)$ strictly defeats A_{i-1}: both the set of all 'odd', and that of all 'even' arguments are preferred extensions, but any 'proof' has to be infinite.

Extending these ideas to the case with defeasible priorities is still to be investigated.

6 Formal models of agents and protocols for dispute

With respect to formal models of agents this paper is relevant as follows. As noted earlier by [23], the dialectical proof theory can serve as the 'logical core' of protocols for disputes in multi-agent decision and negotiation processes (where

the agents can be humans, computers or a combination of both). Such protocols define possible, allowed or obligatory dialogue moves of the agents involved in the dispute, and they define criteria for termination and evaluation of a dispute. Such protocols can be studied as to their degree of rationality (cf. e.g. [10, 8, 23]). The leading idea here is that rationality has a procedural side: an argument is acceptable if it has been successfully defended in a properly conducted dispute. The main aim of this line of research is to find out what makes a dispute proper, i.e. what makes it fair and effective.

A key feature of realistic disputes is that the body of information from which arguments can be constructed is not given in advance, but is constructed dynamically in the course of a debate. Although our dialectical proof theory is relative to a given set of arguments, it can still be embedded in such protocols for dispute. This has, for instance, been done in [18] (cf. also [11, 23]), where the set $Args_{AT}$ is defined as the arguments that are constructible on the basis of the premises that are introduced and not withdrawn at a give stage. Thus the above definitions also apply to disputes where the set of premises is dynamically constructed. Moreover, the soundness and completeness results are thus part of the criteria for fair and effective disputation. This is at least how [23] defines fairness and effectiveness: a protocol is fair if every argument that can be successfully defended against every attack is justified, and it is effective if every justified argument can be successfully defended against every attack.

7 Concluding remarks

In this paper I have discussed three contributions to the formalisation of defeasible argumentation. Firstly, I have, by generalising work of [15], discussed how the abstract framework of [7, 2] can be extended with defeasible priorities. Secondly, I have, by generalising work of [6] and [17], discussed how dialectical proof theories can be defined for this framework and its extension. Finally, I have given an impression of the research questions that arise in the dialectical approach to the proof theory of defeasible argumentation, and I have indicated how this approach is relevant to formal protocols for disputation in multi-agent environments.

As for future research, first of all the preliminary contributions of this paper should, of course, be further developed. Moreover, it would be interesting to investigate in more detail the relation between dialectical proof theories and dialectical protocols for disputation.

References

1. Barth, E.M. & Krabbe, E.C.W.: *From Axiom to Dialogue: a Philosophical Study of Logic and Argumentation.* Walter de Gruyter, New York, 1982.
2. Bondarenko, A., Dung, P.M., Kowalski, R.A. & Toni, F.: An abstract argumentation-theoretic approach to default reasoning. *Artificial Intelligence* 93 (1997), 63–101.

3. Brewka, G.: Reasoning about priorities in default logic. *Proceedings AAAI-94*, 247–260.
4. Brewka, G.: A reconstruction of Rescher's theory of formal disputation based on default logic. *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI-94)*, 366–370.
5. Dung, P.M.: An argumentation semantics for logic programming with explicit negation. *Proceedings of the Tenth Logic Programming Conference*, MIT Press 1993, 616–630.
6. Dung, P.M.: Logic programming as dialogue game. Unpublished paper.
7. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence* 77 (1995), 321–357.
8. Gordon, T.F.: *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice*. Kluwer Academic Publishers, Dordrecht etc. 1995.
9. Hage, J.C.: *Reasoning With Rules. An Essay on Legal Reasoning and Its Underlying Logic*. Kluwer Law and Philosophy Library, Dordrecht etc. 1997.
10. Loui, R.P.: Process and policy: resource-bounded non-demonstrative reasoning. Report WUCS–92–43, Washington-University-in-St-Louis, 1993. To appear in *Computational Intelligence*.
11. Loui, R.P. & and Norman, J.: Rationales and argument moves. *Artificial Intelligence and Law* 3: 159–189, 1995.
12. Pollock, J.L.: Defeasible reasoning. *Cognitive Science* 11 (1987), 481–518.
13. Prakken, H.: A semantic view on reasoning about priorities (extended abstract). *Proceedings of the Second Dutch/German Workshop on Nonmonotonic Reasoning*, Utrecht 1995, 152–159.
14. Prakken, H.: *Logical Tools for Modelling Legal Argument. A Study of Defeasible Reasoning in Law*. Kluwer Law and Philosophy Library, Dordrecht etc. 1997.
15. Prakken, H. & Sartor, G.: A system for defeasible argumentation, with defeasible priorities. *Proceedings of the International Conference on Formal and Applied Practical Reasoning (FAPR'96)*, Bonn 1996. Springer Lecture Notes in AI 1085, Springer Verlag, 1996, 510–524.
16. Prakken, H. & Sartor, G.: A dialectical model of assessing conflicting arguments in legal reasoning. *Artificial Intelligence and Law* 4 (1996), 331–368. Reprinted in H. Prakken & G. Sartor (eds.): *Logical Models of Legal Argument*. Kluwer Academic Publishers, Dordrecht etc. 1996, 175–212.
17. Prakken, H. & Sartor, G.: Argument-based extended logic programming with defeasible priorities. *Journal of Applied Non-classical Logics* 7 (1997), 25–75.
18. Prakken, H. & Sartor, G.: Reasoning with precedents in a dialogue game. *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, 1–9. ACM Press, New York, 1997.
19. Rescher, N.: *Dialectics: a Controversy-oriented Approach to the Theory of Knowledge*. State University of New York Press, Albany, 1977.
20. Royakkers, R. & Dignum, F.: Defeasible reasoning with legal rules. In M.A. Brown and J. Carmo (eds.) *Deontic Logic, Agency and Normative Systems*. Springer, Workshops in Computing, London etc. 1996, 174–193.
21. Simari, G.R. & Loui, R.P.: A mathematical treatment of defeasible argumentation and its implementation. *Artificial Intelligence* 53 (1992), 125–157.
22. Vreeswijk, G.: Defeasible dialectics: a controversy-oriented approach towards defeasible argumentation. *Journal of Logic and Computation*, 1993, Vol. 3, No. 3., 317–334.

23. Vreeswijk, G.: Representation of formal dispute with a standing order. *Research Report MATRIX, University of Limburg, 1996.* Also presented at the *Workshop Computational Dialectics* of the International Conference on Formal and Applied Practical Reasoning (FAPR'96), Bonn 1996.
24. Vreeswijk, G.: Abstract argumentation systems. *Artificial Intelligence* 90 (1997), 225–279.