Relating Protocols for Dynamic Dispute with Logics for Defeasible Argumentation

Henry Prakken*
Department of Computer Science, Utrecht University
Utrecht, The Netherlands
henry@cs.uu.nl, http://www.cs.uu.nl/staff/henry.html

December 14, 1999

Abstract

This article investigates to what extent protocols for dynamic disputes, i.e., disputes in which the information base can vary at different stages, can be justified in terms of logics for defeasible argumentation. First a general framework is formulated for dialectical proof theories for such logics. Then this framework is adapted to serve as a framework for protocols for dynamic disputes, after which soundness and fairness properties are formulated for such protocols relative to dialectical proof theories. It then turns out that certain types of protocols that are perfectly fine with a static information base, are not sound or fair in a dynamic setting. Finally, a natural dynamic protocol is defined for which soundness and fairness can be established.

1 Introduction

This paper studies the exchange of arguments and counterarguments in dynamic disputes, i.e., in disputes where the available information can change during the dispute. The research is motivated by two recent developments in Artificial Intelligence: research on logical systems for defeasible argumentation, and research on the use of argumentation in multi-agent interaction, such as in negotiation, group decision making and dispute mediation.

Logics for defeasible argumentation (e.g. Pollock 1992; Dung 1995; Vreeswijk 1997 and, for a survey, Prakken and Vreeswijk 2000) are one approach to the formalisation of so-called defeasible, or nonmonotonic reasoning. This is reasoning where tentative conclusions are drawn on the basis of uncertain or incomplete information, which might have to be withdrawn if more information becomes available. Logical argumentation systems formalise this kind of reasoning in terms of the interactions between arguments for alternative conclusions.

*I thank Gerd Brewka, Tom Gordon and Gerard Vreeswijk for interesting discussions on the topic of this article, and Koen Hindriks for his useful comments on an earlier version of the article.

Nonmonotonicity arises since arguments can be defeated by stronger counterarguments.

In Artificial Intelligence the exchange of arguments and counterarguments has also been studied in the context of multi-agent interaction. (See also Walton [1999], for an argumentation-theorist interested in AI and multi-agent research.) For instance, Kraus *et al.* [1998] and Parsons *et al.* [1998] have studied argumentation as a component of negotiation protocols, where arguments for an offer should persuade the other party to accept the offer. Loui [1994] has also studied the role of argumentation in negotiation. Argumentation is also part of some recent formal models and computer systems for dispute mediation [Gordon, 1995, Gordon *et al.*, 1997, Brewka, 1999], and it has been used in computer programs for intelligent tutoring: for instance, in a system (Belvedere) that teaches scientific reasoning [Suthers *et al.*, 1995] and in a system (CATO) that teaches disputation skills to law students [Aleven and Ashley, 1997]. In sum, there are computer systems that perform argumentation, systems that mediate it, and systems that teach it.

Such computer applications raise the issue against which formal standards their argumentation aspects can be evaluated. In other words, are there rational principles governing the exchange of arguments and counterarguments in disputational dialogues, and if so, can these principles be formalised? The purpose of this paper is to give a (partial) answer to these questions. In particular, I shall investigate to what extent these principles can be formulated in terms of logics for defeasible argumentation.

Such logics seem very suitable for this purpose, not only since they are about the interaction between arguments and counterarguments, but also since they can be cast in the dialectical style of an argument game, as a dispute between a proponent and opponent of a claim. The proponent starts with an argument for this claim, after which each player must attack the other player's previous argument with a counterargument of sufficient strength. The initial argument is provable if the proponent has a winning strategy, i.e., if he can make the opponent run out of moves in whatever way she attacks. This setup fits well with the just-mentioned multi-agent applications, which often incorporate their argumentation features in protocols for dialogue. Accordingly, the way in which in this paper the connection between defeasible reasoning and disputation will be made, is by reinterpreting dialectical proof theories as part of the discourse rules, or 'protocols' for disputational dialogues. (Such protocol rules are, in terms of Hintikka [1999], definitory instead of strategic rules: i.e., they define the allowed disputational moves instead of when such moves are good or bad.)

However, some research problems have to be solved before this connection can be made. The first is that, while in proof-theoretical disputes all arguments are constructed from a *given* body of information, in disputes between real agents this body of information is usually constructed dynamically, during the dispute, since the participants can at any time supply new or withdraw old information. One consequence of this is that the outcome of a dispute is relative to the stages of the dispute: a certain outcome can always be overturned by supplying new information. A second, and related problem is that while in dialectical

proof theories the focus is on the *possibility* to win a dispute (since this implies provability), in protocols for 'real' disputes the focus is on the result of a given dispute as it has actually happened. And in such an actual dispute it is, for instance, possible that possible arguments have not been advanced. Therefore, in disputes, unlike in proof theories, it does not always make sense to evaluate the outcome of a dispute with respect to all possible arguments; sometimes only those arguments that have actually been stated are relevant. A clear example is provided by several legal procedures in civil cases, where possible arguments that have not been stated by one of the parties are legally irrelevant.

These observations lead to such questions as the following:

1. Is a given protocol for dispute *sound*, in the sense that if the proponent wins a dispute, its initial argument is defeasibly provable on the basis of what has been said in the dispute?

2. Is a given protocol for dispute *fair* (or complete), in the sense that if a certain argument becomes defeasibly provable on the basis of what has been said in the dispute, the proponent (opponent) can win any continuation of the dispute in which no new information is introduced?

These questions will be studied in this paper for several natural protocols for dispute. One of the main results will be that certain protocols that are perfectly acceptable in a static setting, have problems when the information base is built dynamically.

The structure of this paper is as follows. In Section 2 I put the study of disputes in context by discussing their place in dialogical models of argumentation. In Section 3 I briefly outline the general ideas behind logics for defeasible argumentation, and present a formal framework for their dialectical proof-theories. This dialectical framework is then reinterpreted in the rest of this paper as a protocol for dispute: in Section 4 the general ideas are explained, and in Sections 5 and 6 some simpler and more complex protocols are investigated on soundness and fairness. Section 7 then discusses related research, after which Section 8 concludes.

2 Argumentation and dialogue: the place of dispute

In this section I discuss the place of disputes in dialogical models of argumentation. In argumentation theory, several types of dialogues are distinguished, such as information seeking, negotiation, persuasion or critical discussions, and deliberation or group decision making [Walton, 1990, Walton and Krabbe, 1995]. Each type of dialogue is characterised by an initial situation and a goal. The protocols to be studied in the present paper are for a subtype of persuasion dialogues. In persuasion, the initial situation is a conflict of opinion, and the goal is to resolve this conflict by verbal means. The proponent of a claim aims at making the opponent concede his claim; the opponent instead aims at making the proponent withdraw his claim. Logic governs the dialogue in various ways. For instance, if a participant is asked to give grounds for a claim, these grounds

have to logically imply the claim (in some models, if this is not the case, then the 'hidden premise' is implicitly made part of the argument.) Or if a proponent's claim is logically implied by the opponent's concessions, the opponent is forced to accept the claim, or else withdraw some of her concessions.

In current argumentation-theoretic models of persuasion (e.g. Hamblin 1971; MacKenzie 1990; Walton and Krabbe 1995), the underlying logic is standard deductive logic. However, in this paper I shall study the case where the underlying logic is defeasible, resulting from the fact that the parties can exchange not only arguments but also counterarguments. So support for a claim may be defeasible (e.g. inductive or analogical) instead of watertight, and concession of a claim is forced if the arguments for the claim defeat the arguments against it. I shall call this subtype of persuasion dialogues 'disputes'. This notion comes close but is not equal to Walton and Krabbe's [1995] notion of dispute, which is a persuasion dialogue that starts with inconsistent claims by both parties. In their disputes, arguments must still be deductive, and counterarguments still have no place.

In fact, I shall study only one aspect of disputes, viz. the exchange of arguments and counterarguments. I shall ignore the speech act aspects of disputes (such as asserting, challenging, conceding and withdrawing a claim), assuming that these aspects are regulated by a 'standard', Hamblin-MacKenzie-type protocol. Combining such a protocol with a protocol for dispute will then yield an overall account of persuasion dialogues (such a combination is in fact an issue for future research).

The following example illustrates how a dispute can be part of a persuasion dialogue.

`Paul:` My car is safer than your car. (*persuasion: making a claim*)
`Olga:` Why is your car safer? (*persuasion: asking grounds for a claim*)
`Paul:` Since it has an airbag. (*persuasion: offering grounds for a claim; dispute, stating an initial argument*)
`Olga:` It is true that your car has an airbag (*persuasion: conceding a claim*) but I disagree that this makes your car safe: the newspapers recently had several reports on cases where airbags expanded without cause. (*dispute: stating a counterargument*)
`Paul:` I also read that report (*persuasion: conceding a claim*) but a recent scientific study showed that cars with airbags are safer than cars without airbags, and scientific studies are more reliable than sporadic newspaper reports. (*dispute: rebutting a counterargument, and arguing about strength of conflicting arguments*)
`Olga:` OK, I admit that your argument is stronger than mine. (*persuasion: conceding a claim*) However, your car is still not safer, since its maximum speed is much higher. (*dispute: alternative counterargument*)

This shows how disputes can be embedded in persuasion dialogues. However, as said above, this paper confines itself to protocols for dispute, leaving their embedding in protocols for persuasion for future research. Accordingly, in the above example, my focus will be on its following 'disputational core':

`Paul:` My car is safer than the other car since it has an airbag.
`Olga:` An airbag does not make your car safe: the newspapers recently had several reports on cases where airbags expanded without cause.
`Paul:` But a recent scientific study showed that cars with airbags are safer than cars without airbags, and scientific studies are more reliable than sporadic newspaper reports.
`Olga:` Your car is still not safer, since its maximum speed is much higher.

3 Basics of logics for defeasible argumentation

In this section I introduce the basics of logics for defeasible argumentation (or 'argumentation systems' for short). After a brief sketch of the main ideas, the larger part will be devoted to a formal framework of dialectical formulations of argumentation systems. A detailed overview of the field can be found in [Prakken and Vreeswijk, 2000].

3.1 Logics for defeasible argumentation: the general idea

Argumentation systems are one way to formalise nonmonotonic reasoning, viz. as the construction and comparison of arguments for and against certain conclusions. The idea is that the construction of arguments is monotonic, i.e., an argument stays an argument if more premises are added. Nonmonotonicity, or defeasibility, is explained in terms of the interactions between conflicting arguments: it arises from the fact that new premises may give rise to counterarguments that defeat the original argument. Accordingly, the input of an argumentation system is a set of arguments and a relation of strength among arguments, and the output is an assignment of a status to arguments. Typically this status is defined in terms of three classes: the 'winning' arguments, the 'losing' arguments, and the 'ties', i.e., the arguments that are involved in an irresolvable conflict. It is important to note that, as shown by Dung [1995] and Bondarenko *et al.* [1997], argumentation systems can also be seen as a general framework for nonmonotonic logics. Therefore, this paper's focus on argumentation systems is not a serious limitation.

Argumentation systems can be stated both in a 'semantic' and in a 'proof-theoretic' form. The semantics assigns a status to all arguments in a given set, on the basis of their mutual defeat relations. It does so without any regard for how this status could be computed. Typically, the semantics has the form of a fixed-point definition. In their proof-theoretic form, argumentation systems define the notion of a (defeasible) proof that a particular argument has a certain status. The most common proof-theoretic form is that of a dispute, as described in the introduction and further explained below. In this paper I shall focus on such 'dialectical' proof theories. However, my results will also be relevant for argument-based semantics, insofar as a dialectical proof theory has such a semantics.

First, however, it is useful to say something about the origin and structure

of arguments. In the logical study of defeasible argumentation the main focus is not on the structure and validity of individual arguments but on the interaction of conflicting arguments. Accordingly, the idea of an argumentation system is compatible with almost any view on arguments. (This is also one of the reasons why other nonmonotonic logics can be translated into argumentation systems.) The individual arguments could be required to be deductive, but they could also be allowed to be 'ampliative', such as analogical, inductive or abductive arguments. They could even be formed by links between unanalysed pieces of text. The latter is especially relevant for applications in mediation systems, where arguments could, for instance, have the form of email messages written by humans. See [Gordon *et al.*, 1997, Gordon and Karaçapilidis, 1997] for such a mediation system. Finally, even if arguments are defined in a formal logic, subtle distinctions can be made. At first sight, one might think that the 'input' set of arguments of an argumentation system is simply the set of all arguments that are logically enabled by a given set of premises. However, computer programs cannot be guaranteed to find arguments in reasonable time, and sometimes not even in finite time. Therefore, some argumentation systems leave room for resource bounds on the computation of arguments (see especially Pollock 1992; Loui 1998). Then the 'input' set of arguments is a subset of all arguments that are logically enabled by the premises. For these reasons, I shall in the present paper leave the origin and internal structure of arguments unspecified.

The same holds for the origin and structure of the relation of strength among arguments. For some time, researchers in nonmonotonic reasoning thought that the specificity principle would be a useful general commonsense principle for comparing arguments. (That is, the more specific the information on which an argument is based, the stronger it is.) However, it is now widely acknowledged that many domains have their own standards, and that often these standards are themselves the subject of debate (as illustrated by Paul's third move in the example of Section 2). Accordingly, most argumentation systems leave the origin of the strength criteria undefined, and some systems even make them defeasibly derivable within the system. Therefore, to capture the various options I shall assume as little as possible about these criteria.

3.2 A framework for dialectical logics for defeasible argumentation

I now describe a framework for dialectical proof theories for defeasible argumentation, which summarises and abstracts existing proof theories of e.g. Simari and Loui [1992], Dung [1994], Vreeswijk [1995] and Prakken and Sartor [1997]. This framework will be reinterpreted in the following section as a framework for protocols of dispute. The advantage of using a general framework is that the results of the following sections apply to any proof theory of a certain format. A disadvantage is perhaps that it is harder for the reader to see how dialectical proof theories work. Therefore, I shall end this section with an example.

The work in this section is inspired by an earlier framework of this kind presented by Ronald Loui in his groundbreaking paper (cf. Loui 1998). However,

there are some differences in focus and formalisation, on which more below in Section 7. The present framework also incorporates some elements of [Jakobovits and Vermeir, 1999].

I shall first informally discuss the elements of the framework, and then define it formally.

3.2.1 Informal ideas

As already explained in the introduction, the idea of a dialectical proof theory is that a proof of a proposition has the form of an argument game between a proponent and opponent of the proposition. The proponent starts with an argument for it, after which each player must attack the other player's previous argument with a counterargument of sufficient strength. The initial argument and thereby its conclusion is provable if the proponent has a winning strategy, i.e., if he can make the opponent run out of moves in whatever way she attacks.

Accordingly, dialectical proof theories for defeasible argumentation have the following elements.

- *Players*, viz. a proponent and an opponent of an initial argument. Other possible players, such as referees, chairs, judges or mediators, belong not to disputes but to different types of protocols.

- A set *Args* of well-formed, or valid *arguments*, according to some monotonic (but not necessarily standard) notion of logical consequence. For reasons stated above, the structure and origin of this set are irrelevant for present purposes.

- A relation of relative *strength of arguments*. An argument's strength is one of the grounds on which legality of a move is determined. I shall leave unspecified how this is done.

- *Well-formed moves*. A move has a player, i.e., the one who moves, a main content, which is an argument, and an indication of the move to which it replies. The set of well-formed moves is completely determined by the set *Args*.

- A function *PlayerToMove*, which for each stage of a dialogue determines which player is to move next. The proponent is always the first player to move. In all current dialectical proof theories, the players take turns after each move. However, for disputes protocols are conceivable where this is otherwise.

- A *legal-move function*. Well-formed moves are not always legal. The function *Legal* specifies for each stage of a dialogue which moves are legal at that stage. The legality of a move is completely determined by the dialogue thus far. In all dialectical proof theories a necessary condition for legality of moves is that they reply to the preceding move. The definition of legality must be completed by particular proof theories. Above all, they

must state the required strength of counterarguments, and regulate when repetition of moves is allowed.

- A notion of a *dialogue*, being a series of legal moves made by the player to move.
- A *winning* criterion, being that a dialogue is won by a player iff the other player cannot move.
- A *provability* criterion, being that an argument is provably tenable iff the proponent has a winning strategy in a dialogue that begins with moving this argument. In other words, an argument is provably tenable iff the proponent can make the opponent run out of moves against every way of attack.

3.2.2 The formal framework

I now formalise the informal description of the previous subsection. I start with the notion of a dialectical framework, which is a given ordered body of information plus a dialectical proof theory that is to be applied to it.

Definition 3.1 [Dialectical frameworks.] A *dialectical framework* is a triple $(Args, \preceq, T)$, where

- $Args$ is a set of arguments.
- $\preceq \subseteq Args \times Args$ is a relation of strength among arguments.[1] $A \preceq B$ means that B is not weaker than A. $A \prec B =_{df} A \preceq B$ and $B \not\preceq A$.
- T is a dialectical proof theory for $(Args, \preceq)$.

Definition 3.2 [Dialectical proof theories.] A *dialectical proof theory* for a pair $(Args, \preceq)$ is a tuple (*Players, Moves, PlayerToMove, Legal, Dialogues, Winner*), where:

- $Players = \{P, O\}$. $\overline{Player} = O$ if $Player = P$, and P iff $Player = O$.
- $Moves$ is the set of all well-formed moves. As for notation and terminology, all moves are initial moves or replying moves. An *initial move* is of the form $M_1 = (Player, Arg)$, and a *replying move* is of the form $M_i = (Player, Arg, Move)$ $(i > 1)$. The first element of a move M_i is denoted by $Player(M_i)$, its second element by $Arg(M_i)$, and its third element by $Move(M_i)$. If $Move(M_i) = M_j$, we say that M_i is a *reply to*, or *replies to* M_j.

  Now the set $Moves$ is recursively defined as the smallest set satisfying the following conditions:

[1] For systems in which the ordering is itself defeasibly derived, this must be refined. However, since my framework leaves the precise use of the ordering unspecified, I shall ignore this complication.

- If $Arg \in Args$ and $Player \in Players$, then $(Player, Arg) \in Moves$.
- If $Player \in Players$, $Arg \in Args$ and $M_i \in Moves$, then $(Player, Arg, M_i) \in Moves$.

- *PlayerToMove* is a function that determines the player to move at each stage of a dialogue. Let $Pow^*(Moves)$ be the set of all finite sequences of subsets of $Moves$. Then

  $PlayerToMove$: $Pow^*(Moves) \longrightarrow Players$

  such that $PlayerToMove(D) = P$ iff D is of even length, and $PlayerToMove(D) = O$ iff D is of odd length.

- *Legal* is a function that at each point in a dialogue defines the moves that can be made at that point:

  $Legal$: $Pow^*(Moves) \longrightarrow Pow(Moves)$.

  such that

  1. $M_i \in Legal(\emptyset)$ iff M_i is an initial move;
  2. If $M_i \in Legal(M_1, \ldots, M_{i-1})$, $(i > 1)$, then M_i replies to M_{i-1}.

- *Dialogues* is the set of all sequences of moves $M_1, \ldots, M_n$ of moves such that for all i:

  1. $Player(M_i) = PlayerToMove(M_1, \ldots, M_{i-1})$,
  2. $M_i \in Legal(M_1, \ldots, M_{i-1})$.

  A member of $Dialogues$ will be called a dialogue based on F, or a T-dialogue based on $Args$ or, when T is clear from the context, simply 'based on $Args$'.

$Winner$ is a (partial) function that determines the winner of a dialogue, if any:

$Winner$: $Dialogues \longrightarrow Players$

such that player p wins a dialogue D iff $PlayerToMove(D) = \overline{p}$ and $Legal(D) = \emptyset$.

This completes the definition of a dialectical proof theory. Note that the only elements that are not completely defined are $Args$, $\preceq$ and $Legal$. So, dialectical proof theories can differ only on these three points.

I now turn to the notion of defeasible provability, which is the same for all dialectical proof theories. It is defined in terms of the notion of a strategy. A strategy for a player has the form of a tree of dialogues that for each possible move of the other player specifies a unique reply.

Definition 3.3 [Strategies.] A *strategy* for player p in a dialectical framework F is a tree of dialogues based on F only branching after p's moves, and containing all legal replies of $\overline{p}$.

It is easy to see that a winning strategy for the proponent is a strategy in which all branches end with a move by the proponent.

Defeasible provability is now defined as follows. Note first that the idea behind the *Winner* function is that if, say, P's last argument in a dialogue remains unchallenged, it also 'reinstates' all other of P's arguments in the dialogue, in particular the initial argument. Reversely, if O's last argument is unchallenged, it not only discredits P's last argument, but indirectly also all other of P's arguments, in particular the initial argument. Now P's initial argument is shown to be acceptable if in any line of attack by O, P can eventually have the last word.

Definition 3.4 [Provability.] An argument A is *defeasibly provable* in a dialectical framework F iff the proponent has a winning strategy in F with as root the move (P, A).

3.3 An illustration

To illustrate how this framework can capture dialectical proof theories, I now discuss one example, viz. Dung's [1994] dialectical proof theory for his own grounded sceptical semantics, in the (equivalent) reformulation of Prakken and Sartor [1997].

As noted above, specific dialectical proof theories are determined by a specific definition of the set $Args$, the strength relation $\preceq$ and the function $Legal$. Now, $Args$ is, as above, a set of elements with unspecified structure, while $\preceq$ is a binary relation of "attack" on $Args$ of which no properties are assumed. And for each dialogue $D = M_1, \ldots, M_{i-1}$ a move M_i is legal iff in addition to the legality conditions of Definition 3.2 it holds that

- If $Player(M_i) = P$, then $Arg(M_i)$ attacks $Arg(M_{i-1})$ while $Arg(M_{i-1})$ does not attack $Arg(M_i)$; and
- If $Player(M_i) = O$, then $Arg(M_i)$ attacks $Arg(M_{i-1})$.

Thus the burden of proof is put on the proponent: his arguments must be stronger than the opponent's counterarguments, while opponent's arguments only have to cast doubt on proponent's arguments.

Prakken and Sartor [1997] add an extra condition on legality of P's moves, viz. that P may not repeat one of his earlier arguments. This condition does not change provability of any argument (since O will have a reply the second time iff she had a reply the first time), but it avoids infinite dialogues if $Args$ is finite, which is especially convenient for actual disputes.

As shown by Dung [1994], this proof theory is sound in general, and complete for the case that each argument is attacked by at most a finite number of arguments. Prakken and Sartor [1997] generalise this result for dialogues with reasoning about the strength of arguments. Note 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.

To give an example, consider the two trees of dialogues in Figure 1. The tree on the left is based on a dialectical framework F_1 with $Args = \{A, B, C, D, E, F, G\}$ and $\preceq$ as shown by the arrows. Here P has a winning strategy, since in all dialogues O eventually runs out of moves; so argument A is provable in F_1. The tree on the right is based on an extension of F_1 into F_2 by adding H, I and J to $Args$ and adding new relations to $\preceq$ corresponding to the new arrows (the extension is shown inside the dotted box). Here P does not have a winning strategy, since one dialogue ends with a move by O; so A is not provable in F_2.

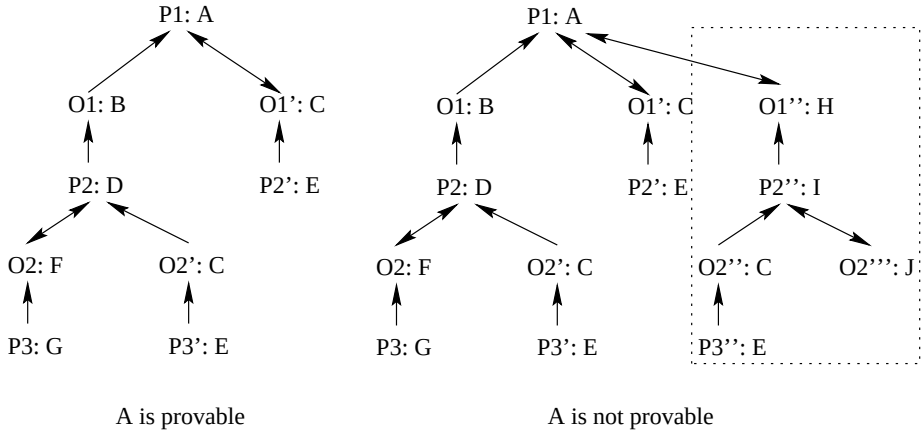

Figure 1: Two trees of proof-theoretical dialogues.

For a (partial) natural-language version of the example, recall the dispute between Paul and Olga from Section 2. Assume that the defeat relations between the various arguments are as shown in the figure by the arrows, and note that the formalisation of the individual arguments depends on the monotonic logic that underlies the argumentation system. Now let A be Paul's first argument that his car is safer because it has an airbag. B is then Olga's first counterargument based on the newspaper report with exploding airbags, and D is Paul's attack on Olga's argument based on a scientific study; note that D includes the reason why Paul thinks that D defeats B. Argument C is Olga's second attack on Paul's first argument, saying that his car is not safer since it has a higher speed limit. Then E could be a reply by Paul that a higher speed limit does not yet make a car unsafer, since drivers of fast cars can still drive slowly.

4 Protocols for dispute: general setup

The previous section abstracted the state-of-the art in dialectical proof theory, resulting in a formal framework for such theories. This section starts the discussion of the main topic of this paper, by reinterpreting this framework as the basis for protocols for dispute. The general notions of a dispute and of protocols for dispute will be defined, as well as some types of dispute.

As said in the introduction, there are two important differences between

dialectical proof theories and protocols for dispute: a focus on possible vs. on actually evolved dialogues, and a fixed vs. dynamic information base.

As for the first difference, in proof-theories our focus was only on the existence of winning strategies, i.e. on the *possibility* to win a dialogue. The actual construction of a dialectical proof (which would have the form of a *tree* of dialogues as defined in Definition 3.3) was outside our scope. In protocols for dispute, however, the focus is on evaluating disputes as they are actually evolving, so now we must also focus on how such a tree of dialogues can be constructed during a dispute. This adds a level of complexity to protocols of disputes: the legal-move function now not only specifies how to build one dialogue, but also how to build a tree of dialogues. (Actually, this must also be added to dialectical theorem provers, i.e., to methods for actually finding dialectical proofs).

Now protocols for dispute can vary in several ways (cf. Loui 1998; Vreeswijk 2000). For instance,

- Players can reply just once to the other player's moves, or may try alternative replies (*unique vs. multi-response disputes*).
- Players can make just one or may make several moves per turn (*unique- vs. multi-move disputes*).

As observed by both Loui and Vreeswijk, which protocol is best will depend on the circumstances. For instance, when a quick decision has to be reached, a unique-response dispute may be appropriate, since it forces the players to play their strongest arguments without wasting time on less promising choices. But when the quality of the outcome is more important than the time spent on it, multi-response protocols will be better.

In this paper I shall confine myself to unique-move disputes, which may be unique- or multi-response. Extending the analysis to multi-move disputes is left for future research.

The second important difference between proof-theoretical dialogues and disputes was that in the latter there is no fixed basis for discussion, since during a dispute the parties can claim new information. (I shall assume that information can only be added, not withdrawn; I leave protocols in which arguments can be withdrawn for future research.) How can we cope with this dynamic aspect of disputes? Perhaps surprisingly, I shall still assume that they are based on a fixed set *Args* of arguments. However, I shall give this set a different intended use. *Args* now does not stand for the arguments enabled by certain information, but for all arguments that could possibly be stated in a certain language. The idea is that these arguments are only relevant for the outcome if they are enabled by what has actually been said in the dispute.

We can now define the notion of a protocol for (unique-move) dispute. The first thing to note is that a protocol for dispute is parametrised by a dialectical framework F, i.e., by a given ordered set of arguments and a dialectical proof theory. The framework for disputational protocols inherits all the notions of that for dialectical proof theories, except the notion of *Dialogues*, which is replaced by the notion of *Disputes* (allowing for trees of dialogues). Furthermore, one new

notion is added, viz. *InitialArgs*, which captures the information on which the players have agreed or to which they are bound from the start of the discussion. From the retained notions, *Players*, *Moves* and *PlayerToMove* are the same as in F, but *Legal* is different, since it must now capture the various options for building a tree of dialogues; however, the *Legal* function still incorporates its counterpart in F for individual F-dialogues. Finally, the notion of a *Winner* is also defined differently, to account for both the dynamic and the 'actual' nature of disputes. However, the definition of this notion requires a separate discussion in the following section.

Below I shall, to avoid confusion, add a subscript R to the notions retained from dialectical proof theories, and add a subscript T to their proof-theoretical counterparts. I also use the following notation for identifying dialogues in a dispute.

Notation 4.1 For any sequence of moves $D = M_1, \ldots, M_n$ (where M_1 is an initial move) L_j is the dialogue (according to Definition 3.2) $M_1, \ldots, M_j$ contained in D.

Definition 4.2 [Protocols for unique-move dispute.] A *protocol for unique-move*[2] *dispute* R is a tuple (F, *InitialArgs*, $Players_R$, $Moves_R$, $PlayerToMove_R$, $Legal_R$, *Disputes*, $Winner_R$), where:

- $F = (Args, \preceq, T)$ is a dialectical framework;
- *InitialArgs* is a, possibly empty, subset of $Args$;
- $Moves_R = Moves_T$;
- $Players_R = Players_T$;
- $PlayerToMove_R = PlayerToMove_T$;
- $Legal_R$: $Pow^*(Moves_R) \longrightarrow Pow(Moves_R)$

  such that for all sequences of moves $D = M_1, \ldots, M_i$:

  1. $M_i \in Legal_R(\emptyset)$ iff M_i is an initial move;
  2. If $M_{i+1} \in Legal_R(D)$, then $Move(M_{i+1}) \in D$ and $Player(Move(M_{i+1})) = \overline{Player(M_{i+1})}$;
  3. (a) If $M_{i+1} \in Legal_T(L_i)$, then $M_{i+1} \in Legal_R(D)$;
     (b) If $M_{i+1} \in Legal_R(D)$ and $Move(M_{i+1}) = M_k$, then $M_{i+1} \in Legal_T(L_k)$.
  4. If M_{i+1} and M_j $(j < i)$ are both replies to M_k, $M_j \in D$ and $M_{i+1} \in Legal_R(M_1, \ldots, M_i)$, then $Arg(M_{i+1}) \neq Arg(M_j)$.

- *Disputes* is the set of all finite sequences of moves $M_1, \ldots, M_n$ such that for all i:

[2]Below the term 'unique-move' will be left implicit.

1. $Player_R(M_i) = PlayerToMove_R(M_1, \ldots, M_{i-1})$;
2. $M_i \in Legal_R(M_1, \ldots, M_{i-1})$.

A member of *Disputes* is called a dispute *based on* R, or an R-*dispute*. For any dispute D_i, any L_j $(j \leq i)$ is called a *dispute line* of D_i.

- $Winner_R$: $Disputes \longrightarrow Players_R$ is a (possibly partial) function that determines the winner of a dispute at each stage.

Let us look in more detail at the function $Legal_R$. The first two conditions are taken from $Legal_T$, and say that a dispute starts with an initial move, and that if a move replies to another move, the replied-to move is indeed part of the dispute so far, and was moved by the other player. Condition 3 is the crucial condition; it incorporates the proof rules of the dialectical proof theory T associated with the protocol: clause 3a says that a T-legal reply to the last move in a dispute is always legal according to R, and clause 3b says that all R-legal moves must, as a reply, be legal according to T. The final condition on $Legal_R$ forbids two subsequent replies to the same move to have the same content.

It is important to note that winning a dispute is independent of its termination, since termination is assumed to be regulated outside the protocol for dispute, by a protocol for, for instance, negotiation or deliberation. Accordingly, winning is determined relative to what has been said in a dispute up to a certain point: even if a player is winning at some point, the other player might be able to reverse the outcome by introducing new information, if the surrounding protocol allows her to do so.

To conclude this section, I give a more precise definition of two natural types of dispute by formulating further conditions on the function *Legal*.

Definition 4.3 [Unique-response and backtracking disputes.]

1. A dispute is *unique-response* iff: if $M_i \in Legal(D)$, then D contains no reply to $Move(M_i)$.

2. A dispute is *backtracking* iff: if $M_i \in Legal(D)$, then $Move(M_i)$ is an ancestor of M_i, i.e., then $Move(M_i)$ is part of the dispute line $M_1, \ldots, M_i$ of D.

In unique-response disputes, each player may reply just once to each move of the other player. Thus no tree of dialogues is created but just one dialogue. An obvious rationale for this is the existence of time limitations. However, mistakes can be fatal, in the sense that a player with a winning strategy can still lose the game if he does not follow this strategy. By contrast, backtracking disputes allow for corrections of a mistake, although in a limited way: they only allow to backtrack higher up in the 'current' dispute line (i.e. in the line ending with the last move by the other player); they do not allow 'jumps' from the current branch in the dispute tree to an earlier branch. A rationale for this restriction is that too much freedom might cause a less focussed and so more confusing debate with perhaps less quality.

5 Soundness and fairness: general observations

We now come to the main purpose of this paper, defining logically acceptable protocols for (unique-move) dispute. Note that the just-discussed protocol types are not yet complete protocols for dispute, since their legal-move functions are only partially specified, and their winning functions are not defined at all. To obtain particular protocols, these notions must be completely defined. The main problem is, given a certain definition of $Legal_R$, how to define a sound and fair notion of winning. First I make some general observations on this problem.

5.1 Which arguments are relevant

The first question is: when determining the winner of a dispute, which arguments should be taken into account? Can we simply use the winning criterion for dialectical proof theories, being that a player wins iff the other player cannot move based on $Args$? No, this is clearly inadequate, because of the different uses of the set $Args$ in proof theories and in disputes. In dialectical proof theories the idea is that this set is given by a fixed body of known information, the 'premises', and for proof-theoretic winning it should obviously be required that the premises enable no reply. However, in disputes $Args$ is given by everything that could possibly be introduced during a dispute, i.e., by the set of all well-formed formulas of the underlying language. With such a content of $Args$, it is clearly much too strong to require for winning that no reply based on $Args$ is possible. The winning criterion should restrict itself to arguments on the basis of what has actually been said at a certain stage of the dispute.

Our first task, then, is to identify 'what has been said' in a dispute. At first sight, this would simply seem to be the set of all arguments advanced in the dispute. As for notation:

Notation 5.1 For any dispute D, $Args(D)$ denotes the set of all arguments moved in D.

However, things are not that simple. As explained above in Section 3.1, the set $Args$ of a dialectical framework will often be closed under rules of inference (perhaps limited by resource bounds). However, even if this is so, a set $Args(D)$ for a dispute D is not always closed under the same rules, since the players will rarely state all arguments that are enabled by their statements in the dispute.

Must, when determining the winner, these 'implicit arguments' also be taken into account? I think that this cannot be answered in general but depends on the application. Therefore, I shall define a closure function Cl on sets of arguments, but assume only very weak properties, to allow for the whole spectrum between 'only the actually moved arguments count' via 'only those arguments count that can be computed within certain resource bounds given the joint premises of $Args(D)$', to 'all arguments count that are logically possible given the joint premises of $Args(D)$'.

Definition 5.2 [Closure of a set of arguments.] For any set T of arguments the *closure function* $Cl(T)$ returns a set of arguments such that:

- $T \subseteq Cl(T)$; and
- If $T' \supseteq T$, then $Cl(T') \supseteq Cl(T)$ (monotonicity).
- If D is a dispute based on $Args$ and $T = Args(D)$, then $Cl(T) \subseteq Args$.

A protocol for dispute in which $Cl(Args(D)) = Args(D)$ for all D is called a *protocol for dispute without computation*, otherwise it is for dispute *with computation*. Moreover, a protocol with computation is *with full computation* iff for all D, $Cl(Cl(Args(D))) = Cl(Args(D))$.

Other properties of Cl could be stated: for instance, if arguments have subarguments, closure under subarguments could be imposed. However, for present purposes this is irrelevant. Note that we cannot in general require that $Cl(Cl(T)) = Cl(T)$, since we must allow for partial computation.

Now we are ready to define 'what has been said' in a dispute. In doing so, we must also determine the role of the initial basis for discussion, captured by the set *InitialArgs*. Many disputes start with partial agreement on the basis for dispute. For instance, a scientific dispute could assume for granted certain parts of a scientific theory, or a legal dispute could be bound by certain relevant legislation or 'generally known facts'. The arguments enabled by such an initial basis must also be included in 'what has been said'. Therefore, the definition of the information base of a dispute is as follows.

Definition 5.3 [Information base of a dispute.] The information base of a dispute D, $Info(D)$, is defined as $Cl(Args(D) \cup InitialArgs)$.

I can now give a precise definition of the notions of soundness and fairness of a protocol. In fact, for fairness a caveat must be made. Since only finite disputes can be won, we cannot take provable arguments into account for which all of P's winning strategies in the relevant proof theory have an infinite number of branches. Therefore, I restrict the notion of fairness to finitely provable arguments.

Definition 5.4 An argument A is *finitely provable* in a dialectical framework F iff it is defeasibly provable in F and P's winning strategy has a finite number of dialogues.

Definition 5.5 [Soundness and fairness.] For any protocol R for dispute with dialectical framework $(Args, T)$ it holds that

1. R is *sound* iff for any R-dispute D, if D is won by P, then $Arg(M_1)$ of D is defeasibly provable in $(Info(D), T)$.
2. R is *fair* iff for any R-dispute D_i such that $Arg(M_1)$ is finitely provable in $(Info(D_i), T)$ but not in $(Info(D_{i-1}), T)$, P can win any continuation of D_i that is based on $Info(D_i)$.

Note that fairness does not require that P can win at *all* stages where his main argument is provable, but only that P can win at the *first* stage where this happens. Note also that fairness does not require that if the main argument becomes provable at a certain stage, P is already winning at that stage. To illustrate this, consider the following example.

Example 5.6 Assume a protocol incorporating Dung's [1994] sceptical proof theory (see Section 3.3), and with as winning criterion that a player wins at a certain stage if the other player cannot move with an argument from the information base at that stage. Assume next that $\mathit{InitialArgs} = \{A, B, C\}$, that as long as no new information is introduced, these are all the arguments that can be moved, and that $A \prec B$, $B \prec C$ and $A \prec C$. Now if P moves A at M_1, then A is provable in $\mathit{Info}(M_1)$ but P is not winning after M_1, since O can reply with B. All that fairness requires is that after M_1 P has a winning strategy provided that O does not introduce new information. Here this is the case, since P can reply to B with C, after which O has no moves based on $\mathit{Info}(M_1)$.

5.2 How should the relevant arguments be taken into account?

5.2.1 Trivial vs. nontrivial winning criteria

Now that we know which arguments must be taken into account, how can a logically acceptable notion of winning be defined, i.e. a winning condition that makes a protocol sound and fair? The most trivial solution is to say that the proponent (opponent) has won a dispute D iff its initial argument is (is not) defeasibly provable in $\mathit{Info}(D)$. This criterion is actually used by Gordon [1995], Loui [1998], Lodder [1999] and Brewka [1999]. Thus soundness and fairness of the protocol are made to hold by definition. However, there are reasons not to use this approach.

The first reason is computational, viz. that this definition of winning requires double work: first the dispute develops according to the dialogue rules of some proof theory (since these rules are incorporated in the protocol), and then the proof theory must be used again from scratch to compute the defeasible status of the initial argument, with a new proof-theoretic dialogue. It is more efficient to exploit the incorporation of a proof theory in protocols for dispute, by replacing the extra computation at each stage with inspection of the dispute as it has actually evolved.

A second reason is of a pragmatic nature, and especially applies to tutoring and mediation applications of dispute systems. In general, such systems might be more easily accepted by humans if they are based on natural and simple concepts. Now the problem with the 'trivial' winning condition is that it is of a technical-logical nature, and not very transparent for the human users of tutoring or mediation systems (likewise, Gordon and Karaçapilidis 1997).

Another reason is more philosophical. If in identifying the winner the structure of the actually evolved dispute is ignored and only its information base is used, why should a dispute still incorporate the legality conditions of the dialectical proof theory? Why not instead allow the players to 'shout' anything at

each other? Although in some contexts this might indeed be a sensible protocol, in many other contexts it surely is not. The underlying idea of protocols for dispute is the idea of procedural rationality, viz. the idea that regulating disputes is a way to increase their rationality and efficiency. Now the purpose of the present research is to investigate to what extent certain protocol rules, viz. the rules taken from a dialectical proof theory, contribute to this aim. It is with this purpose in mind that below I define some protocols with 'non-trivial' winning criteria, and investigate their soundness and fairness.

From this a further pragmatical reason can be derived. If the protocol leaves the strength of an argument completely unregulated, the players are not required to think about why their arguments are better than those of the other player. As a result, many disputes will have a 'flat' structure, in which it remains unclear why arguments should defeat each other. If, on the other hand, protocol rules are used which require arguments to have a certain strength, then the players are forced to think about the strength of arguments, which may enhance the quality of the discussion. (Recall that several argumentation systems, e.g. [Prakken and Sartor, 1997], allow for arguments about the strength of other arguments).

Summarising, there are good reasons to investigate 'non-trivial' winning criteria for disputes. I do not claim that such criteria are better in all applications, but they certainly deserve to be investigated.

5.2.2 A nontrivial winning criterion

Now if we want to exploit the structure of an actually evolved dispute, then an obvious candidate winning criterion is that a player wins a dispute D iff s/he made the last move, and the other player cannot move on the basis of $\mathit{Info}(D)$. Such a definition is very natural and is therefore, if sound and fair, very suitable for tutoring or mediation systems. Unfortunately, however, for several natural types of dispute, in particular for unique-move and backtracking disputes, this winning criterion is neither sound nor fair. I shall show this with some examples. In all of them it is crucial that the information base is constructed dynamically; if it were given from the start, the problems would not arise.

Example 5.7 Consider the following unique-move dispute: (As for notation, moves are shown as *Player*: *Argument*. Each move is assumed to reply to its predecessor, and the information base after each move is shown between brackets).

P_1: A $\{A\}$ $\quad\quad$ O_1: B $\{A, B\}$
P_2: C $\{A, B, C, D\}$

Note that $Cl(\{A, B, C\}) = \{A, B, C, D\}$. Assume now that D is a valid alternative reply to P_1, that $Cl(\{A, B, C, D\}) = \{A, B, C, D\}$ and that P has no reply to D on the basis of $\{A, B, C, D\}$. Then A is not defeasibly provable in $\mathit{Info}(P_1, O_1, P_2)$. However, in a unique-response dispute it is too late for O to move this reply, so our winning criterion makes P the winner, and makes such a protocol unsound.

This example illustrates the difference between (dynamic) disputes and (static) proof-theoretical dialogues: in the latter the newly enabled argument D is available from the start, so that O can move it as soon as possible, viz. immediately after P_1.

Note also that backtracking protocols do not avoid the problem: it is easy to construct examples in which the newly enabled argument is a reply to a move in an earlier line of the dispute (cf. Example 5.9 below).

The following example illustrates how an argument can be included in the information base without being explicitly moved.

Example 5.8 Consider an argumentation system in the style of e.g. [Pollock, 1992, Simari and Loui, 1992], in which first-order formulas can be combined with a metalinguistic connective $\Rightarrow$, expressing 'is a defeasible reason for', and in which arguments can be formed according to first-order logic or by chaining defeasible reasons. Then assume that the protocol for dispute is for full computation and incorporates Dung's [1994] proof theory for grounded (sceptical) semantics (cf. Section 3.3) and assume that all arguments below satisfy the strength requirements of this proof theory.

The example concerns a diagnostic problem, where the observations are supplied dynamically, but where the causal and evidential rules are agreed upon from the start (the pragmatic role of a rule is expressed with subscripts c and e). Accordingly, the set *InitialArgs* contains all rules stated in the example. Assume, finally, that alternative causal explanations of the same finding are made incompatible by the appropriate formulas in *InitialArgs*.

P_1: a, $a \Rightarrow_e b$, so b, $b \Rightarrow_c c$, so c $\quad$ O_1: a, $a \Rightarrow_e b$, so b, $b \Rightarrow_c d$, so d

P_2: e, $e \Rightarrow_c \neg d$, so $\neg d$

Assume now that *InitialArgs* also contains the evidential rule $e \Rightarrow_e \neg b$. Then the information base of this dispute enables a new reply for O against P_1, viz.

O_1': e, $e \Rightarrow_e \neg b$, so $\neg b$.

However, again a unique-move protocol does not allow this move.

Similar problems arise with fairness, as illustrated by the following example, with the same underlying logic. Actually, this example illustrates the problems for backtracking protocols and, moreover, it shows that even for disputes without computation the problems can arise.

Example 5.9 Consider the following backtracking dispute (the move to which a move replies is now indicated between parentheses):

P_1:	$\Rightarrow a$, so a.	$O_1(P_1)$:	b, $b \Rightarrow \neg a$, so $\neg a$
$P_2(O_1)$:	c, $c \Rightarrow \neg b$, so $\neg b$.	$O_2(P_2)$:	d, $d \Rightarrow \neg c$, so $\neg c$
$P_3(O_1)$:	d, $d \Rightarrow \neg b$, so $\neg b$.	$O_3(P_3)$:	e, $e \Rightarrow \neg d$, so $\neg d$

The information base of this dispute makes P_1's argument defeasibly provable, since on this basis, P could have moved at P_3 with O_3's argument. However, backtracking protocols do not allow P to jump back at P_4 to the dispute line $P_1, \ldots, O_2$, so P cannot win this dispute without introducing new information.

Summarising, problems with soundness and fairness arise if newly enabled arguments can be moved as alternative replies to earlier moves, but if the protocol does not allow this move. These problems arise because of the dynamic nature of disputes, in which new information can be supplied at any time.

To solve these problems, in principle two ways are open. The first is to define a stricter winning criterion for unique-move and backtracking disputes, and the second is to define a more liberal kind of dispute in which a newly enabled relevant reply to an earlier move is always legal. I shall follow only the second way, since the first seems less attractive. The problem is not so much with soundness as with fairness. Fairness is not about the past of a dispute but about its future, so if a protocol disallows moving a newly enabled reply to an earlier move, it is inevitably unfair. Therefore, it seems that in unique-response and backtracking disputes the only winning criterion that ensures fairness is the trivial winning condition of Section 5.2.1.

6 A sound and fair protocol: liberal disputes

6.1 Defining the protocol

I shall now define a protocol for which the nontrivial winning criterion from Section 5.2.2 is both sound and fair. We have just seen that a problem of unique-response and backtracking protocols is that they sometimes disallow relevant moves. Accordingly, one way to solve the problem is to give a natural definition of when a move is relevant: then we can simply require that every move must be relevant. So, postponing the definition of relevance for a moment, the new protocol is defined as follows.

Definition 6.1 [Liberal disputes] A protocol is for *liberal dispute* iff for any dispute D and move M it holds that $M \in Legal_R(D)$ iff

1. M satisfies the legality conditions of Definition 4.2; and
2. M is relevant in D.

Liberal disputes in a sense generalise backtracking disputes: both types of dispute jointly build a tree of dispute lines, but in liberal disputes the players can freely jump from the current to earlier branches of the tree, at least if the jump is relevant. As remarked above, this might induce less focussed disputes and therefore decrease their quality. However, this is the price to be paid for fairness.

Now, anticipating a suitable definition of relevance, we can define winning a liberal dispute D simply as the situation where the other player cannot move on the basis of $Info(D)$.

Definition 6.2 [Winning liberal disputes.] If D is a liberal dispute, then $Winner(D) = p$ iff:

1. $PlayerToMove(D) = \overline{p}$; and
2. For any $M \in Legal_R(D)$, $Arg(M) \notin Info(D)$.

The main job left to be done is defining when a move is relevant. I first illustrate this notion with an example.

Example 6.3 Figure 2 displays the dispute $D = P_1, O_1, P_2, O_2, P_3, O_3, P_4$ in tree form. Suppose that $Info(D)$ enables a move O_4 against P_3. Is O_4 relevant for the outcome of the dispute? No, it is not, since, although the dispute line $P_1, \ldots, P_3$ ends with a move by P, P has backtracked from this line and is now pursuing the alternative line P_1, O_1, P_4.

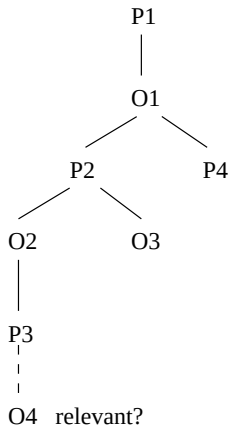

Figure 2: Relevance of moves.

How can this notion of relevance be defined? In fact, this can be done quite elegantly with a labelling function for dialectical trees that was earlier used by Gordon and Karaçapilidis [1997] and Garcia *et al.* [1998] (although they use it for different purposes, as explained below in Section 7). I shall use it for defining the 'disputational status' of all moves in a dispute D. This is the status that can be assigned to a move given only the moves contained in D, so disregarding the moves enabled by $Info(D)$ but not moved in D. The idea is very simple. In terms of a tree of moves: a node is in iff all its children are out.

Definition 6.4 [Disputational status of dispute moves] A move M of a dispute D is *in* in D iff all moves in D that reply to it are out in D. Otherwise M is *out* in D.

A move is then relevant iff it changes the status of the initial move of the dispute.

Definition 6.5 [Relevance.] A move is *relevant* in a dispute D iff it changes the disputational status of D's initial move. For any relevant move M, $Move(M)$ is called a *relevant target* for $Player(M)$.

To illustrate these definitions with the dispute of Example 6.3, consider figure 3. The dispute tree on the left is the situation after P_4. The tree in the middle shows the labelling when O has continued after P_4 with O_4, replying to P_3: this move does not affect the status of P_1, so O_4 is irrelevant. Finally, the tree on the right shows the situation where O has instead continued after P_4 with O'_4, replying to P_4: then the status of P_1 has changed, so O'_4 is relevant.

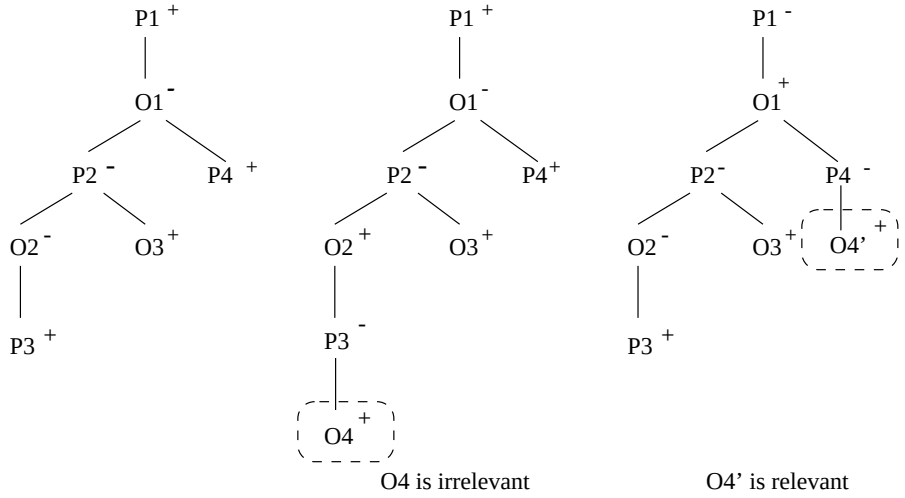

Figure 3: Disputational status of moves.

6.2 Soundness of protocols for liberal dispute

I now prove that protocols for liberal dispute are sound. The idea is that if a dispute D is won by P, the tree of dialogues contained in D can be pruned to a winning strategy for P in a proof-theoretical dialogue on the basis of $Info(D)$. This is possible since P's last move has made the first move in, so that all its replies in D are out. But then all these replies have one reply by P in D that is in, and so on. Eventually, this ends with a move by P.

To prove this, first a lemma is needed about the effects of P's and O's moves on the disputational status of the main claim.

Lemma 6.6 For any dispute $D = M_1, \ldots, M_n$,

1. If $M_{n+1} \in Legal_R(D)$, then $Move(M_{n+1})$ is in in D.
2. M_1 is in in D iff $PlayerToMove(D) = O$.

Proof: Proof of (1): obvious. Proof of (2): Trivially, P's first move makes itself in, and O's first move makes it out. Then assume that M_i, moved by O, makes M_1 out. Then clearly, any move M_{i+1} is only legal if it makes M_1 in. Reversely when M_i is moved by P, so the lemma has been proven by induction. □

Theorem 6.7 Any protocol for liberal dispute is sound.

Proof: It must be shown that if the condition holds, then P has a winning strategy in a proof-theoretical dialogue of T that starts with his first move in D. The proof exploits the definition of disputational status. The idea is that P has a strategy such that all legal replies by O are moves that are out in D, after which P can reply with a move that is in in D. Eventually, P will thus reply with a move that has no children in D, so that he has made O run out of replies.

This can be written more formally as follows. Let M_1 be the initial move of D and consider any proof-theoretical dialogue L_D based on $Info(D)$ that starts with M_1. It will be shown by induction that if P plays moves that are in in D whenever possible, any such dialogue ends with a move by P, which proves that P has a winning strategy.

As for the base case, since P has made the last move in D, by Lemma 6.6 M_1 is in in D. Then all its children in D are out in D. By Definition 4.2 (3a) it holds that among them are all moves $M_2 \in Legal_T(M_1)$. Moreover, any such M_2 has a child $M_3 \in D$ (moved by P) that is in in D. By Definition 4.2 (3b) it holds that $M_3 \in Legal_T(M_1, M_2)$. This proves that in any L_D, for any reply by O to M_1, P can reply with a move that is in in D.

In the induction step it can be assumed for any M_i moved by O that $Move(M_i)$ is in in D. Then the above line of reasoning can be repeated. Now since D is finite and all moves in L_D have to reply to each other, this repetition will eventually lead to a move by P that has no children in D. Then, since by Definition 4.2 (3a) any T-legal reply to such a move is an R-legal move, O has no T-legal reply to it in L_D on the basis of $Info(D)$. □

6.3 Fairness of protocols for liberal dispute

For fairness, it must be shown that if at a certain stage an argument becomes (finitely) defeasibly provable on the basis of what has been said in a dispute, then P can win any continuation of the dispute from that stage. In fact, I shall prove the stronger result that P can win at *any* stage where his initial argument is provable. This is possible since a liberal protocol always allows a player to recover from a mistake.

The idea of the proof of fairness is as follows. Note that if an argument is (finitely) defeasibly provable, then P has a (finite) winning strategy S for M_1 in $(Info(D), T)$. Now the idea is to show that at any point in the continuation of D, P can make a move from S (possibly by retreating from a line of D that is not in S). Then, although O might then continue with a reply not to an S-move but to a relevant target for O not in S, if P continues moving S-moves, O will at some point have exhausted all its legal non-S-moves, since all the P-moves of D not in S have been made out and have thus ceased to be relevant targets for O. But then O must after this point always reply with an S-move, and then P can continue the dispute according to S.

Of course, for this to work, it must be shown that P does not exhaust his S-moves before O exhausts her non-S-moves. This will be shown as Lemma 6.9. But first a useful lemma is proven which says that each move that is out in a

liberal dispute, has exactly one child that is in.

Lemma 6.8 All moves that are out in a liberal dispute D have exactly one child that is in in D.

Proof: By Lemma 6.6 (1), a reply M_j to a move M_i is only legal if M_i is in. But if M_i has a child that is in, M_i is out, so M_j is not legal. □

Next I prove that if P has a proof-theoretical winning strategy in $(\mathit{Info}(D), T)$, P can always use it in D.

Lemma 6.9 For any liberal dispute D and dialectical proof theory T, if P has a finite winning strategy S for M_1 in $(\mathit{Info}(D), T)$, then P can at any continuation of D that is based on $\mathit{Info}(D)$ move with a move from S.

Proof: Suppose O has made the last move in D, and consider the subdispute D^* of D in which in all lines L, all moves have been omitted from the first P-move in L that is not in S. Since M_1 was made out by O's last move, D^* contains at least one line ending with a move by O. It must be shown that any such O-move is a relevant target for P. Since by its construction D^* does not branch after O-moves and since its root is out in D, in all such lines, all P-moves are out and all O-moves are in. Then by Lemma 6.8, no O-move in any such line has a sibbling in D that is in. But then all such O-moves are relevant targets for P in D. Observe now that any legal reply from S to such moves is legal in D by Definition 4.2 (3a). It is left to show that P can reply to at least one such move with a move in S. Assume for contradiction that this is not the case (i.e., all possible P-replies from S are already in D^*). Then, since S contains all possible O-replies to any P-move in S, and since by construction of D^* all P-moves in D^* are in S, S contains a subtree of which not all branches end with a P-move. But then S is not a winning strategy for P. Contradiction. □

Now the main theorem can be proven.

Theorem 6.10 Any protocol for liberal dispute is fair.

Proof: Suppose D's first argument is finitely provable in $(\mathit{Info}(D), T)$, let S and D^* be defined as in the proof of Lemma 6.9 and consider any continuation D' of D that is based on $\mathit{Info}(D)$. By Lemma 6.9, P can at any point in D' make a move from S. It must be shown that there comes a point after which O can only play moves from S, after which P can win by following S, which is a winning strategy.

Consider for any stage in D' the set S^- of relevant targets for O that are not in S. If this set is empty, we are done. Otherwise, it needs to be shown that at any following stage, there are fewer moves in S^- to which O can reply, so that there comes a stage at which S^- is empty. The idea of the proof is that P can always continue with a move from S, and that no such move makes any P move from S^- in.

More precisely, consider any P-move $M_j \in S^-$ and assume that O replies to it with $M_i \in D' - D$. Then M_j is made out by M_i. By Lemma 6.9 P can continue at M_{i+1} with at least one move from S. Consider any such $M_{i+1} \in S$. M_{i+1} does not make M_j in, since M_{i+1} replies to an O-move in D^* and since all dispute lines in D' are cut off in D^* at M_j. But then M_j is not a relevant target for O at $i+2$, which shows that S^- becomes smaller with any move by O. But then by finiteness of D there comes a point that S^- is empty, after which O must reply to a move in S, and P can win by following S. □

7 Related research

Probably the first detailed study of protocols for dispute was carried out by Loui [1998]. The present framework for disputes is inspired by Loui's work. However, there are also differences.

A difference in focus is that Loui is mainly interested in partial computation from a static information base, while I am interested in any form of computation from a 'truly' dynamic information base. In other words, in Loui's framework dynamics is caused by partial computation, while in my approach it is caused by supply of new information.

There are also differences in formalisation. Some of them are simplifications, which I have made because of the difference in focus, or in order not to distract from the main points of this paper. This concerns, for instance, Loui's study of multi-move disputes, his explicit notion of *resources*, and his study of the case where players have partial information about the dispute.

However, one difference reflects disagreement on how protocols for dispute should be designed. This concerns the fact that I have no exact counterpart of Loui's function *current.opinion*. This function determines for each stage of a dispute the logical status of the proponent's main claim, in terms of an assumed underlying logic for defeasible argumentation. In particular, this logic is applied to everything that has been said in the dispute up to that stage. Loui's main use of the function *current.opinion* is a protocol rule, adopted in most of his protocols, that each move must alter current opinion. Although the intuition behind this rule is correct, Loui's formalisation of it in terms of *current.opinion* has some problems. One of them is that this notion actually is the 'trivial' winning condition criticised above in Section 5.2. This also explains why Loui does not study the issues of soundness and fairness in the way they are defined in the present article.

In Section 6 I have in fact given an alternative formalisation of Loui's intuition that each move must change current opinion: the *current.opinion* function should not stand for defeasible provability in the underlying logic, but it should capture the disputational status of a move given the moves that have been made in a dispute. The examples of Section 5.2 have shown that this disputational status of a move does not always coincide with the proof-theoretical status of its argument. We have seen that the notion of disputational status can very well be used for capturing relevance of moves, without referring to defeasible

provability.

It is interesting to note, however, that the above definition of disputational status can also be used for defining the notion of a defeasible proof (as is done by Gordon and Karaçapilidis [1997] and Garcia *et al.* [1998]). However, this is only possible if a dialogue tree contains all possible moves of the opponent, i.e., all moves that can be made on the basis of the input information. And we have seen that in the context of disputes this condition does not always hold: for $Args$ it rarely holds, and even for $Info(D)$ of a dispute D it does not always hold.

Vreeswijk (e.g. 1995, 2000) has also formalised protocols for dispute. He assumes a fixed basis for discussion, which makes his work essentially a study of dialectical proof procedures. However, in [Vreeswijk, 2000], he also studies 'self-modifying' protocols, which allow arguing within the protocol about whether to change it. This topic is potentially very relevant for multi-agent and dispute-mediation applications.

Other work on frameworks for (static) dialectical proof theories has been done by Jakobovits and Vermeir [1999], who have developed such a framework for Dung-style argumentation systems [Dung, 1995], and applied it to two semantics for such systems. Their framework contains counterparts of my functions $PlayerToMove_T$ and $Legal_T$.

In the work discussed so far, the information base of a dispute is static. However, there also is work in which the information base can change due to supply of new information into the dispute.

In the field of Artificial Intelligence and Law, several researchers have proposed disputational models of legal procedure with this feature [Hage *et al.*, 1994, Gordon, 1995, Bench-Capon, 1998, Lodder, 1999]. This is not surprising, since the law is a prime example of a domain where the dynamic establishment of a basis for discussion (and then decision) is regulated by protocols for dispute. These AI & Law models incorporate the possibility of counterargument in what are essentially Hamblin-MacKenzie-style protocols for persuasion. However, the relation with dialectical proof theory is not investigated. Hage *et al.*, Gordon and Lodder define winning as Loui does, viz. in the 'trivial' way that refers to an unspecified nonmonotonic logic, while Bench-Capon does not refer to an underlying logic at all. Gordon also pays much attention to defining when moves are relevant. However, since he does so for a multi-move unique-response protocol, the problem is not the same as in the present framework.

Finally, Brewka [1999] has developed a formal model for argumentation processes that combines nonmonotonic logic with protocols for dispute. His framework is complementary to the present one. He does not study the details of arguing and counterarguing but simply applies an unspecified nonmonotonic logic to everything that has been said in a dispute (thus using the 'trivial' winning criterion). On the other hand, Brewka pays more attention to the speech act aspects of disputes: in order to capture the effects of speech acts, he formalises disputational protocols in situation calculus (in the form of [Reiter, 1999]). Such a logical formalisation of protocols allows him to define protocols in which the legality of a move can be disputed. A final difference with the present work is

that Brewka admits elements from deliberation in his model, by allowing the players to be any type of actor, including referees or 'determiners'. By contrast, my approach was to define separate protocols for distinct types of dialogue with distinct initial situations and goals, and then to study their combination.

8 Conclusion

The contribution of this paper has been threefold. Firstly, it has provided general frameworks for dialectical proof theories and disputational protocols, extend and adapting earlier work of others. Secondly, it has resulted in logical foundations for dynamic protocols for dispute; to my knowledge this is the first contribution of this kind where the dynamic aspects of disputes taken into account. Finally, since the formal study of protocols for dynamic dispute is not yet wide-spread, my specification of such protocols has also contributed to the study of their design.

The results can be summarised in more detail as follows.

- I have shown how (static) dialectical proof theories for defeasible reasoning can be reinterpreted as parts of (dynamic) protocols for dispute.
- I have given a precise definition of two desirable properties of dynamic protocols for dispute, viz. soundness and fairness.
- It has turned out that some natural protocols for dispute that are sound and fair with a static information base, can lose these properties in a dynamic setting.
- I have developed another protocol that preserves soundness and fairness in a dynamic setting.

It should be noted that these results are quite general, applying to a wide range of dialectical systems and protocols. This has been achieved in two ways: by parametrising protocols for dispute with dialectical proof theories, and by using an unspecified closure function on the arguments advanced in a dispute.

In future research the above analysis should be extended to multi-move disputes. It would also be interesting to study whether restrictions on the language of arguments or on their attack relations can ensure soundness and fairness for unique-response and backtracking disputes. And it is important to study the case where information can not only be added but also withdrawn. Such a study would pave the way for combining disputational protocols with Hamblin-MacKenzie-style protocols for persuasion. Finally, the ultimate goal of this research is to combine protocols for dispute and persuasion with protocols for other types of dialogue, such as negotiation and group decision making.

References

[Aleven and Ashley, 1997] V. Aleven and K.D. Ashley. Evaluating a learning environment for case-based argumentation skills. In *Proceedings of the Sixth*

International Conference on Artificial Intelligence and Law, pages 170–179, New York, 1997. ACM Press.

[Bench-Capon, 1998] T.J.M. Bench-Capon. Specification and implementation of Toulmin dialogue game. In *Legal Knowledge-Based Systems. JURIX: The Eleventh Conference*, pages 5–19, Nijmegen, 1998. Gerard Noodt Instituut.

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

[Brewka, 1999] G. Brewka. Dynamic argument systems: a formal model of argumentation processes based on situation calculus. *Journal of Logic and Computation*, 1999. To appear.

[Dung, 1994] P.M. Dung. Logic programming as dialog game. Unpublished paper, Division of Computer Science, Asian Institute of Technology, Bangkok, 1994.

[Dung, 1995] P.M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n–person games. *Artificial Intelligence*, 77:321–357, 1995.

[Garcia *et al.*, 1998] A.J. Garcia, G.R. Simari, and C.I. Chesñevar. An argumentative framework for reasoning with inconsistent and incomplete information. In *Proceedings of the ECAI'98 Workshop on Practical Reasoning and Rationality*, Brighton, UK, 1998.

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

[Gordon *et al.*, 1997] T.F. Gordon, N. Karaçapilidis, H. Voss, and A. Zauke. Computer-mediated cooperative spatial planning. In H. Timmermans, editor, *Decision Support Systems in Urban Planning*, pages 299–309. E & FN SPON Publishers, London, 1997.

[Gordon, 1995] T.F. Gordon. *The Pleadings Game. An Artificial Intelligence Model of Procedural Justice.* Kluwer Academic Publishers, Dordrecht/Boston/London, 1995.

[Hage *et al.*, 1994] J.C. Hage, R.E. Leenes, and A.R. Lodder. Hard cases: a procedural approach. *Artificial Intelligence and Law*, 2:113–166, 1994.

[Hamblin, 1971] C.L. Hamblin. Mathematical models of dialogue. *Theoria*, 37:130–155, 1971.

[Hintikka, 1999] J. Hintikka. Is logic the key to all good reasoning? In *Inquiry as Inquiry: a Logic of Scientific Discovery*, volume 5 of *Jaakko Hintikka Selected*

Papers, pages 1–24. Kluwer Academic Publishers, Dordrecht/Boston/London, 1999.

[Jakobovits and Vermeir, 1999] H. Jakobovits and D. Vermeir. Dialectic semantics for argumentation frameworks. In *Proceedings of the Seventh International Conference on Artificial Intelligence and Law*, pages 53–62, New York, 1999. ACM Press.

[Kraus *et al.*, 1998] S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104:1–69, 1998.

[Lodder, 1999] A.R. Lodder. *DiaLaw. On Legal Justification and Dialogical Models of Argumentation.* Kluwer Academic Publishers, Dordrecht/Boston/London, 1999.

[Loui, 1994] R.P. Loui. Argument and arbitration games. In *Working notes of the AAAI-94 workshop on Computational Dialectics*, 1994.

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

[MacKenzie, 1990] J.D. MacKenzie. Four dialogue systems. *Studia Logica*, 51:567–583, 1990.

[Parsons *et al.*, 1998] S. Parsons, C. Sierra, and N.R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8:261–292, 1998.

[Pollock, 1992] J.L. Pollock. How to reason defeasibly. *Artificial Intelligence*, 57:1–42, 1992.

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

[Prakken and Vreeswijk, 2000] H. Prakken and G.A.W. Vreeswijk. Logical systems for defeasible argumentation. In D. Gabbay, editor, *Handbook of Philosophical Logic*. Kluwer Academic Publishers, Dordrecht/Boston/London, 2000. Second edition, to appear.

[Reiter, 1999] R. Reiter. *Knowledge in Action: Logical Foundations for Describing and Implementing Dynamical Systems.* 1999. Book draft, available at http://www.cs.toronto.edu/cogrobo/.

[Simari and Loui, 1992] G.R. Simari and R.P. Loui. A mathematical treatment of defeasible argumentation and its implementation. *Artificial Intelligence*, 53:125–157, 1992.

[Suthers *et al.*, 1995] D. Suthers, A. Weiner, J. Connelly, and M. Paolucci. Belvedere: engaging students in critical discussion of science and public policy issues. In *Proceedings of the Seventh World Conference on Artificial Intelligence in Education*, pages 266–273, 1995.

[Vreeswijk, 1995] G.A.W. Vreeswijk. The computational value of debate in defeasible reasoning. *Argumentation*, 9:305–341, 1995.

[Vreeswijk, 2000] G.A.W. Vreeswijk. Representation of formal dispute with a standing order. *Artificial Intelligence and Law*, 8(2), 2000. To appear.

[Walton and Krabbe, 1995] D.N. Walton and E.C.W. Krabbe. *Commitment in Dialogue. Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, 1995.

[Walton, 1990] D. Walton. What is reasoning? What is an argument? *Journal of Philosophy*, 87:399–419, 1990.

[Walton, 1999] D. Walton. Applying labelled deductive systems and multi-agent systems to source-based argumentation. *Journal of Logic and Computation*, 9:63–80, 1999.