Some preliminary steps towards a meta-theory for formal inter-agent dialogues

Simon Parsons[1,2], Peter McBurney[2], Michael Wooldridge[2]

[1] Department of Computer and Information Science, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, New York, NY 11210, USA
parsons@sci.brooklyn.cuny.edu

[2] Department of Computer Science, University of Liverpool, Chadwick Building, Liverpool L69 7ZF, UK.
{p.j.mcburney,m.j.wooldridge}@csc.liv.ac.uk

Abstract. This paper investigates the properties of argumentation-based dialogues between agents. It takes a previously defined system by which agents can trade arguments, and examines how different classes of protocols for this kind of interaction can have profoundly different outcomes. Studying such classes of protocol, rather than individual protocols as has been done previously, allows us to start to develop a *meta-theory* of this class of interactions.

1 Introduction

Research into the theoretical properties of protocols for multi-agent interaction can be crudely divided into two camps. The first camp is broadly characterised by the application of game and economic theory to understanding the properties of multi-agent protocols; this camp includes, for example, research on auction protocols and algorithmic mechanism design. The second camp may be broadly characterised by an understanding of agents as practical reasoning systems, which interact in order to to resolve differences of opinion and conflicts of interest; to work together to resolve dilemmas or find proofs; or simply to inform each other of pertinent facts. As work in the former camp has been informed by game and economic theory, so work in this latter camp has been informed in particular by research in the area of *argumentation* and *dialogue games*. Examples of argumentation-based approaches to multi-agent dialogues include the work of Dignum *et al.* [4], Kraus [12], Reed [19], Schroeder *et al.* [20] and Sycara [21].

The work of Walton and Krabbe has been particularly influential in argumentation-based dialogue research [22]. They developed a typology for inter-personal dialogue which identifies six primary types of dialogues and three mixed types. The categorization is based upon: what information the participants each have at the commencement of the dialogue (with regard to the topic of discussion); what goals the individual participants have; and what goals are shared by the participants, goals we may view as those of the dialogue itself. This *dialogue game* view of dialogues overlaps with work on conversation policies (see, for example, [3, 6]), but differs in considering the entire dialogue rather than dialogue segments. As defined by Walton and Krabbe, the three types of dialogue we have considered in our previous work are: *Information-Seeking Dialogues*

(where one participant seeks the answer to some question(s) from another participant, who is believed by the first to know the answer(s)); *Inquiry Dialogues* (where the participants collaborate to answer some question or questions whose answers are not known to any one participant); and *Persuasion Dialogues* (where one party seeks to persuade another party to adopt a belief or point-of-view he or she does not currently hold). Persuasion dialogues begin with one party supporting a particular statement which the other party to the dialogue does not, and the first seeks to convince the second to adopt the proposition. The second party may not share this objective.

Our previous work investigated capturing these types of dialogue using a formal model of argumentation [2], and the basic properties and complexity of such dialogues [15]. Most recently, we have looked at how the outcomes of these dialogues can depend upon the order in which agents make utterances [16]. Here we extend this investigation, by moving from the study of particular protocols to the study of *classes of protocols*, and the properties of those classes. These results, then, are (very preliminary) results about the *meta-theory* of argumentation-based dialogues. The advantage of this change in perspective is that our results are robust—they hold for a wider range of possible dialogues—and more wide-reaching that we have been able to obtain hitherto, permitting a more complete analysis of argumentation-based dialogues. Note that, despite the fact that the types of dialogue we are considering are drawn from the analysis of human dialogues, we are only concerned here with dialogues between *artificial* agents. Unlike Grosz and Sidner [10] for example, we choose to focus in this way in order to simplify our task—dealing with artificial languages avoids much of the complexity of natural language dialogues.

2 Background

In this section, we briefly introduce the formal system of argumentation that underpins our approach [1], a system that extends Dung [5] with preferences. We start with a (possibly inconsistent) knowledge base Σ with no deductive closure. We assume Σ contains formulas of a propositional language $\mathcal{L}$, that $\vdash$ stands is the classical inference relation, and $\equiv$ stands for logical equivalence. An argument is a proposition and the set of formulae from which it can be inferred:

Definition 1. An *argument* is a pair $A = (H, h)$ where h is a formula of $\mathcal{L}$ and H a subset of Σ such that:
1. H is consistent;
2. $H \vdash h$; and
3. H is minimal, so no proper subset of H satisfying both (1) and (2) exists.

H is called the *support* of A, written H = Support(A) and h is the *conclusion* of A, written h = Conclusion(A).

We thus talk of h being *supported* by the argument (H, h)

In general, since Σ is inconsistent, arguments in $\mathcal{A}(\Sigma)$, the set of all arguments which can be made from Σ, will conflict, and we make this idea precise with the notion of *undercutting*:

Definition 2. Let A_1 and A_2 be two arguments of $\mathcal{A}(\Sigma)$. A_1 *undercuts* A_2 iff $\exists h \in$ *Support*(A_2) such that $h \equiv \neg$*Conclusion*(A_1).

In other words, an argument is undercut iff there is another argument which has as its conclusion the negation of an element of the support for the first argument.

To capture the fact that some facts are more strongly believed than others, we assume that any set of facts has a preference order over it. We suppose that this ordering derives from the fact that the knowledge base Σ is stratified into non-overlapping sets $\Sigma_1, \ldots, \Sigma_n$ such that facts in Σ_i are all equally preferred, and are more preferred than those in Σ_j where $j > i$. The preference level of a nonempty subset H of Σ, *level*(H), is the number of the highest numbered layer which has a member in H.

Definition 3. Let A_1 and A_2 be two arguments in $\mathcal{A}(\Sigma)$. A_1 is *preferred* to A_2 according to *Pref*, *Pref*(A_1, A_2), iff *level*(*Support*(A_1)) $\leq$ *level*(*Support*(A_2)).

By $\gg^{Pref}$, we denote the strict pre-order associated with *Pref*. If A_1 is preferred to A_2, we say that A_1 is *stronger* than A_2[1]. We can now define the argumentation system we will use:

Definition 4. An *argumentation system* (AS) is a triple $\langle \mathcal{A}(\Sigma), \mathit{Undercut}, \mathit{Pref} \rangle$ such that:
• $\mathcal{A}(\Sigma)$ is a set of the arguments built from Σ,
• *Undercut* is a binary relation representing the defeat relationship between arguments, *Undercut* $\subseteq \mathcal{A}(\Sigma) \times \mathcal{A}(\Sigma)$, and
• *Pref* is a (partial or complete) preordering on $\mathcal{A}(\Sigma) \times \mathcal{A}(\Sigma)$.

The preference order makes it possible to distinguish different types of relation between arguments:

Definition 5. Let A_1, A_2 be two arguments of $\mathcal{A}(\Sigma)$.
• If A_2 undercuts A_1 then A_1 *defends itself* against A_2 iff $A_1 \gg^{Pref} A_2$. Otherwise, A_1 *does not defend itself.*
• A set of arguments $\mathcal{S}$ *defends* A iff: $\forall$ B undercuts A and A does not defend itself against B then $\exists$ $C \in \mathcal{S}$ such that C undercuts B and B does not defend itself against C.

We write $C_{\mathit{Undercut},\mathit{Pref}}$ to denote the set of all non-undercut arguments and arguments defending themselves against all their undercutting arguments. The set $\underline{\mathcal{S}}$ of acceptable arguments of the argumentation system $\langle \mathcal{A}(\Sigma), \mathit{Undercut}, \mathit{Pref} \rangle$ is the least fixpoint of a function $\mathcal{F}$ [1]:

$$\mathcal{S} \subseteq \mathcal{A}(\Sigma)$$
$$\mathcal{F}(\mathcal{S}) = \{(H,h) \in \mathcal{A}(\Sigma) \mid (H,h) \text{ is defended by } \mathcal{S}\}$$

Definition 6. The set of *acceptable* arguments for an argumentation system $\langle \mathcal{A}(\Sigma), \mathit{Undercut}, \mathit{Pref} \rangle$ is:

$$\begin{aligned} \underline{\mathcal{S}} &= \bigcup \mathcal{F}_{i \geq 0}(\emptyset) \\ &= C_{\mathit{Undercut},\mathit{Pref}} \cup \left[\bigcup \mathcal{F}_{i \geq 1}(C_{\mathit{Undercut},\mathit{Pref}}) \right] \end{aligned}$$

[1] We acknowledge that this model of preferences is rather restrictive and in the future intend to work to relax it.

An argument is *acceptable* if it is a member of the acceptable set, and a proposition is *acceptable* if it is the conclusion of an acceptable argument.

Definition 7. If an agent A has an acceptable argument for a proposition p, then the *status* of p for that agent is *accepted*, while if the agent does not have an acceptable argument for p, the status of p for that agent is *not accepted*.

An acceptable argument is one which is, in some sense, proven since all the arguments which might undermine it are themselves undermined.

3 Locutions and attitudes

As in our previous work, agents put forward propositions and accept propositions put forward by other agents based on their acceptability. The exact locutions and the way that these locutions are exchanged define a formal *dialogue game* which agents engage in.

Dialogues are assumed to take place between two agents, for example called P (for "pro") and C ("con"). Each agent $i \in \{P, C\}$ has a knowledge base, Σ_i, containing its beliefs. In addition, each agent i has a further knowledge base CS_i, visible to both agents, containing *commitments* made in the dialogue. We assume an agent's *commitment store* is a subset of its knowledge base. Note that the union of the commitment stores can be viewed as the state of the dialogue at a given time. Since each agent has access to their private knowledge base and both commitment stores, agent i can make use of $\langle \mathcal{A}(\Sigma_i \cup CS(j)), Undercut, Pref \rangle$ where $i, j \in \{P, C\}$ and $i \neq j$.

All the knowledge bases contain propositional formulas and are not (necessarily) closed under deduction, and moreover all are stratified by degree of belief as discussed above. Here we assume that these degrees of belief are static and that both the players agree on them (acknowledging that this is a limitation of this approach).

With this background, we can present a set of dialogue moves, based on those first introduced in [15], and then modified in [14]. Each locution has a rule describing how to update commitment stores after the move, and groups of moves have conditions under which the move can be made—these are given in terms of the agents' assertion and acceptance attitudes (defined below). For all moves, player P addresses the ith move of the dialogue to player C.

assert(p) where p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \cup \{p\} \text{ and } CS_i(C) = CS_{i-1}(C)$$

Here p can be any propositional formula, as well as the special character $\mathcal{U}$, discussed below. This makes a statement that the agent is prepared to back up with an argument.

assert(S) where S is a set of formulas representing the support of an argument.

$$CS_i(P) = CS(P)_{i-1} \cup S \text{ and } CS_i(C) = CS_{i-1}(C)$$

accept(p) p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \cup \{p\} \text{ and } CS_i(C) = CS_{i-1}(C)$$

This explicitly notes that P agrees with something previously stated by C.

reject(p) p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \text{ and } CS_i(C) = CS_{i-1}(C)$$

This explicitly notes that P disagrees with something previously stated by C.

challenge(p) where p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \text{ and } CS_i(C) = CS_{i-1}(C)$$

A challenge is a means of making the other player explicitly state the argument supporting a proposition that they have previously asserted[2]. In contrast, a question can be used to query the other player about any proposition.

question(p) where p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \text{ and } CS_i(C) = CS_{i-1}(C)$$

question is used to start an information-seeking dialogue. The last two locutions are used to start particular types of dialogue [14]:

know(p) where p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \text{ and } CS_i(C) = CS_{i-1}(C)$$

know(p) is a statement akin to "do you know that p is true", which kicks off a persuasion dialogue.

prove(p) where p is a propositional formula.

$$CS_i(P) = CS_{i-1}(P) \text{ and } CS_i(C) = CS_{i-1}(C)$$

prove(p) is an invitation to start an inquiry dialogue to prove whether p is true or not. This is the set of moves, $\mathcal{M}_{DC}^{PK}$ from [14], an expansion of those in [15] that allows for more elegant dialogues[3].

The way in which these locutions are used will be determined by the protocol used (examples of which are given below) and the *attitudes* which control the assertion and acceptance of propositions. Following our previous investigation [15, 16], we deal with

[2] In this system it is only possible to issue a challenge for a proposition p following an *assert*(p) by the other agent.

[3] The locutions in $\mathcal{M}_{DC}^{PK}$ are similar to those discussed elsewhere, for example [7, 18], though there is no *retract* locution

"thoughtful/skeptical" agents that can assert any proposition p for which they can construct an acceptable argument, and will accept any proposition p for which they can construct an acceptable argument. Whatever the protocol, no agent is allowed to repeat exactly the same locution (down to the proposition or propositions that instantiate it) without immediately terminating the dialogue.

We refer to the system described here as $\mathcal{DG}$, irrespective of the protocol that controls the exchange of locutions.

4 Types of dialogues

Previously [15], we defined three basic protocols for information seeking, inquiry and persuasion dialogues. These were subsequently updated in [14], and despite their apparent simplicity, have proved to be theoretically very rich.

4.1 Information-seeking

The following protocol, denoted $\mathcal{IS}$, is unchanged from [15] and captures basic information seeking:

1. A asks *question*(p).
2. Depending upon the contents of its knowledge-base and its assertion attitude, B replies with either *assert*(p), *assert*$(\neg p)$, or *assert*$(\mathcal{U})$, where $\mathcal{U}$ indicates that, for whatever reason, B cannot give an answer.
3. A either *accept*s B's response, if its acceptance attitude allows, or *challenge*s. $\mathcal{U}$ cannot be *challenge*d, and as soon as it is asserted, the dialogue terminates without the question being resolved.
4. B replies to a *challenge* with an *assert*(S), where S is the support of an argument for the last proposition challenged by A.
5. Go to (3) for each proposition in S in turn.

When the dialogue terminates with A *accept*ing the subject of the dialogue, the dialogue is said to be *successful*.

Note that A *accept*s whenever possible, only being able to *challenge* when unable to *accept*.

4.2 Inquiry

The inquiry protocol $\mathcal{I}''$ from [14] is:

1. B proffers *prove*(p), inviting A to join it in the search for a proof of p.
2. A asserts $q \rightarrow p$ for some q or $\mathcal{U}$.
3. B accepts $q \rightarrow p$ if its acceptance attitude allows, or *challenge*s it.
4. A replies to a *challenge* with an *assert*(S), where S is the support of an argument for the last proposition challenged by B.
5. Go to (2) for each proposition $s \in S$ in turn, replacing $q \rightarrow p$ by s.
6. B *assert*s q, or $r \rightarrow q$ for some r, or $\mathcal{U}$.

7. If $\mathcal{A}(CS(A) \cup CS(B))$ includes an argument for p that is acceptable to both agents, then first A and then B *accept* it and the dialogue terminates successfully.
8. If at any point one of the propositions is not acceptable to an agent, it issues a *reject*, and the dialogue ends unsuccessfully.
9. Go to 6, reversing the roles of A and B and substituting r for q and some t for r.

This protocol has some core steps in common with $\mathcal{IS}$ dialogues, and we discuss these below.

4.3 Persuasion

The persuasion protocol $\mathcal{P}'$ from [14] is:

1. A issues a $know(p)$, indicating it believes that p is the case.
2. A *assert*s p.
3. B *accept*s p if its acceptance attitude allows, else B either *assert*s $\neg p$ if it is allowed to, or else *challenge*s p.
4. If B asserts $\neg p$, then go to (2) with the roles of the agents reversed and $\neg p$ in place of p.
5. If B has *challenge*d, then:
   (a) A asserts S, the support for p;
   (b) Go to (2) for each $s \in S$ in turn.
6. If B does not *challenge*, then it issues either $accept(p)$ or $reject(p)$, depending upon the status of p for it.

Note that this kind of persuasion dialogue does not assume that agents necessarily start from opposite positions, one believing p and one believing $\neg p$. Instead one agent believes p and the other may believe $\neg p$, but also may believe neither p nor $\neg p$. This is perfectly consistent with the notion of persuasion suggested by Walton and Krabbe [22].

Protocols $\mathcal{IS}$, $\mathcal{I}''$, and $\mathcal{P}'$ define a range of possible sequences of locutions, and we call these sequences *dialogues* (the relationship between the two is exploreed more in [14]). Here a protocol is a blueprint for many different dialogues, depending on the beliefs of the agents who use the protocol. We will refer to any dialogue under the X protocol as an "X dialogue".

5 Classes of protocol

We have previously [15, 16] studied the properties of these three individual protocols. Here we extend this work, investigating whether there are properties, especially properties related to the outomes of dialogues under these protocols, that are determined by the structure of the dialogues.

A: *question*(p)
B: *assert*(p)
A: *challenge*(p)
B: *assert* $\left(\bigcup_i \{s_i\}_{i=1\ldots n}\right)$
A: *challenge*(s_1)
B: *assert*($\{s_1\}$)
A: *accept*(s_1)
A: *challenge*(s_2)
B: *assert*($\{s_2\}$)
A: *accept*(s_2)
$\vdots$
A: *challenge*(s_n)
B: *assert*($\{s_n\}$)
A: *accept*(s_n)
A: *accept* $\left(\bigcup_i \{s_i\}_{i=1\ldots n}\right)$
A: *accept*(p)

Fig. 1. An example information-seeking dialogue

5.1 The general shape of dialogues

.

We start by considering the structure of an $\mathcal{IS}$ dialogue, the general form of which will be as in Figure 1. The dialogue is written to emphasize that one way to think of it is as a set of sub-dialogues. There is an outer dialogue of three locutions, inside that there is another 3 locution dialogue, which in turn has a sequence of three-locution dialogues inside it. Looking at the other kinds of dialogue defined above reveals that they not only do they have a similar structure [14], but that the sub-dialogues they contain have the same structure. We can exploit this structure to obtain general results about dialogues constructed in this way.

We can consider the repeated sub-dialogue in Figure 1 to be an *atomic protocol*[4], which, along with some additional ones identified in [14] (along with a set of rules for combining them) are sufficient to construct the protocols given above. These are similar in concept to conversation policies [8], being fragments from which a dialogue can be created. The atomic protocol distilled from the repeated sub-dialogue in Figure 1 we call A. This starts following an *assert*(X) and runs:

A: *challenge*(X)
B: *assert*(Y)
A: *accept*(X) or *reject*(X)

where X and Y are variables, and Y is the support for whatever proposition instantiates X. By analogy with the $\mathcal{IS}$ dialogue, we say that an A dialogue is *successful* if it concludes with an *accept*.

[4] In the sense that it cannot be broken down further and yield a recognisable protocol.

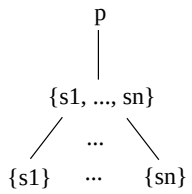

Fig. 2. An A dialogue.

Additional A dialogues may be nested inside the dialogue generated by this protocol, and typically we will have a series of such dialogues after the *assert* (just as in Figure 1). This corresponds to the construction of a *proof tree* for X. Thus if the X is instantiated with p and Y with $S = \{s_1, \ldots s_n\}$, then the proof tree unfolded by the instance of A above, and subsequent A dialogues about each s_i will build the proof tree in Figure 2. This figure denotes that the set $\{s_1, \ldots, s_n\}$ is the set of grounds for p, and that each s_i has a set of grounds $\{s_i\}$.

Definition 8. The *subject* of a dialogue is p iff the first locution in the dialogue concerns p.

Definition 9. Consider two dialogues D and E. D is said to be *embedded* in E if the sequence of locutions that make up D is a subsequence of those that make up E.

Definition 10. Consider two dialogues D and E. D is said to be *directly embedded* in E if D is embedded in E and there is no dialogue F such that D is embedded in F and F is embedded in D.

If D is embedded in E but is not directly embedded in E, then there are one or more *intermediate* dialogues F, such that D is embedded in F and F is embedded in E. In such a case every F is said to be *between* D and E. In Figure 1, the dialogue:

A: *challenge*(s_1)
B: *assert*$(\{s_1\})$
A: *accept*(s_1)

is embedded in the dialogue:

A: *question*(p)
B: *assert*(p)
$\vdots$
A: *accept*(p)

and directly embedded in the A dialogue:

A: *challenge*(p)
B: *assert* $\left(\bigcup_i \{s_i\}_{i=1\ldots n}\right)$

$\vdots$

A: *accept* $\left(\bigcup_i \{s_i\}_{i=1\ldots n}\right)$

If both D and E are carried out under A then the only reasonable ways to embed D in E is to have D follow the *assert* in E, or to follow another dialogue F that is already embedded in E.

Definition 11. Consider two dialogues D and E, where D is directly embedded in E. If E has a *level of embedding* of n, then D has a level of embedding of $n+1$. A dialogue that is not embedded in another has a level of embedding of 0.

We can then show:

Proposition 12. *If E is an A dialogue with subject p and a level of embedding n, and D is an A dialogue embedded in E such that all intermediate dialogues between D and E are A dialogues, then the maximum level of embedding of D is $n+1$.*

Proof. The maximum level of embedding will occur when dialogues are nested as deeply within one another as possible, so we proceed by constructing the deepest possible nesting. If E has subject p, then the second locution of E will be the assertion of the grounds for p. This will be some set of propositions S which are a subset of the knowledge base of the agent replying to the assertion (by definition). Each member of this set can then be challenged by a new dialogue D_i with subject $s_i \in S$. The only possible response to such a challenge is to assert $\{s_i\}$ (the agent that asserts this has nothing else to back s_i with), and either D_i will end without another A dialogue being embedded in it, or E will terminate because of repetition. Either way there will be no A dialogues embedded in D_i. □

In other words we can only have two levels of direct embedding of A dialogues. With this result, we are ready to start analysing combinations of atomic protocols.

5.2 Simple dialogues

We will start by just considering combinations of A dialogues. Since we can only have two levels of direct embedding of A dialogue, a dialogue under $\mathcal{IS}$ will never end up building a proof tree deeper that in Figure 2. This is the reason we can obtain termination results like those in [17]—the dialogue must terminate once the elements of the tree have been enumerated.

What do the proof trees look like for other kinds of dialogue? Well, dialogues conducted under $\mathcal{I}''$ will consist of a sequence of $\mathcal{IS}$ dialogues linked by their subject. If the subject of the nth dialogue is $r \to q$, then the subject of the $n+1$th is r or $s \to r$. The subject of the first dialogue is $q \to p$, for some q, where p is the subject of the $\mathcal{I}''$ dialogue. This creates a structure like that in Figure 3. In an $\mathcal{IS}$ dialogue, the key thing is the acceptance, or otherwise, of the subject of the dialogue and hence the subject of the top-level A dialogue. In an $\mathcal{I}''$ dialogue, the focus is much more on whether it is possible to prove something about the subject of the dialogue. In other words, for a dialogue with subject p, we are interested in whether $\cup_i \{a_i\} \vdash p$ where a_i is the subject

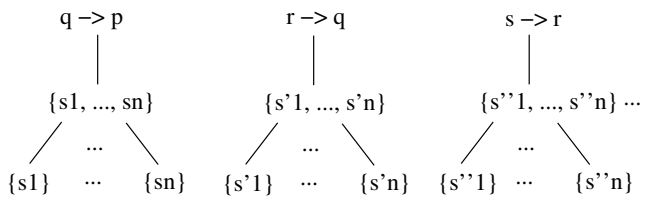

Fig. 3. An $\mathcal{I}''$ dialogue.

of the *i*th top-level A dialogue. We refer to all logically distinct and non-tautological propositions like *p* that can be inferred from things that have been the subject of a successful A dialogue as being *agreed conclusions* of the dialogue. Obviously the subjects of all successful A dialogues are themselves agreed conclusions. The following result justifies the name:

Proposition 13. *Given a dialogue D between agents F and G, where D consists of one or more A dialogues, and where p is an agreed conclusion of D, then both agents have an acceptable argument for p.*

Proof. The subject of each A dialogue that has the status of agreed conclusion is acceptable to both agents by definition—any proposition that is not acceptable will have been *reject*ed. Any agreed conclusion p is a logical consequence of these subjects a_i, and therefore an agent can build an argument $(\cup_i\{a_i\}, p)$. Because the a_i are acceptable, there are no acceptable undercutting arguments for the a_i, and hence none for $(\cup_i\{a_i\}, p)$. So both agents have an acceptable argument for p. □

The idea of agreed conclusions allows us to talk about outcomes other than those considered in [16]. There, we focused on *acceptance outcomes*—those propositions which one agent *assert*ed and the other later *accept*ed. Such acceptance outcomes include all the propositions in Figure 2 and 3.

The relationship between acceptance outcomes and agreed conclusions is captured by the following results.

Proposition 14. *For any dialogue under a protocol which permits only one A dialogue, the set of agreed conclusions is exactly the set of acceptance outcomes.*

Proof. The subject p of the A dialogue can be an acceptance outcomes, and if so the only acceptance outcome—since the grounds for p that are asserted are not accepted if there is only one A dialogue they can't be accepted. If is an acceptance outcomes, then p is also an agreed conclusion, and if p is not an acceptance outcome, there are no agreed conclusions, so the result holds. □

Proposition 15. *Given any dialogue between agents F and G that has two A dialogues D and E embedded in it, such that D is directly embedded in E, or so that D and E are in sequence, then the set of acceptance outcomes is a subset of the agreed conclusions of the dialogue.*

Proof. Consider D and E in sequence and imagine both are successful. For both dialogues, Proposition 14 tells us that the acceptance outcomes are exactly the set of agreed conclusions. Let's call these acceptance outcomes p and q. Then $p \wedge q$, which need not be an acceptance outcome, is an agreed conclusion and the result holds for this case. Exactly the same argument holds if one of D and E is embedded in the other. If either, or both, of D and E are not successful, then the the set of agreed conclusions is exactly the set of acceptance outcomes for this dialogue, $\emptyset$, and the result holds. □

So, if there is only one A, then acceptance outcomes and agreed conclusions coincide; but if a second A is included in the dialogue, then the set of agreed conclusions expands beyond the acceptance outcomes.

The reason that agreed conclusions and the A protocol are important ideas is that they give us a route to producing meta-theoretic results about the kinds of dialogue system we have been studying in [15, 16] that relate to dialogue structure. The above results are results about general classes of protocol—those that do and do not allow multiple A dialogues—rather than results about particular protocols. These are the kind of first, tentative, steps towards a meta-theory that we make in this paper.

The previous results suggest that it makes sense to classify protocols by the number of A dialogues that they permit. Since protocols that permit at most one A dialogue are not very interesting, we won't consider these to be a separate class. Instead we will classify protocols into those that do and do not permit sequences of A at the lowest level of embedding of such dialogues. (This is the only level at which it makes sense to discuss protocols which do not allow sequences—as soon as a set of grounds are asserted, as they must be in a A protocol, it does not make sense to prevent an embedded sequence of As testing the validity of the propositions in the grounds—so there is no point in considering restrictions on A dialogues at higher levels of embedding.)

Protocols like $\mathcal{I}''$ that allow sequences of A dialogues at the top level we will call *A-sequence* protocols and those like $\mathcal{IS}$ that do not allow such sequences of A dialogues we will call *A-singleton* protocols. Note that classifying a dialogue as A-singleton says nothing about whether it has embedded A dialogues. An A-sequence dialogue will in general generate more agreed conclusions than an A-singleton dialogue.

5.3 More complex dialogues

We are now ready to consider combinations of A with other atomic protocols, and will start by looking at the $\mathcal{P}'$ dialogue (since this neatly introduces another atomic protocol). There are two ways that a $\mathcal{P}'$ dialogue with subject p can unfold. In one, which in [14] we called *persuasion*$_1$, the initial combination of *know*, *assert* is followed by a single A dialogue. In the other, which in [14] we called *persuasion*$_2$, *know*(p), *assert*(p) is followed by *know*$(\neg p)$, the assertion of $\neg p$ and then by a A dialogue with subject $\neg p$. Clearly, then $\mathcal{P}'$ is an A-singleton protocol (though it can still have a set of agreed conclusions which is a superset of its set of acceptance outcomes). Since the atomic protocol:

A: *know*(x)
A: *assert*(x)
B: *reject*(x) or *accept*(x)

was called K in [14], we will classify protocols like $\mathcal{P}'$ which have K and A protocols embedded in K-protocols (but no K protocols embedded in the As, and no sequences of Ks) as K-embedded protocols. Such protocols are rather limited. If the sequence of embedded K protocols concern the same proposition *p*, and so start with *know*(p) then *know*($\neg p$), and so on we will call this a K(p)-embedded dialogue. Clearly the rule about repetition in $\mathcal{DG}$ implies that in practice there is no "and so on":

Proposition 16. *In $\mathcal{DG}$, K(p)-embedded dialogues can be composed of at most two K dialogues.*

Although this limiting result—which restricts K(p)-embedded dialogues to basically be identical to $\mathcal{P}'$—doesn't hold for other kinds of K-embedded dialogue, it isn't clear that such dialogues makes sense—they would involve a *know*/*assert* pair about two unconnected propositions (they might, however, be a basis for eristic dialogues—quarrels).

Since $\mathcal{P}'$ summarises all the possibilities for K(p)-embedded dialogues, we have:

Proposition 17. *A K(p)-embedded dialogue where the lowest level of embedding of K is n has the same set of agreed outcomes as an A-singleton dialogue with a level of embedding of $n+1$ and a subject of p, or an A-singleton dialogue with a level of embedding of $n+2$ and a subject of $\neg p$.*

Proof. Follows immediately from the unfolding of a dialogue under $\mathcal{P}'$. □

Thus $\mathcal{P}'$ and the whole class of K(p)-embedded dialogues capture a much narrower range of interactions than A-sequence dialogues.

It is possible to extend $\mathcal{P}'$ to obtain a similar kind of dialogue that is in the A-sequence class, but only in a limited way. Consider a dialogue that is a hybrid of *persuasion*$_1$ and *persuasion*$_2$ (which isn't possible under $\mathcal{P}'$, but would be under a close relative of it) with subject *p* in which the assertion of *p* is followed by the same A dialogue as in *persuasion*$_1$, but which doesn't stop[5] once the grounds for *p* have been found acceptable by both agents. Instead, the agent to which the initial *assert*(p) was addressed is now allowed to *assert* $\neg p$, and there is another A dialogue about the grounds for $\neg p$. The result is the construction of the proof tree in Figure 4. At this point, both agents judge the overall acceptability of *p* and $\neg p$ (which will depend in the limit on the strengths with which propositions are believed) and one will *accept*(p) or the other will *accept*($\neg p$). This new persuasion dialogue will be called e$\mathcal{P}$.

We will classify protocols like e$\mathcal{P}$—protocols in which there are successive K dialogues at a level of embedding of 1—we will relax this restriction later—as K-sequence protocols. Such protocols are allowed to have A protocols embedded in the K protocols, just as in $\mathcal{P}$, and maybe other protocols around the K-protocols.

It turns out that it is useful to distinguish K-sequence protocols in which successive K dialogues start with *know*(p) then *know*($\neg p$), and so on. We call such dialogues K(p)-sequence dialogues. Clearly the limitation on repetition in $\mathcal{DG}$ again means that:

Proposition 18. *In $\mathcal{DG}$, K(p)-sequence dialogues can have at most two K dialogues at a level of embedding of 1.*

[5] What we are describing here is the fullest extent of a dialogue under such a protocol—what [14] calls the *extensive form*. Clearly, a dialogue under this protocol might stop at this point.

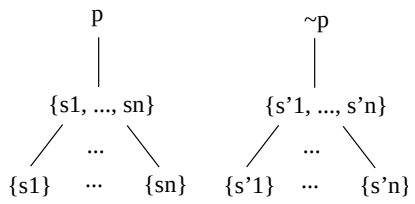

Fig. 4. An extended $\mathcal{P}$ dialogue.

We this notation, we can study the outcomes of dialogues like e$\mathcal{P}$. $\mathsf{K}(p)$-sequence dialogues are rather different to $\mathcal{P}'$ dialogues. A *persuasion*$_1$ dialogue between F and G in which F utters the first locution will result in G either accepting or not accepting p, but there will be no change in F's beliefs about p. Similarly, a *persuasion*$_2$ dialogue will either result in F accepting $\neg p$ or not accepting $\neg p$, but there will be no change in G's beliefs about $\neg p$. In an e$\mathcal{P}$ dialogue, either of the agents may change the status of p, but we can't tell which from the form of the dialogue. Indeed we won't be able to say anything about the outcome of the dialogue until the end. However, we do know that both agents cannot change their minds in this way:

Proposition 19. *In $\mathcal{DG}$, an $\mathsf{K}(p)$-sequence dialogue between agents F and G under a protocol in which the only dialogues at a level of embedding of 1 are K dialogues cannot result in one agent changing the status of p and the other changing the status of $\neg p$.*

Proof. For both agents to persuade the other to change the status of p we need the following scenario, or some symmetric variant, to take place. Before the dialogue, p is acceptable to F and $\neg p$ is acceptable to G. F starts a K dialogue with subject p and has p as an acceptance result. G has then changed status. G now has to get F to change the status of p. Consider the course of the dialogue unfolding in the best way to allow both agents to change the status of p. F asserts p, and may need to support this, and G accepts. The only remaining sub-dialogue requires that G assert $\neg p$ at this point, which it cannot do thanks to F's argument. The only time G can succeed in its persuasion is when F fails to make G change the status of p. □

This result hinges on the fact that both K dialogues are about the same proposition, and a G that has been persuaded that p is the case cannot then turn around and persuade F that $\neg p$ is the case. More general K-sequence dialogues, in which sucessive persuasions are about different propositions, can result in both agents changing the status of the subjects of successive dialogues.

We can extend the kinds of dialogue we can assemble with the K dialogue, by allowing K dialogues to be embedded in A dialogues. Denoting protocols that allow K dialogues within A dialogues as well as A dialogues within K dialogue as *AK-embedded* protocols, it is no surprise to find that:

Proposition 20. *Every K-sequence protocol is an AK-embedded protocol. Some AK-embedded protocols are not K-sequence protocols.*

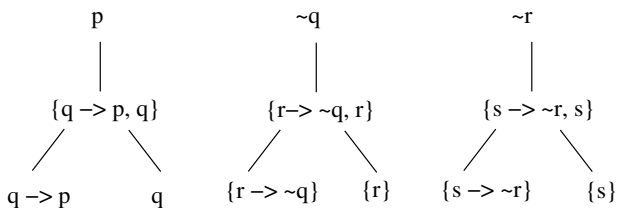

Fig. 5. An AK-embedded dialogue.

Proof. Immediate from the definition of K-sequence and AK-embedded protocols. □

However, the range of additional dialogues that are enabled by this extra embedding is maybe startling:

Proposition 21. *The class of AK-embedded protocols can generate dialogues which include embedded dialogues at arbitrarily large levels of embedding.*

Proof. Since K dialogues are allowed to be embedded in A dialogues, we can keep deepening the proof tree (if the knowledge bases of the agents suffice) by answering every *assert*(p) in a K dialogue with an A dialogue with subject p, and then meeting the assertion of one of the grounds s of the argument for p with an K dialogue that begins *know*($\neg s$). □

In other words, the argument can now continue as long as the participants have something new to say.

Such dialogues now make a new kind of persuasion possible—A can propose p, B can come up with an undercutter (attacking the grounds of p), but this can then be overruled by another argument from A which is undefeated and undercuts the undercutter. The proof tree for such a dialogue is given in Figure 5. However, despite the fact that they support this new kind of persuasion, AK-embedded protocols still have significant commonality with K-sequence dialogues:

Proposition 22. *Consider two agents F and G, with databases Σ_F and Σ_G. If F and G engage in a K-sequence dialogue, their agreed conclusions will be a subset of their agreed conclusions under a AK-embedded dialogue.*

Proof. The result holds because K-sequence and AK-embedded dialogues start out in the same way—they only differ in terms of assertions (which are the locutions that give rise to agreed conclusions) once the dialogue gets to the first embedded K-dialogue. So while AK-embedded dialogues may have agreed conclusions that aren't achieved by K-sequence dialogues, they will have all the agreed conclusions (which may be the empty set of agreed conclusions) of the K-sequence dialogue up to that first embedded K-dialogue. □

At this point it makes sense to ask whether we have a kind of monotonicity result for AK-embedded dialogues that says, just as Proposition 19 does for K(p)-sequence

dialogues, that once both agents agree on a proposition, it remains agreed throughout the dialogue. In fact, we can show the opposite of Proposition 19 for AK-embedded dialogues:

Proposition 23. *A dialogue between agents F and G under an AK-embedded protocol can result in one agent changing the status of p and the other changing the status of* $\neg p$.

Proof. For this result we only need an existence proof. An instance occurs in following scenario, or some symmetric variant. Before the dialogue, p is acceptable to F and $\neg p$ is acceptable to G. F starts a K dialogue with subject p and has p as an acceptance result. G has then changed status. G now has to get F to change the status of p. It can't do this by asserting $\neg p$, since it no longer has an acceptable argument for $\neg p$, but it can now assert some q (if there is such a proposition) that allows F to create an acceptable argument for $\neg p$. If this q does not, so far as G knows, bear upon p or $\neg p$, then G remains convinced of the acceptability of p and both agents have changed status as required by the result. □

This is a critical point, and it is worth considering it in more detail. As an example of how we can have the kind of situation in the proof of Proposition 23, consider the dialogue outlined in Figure 5. Consider further that F starts the dialogue by stating p, G challenges, F replies with $\{q \rightarrow p, q\}$ and so on. By the time that the dialogue finishes with the statement of $\{s\}$, G has an acceptable argument for p and so changes status. However, a later assertion by G (and such an assertion is not ruled out in an AK-embedded dialogue), t, which is unrelated to the proof tree in Figure 5 provides the final piece of a convincing argument from Σ_F (and thus invisible to G) against p. Then F will change the status of p.

Note that t cannot be part of the chain of argument about p. If it were, if t was part of the grounds for $\neg q$, say, and also a crucial part of some argument against p the rest of which was only known to F, then this argument would also be an argument against t and so be objected to by F. If it were able to cause F to find p not acceptable, then it would also prevent G changing the status of $\neg p$.

The important thing that is happening here is that, unlike what happens in the simple dialogues we have been studying up until now, both agents are making assertions and then further assertions in their defence, and later assertions need not be directly related—that is related in a way that is visible to both agents—to earlier ones. As the commitment stores grow, the set of new arguments that both agents can make as a result of the dialogue is growing, and, in particular, the non-overlapping part of this is growing. As this happens, the non-monotonicity of the notion of acceptability is coming to the fore. An obvious question then is, doesn't Proposition 19 contradict Proposition 23? Doesn't the non-monotonicity of the agreed conclusions (they are non-monotonic because they are determined by acceptability) mean that two agents can have an K-sequence dialogue about p and obtain agreed outcomes that are not agreed outcomes of an AK-embedded dialogue about p between the same two agents?

The answer is that the result of Proposition 19 holds *across the course of the dialogue* rather than *at the end of the dialogue*. In other words, it is possible for those agents to have an AK-embedded dialogue about p that ends up with a set of agreed outcomes that do not include the agreed outcomes of a K-sequence dialogue about p, but

along the way they will have agreed on exactly the same outcomes, only to later reject them when they considered additional information.

The notion that we have to consider results across the course of the dialogue, and so take the non-monotonicity of the agreed outcomes properly into account, will be the focus of our future work.

6 Conclusions

This paper has extended the analysis of formal inter-agent dialogues in [14–16]. The main contribution of this extension has been to begin to provide a meta-theory for such dialogues based on structural classification, making it possible to establish results for whole classes of dialogue protocol. This, in turn, allows us to classify the whole space of possible protocols, establishing relations between them, and giving us ways of identifying good and bad classes. An early attempt in this direction was a second major contribution of this paper—giving a more extensive analysis of the relation between types of protocol and the outcome of dialogues under different protocols than has previously been possible [16].

In this paper we have only scratched the surface of the work that needs to be done in this area. There are a number of future directions that we are taking. First, we are deepening the analysis in this paper, extending the work to handle the notion of "across the course of the dialogue", and investigating other kinds of dialogue, such as the deliberation (in the terminology of [22]) dialogues [9]. Second, we are looking to strengthen our meta-theory using techniques from dynamic logic [11], to come up with tools that allow us to analyse dialogues in a way analogous to that in which dynamic logic is currently used to analyse program correctness. From this perspective we can think of each locution as a "program" in the usual program correctness sense, and then identify the effect of combinations of these. Finally, we are developing a denotational semantics for our dialogues using category theory [13]. This allows us to talk about properties of dialogues at a very abstract level.

Acknowledgments: This work was partially supported by NSF REC-02-19347 and NSF IIS-0329037. Thanks are due to Frank Dignum for suggesting we look at proof trees.

References

1. L. Amgoud and C. Cayrol. On the acceptability of arguments in preference-based argumentation framework. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 1–7, 1998.
2. L. Amgoud, N. Maudet, and S. Parsons. Modelling dialogues using argumentation. In E. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems*, pages 31–38, Boston, MA, USA, 2000. IEEE Press.
3. B. Chaib-Draa and F. Dignum. Trends in agent communication language. *Computational Intelligence*, 18(2):89–101, 2002.

4. F. Dignum, B. Dunin-Kęplicz, and R. Verbrugge. Agent theory for team formation by dialogue. In C. Castelfranchi and Y. Lespérance, editors, *Seventh Workshop on Agent Theories, Architectures, and Languages*, pages 141–156, Boston, USA, 2000.
5. 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.
6. R. A. Flores and R. C. Kremer. To commit or not to commit. *Computational Intelligence*, 18(2):120–173, 2002.
7. T. F. Gordon. The pleadings game. *Artificial Intelligence and Law*, 2:239–292, 1993.
8. M. Greaves, H. Holmback, and J. Bradshaw. What is a conversation policy? In F. Dignum and M. Greaves, editors, *Issues in Agent Communication*, Lecture Notes in Artificial Intelligence 1916, pages 118–131. Springer, Berlin, Germany, 2000.
9. B. J. Grosz and S. Kraus. The evolution of sharedplans. In M. J. Wooldridge and A. Rao, editors, *Foundations of Rational Agency*, volume 14 of *Applied Logic*. Kluwer, The Netherlands, 1999.
10. B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
11. D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, 2000.
12. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104(1–2):1–69, 1998.
13. P. McBurney and S. Parsons. A denotational semantics for deliberation dialogues. In *3rd International Conference on Autonomous Agents and Multi-Agent Systems*. IEEE Press, 2004.
14. S. Parsons, P. McBurney, and M. Wooldridge. The mechanics of some formal inter-agent dialogue. In F. Dignum, editor, *Advances in Agent Communication*. Springer-Verlag, Berlin, Germany, 2003.
15. S. Parsons, M. Wooldridge, and L. Amgoud. An analysis of formal inter-agent dialogues. In *1st International Conference on Autonomous Agents and Multi-Agent Systems*. ACM Press, 2002.
16. S. Parsons, M. Wooldridge, and L. Amgoud. On the outcomes of formal inter-agent dialogues. In *2nd International Conference on Autonomous Agents and Multi-Agent Systems*. ACM Press, 2003.
17. S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation*, 13(3):347–376, 2003.
18. H. Prakken. Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese*, 127:187–219, 2001.
19. C. Reed. Dialogue frames in agent communications. In Y. Demazeau, editor, *Proceedings of the Third International Conference on Multi-Agent Systems*, pages 246–253. IEEE Press, 1998.
20. M. Schroeder, D. A. Plewe, and A. Raab. Ultima ratio: should Hamlet kill Claudius. In *Proceedings of the 2nd International Conference on Autonomous Agents*, pages 467–468, 1998.
21. K. Sycara. Argumentation: Planning other agents' plans. In *Proceedings of the Eleventh Joint Conference on Artificial Intelligence*, pages 517–523, 1989.
22. 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.