On the Outcomes of Formal Inter-agent Dialogues

Simon Parsons
Department of Computer and Information Science
Brooklyn College,
Brooklyn, NY 11210, USA.
parsons@sci.brooklyn.cuny.edu

Michael Wooldridge
Department of Computer Science,
University of Liverpool,
Liverpool L69 7ZF, UK.
m.j.wooldridge@csc.liv.ac.uk

Leila Amgoud
IRIT, 118 route de Narbonne,
31062 Toulouse Cedex,
France.
leila.amgoud@irit.fr

ABSTRACT

This paper studies argumentation-based dialogues between agents. It takes a previously defined system by which agents can trade arguments and examines the outcomes of the dialogues this system permits. In addition to providing a first characterisation of such outcomes, the paper also investigates the extent to which outcomes are dependent on tactical play by the agents, and arguing that this violates principles of mechanism design, identifies how to prevent tactics having an effect.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Coherence and co-ordination; multiagent systems.*

General Terms

Languages, theory.

Keywords

Agent communication, dialogue games, argumentation.

1. INTRODUCTION

When building multi-agent systems, we take for granted the fact that the agents which make up the system will need to communicate: 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. Many of these communication requirements cannot be fulfilled by the exchange of single messages. Instead, the agents concerned need to be able to exchange a sequence of messages which all bear upon the same subject. In other words they need the ability to engage in dialogues. As a result of this requirement, there has been much work on providing agents with the ability to hold such dialogues. Recently some of this work has considered argument-based approaches to dialogue, for example the work by Dignum *et al.* [5], Parsons and Jennings [16], Reed [22], Schroeder *et al.* [23] and Sycara [24].

Reed's work built on an influential model of human dialogues due to argumentation theorists Doug Walton and Erik Krabbe [25], and we also take their dialogue typology as our starting point. Walton and Krabbe set out to analyze the concept of commitment in dialogue, so as to "provide conceptual tools for the theory of argumentation" [25, page ix]. This led to a focus on persuasion dialogues, and their work presents formal models for such dialogues. In attempting this task, Walton and Krabbe recognized the need for a characterization of dialogues, and so they present a broad typology for inter-personal dialogue. They make no claims for its comprehensiveness.

Their categorization 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, revived by Hamblin [11] and extending back to Aristotle, overlaps with work on conversational policies (see, for example, [4, 7]), but differs in considering the entire dialogue rather than dialogue segments.

As defined by Walton and Krabbe, the three types of dialogue we consider here are:

Information-Seeking Dialogues: 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: The participants collaborate to answer some question or questions whose answers are not known to any one participant.

Persuasion Dialogues: One party seeks to persuade another party to adopt a belief or point-of-view he or she does not currently hold. These 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 properties and complexity of such dialogues [19]. Here we extend this investigation, turning to consider the questions "how can we characterise the outcomes of dialogues",

AAMAS'03, July 14-18, 2003, Melbourne, Australia.

and "to what extent are the outcomes of dialogues predetermined?" In other words, how does the knowledge that agents have affect the final result of the dialogue, and does it matter what locutions agents utter, and in what order they utter illocutions, or are the results of the dialogue entirely determined by what the agents know and the protocols they use?

One of the interesting things about this question is that the kind of answer we would like to the above question depends on our perspective. On one hand, it seems attractive for agents to be able to "control their own destiny", and have the ability to reach different outcomes[1] depending on how they act within the constraints of a dialogue protocol. On the other hand, from a mechanism design [12] perspective, it is attractive for the protocol to ensure that agents with the same knowledge coming into a dialogue will always find the same result—that way the protocol can be seen to stop one agent misleading another. As we shall see, our formal framework makes both answers to the question possible under different conditions.

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 inherent in natural language dialogues.

2. BACKGROUND

In this section we briefly introduce the formal system of argumentation which forms the backbone of our approach. This is inspired by the work of Dung [6] but goes further in dealing with preferences between arguments. Further details are available in [1]. We start with a possibly inconsistent knowledge base Σ with no deductive closure. We assume Σ contains formulas of a propositional language $\mathcal{L}$. $\vdash$ stands for classical inference and $\equiv$ 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 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 $h \equiv \neg Conclusion(A_1)$.*

[1] In the sense of what agents end up believing.

In other words, an argument is undercut if and only if 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[2] 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[3]. We can now define the argumentation system we will use:

DEFINITION 4. *An* argumentation system *(AS) is a triple $\langle \mathcal{A}(\Sigma), Undercut, Pref \rangle$ 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.*

Henceforth, $C_{Undercut,Pref}$ will gather all non-undercut arguments and arguments defending themselves against all their undercutting arguments. In [1], it was shown that the set $\underline{\mathcal{S}}$ of acceptable arguments of the argumentation system $\langle \mathcal{A}(\Sigma), Undercut, Pref \rangle$ is the least fixpoint of a function $\mathcal{F}$:

$$\begin{aligned} \mathcal{S} &\subseteq \mathcal{A}(\Sigma) \\ \mathcal{F}(\mathcal{S}) &= \{(H, h) \in \mathcal{A}(\Sigma) | (H, h) \text{ is defended by } \mathcal{S}\} \end{aligned}$$

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

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

An argument is acceptable *if it is a member of the acceptable set.*

[2] Here we only deal with beliefs, though the approach can also handle desires and intentions as in [18] and could be extended to cope with other mental attitudes.

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

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 decide what they know by determining which propositions they have acceptable arguments for. They trade propositions for which they have acceptable arguments, and accept propositions put forward by other agents if they find that the arguments are acceptable. The exact locutions and the way that they are exchanged define a formal dialogue game which agents engage in.

Dialogues are assumed to take place between two agents, for example called P and C. Each agent has a knowledge base, Σ_P and Σ_C respectively, containing their beliefs. In addition, each agent has a further knowledge base, accessible to both agents, containing commitments made in the dialogue[4]. These commitment stores are denoted $CS(P)$ and $CS(C)$ respectively, and in this dialogue system an agent's commitment store is just 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. Each agent has access to their own private knowledge base and both commitment stores. Thus P can make use of $\langle \mathcal{A}(\Sigma_P \cup CS(C)), Undercut, Pref \rangle$[5] and C can make use of $\langle \mathcal{A}(\Sigma_C \cup CS(P)), Undercut, Pref \rangle$.

All the knowledge bases contain propositional formulas and are not closed under deduction, and 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, though it is possible [3] to combine different sets of preferences, and it is also possible to have agents modify their beliefs on the basis of the reliability of their acquaintances [15].

With this background, we can present the set of dialogue moves first introduced in [19]. 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.

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

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

The counterpart of these moves are the acceptance moves. They can be used whenever the protocol and the agent's acceptance attitude allow.

accept(p). p is a propositional formula.

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

accept(S). S is a set of propositional formulas.

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

There are also moves which allow questions to be posed.

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. In contrast, a question can be used to query the other player about any proposition.

question(p). where p is a propositional formula.

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

We refer to this set of moves as the set $\mathcal{M}'_{\mathcal{DC}}$. The locutions in $\mathcal{M}'_{\mathcal{DC}}$ are similar to those discussed in legal reasoning [8, 21] and it should be noted that there is no *retract* locution. Note that these locutions are ones used within dialogues—locutions such as those discussed in [14] would be required to frame dialogues.

We also need to define the attitudes which control the assertion and acceptance of propositions.

DEFINITION 7. *An agent may have one of two* assertion *attitudes.*

- *a* confident *agent can assert any proposition p for which it can construct an argument (S, p).*
- *a* careful *agent can assert any proposition p if it is unable to construct a stronger argument for $\neg p$.*
- *a* thoughtful *agent can assert any proposition p for which it can construct an acceptable argument (S, p).*

DEFINITION 8. *An agent may have one of three* acceptance *attitudes.*

- *a* credulous *agent can accept any proposition p if it is backed by an argument.*
- *a* cautious *agent can accept any proposition p if it is unable to construct a stronger argument for $\neg p$.*
- *a* skeptical *agent can accept any proposition p if there is an acceptable argument for p.*

Since agents are typically involved in both asserting and accepting propositions, we denote the combination of an agent's two attitudes as

$$\langle assertion\ attitude \rangle / \langle acceptance\ attitude \rangle$$

The effects of this range of agent attitudes on dialogue outcomes is studied in [20], and for the rest of this paper we will focus on thoughtful/skeptical agents.

[4] Following Hamblin [11] commitments here are propositions that an agent is prepared to defend.

[5] Which, of course, is exactly the same thing as $\langle \mathcal{A}(\Sigma_P \cup CS(P) \cup CS(C)), Undercut, Pref \rangle$.

4. TYPES OF DIALOGUE

Previously [19] we defined three protocols for information seeking, inquiry and persuasion dialogues. These protocols are deliberately simple, the simplest we can imagine that can satisfy the definitions given by [25], since we believe that we need to understand the behaviour of these simple protocols before we are to able to understand more complex protocols.

Information-seeking. In an information seeking dialogue, one participant seeks the answer to some question from another participant. If the information seeker is agent A and the other agent is B, then we can define the protocol $\mathcal{IS}$ for an information seeking dialogue about a proposition p as follows:

1. A asks $question(p)$.
2. B replies with either $assert(p)$, $assert(\neg p)$, or $assert(\mathcal{U})$. Which will depend upon the contents of its knowledge-base and its assertion attitude. $\mathcal{U}$ 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.

Note that A accepts whenever possible, only being able to challenge when unable to accept—"only" in the sense of only being able to challenge then and *challenge* being the only locution other than *accept* that it is allowed to make. More flexible dialogue protocols are allowed, as in [2], but at the cost of possibly running forever[6].

Inquiry. In an inquiry dialogue, the participants collaborate to answer some question whose answer is not known to either. There are a number of ways in which one might construct an inquiry dialogue (for example see [13]). Here we present one simple possibility. We assume that two agents A and B have already agreed to engage in an inquiry about some proposition p by some control dialogue as suggested in [14], and from this point can adopt the following protocol $\mathcal{I}$:

1. A asserts $q \rightarrow p$ for some q or $\mathcal{U}$.
2. B accepts $q \rightarrow p$ if its acceptance attitude allows, or challenges it.
3. A replies to a *challenge* with an $assert(S)$, where S is the support of an argument for the last proposition challenged by B.
4. Goto 2 for each proposition $s \in S$ in turn, replacing $q \rightarrow p$ by s.
5. B *assert*s q, or $r \rightarrow q$ for some r, or $\mathcal{U}$.
6. If $\mathcal{A}(CS(A) \cup CS(B))$ includes an argument for p which is acceptable to both agents, then first A and then B accept it and the dialogue terminates successfully.
7. Go to 5, reversing the roles of A and B and substituting r for q and some t for r.

This protocol[7] is basically a series of implied $\mathcal{IS}$ dialogues. First A asks "do you know of anything which would imply p were it known?". B replies with one, or the dialogue terminates with $\mathcal{U}$. If A accepts the implication, B asks "now, do you know q, or any r which would imply q were it known?", and the process repeats until either the process bottoms out in a proposition which both agents agree on, or there is no new implication to add to the chain.

Persuasion. In a persuasion dialogue, one party seeks to persuade another party to adopt a belief or point-of-view he or she does not currently hold. The dialogue game DC, on which the moves in [2] are based, is fundamentally a persuasion game, so the protocol below results in games which are very like those described in [2]. This protocol, $\mathcal{P}$, is as follows, where agent A is trying to persuade agent B to accept p.

1. A *assert*s p.
2. B *accept*s p if its acceptance attitude allows, if not B *assert*s $\neg p$ if it is allowed to, or otherwise *challenge*s p.
3. If B asserts $\neg p$, then goto 2 with the roles of the agents reversed and $\neg p$ in place of p.
4. If B has *challenge*d, then:
   (a) A asserts S, the support for p;
   (b) Goto 2 for each $s \in S$ in turn.

If at any point an agent cannot make the indicated move, it has to concede the dialogue game. If A concedes, it fails to persuade B that p is true. If B concedes, then A has succeeded in persuading it. An agent also concedes the game if at any point if there are no propositions made by the other agent that it hasn't accepted.

We should point out 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 [25].

Note that all three of these protocols have the same core steps. One agent *assert*s something, the other *accept*s if it can, otherwise it *challenge*s. A *challenge* provokes the *assert*ion of the grounds, which are in turn either *accept*ed or *challenge*d. The proposition p that is the first assertion, and the central proposition of the dialogue, is said to be the *subject* of the dialogue. This basic framework has been shown [17, 19] to be capable of capturing a range of dialogue types. Here we examine what its consequences are in terms of the outcomes of dialogues that are carried out using this framework.

[6] The protocol in [2] allows an agent to interject with $question(p)$ for any p at several points, making it possible for a dialogue between two agents to continue indefinitely.

[7] Which differs from the inquiry dialogue in [19] in the *accept* moves in step 6.

5. DIALOGUE OUTCOMES

One important set of properties to consider of a dialogue system intended for use between autonomous agents is the extent to which the outcome of the dialogue depends upon the way the dialogue progresses, on the knowledge[8] that agents choose to reveal to one another, rather than on what they believe to be true, and the protocol.

To do this, we need a precise notion of the outcome of a dialogue. There are several things that can be used to measure an outcome. One is in terms of the what agents come to accept during the course of the dialogue:

DEFINITION 9. *Consider two agents F and G engaging in a dialogue. The set of* acceptance outcomes for F *of the dialogue are all the propositions p such that F makes the move assert(p), and subsequently G makes the move accept(p).*

For each acceptance outcome for a given agent, we say that the other agent *has accepted* the proposition in question. Since propositions that are not the subject of the dialogue are often accepted, an acceptance result need not be the subject of the dialogue.

We can also relate outcomes to what agents know. We define:

DEFINITION 10. *Consider two agents F and G that are engaging in a dialogue. Then:*

- *the set of* knowledge outcomes for F $O_k(F|G)$ *is the set of all the propositions p such that p is the conclusion of an argument in $\mathcal{A}(\Sigma_F \cup CS(G))$;*
- *the set of* joint knowledge outcomes $O_k(F \wedge G)$ *is the set of all the propositions p such that p is the conclusion of an argument in $\mathcal{A}(\Sigma_F \cup \Sigma_G)$;*
- *the set of* committed outcomes *for F $O_c(F|G)$ is the set of all the propositions p such that p is the conclusion of an argument in $\mathcal{A}(CS(F))$; and*
- *the set of* joint committed outcomes $O_c(F \wedge G)$ *is the set of all the propositions p such that p is the conclusion of an argument in $\mathcal{A}(CS(F) \cup CS(G))$.*

We can also define the acceptable subsets of these outcomes—the *acceptable knowledge outcomes for F* $O_k^a(F|G)$, *acceptable joint knowledge outcomes* $O_k^a(F \wedge G)$, *acceptable committed outcomes* for F $O_c^a(F|G)$, and *acceptable joint committed outcomes* $O_c^a(F \wedge G)$.

Note that all but the joint knowledge outcomes and acceptable joint knowledge outcomes will change over the course of a dialogue as the contents of the commitment stores alter. Unless otherwise noted, here we will only consider these sets at the end of a dialogue.

Now, it is easy to show that there is an inclusion relationship between knowledge and committed outcomes:

PROPOSITION 1. *For a dialogue between any two agents F and G:*

$$\begin{aligned} O_k(F|G) &\subseteq O_k(F \wedge G) \\ O_c(F|G) &\subseteq O_c(F \wedge G) \\ O_c(F|G) &\subseteq O_k(F|G) \\ O_c(F \wedge G) &\subseteq O_k(F \wedge G) \end{aligned}$$

[8]For now we will just consider this information to be beliefs, though as in [17], we can extend the approach to other mental notions as well.

PROOF. *Since an agent's commitment store is a subset of its knowledge base, $(\Sigma_F \cup CS(G)) \subseteq \Sigma_F \cup \Sigma_G$, and the first relation follows directly from the monotonicity of propositional logic. The remaining relations follow for similar reasons.* □

No such firm relationship exists between acceptable knowledge and committed outcomes:

PROPOSITION 2. *For a dialogue between two agents F and G, the relationships between $O_k^a(F|G)$, $O_k^a(F \wedge G)$, $O_c^a(F|G)$, and $O_c^a(F \wedge G)$ depend upon the knowledge of the agents and the contents of the commitment stores.*

PROOF. *This follows from the non-monotonicity of acceptability. If there are no conflicting arguments in $\mathcal{A}(\Sigma_F)$ or $\mathcal{A}(\Sigma_G)$, then $O_k^a(F|G)$, $O_k^a(F \wedge G)$, $O_c^a(F|G)$, and $O_c^a(F \wedge G)$ will be exactly $O_k(F|G)$, $O_k(F \wedge G)$, $O_c(F|G)$, and $O_c(F \wedge G)$, respectively, and the relationship between the sets of arguments will be as in Proposition 1. However, if between them the agents have a pair of arguments that make each other unacceptable—$(\{a, a \rightarrow c\}, c)$ and $(\{b, b \rightarrow \neg c\}, \neg c)$ (all with the same preference), for example—then if neither has been asserted, $O_k^a(F|G) \not\subseteq O_k^a(F \wedge G)$. If G asserts one of these arguments and F does not, then $O_c^a(F|G) \not\subseteq O_c^a(F \wedge G)$. If F then asserts its argument, then its conclusion will be an acceptable committed outcome but cannot be an acceptable knowledge outcome for F, so $O_c^a(F \wedge G) \not\subseteq O_k^a(F|G)$. Finally, either agent might have asserted something, r say, that is acceptable given what it knows and what the other agent has in its commitment store (and so is an acceptable joint committed outcome), but which the other agent has a stronger argument against (and so cannot be an acceptable joint knowledge outcome) since the dialogue may end before the acceptability of r comes into question. Hence the relationships in question depend on the contents of the agents' knowledge bases and the commitment stores.* □

This result is a positive one. Without it dialogues would be so predictable that we did not even need to consider what the agents knew in order to predict the outcome. What the result means is that—since the content of the commitment stores is critical in determining the outcomes, and since the contents of the commitment stores is determined by what moves the agents make, and the moves are determined in part by the protocol—the protocol has a role in determining the outcome of the dialogues.

Finally, to end the preliminaries, we can relate acceptance outcomes and acceptable knowledge outcomes:

PROPOSITION 3. *If p is an acceptance result for F in a dialogue with G, then p is an acceptable knowledge outcome for F and G. The reverse does not hold.*

PROOF. *For p to be an acceptance result for F, is has to be asserted by F and accepted by G. To be asserted by F, it must be acceptable given all F knows and all G has asserted—thus it has to be an acceptable knowledge outcome for F. To accept p, G has to check p against the contents of its knowledge base and what F has in its commitment store. So if p is accepted, it must be an acceptable knowledge outcome for G.*

For p to be an acceptance result for F, F must assert it and have it accepted by G. Consider a dialogue in which F

asserts q, it is challenged by G, and so F asserts the support of q, which includes p. Now suppose that G finds that the first element of the support, r say, is not acceptable given $\Sigma_G \cup CS(F)$. The dialogue will end without p being accepted, even if p is an acceptable knowledge outcome for G (and it has to be one for F before it can be asserted). □

The fact that the result does not hold in both directions is the reason that we need both notions of outcome in order to characterise the results of dialogues.

Now we are ready to characterise the outcomes of dialogues. Since the proposition $\mathcal{U}$ indicates that a dialogue ends because of a lack of knowledge, we consider dialogues that end with a $\mathcal{U}$ or a repeated locution to have failed in some way. Thus any dialogue that does not end in this way will be said to be *successful*. Information seeking dialogues will end, if successful, with one agent getting the other to accept p or $\neg p$:

PROPOSITION 4. *A dialogue between agents F and G about p under protocol $\mathcal{IS}$, in which F makes the first move, will end either with one of:*

- *G making the move $\mathcal{U}$;*
- *one agent repeating a locution;*
- *p or $\neg p$ being an acceptance result for G.*

PROOF. *If F moves first, G asserts p, $\neg p$, or $\mathcal{U}$. The latter ends the dialogue. Otherwise F accepts, in which case the p or $\neg p$ is an acceptance result for G, or challenges. If F challenges, G asserts the support for p, and F considers each member of the support in turn. As in [19], this results in either the acceptance of the support and thus whichever of p and $\neg p$ was initially asserted, or a repeated locution, when an unacceptable member of the support is challenged twice. Either way the result holds.* □

As a result of Proposition 3, a successful dialogue between agents F and G about p under protocol $\mathcal{IS}$ can result in either p or $\neg p$ being in the set of acceptable knowledge outcomes for both agents.

Persuasion dialogues can end with either agent having an acceptance outcome:

PROPOSITION 5. *A dialogue between agents F and G about p under protocol $\mathcal{P}$, in which F makes the first move, will end with one of:*

- *one agent repeating a locution;*
- *p being an acceptance outcome for F; or*
- *$\neg p$ being an acceptance outcome for G.*

PROOF. *If F moves first, it asserts p. If G does not challenge, the rest of the dialogue plays out like a dialogue about p under $\mathcal{IS}$ that G starts, and ends with a repeated locution or p being an acceptance result for F ($\mathcal{U}$ can only be uttered at an earlier point). If G asserts $\neg p$, either F ends the dialogue by reasserting p, or the rest of the dialogue plays out like a dialogue about $\neg p$ under $\mathcal{IS}$ that F starts. Either way, the result holds.* □

Again, Proposition 3 tells us that a successful persuasion dialogue between agents F and G about p under protocol $\mathcal{P}$ ends with either p or $\neg p$ being in the set of acceptable knowledge outcomes of both agents.

Inquiry dialogues are a little different, since successful inquiry dialogues do not, as defined, have acceptance outcomes that directly relate to the subject of the dialogue. However, we can characterise the outcome as follows:

PROPOSITION 6. *A dialogue between agents F and G about p under protocol $\mathcal{I}$, will end with one agent making the move $\mathcal{U}$, or with p as an acceptable joint committed outcome.*

PROOF. *The dialogue is just a backward search for a proof of p. Agents take it in turns to assert an implication that leads to p. The process ends if either cannot add a new step in the proof that is acceptable to both agents*[9]*, or with the assertion of all the steps in a proof for p that is acceptable to both agents. In the latter case, then all the steps of an acceptable argument for p are in the union of the commitment stores and p is an acceptable joint knowledge outcome.* □

6. PREDETERMINISM

As was shown in [20], if all agents in a dialogue are both thoughtful/skeptical, then one agent cannot deliberately mislead another in the sense of the dialogue having an acceptance outcome for the first agent concerning a proposition if the first agent has a better argument for the negation of that proposition. However, this does not prevent dialogues having acceptance outcomes in situations where intuitively both agents together should be able to figure out that the proposition in question should not be accepted, as the following result shows:

PROPOSITION 7. *There exist dialogues between two agents F and G under protocols $\mathcal{IS}$, $\mathcal{I}$, or $\mathcal{P}$ such that there is an acceptance outcome for one agent for a proposition which is not acceptable given $\Sigma_F \cup \Sigma_G$. This holds even if both agents are thoughtful/skeptical.*

PROOF. *This is an existence result which we prove by example. Let $\Sigma_F = \{\neg c, a, a \rightarrow b\}$ and $\Sigma_G = \{\neg c \rightarrow \neg b\}$ where $\neg c$ and $\neg c \rightarrow b$ have a higher preference level than the other propositions. Then F can assert b, and G will accept it even though $(\{a, a \rightarrow b\}, b)$ is not in $\mathcal{A}(\Sigma_F \cup \Sigma_G)$.* □

Now, there are two ways to read this result. One, which we will call the *tactical dialogue* reading, takes this result as a positive feature of the $\mathcal{IS}$, $\mathcal{I}$, and $\mathcal{P}$ protocols. According to this position, one wishes to build agents that have some kind of rhetorical ability, and can, by clever tactics, get other agents to accept things that they might otherwise not accept. In other words, we want protocols which do not make the outcome predetermined by the agents' knowledge.

The positive view of the $\mathcal{IS}$, $\mathcal{I}$, and $\mathcal{P}$ protocols in the tactical dialogue view is further bolstered by the following result:

PROPOSITION 8. *There exist dialogues between two agents F and G under protocols $\mathcal{IS}$, $\mathcal{I}$, or $\mathcal{P}$ such that the acceptance outcomes of the dialogue depend upon the order in which propositions are asserted. This holds whatever the assertion and acceptance attitudes of the agents.*

[9] More complex inquiry dialogues are considered in [20].

PROOF. *This is another existence result that we prove by example. Let* $\Sigma_F = \{a, a \rightarrow b, b \rightarrow c, a \rightarrow f, f \rightarrow c\}$ *and* $\Sigma_G = \{b, f \rightarrow \neg b, b \rightarrow \neg f\}$ *where all propositions are equally preferred. If* F *asserts* $(\{a, a \rightarrow b, b \rightarrow c\}, c)$ *then the dialogue has the acceptance outcome of* c *for* F*. However, if* F *first asserts* $(\{a, a \rightarrow f, f \rightarrow c\}, c)$*, then* c *will not be an acceptance outcome for* F *even if it subsequently asserts* $(\{a, a \rightarrow b, b \rightarrow c\}, c)$ *because supplying* f *gives* G *arguments that attack both of* F*'s arguments for* C*.* □

A different view of the protocols, what we will call the *dialogue as mechanism* view, suggests that Propositions 7 and 8 are both negative. From this perspective, dialogue protocols should be like economic mechanisms [12]. In the same way that economic mechanism design tries to ensure that the results of, for instance, an auction does not depend upon the order in which the bids are made but only the values participants place upon the good being auctioned, so the outcomes of dialogues should not depend upon the order of agents' locutions, but upon what they know. In other words, the outcomes should be predetermined.

Both these last results are special cases of:

PROPOSITION 9. *Consider two agents* F *and* G *engaging in a dialogue under protocols* $\mathcal{IS}$*,* $\mathcal{I}$*, or* $\mathcal{P}$*, the set of acceptance outcomes for* F *is not necessarily the set of acceptable joint knowledge outcomes.*

PROOF. *By propositions 4, 5 and 6, we know that outcomes of dialogues under protocols* $\mathcal{IS}$*,* $\mathcal{I}$*, or* $\mathcal{P}$ *are acceptable knowledge outcomes for one or other agent (in the case of* $\mathcal{IS}$ *and* $\mathcal{P}$ *dialogues) or acceptable joint committed outcomes (in the case of* $\mathcal{I}$ *dialogues). Proposition 2 tells us that neither acceptable knowledge outcomes or acceptable joint committed outcomes need be acceptable joint knowledge outcomes, and the result follows.* □

This result in turn suggests that equivalence of acceptance outcomes and the acceptable joint knowledge outcomes might be an important consideration in determining whether protocols allow for tactical play or whether they predetermine the outcome. Certainly, the set of acceptable joint knowledge outcomes (as remarked above) is the only one of the outcome measures we are dealing with that can be identified without going through a dialogue. In fact:

PROPOSITION 10. *Consider two agents* F *and* G *engaging in a dialogue under a protocol in which the set of acceptance outcomes for either agent is exactly the set of acceptable joint knowledge outcomes. The acceptance outcomes will not depend upon the order in which propositions are asserted.*

PROOF. *We are told that any* p *that is an acceptance outcome is in* $O^a_K(F \wedge G)$*. Since, by definition,* $O^a_K(F \wedge G)$ *depends only on the contents of* Σ_F *and* Σ_G *rather than the contents of the commitment stores, it is clearly independent of the moves in the dialogue.* □

This, then, gives us a way to test whether a protocol ensures a predetermined outcome—all we have to do is to see whether it allows propositions that are not in the set of acceptable joint knowledge outcomes.

Now, the key thing about the result given above is that the set of acceptable joint knowledge outcomes is in some sense the maximal set of propositions that can be obtained from the dialogue. It is not necessarily the biggest set of acceptable outcomes, the nonmonotonicity of argumentation sees to that, but it is based on all the arguments that might be put forward by both participants. Thus once a proposition makes it into the set of acceptable joint knowledge outcomes, there are no more arguments that might overturn it.

PROPOSITION 11. *Consider two agents* F *and* G *engaging in a dialogue under some protocol. That the set of acceptance outcomes for either agent is exactly the set of acceptable joint knowledge outcomes is a necessary and sufficient condition for the acceptance outcomes to not depend upon the order in which propositions are asserted.*

PROOF. *Proposition 10 shows that if the acceptance outcomes are acceptable joint knowledge outcomes, then the outcomes do not depend on the order of assertion. We thus only need to show that if the acceptance outcomes are not acceptable joint knowledge outcomes, then the outcomes will depend on the order of assertion. Consider* p*, an acceptance outcome for* F *that is not an acceptable joint knowledge outcome. Since* p *is not an acceptable joint knowledge outcome, there is some argument in* $\Sigma_F \cup \Sigma_G$ *that makes* p *not acceptable, and so if the dialogue had preceded differently,* p *would not have been an acceptance outcome. Thus the result follows.* □

This gives us a way that is easy to state, for adapting the protocols $\mathcal{IS}$, $\mathcal{I}$, and $\mathcal{P}$ to make them predetermined. We just ensure that every proposition that will be needed to establish the acceptable joint knowledge outcomes is asserted. Of course, while this is easy to state, it is much harder to establish exactly what the right propositions should be.

One solution would be to simply assert the full set of propositions in Σ_F and Σ_G, ensuring that the full set of necessary propositions has to be asserted. However, this kind of approach will be inefficient in general, both in the communication it requires, and the time and computational resources the agents consume in processing this information. A better solution is to only assert those propositions that will have a bearing on the acceptability of the subject of the dialogue. Since the propositions that need to be asserted must be part of arguments that attack the argument supporting the subject (or attack arguments that attack such arguments, and so on), then it is possible to identify them. However, it is hard to see how to identify them systematically, since their connection could be revealed at any time.

For now the only way we can see to ensure that all necessary propositions are asserted is to modify dialogues along the lines of one of the inquiry dialogue in [20]. Under this protocol, (i) any response to a challenge involves the assertion of the support for *every* argument for the proposition in question, (ii) at every step in the proof, every possible implication that might form the next step of the proof is asserted, and (iii) agents are released from the need to take turns. Future work will consider if there is a notion of relevance that can help to better identify the minimal set of propositions to assert.

Finally, it seems to us that the choice of whether dialogues should have predetermined outcomes or allow for tactical play is one that will depend upon the context in which the dialogues are taking place. The results presented here should make it possible to choose the right kind of dialogue for a given context, but further work is required to establish which contexts require which kind of dialogue.

7. CONCLUSIONS

This paper has extended the analysis of formal inter-agent dialogues in [19]. We have provided the first detailed characterisation of the outcomes of such dialogues, and then we have investigated the extent to which outcomes are dependent on tactical play by the agents. Finding that tactics can have a big effect on the outcome, we then identified how to rule out the effect of tactics, arguing that this is desirable from a mechanism design perspective.

More work, of course, remains to be done in this area in addition to that outlined above. Particularly important are: determining the relationship between the locutions we use in these dialogues and those of agent communication languages such as the FIPA ACL; examining the effect of adding new locutions (such as *retract*) to the language; extending the system with a more detailed model of preferences; and providing an implementation. We are currently investigating these matters along with further dialogue types, such as planning dialogues [9].

Acknowledgments

We would like to thank Peter Stone for first posing the question "to what extent are the outcomes of dialogues predetermined?", and the anonymous referees for their helpful comments.

8. 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] L. Amgoud and S. Parsons. Agent dialogues with conflicting preferences. In J.-J. Meyer and M. Tambe, editors, *Proceedings of the 8th International Workshop on Agent Theories, Architectures and Languages*, pages 1–15, 2001.

[4] B. Chaib-Draa and F. Dignum. Trends in agent communication language. *Computational Intelligence*, 18(2):89–101, 2002.

[5] 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.

[6] 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.

[7] R. A. Flores and R. C. Kremer. To commit or not to commit. *Computational Intelligence*, 18(2):120–173, 2002.

[8] T. F. Gordon. The pleadings game. *Artificial Intelligence and Law*, 2:239–292, 1993.

[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] C. L. Hamblin. *Fallacies*. Methuen and Co Ltd, London, UK, 1970.

[12] M. O. Jackson. Mechanism theory. In *The Encyclopedia of Life Support Systems*. EOLSS Publishers, 2000. URL: `www.eolss.co.uk`.

[13] P. McBurney and S. Parsons. Risk agoras: Dialectical argumentation for scientific reasoning. In C. Boutilier and M. Goldszmidt, editors, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Stanford, CA, USA, 2000. UAI.

[14] P. McBurney and S. Parsons. Games that agents play: A formal framework for dialogues between autonomous agents. *Journal of Logic, Language, and Information*, 11(3):315–334, 2002.

[15] S. Parsons and P. Giorgini. An approach to using degrees of belief in BDI agents. In B. Bouchon-Meunier, R. R. Yager, and L. A. Zadeh, editors, *Information, Uncertainty, Fusion*. Kluwer, Dordrecht, 1999.

[16] S. Parsons and N. R. Jennings. Negotiation through argumentation — a preliminary report. In *Proceedings of Second International Conference on Multi-Agent Systems*, pages 267–274, 1996.

[17] S. Parsons and P. McBurney. Argumentation-based communication between agents. In M.-P. Huget, editor, *Agent Communication Languages*. Springer Verlag, Berlin, 2003.

[18] S. Parsons, C. Sierra, and N. R. Jennings. Agents that reason and negotiate by arguing. *Journal of Logic and Computation*, 8(3):261—292, 1998.

[19] 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.

[20] S. Parsons, M. Wooldridge, and L. Amgoud. Properties and complexity of formal inter-agent dialogues. *Journal of Logic and Computation*, (to appear), 2003.

[21] H. Prakken. Relating protocols for dynamic dispute with logics for defeasible argumentation. *Synthese*, 127:187–219, 2001.

[22] 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.

[23] 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.

[24] K. Sycara. Argumentation: Planning other agents' plans. In *Proceedings of the Eleventh Joint Conference on Artificial Intelligence*, pages 517–523, 1989.

[25] 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.