How Agents Alter Their Beliefs After an Argumentation-Based Dialogue

Simon Parsons and Elizabeth Sklar

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

Abstract. In our previous work on dialogue games for agent interaction, an agent's set of beliefs (Σ) and an agent's "commitment store" (CS) — the set of locutions uttered by the agent — play a crucial role. The usual assumption made in this work is that the set of beliefs is static through the course of a dialogue, while the commitment store is dynamic. While the assumption of static beliefs is reasonable during the progress of the dialogue, it seems clear that some form of belief change is appropriate once a dialogue is complete. What form this change should take is our subject in this paper.

1 Introduction

Finding ways for agents to reach agreements in multiagent systems is an area of active research. One mechanism for achieving agreement is through the use of *argumentation*—where one agent tries to convince another agent of something during the course of some *dialogue*. Examples of argumentation-based approaches to multiagent agreement include the work of Dignum *et al.* [3], Kraus [10], Reed [15], Schroeder *et al.* [16] and Sycara [19].

The work of Walton and Krabbe [21] has been particularly influential in argumentation-based dialogue research. They developed a typology for inter-personal dialogue which identifies six primary types of dialogues including *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 agent seeks to persuade another agent to adopt a belief or point-of-view she does not currently hold). This *dialogue game* [9] view of dialogues overlaps with work on conversation policies (see, for example, [2, 5]), but differs in considering the entire dialogue rather than dialogue segments.

In this paper, we extend the work of [13, 14] by considering how agents alter their beliefs as a result of participating in dialogues. In particular we are interested in the way in which the beliefs of an agent change over the course of several dialogues with another agent. The work described here allows us to obtain results which show that, under certain conditions, the beliefs of a pair of agents will converge over time.

S. Parsons et al. (Eds.): ArgMAS 2005, LNAI 4049, pp. 297–312, 2006.

2 Background

We begin by introducing the components of the formal system of argumentation that underpin our approach, as well as the corresponding terminology and notation, all taken from [1, 4, 13]. This is a bit lengthy, but the material is required in order to obtain the technical results later in the paper.

A dialogue game is a set of interactions that occur between two agents, M and U. Each agent maintains a knowledge base, Σ, containing formulas of a propositional language $\mathcal{L}$ and having no deductive closure. Each agent also maintains a list of utterances, called the "commitment store", CS. We can refer to CS as an agent's "public knowledge", since it contains information that is shared with other agents. In contrast, the contents of Σ are "private". The agent also maintains two Σ-like components: J and Γ. These will be discussed later. For now it suffices to know that such structures exist and are indexed by the name of the agent's dialogue partner.

Note that in the description that follows, we assume that $\vdash$ is the classical inference relation, that $\equiv$ stands for logical equivalence, and we use Δ to denote all the information available to an agent. Thus in a dialogue with U, $\Delta_M = \Sigma_M \cup \Gamma_{M,U} \cup J_{M,U} \cup CS_U$. The commitment store CS_M can be loosely thought of as a subset of Δ_M; according to the rules of the dialogue game, M can only say things it can support (or justify), i.e., using arguments in Δ_M to support propositions in CS_M.

Definition 1 (Argument). *An* argument *is a pair $A = (S, p)$ where p is a formula of $\mathcal{L}$ and S a subset of Δ such that:*

1. *S is consistent;*
2. *$S \vdash p$; and*
3. *S is minimal, so no proper subset of S satisfying both (1) and (2) exists.*

S is called the support *of A, written $S = Support(A)$ and p is the* conclusion *of A, written $p = Conclusion(A)$. Thus we talk of p being* supported *by the argument (S, p).*

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

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

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 beliefs are more strongly held than others, we assume that any set of beliefs has a *preference order* over it. We consider all information available to an agent, Δ, to be stratified into non-overlapping sets $\Delta_1, \ldots, \Delta_n$ such that beliefs in Δ_i are all equally preferred and are preferred over elements in Δ_j where $i < j$. This could be thought of as saying that an agent's first choice(s) are contained in Δ_1, second choices in Δ_2, and so on. The *preference level* of a nonempty subset $S \subset \Delta$, where different elements $s \in S$ may belong to different layers Δ_i, is valued at the highest numbered layer which has a member in S and is referred to as $level(S)$.

In other words, S is only as strong as its weakest member. Note that the strength of a belief as used in this context is a separate concept from the notion of support discussed earlier. That is, a strong belief does not necessarily mean that there are many arguments supporting that belief.

Definition 3 (Preference). *Let A_1 and A_2 be two arguments in $\mathcal{A}(\Delta)$. A_1 is* preferred *to A_2 according to Pref and following the strict pre-order associated with it. In other words, $A_1 \gg^{Pref} A_2$, iff $level(Support(A_1)) \leq level(Support(A_2))$. If A_1 is preferred to A_2, we say that A_1 is* stronger *than A_2.*

We can now define the argumentation system we will use:

Definition 4 (Argumentation System). *An* argumentation system *(AS) is a triple $\langle \mathcal{A}(\Delta), Undercut, Pref \rangle$ such that:*

- *$\mathcal{A}(\Delta)$ is a set of the arguments built from Δ,*
- *Undercut is a binary relation representing the defeat relationship between arguments, $Undercut \subseteq \mathcal{A}(\Delta) \times \mathcal{A}(\Delta)$, and*
- *Pref is a (partial or complete) pre-ordering on $\mathcal{A}(\Delta) \times \mathcal{A}(\Delta)$.*

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

Definition 5 (Defense). *Let A_1, A_2 be two arguments of $\mathcal{A}(\Delta)$.*

- *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 $\mathcal{A}$* defends *A_1 iff: $\forall$ A_2 undercuts A_1 and A_1 does not defend itself against A_2 then $\exists$ $A_3 \in \mathcal{A}$ such that A_3 undercuts A_2 and A_2 does not defend itself against A_3.*

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

$$\mathcal{A} \subseteq \mathcal{A}(\Delta)$$
$$\mathcal{F}(\mathcal{A}) = \{(S,p) \in \mathcal{A}(\Delta) \mid (S,p) \text{ is defended by } \mathcal{A}\}$$

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

$$\underline{\mathcal{A}}(\Delta) = \bigcup \mathcal{F}_{i \geq 0}(\emptyset)$$
$$= A_{Undercut,Pref} \cup \left[\bigcup \mathcal{F}_{i \geq 1}(A_{Undercut,Pref}) \right]$$

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 (Status). *If an agent M 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, Attitudes and Protocols

The basis for our work is the dialogue system $\mathcal{DG}$, presented in [12] (which is a modest extension of that in [13, 14]), modified with some features from the dialogue system in [17]. Here we present as brief a summary of the combined system as we can give.

As described above, dialogues are assumed to take place between two agents, for example called M (for "me") and U ("you"). Each agent $i \in \{M, U\}$ has a knowledge base, Σ_i, containing its beliefs. We assume that this knowledge base is consistent in a certain sense — we assume that an agent only has propositions in its knowledge base for which it has an acceptable argument (the grounds of this argument may be just the proposition itself, so that, for example, an agent may have in its knowledge base p supported by the acceptable argument $(\{p\}, p)$.

In addition [9], 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. Following [17], we also assume that each agent i has a knowledge base $\Gamma_{i,j}$ where $j \in \{M, U\}, j \neq i$ which represents i's model of j's beliefs, and a set $J_{i,j}$ which records *lies* that i has told j—propositions p for which $\neg p$ is in Σ_i. Since each agent has access to their private knowledge bases and both commitment stores, agent M can potentially make use of $\langle \mathcal{A}(\Sigma_M \cup \Gamma_{M,U} \cup J_{M,U} \cup CS_U), Undercut, Pref \rangle$. For most of this paper we will assume that $\Gamma_{M,U}$ and $J_{M,U}$ are empty and so only consider Σ_M and CS_U, but towards the end we will deal with non-empty $\Gamma_{M,U}$ and $J_{M,U}$.

All the knowledge bases contain propositional formulas, 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).

During the dialogue the players put forward propositions and accept propositions put forward by other agents based on their acceptability. The exact locutions we adopt are those of [12], but for our purposes here we need only know that propositions are put forward using an *assert* locution (all the other locutions are signalling, *assert* is the only one which transmits data). The axiomatic semantics [20] of *assert* are given in Table 1. The important thing to note is that the subject of an *assert* is something that an agent either has in its knowledge base, or has an acceptable argument for, and that asserting something places it in the agent's commitment store. The *subject* of a dialogue is the argument of the first *assert* to be made—this is the proposition about which the dialogue revolves.

Table 1. Operational semantics for *assert*

assert

LOCUTION:

- $M \rightarrow U : assert(p)$

PRE-CONDITIONS:

1. $(S, p) \in \underline{\mathcal{A}}(\Sigma_M \cup CS_U)$

POST-CONDITIONS:

1. $CS_{M,i} = CS_{M,i-1} \cup \{p\}$ (update)
2. $CS_{U,i} = CS_{U,i-1}$ (no change)

The process by which a dialogue is carried out is determined by a *protocol*. An example is the protocol $\mathcal{P}''$, an extension of $\mathcal{P}'$ in [12] in which M tries to persuade U that p is the case:

1. M issues a $know(p)$, indicating it believes that p is the case.
2. M *assert*s p.
3. U *accept*s p if it has an acceptable argument for it, or U *assert*s $\neg p$ if it has an acceptable argument for that, or U *challenge*s p, or U rejects p.
4. If U asserts $\neg p$ in (3), then go to (3) with the roles of the agents reversed and $\neg p$ in place of p.
5. If U challenges in (3) then M asserts, in turn, every $s \in S$, where S is the support for p and go to (3) for each s in turn in place of p.

The "signal" locutions used here are *know*, which indicates the start of a persuasion dialogue, *challenge*, which indicates that one agent requires the other to present the support for the proposition just asserted, and *accept* and *reject*, which indicate that the agent finds (respectively, does not find) that the previously asserted proposition is supported by an acceptable argument. A signal of *accept* also indicates that the agent that issues it is no longer disputing that proposition and either the dialogue ends (if the subject of the *accept* is the subject of the dialogue), or the dialogue can pass onto the next proposition (if the subject of the *accept* is another proposition and the dialogue is the recursive phase following step 5). A signal of *reject* similarly indicates that the dialogue can pass on to the next proposition (albeit without the former propsition being accepted), and the rejection of the subject of the dialogue is the other way that a dialogue can end.

Note that, in common with previous work on this kind of system, agents are not allowed to repeat exactly the same locution in a dialogue. If the only legal move available to an agent under the protocol is to repeat itself in this way, then the dialogue terminates. This is to prevent infinite dialogues in which one agent, for example, repeatedly asserts p. By "exactly the same" we mean the same locution instantiated with a logically equivalent proposition, so that $assert(p)$ and $assert(p \wedge p)$ are considered the same locution, precisely with preventing infinite dialogues in mind (since $p \wedge p$ contains no more

information that p we assume a rational agent would not *assert* both). The only exception we allow to this rule is that an agent can assert a proposition as its own grounds. Thus, as is often the case, p can be asserted as support for the previous assertion p if there is no other argument for it and p is present in the agent's knowledge base.

Note also that, for now, we don't specify how U makes the decision in step 3 of the protocol. Later we will disinguish between different ways the decision might be made and see how these relate to different outcomes.

Example 1. As an example of a dialogue that can be held under $\mathcal{P}''$, consider the following.

$\Sigma_M = \{p, p \rightarrow q\}$	M know p	
$\Sigma_U = \{p\}$	M *assert* q	$CS_M = \{q\}$
	U *challenge* q	
	M *assert* p	$CS_M = \{p, q\}$
	U *accept* p	U already has an acceptable argument for p
	M *assert* $p \rightarrow q$	
	U *challenge* $p \rightarrow q$	
	M *assert* $p \rightarrow q$	this is allowed under the exception to the repetition rule.
		$CS_M = \{p, q, p \rightarrow q\}$
	U accept $p \rightarrow q$	
	U accept q	

4 How Beliefs Change over Time

Previous work on argumentation-based dialogues has typically concentrated on what happens *during* a *single* dialogue — this is certainly true of the work in [12, 13, 14] — and has not contemplated what happens after a dialogue is complete, or what happens over the course of several dialogues. In contrast, our interest here is in the process by which an agent adapts its beliefs after a dialogue is ended, and what effect this process has over time. Indeed, the only related work we are aware of in an argumentation context is [11] which studies the way that beliefs change during a single argumentation-based dialogue.

4.1 Changes in Belief After a Single Dialogue

Now, without having to commit ourselves to a specific dialogue protocol, we can determine the situation that must hold at the end of a dialogue. Both of the agents engaged in the dialogue will have *assert*ed some propositions, and these will have become, in some sense, common knowledge between the two agents. Furthermore, it is clear that some of these propositions will be acceptable (in the sense of being supported by an acceptable argument) to one or both agents, and that there may be propositions p that were acceptable to an agent before a dialogue that are now no longer acceptable (because, for example, the dialogue has established that $\neg p$ is acceptable):

Proposition 1. *For any proposition p, the status of p for an agent M may change as a result of a dialogue that M has with another agent U.*

Proof. We have four cases to consider—that p is initially acceptable or not acceptable, and that p is a proposition in Σ_M or is the conclusion of an argument from Σ_M. For the result we simply have to show how the change in status may occur.

Let us assume that p is initially acceptable because it is the conclusion of an acceptable argument (S,p) where $S \subseteq \Sigma_M$ and $p \notin \Sigma_M$. The dialogue may result in U asserting an argument that undercuts the argument for p, that is an argument with conclusion $\neg s$ for some $s \in S$, and if (S,p) cannot defend itself against this argument, the status of p will change from acceptable to not acceptable.

The case for which p is initially acceptable and $p \in \Sigma_M$ is very similar. Here p is supported by the argument $(\{p\},p)$, and will change status if U asserts an argument with conclusion $\neg p$ which is preferred to $(\{p\},p)$.

If p is initially not acceptable, this is either because there is no argument that supports it, or because the supporting argument is undercut by some argument A that the supporting argument cannot defend itself against, and is not defended against by any other argument. This situation can easily change, for example if A is undercut by some newly asserted argument, and this can happen for both the case in which $p \in \Sigma_M$ and the case in which p is the conclusion of an argument (S,p) where $S \subseteq \Sigma_M$ and $p \notin \Sigma_M$.

These changes come about because the notion of acceptability is *non-monotonic*. As a dialogue between M and U proceeds, the set of propositions Δ_M that M uses to construct arguments increases monotonically (since no locutions remove propositions from the commitment store), but the set of acceptable arguments can both increase or decrease. (This is proved in [14][1].)

In many situations, it seems sensible for an agent to want to remember the status of the propositions that are interesting to it at the end of the dialogue. This is appropriate, for example, in our learning scenario. It might be considered less appropriate in a purchasing scenario—security might dictate that an agent should not remember sensitive data beyond the end of a dialogue. Our concern here is not on when it is appropriate to remember, but to identify mechanisms for doing so, and to explore their consequences.

There are four obvious ways to ensure that an agent M recalls the status of a proposition following a dialogue with U and these are given below. For now, we will only consider information in Σ_i and CS_i—we will come back to $\Gamma_{i,j}$ and $J_{i,j}$ later.

Definition 8 (Update Mechanisms). *We define the following* mechanisms for updating *Σ_M at the end of a dialogue between agents M and U.*

W1: Expand Σ_M to become $\Sigma_M \cup CS_U$.

W2: Expand Σ_M with all s for which there exists a p such that $(S,p) \in \underline{A}(\Sigma_M \cup CS_U)$, $s \in S$ and $s \notin \Sigma_M$.

[1] And can be easily seen in the following example. M initially has just one argument $(\{q, q \rightarrow p\}, p)$ for p, and by definition this is acceptable. U then puts forward the argument $(\{r, r \rightarrow \neg p\}, \neg p)$ for $\neg p$. Both agents only have knowledge bases that consist of the support of their arguments, and all propositions are equally preferred. After the second argument is asserted, neither argument is acceptable to either agent, and so M's set of acceptable arguments has shrunk while its set of arguments has grown.

W3: Expand Σ_M with all logically distinct p such that $(S,p) \in \underline{\mathcal{A}}(\Sigma_M \cup CS_U)$ and $S \not\subseteq \Sigma_M$.

W4: Replace any $p \in \Sigma_M$ such that $(S,\neg p) \in \underline{\mathcal{A}}(\Sigma_M \cup CS_U)$ with $\neg p$.

Of course, though we have stated the update mechanisms for M alone, there are symmetrical mechanisms for U.

In other words, these mechanisms are as follows: (1) add everything in U's commitment store to M's knowledge base[2]; or (2) add those elements of the support of propositions p for which M only has an acceptable argument *after* the dialogue; or (3) add just the propositions p for which M only has an acceptable argument *after* the dialogue; or (4) replace any propositions in Σ_M whose negations are now acceptable with those negations.

In conjunction with Definition 8, we need to define what constitutes a good mechanism for this updating. It seems reasonable to insist that the update is to ensure that the agent in question keeps a record of just those new propositions that it finds acceptable.

Definition 9 (Update Criteria). *We define the following* criteria for updating *the knowledge-based Σ_M of agent M after a dialogue:*

C1. Updating should cause the addition to Σ_M of exactly those propositions that are acceptable at the end of the dialogue but were not acceptable before the dialogue began.

C2. After updating, $\underline{\mathcal{A}}(\Sigma_M)$ should include all those arguments that are acceptable at the end of the dialogue.

We can use these criteria to identify which mechanism for updating should be adopted, but first we need:

Lemma 1 (from [13]). *If $(S,p) \in \underline{\mathcal{A}}(\Sigma_M)$ then $(S',s) \in \underline{\mathcal{A}}(\Sigma_M)$ for every $s \in S$.*

In other words, every element of the support of an acceptable argument is itself the conclusion of an acceptable argument.

Corollary 1. *If an updating mechanism satisfies C1, then it satisfies C2.*

Proof. Immediate from the definition of C1 and C2, and Lemma 1.

Thus C1 is a stronger criterion than C2 since it specifies that no additional propositions other than those that have become newly acceptable should be added. C2 allows for the addition of propositions that result in Σ_M generating arguments after the updating that are not acceptable so long as all arguments that were acceptable at the end of the dialogue can be constructed. Thus C2 does not imply C1[3].

Proposition 2. *Mechanisms W1 and W2 satisfy C2, mechanism W3 satisfies C1, and mechanism W4 fails to satisfy either criterion.*

[2] This is just the simplest update rule we can imagine, rather than one we think would be adopted by a rational agent, but would be a possible update rule for the *credulous* agents discussed in [13].

[3] We see no way of tightening C2 to make $\underline{\mathcal{A}}(\Sigma_M)$ generate exactly the arguments acceptable at the end of the dialogue without losing valuable information.

Proof. We examine each mechanism in turn, considering the case of updating Σ_M after agent M has completed a dialogue with agent U.

W1 updates by adding every proposition in CS_U to Σ_M. If M has asserted some proposition that M does not find acceptable, then this will be added to Σ_M (since all propositions asserted by U end up in CS_U whether or not M finds them acceptable). W1 thus fails to meet C1 by including propositions that M does not find acceptable, but by adding everything that was asserted by U satisfies C2—all new arguments, including all the acceptable ones, can be constructed.

W2 updates by including the grounds for every p that has become acceptable as a result of the dialogue, and so satisfies C2. It fails to satisfy C1, however, because it does not include the p themselves (unless they are in the grounds of other acceptable arguments).

W3 updates by adding to Σ_M every logically distinct conclusion of every acceptable argument whose support is not already wholly in Σ_M. Since Lemma 1 tells us that every element of the support of such arguments will also be the conclusion of an acceptable argument, the result will be to include all formulae that are acceptable after the dialogue but were not before, which exactly satisfies C1.

W4 updates by replacing every p in Σ_M that was acceptable before the dialogue but is not afterwards by $\neg p$. This is in line with C1 for those propositions which were acceptable before the dialogue and have become unacceptable as a result of it, but fails to deal with propositions for which there was no argument before the dialogue. W4 thus satisfies neither C1 nor C2.

Given this result, the most suitable of these procedures for revision seems to be W3, since it satifies the strongest of the conditions, though examining the proof of Proposition 2 shows that W2 is very nearly as good.

As an illustration of how W3 works, consider the following.

Example 2. After the dialogue in Example 1, U will add $p \rightarrow q$ and q to Σ_U since there are acceptable arguments for these, and the grounds for the argument were not all previously in Σ_U. M will add nothing to Σ_M since U asserted no propositions, and so there are no new arguments that are acceptble to M — note that M does not add q even though it is not part of its original knowledge base.

4.2 Changes in Belief over Several Dialogues

Our primary interest in this paper is to examine how the knowledge-bases of agents develop over time, which we measure in terms of a series of dialogues. To track this development we need the following definition:

Definition 10 (Degree of Agreement). *The* degree of agreement *DA between two sets of formulae S_1 and S_2 is:*

$$DA(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Thus we define the agreement between two knowledge bases by looking at the proportion of formulae they have in common. Two knowledge bases which share no formulae will have a *DA* of 0, and two knowledge bases which contain exactly the same set of

formulae will have a *DA* of 1. Note that the measure as defined is symmetrical and makes no attempt to identify whether one knowledge base is contained in another, a situation that could be considered another form of agreement. We acknowledge that more sophisticated measures of agreement can be established, but this seems to suit our requirements for now.

Given Definition 10 we can establish how a given dialogue changes the extent to which two agents agree. It is simple to show that:

Proposition 3. *If M and U engage in repeated $\mathcal{P}''$ dialogues and update using W3 after each, then the degree of agreement between Σ_M and Σ_U may not increase.*

Proof. For this proof it suffices to show that there is a way for the degree of agreement to not increase.Consider that M starts a dialogue by asserting p, U challenges, and M asserts the support for its argument (S, p). If U rejects the first $s \in S$, then at the end of the dialogue neither agent has anything to add to its knowledge base. This same process can happen for every dialogue, and the degree of agreement between Σ_M and Σ_U will not increase.

There are several comments to make about this result. The first is that the result captures an extreme case—over many dialogues it seems likely that at least one proposition will be accepted by one agent, and so the degree of agreement will increase a little. However, the point the proposition makes is that there is no guarantee that it will. The second comment is that this can be viewed as a good thing. Gabbay and Woods in their discussion of non-coooperation dialogues [6, 7] give the example of a police interrogation, where it may very much be in an agent's interests not to be persuaded that something is true (that one committed a crime about which one has no knowledge for example).

The main comment to make about this result is that though it is weak — it just says that after some dialogues the agents might not be any closer to agreement — the reason behind it suggests the subject deserves more investigation. The reason that agents might not have a greater degree of agreement after a dialogue is, as it is easy to see from the proof of Proposition 3, that if U finds s, that is some part of M's support for (S, p), unacceptable, it can just *reject* and end the dialogue. This can happen even if M has information that would overturn U's objection to s if it were stated. It is this latter possibility that seems worthy of elucidation, especially when we realise that the property of resisting an increase in agreement is not just a property of $\mathcal{P}''$, but also of the various kinds of dialogue introduced in [13].

If we define:

Definition 11 (Closed Mouth Dialogue). *A dialogue between two agents M and U is a* closed mouth dialogue *if either agent replies to an "assert(p)" with an immediate "accept(p)" or "reject(p)" during the course of the dialogue.*

Definition 12 (Open Mouth Dialogue). *A dialogue between two agents M and U is an* open mouthed dialogue *if both agents can only reply to an "assert(p)" with "challenge(p)" or "assert(¬p)" before "reject(p)".*

As introduced the protocol $\mathcal{P}''$ can generate both open-mouthed and a closed-mouth dialogues, but we can devise open and closed mouth versions of $\mathcal{P}''$ that can, respectively, only generate open and closed mouth dialogues. One closed-mouth variant of

$\mathcal{P}''$, denoted $\mathcal{P}''_{CM}$, rejects *whenever* the asserted proposition is not acceptable[4] accepting otherwise. The open-mouthed variant, denoted $\mathcal{P}''_{OM}$, challenges whenever the asserted proposition is not acceptable unless such a challenge would be a repetition. When it cannot challenge, the agent asserts the negation of the asserted proposition if that is possible, and can only accept or reject when such an assertion is impossible. Finally $\mathcal{P}''_{OM}$ accepts if the proposition is acceptable and rejects otherwise.

We are now nearly at a point where we can relate the form of the dialogue, open or closed-mouth, to degree of agreement. Before we can make such a relation, however, we need to consider that each agent "updates" its knowledge base Σ_i with the conclusions p of all acceptable arguments (S, p) that can be made from Σ_i (in other words the agents add every p such that $(S, p) \in \underline{\mathcal{A}}(\Sigma_i)$), doing a kind of pre-emptive W3 update.

With this condition, then, we have:

Corollary 2. *If M and U engage in any series of dialogues under $\mathcal{P}''_{CM}$ and update using W3 after each, then the degree of agreement between Σ_M and Σ_U will not increase.*

Proof. The proof follows quickly from 3. If U rejects whenever the proposition is unacceptable, the only time it can possibly accept is if the proposition is immediately acceptable, but in that case U must have an acceptable argument for it before the dialogue starts, and so the degree of agreement will not increase.

which makes the point that some closed mouth dialogues (the example we have the result for is only one example of a closed mouth dialogue) prevent two agents increasing their degree of agreement. If we didn't add the condition on the knowledge bases before the dialogue, of course, then the degree of agreement would increase if M's assertion made U "realise" that it had grounds to support p all along but just hadn't generated an argument for p.

The key thing about an open-mouthed dialogue is that each agent has to explain why it finds a proposition p unacceptable, challenging if it doesn't have enough information to construct a support for it, and asserting $\neg p$ if it has an argument against it. This results in:

Proposition 4. *At the end of a dialogue about p under $\mathcal{P}''_{OM}$ between agents M and U, p must have the same status for M and U.*

Proof. By definition, in an open mouthed dialogue, if one agent does not have an acceptable argument for a proposition p asserted by the other, it has to either challenge, which will lead to the assertion of other propositions, or assert $\neg p$, which will result in a challenge and the assertion of the grounds for $\neg p$. This process will recurse until neither agent has anything more challenges or assertions to make, and all the information which either agent can bring to bear on the subject has been deployed. At this point both agents have access to the same set of arguments concerning every p asserted by both agents (otherwise the recursion would not have stopped), and both agents will have to grant every p that has been asserted the same status.

This result takes us close to being able to identify open-mouthed dialogues with increases in the degree of agreement, but we first have to consider cases like that in the following example:

[4] Other closed mouth variants of $\mathcal{P}''$ may immediately reject some assertions and not others.

Example 3. All propositions have the same preference level:

$$\begin{array}{lll} \Sigma_M = \{p \wedge q\} & M \text{ know } p & \\ \Sigma_U = \{p \wedge \neg q\} & M \text{ assert } q & CS_M = \{q\} \\ & U \text{ challenge } q & \\ & M \text{ assert } p \wedge q & \\ & U \text{ challenge } p \wedge q & \\ & M \text{ assert } p \wedge q & \\ & U \text{ assert } \neg q & CS_U = \{\neg q\} \\ & M \text{ challenge } \neg q & \\ & U \text{ assert } p \wedge \neg q & \\ & M \text{ reject } \neg q & \end{array}$$

Here agreement on the status of q means both find q unacceptable — M has an argument for q, but it is undercut by the $\neg q$ in CS_U, U has an argument for $\neg q$ but this is undercut by the q in CS_M — and so neither will update its Σ. This is the kind of situation in which, in human argumentation, we say "we must agree to disagree". Both sides have heavily entrenched beliefs that lead to inconsistent positions that cannot be resolved. We capture this in the notion of *deadlock*:

Definition 13 (Deadlock). *Two agents M and U are* deadlocked *over p if $(S,p) \in \underline{\mathcal{A}}(\Sigma_M)$ or $(S,p) \in \underline{\mathcal{A}}(\Sigma_U)$, but $(S,p) \notin \underline{\mathcal{A}}(\Sigma_M \cup \Sigma_U)$.*

The notion of deadlock captures exactly the case in the example above as well as the case where both M and U initially have an acceptable argument for p, but these arguments are built on contradictory grounds — the grounds will be exposed by the dialogue, and neither agent ends up finding the subject of the dialogue acceptable (of course, in such a case one might want to revise not just by W3, but by removing some propositions, but we will leave such considerations for future work — the system we deal with here would just end the dialogue with the contradiction unresolved in such a case).

Proposition 4 captures the limits of persuasive argumentation, at least as far as open-mouthed dialogues are concerned. In an open-mouth dialogue each agent says all that it has to say relating to a subject, but that does not guarantee to create agreement. However, we have a more "agreeable" result if agents are not deadlocked:

Proposition 5. *If two agents M and U engage in a dialogue under $\mathcal{P}''_{OM}$ with subject p, and update using W3 after, then the only cases in which the $DA(\Sigma_M, \Sigma_U)$ does not increase as a result of the dialogue is when either (1) the agents both initially have the same acceptable argument for p or (2) the agents are deadlocked over p and all the grounds for p that M asserts.*

Proof. Consider the progress of an open mouth dialogue as sketched in the proof of Proposition 4. There are only two ways that this process will not lead to some new propositions being accepted by one of the agents, thus increasing $DA(\Sigma_M, \Sigma_U)$. One way is if every assertion is met with an accept. For this to be the case, the two agents must have exactly the same argument for p (and it must be acceptable or otherwise it could not be asserted by either). The other way is if every assertion is ultimately met

with a reject, and that can only happen if the agents are deadlocked on every proposition that is asserted — p and every proposition that is in the grounds for p that are asserted by M.

Thus over many dialogues, we can say that the knowledge bases of the two agents will converge—if they talk for long enough, then they will agree:

Proposition 6. *If M and U engage in n successive dialogues under $\mathcal{P}''_{OM}$ with different subjects, update using W3 after, and are not deadlocked about any of the assertions made during the dialogues, then:*

$$\lim_{n\to\infty} DA(\Sigma_M, \Sigma_U) = 1$$

Proof. Under the conditions stated, Proposition 5 tells us that for each dialogue, either the degree of agreement will increase after that dialogue, or the agents already had the same acceptable argument for the subject of the dialogue. Since the subject changes after each dialogue, this means that as $n \to \infty$, either the degree of agreement increases monotonically, or the agents had exactly teh same set of propositions to begin with (and so had the same acceptable argument for every *subject). In the former case the degree of agreement increases to 1, in the latter case it was 1 to begin with.*

We need the condition about the dialogues having different subjects to prevent the case in which the agents keep having the same dialogue (or small finite set of dialogues) and the degree of agreement never moves beyond some value $\epsilon < 1$. In addition, as the proof points out, there is a degenerate case of "convergence" in which the two agents started out with identical knowledge bases. However, except for this case the convergence is real, and seems likely to be quick. Given Proposition 4, we know that the degree of agreement of the agents will increase by at least one proposition (the subject of the dialogue) each time, and so convergence will require at most N rounds of dialogue, where $N = |\Sigma_M \cup \Sigma_U|$[5]. Finally, we should mention that the condition on deadlock is required for the theorem as stated, but might be relaxed without serious effect on what happens in real dialogues — if the agents are deadlocked on some set of propositions, but this set is small compared with $|\Sigma_M \cup \Sigma_U|$, then the degree of agreement will approach 1.

4.3 Lying and Modelling Other Agents

The results so far concentrate on changes to Σ_M and Σ_U. We can also derive convergence results for the sets of lies each agent has told, $J_{M,U}$ and $J_{U,M}$, and for the models each agent has of the other, $\Gamma_{M,U}$ and $\Gamma_{U,M}$. Let's start by considering $\Gamma_{M,U}$ and $\Gamma_{U,M}$, and extend our update procedure W3 so that at the end of a dialogue with U, M not only updates its knowledge base Σ_M with all the propositions p for which it has an acceptable argument, but also updates its explicit model of U with information it knows that U now accepts. With this additional information $\Delta_M = \Sigma_M \cup \Gamma_{M,U} \cup CS_U$.

[5] Note though that convergence will require both agents to carry out some persuasion — recall that in Example 2, M did not add q to its knowledge base. For q to be accepted by M, U would have to assert q in some later dialogue.

We need some additional definitions:

Definition 14 (Sound Model). *If $\Gamma_{M,U}$ is the model M has of the beliefs of U, then it is a* sound model *of U if $p \in \Gamma_{M,U}$ iff $p \in \Sigma_U$.*

Definition 15 (Complete Model). *If $\Gamma_{M,U}$ is the model M has of the beliefs of U, then it is a* complete model *of U if $p \in \Sigma_U$ iff $p \in \Gamma_{M,U}$.*

With these we can extend Proposition 6 to get:

Proposition 7. *If M and U engage in n successive dialogues under $\mathcal{P}''_{OM}$ with different subjects, and update using W3 after, then as $n \rightarrow \infty$, $\Gamma_{M,U}$ becomes a sound and complete model of U.*

Proof. *Clearly $\Gamma_{M,U}$ is sound and complete if $DA(\Gamma_{M,U}, \Sigma_U) = 1$. Since updating $\Gamma_{M,U}$ takes place in the same way as updating Σ_M, the result follows directly from Proposition 6.*

Thus if they talk for long enough, one agent will converge on a sound and complete model of the other's beliefs. As our discussion of Proposition 6 argues, the number of dialogues required for this convergence is linear in the size of the agents' knowledge bases.

Finally, for agents that are lying, we need to add in the $J_{i,j}$ so that $\Delta_M = \Sigma_M \cup J_{M,U} \cup CS_U$. Recall that the idea of $J_{M,U}$ is that it records things that M believes are false, but uses to build arguments that it seeks to persuade U with — the arguments are not acceptable to M (and in [18] we introduce new semantics for *assert* to deal with this) — and records in order to attempt to only assert things to U that are consistent with the contents of $J_{M,U}$. In such a situation, what M wishes to avoid is being caught in a lie:

Definition 16 (Caught in a Lie). *An agent is* caught in a lie over *p if it is forced to assert both p and $\neg p$ in the same dialogue.*

We have to define being caught in a lie like this, rather than, for example, as the assertion of p and $\neg p$ in different dialogues, since an agent may do this innocently, having changed the status of p in between.

Proposition 8. *If M and U engage in a n successive dialogues under $\mathcal{P}''_{OM}$ with different subjects, then if M lies to U about p and the probability of M being caught in a lie over p is denoted by $\Pr(c(p))$, then:*

$$\lim_{n \rightarrow \infty} \Pr(c(p)) = 1$$

Proof. *If M is in an open-mouthed dialogue with U, M always has to back up its position on every proposition p, and this involves stating the support S, where S may be drawn from Σ_M or $J_{M,U}$. Given what U utters, there is some probability that a given proposition k will be required to be asserted as such support, P_k (we allow this to vary from proposition to proposition). Assuming the probabilities of needing to assert p and $\neg p$ are independent, the probability that M will be caught in a lie is thus $\Pr(c(p)) = P_p \cdot P_{\neg p}$ (which may be very small), and so the probability of not being caught is $1 - P_p \cdot P_{\neg p}$, which is, by definition, less than 1. After n dialogues, the probability of not being caught, $1 - \Pr(c(p)) = (1 - P_p \cdot P_{\neg p})^n$, and this will converge to 0 as n tends to ∞. Thus the result holds.*

Indeed, the result holds even if P_p and $P_{\neg p}$ are not independent—simply replace $P_p \cdot P_{\neg p}$ with $P_{p,\neg p}$, and so long as this is not zero, as long as it is possible that M will be caught, the probability of being caught converges to 1 as the number of dialogues increases.

In other words, the more dialogues that M and U engage in, the greater the chance that M will be caught in a lie. This result depends only on the properties of $\mathcal{P}''_{OM}$ (in a dialogue under $\mathcal{P}''_{CM}$, M would not have to produce grounds) and not the properties of any update operator.

5 Conclusions

This paper has extended the work of [14], which identified the range of possible outcomes of argumentation-based dialogues. Here we have considered what happens at the end of a dialogue—that is what mechanisms are suitable for altering an agent's record of what it believes as a result of a dialogue—and how what happens at the end of a dialogue impacts how an agent's beliefs change after a sequence of dialogues. Our main result is that the way the beliefs change over this sequence depend on the properties of the dialogues themselve, and under certain circumstances, the beliefs of two agents tend to converge as the number of dialogues they engage in grows.

There are three ways that we intend to pursue extensions to this work. One is to consider the mechanisms we have for updating beliefs at the end of a dialogue from the perspective of belief revision [8]. The mechanism we proposed here can clearly be considered as a belief revision mechanism, the question is whether it conforms to the standard properties for such a mechanism. The second extension we plan is to work back towards the results obtained in [14]. That work, in contrast to ours, considered the results of just a single dialogue, and made precise predictions about the outcome based on the contents of the participating agents' knowledge bases. Our work looks at the outcomes of a sequence of dialogues in very general terms, and we would like to see if we can make more precise predictions if we look at the contents of the particpants' knowledge bases in more detail. Finally we intend to look at other forms of open and closed mouth dialogues — the ones we have considered here are two variants of a single protocol — seeking to identify what properties hold for open and closed mouth dialogues in general.

Acknowledgments. This work was partially supported by NSF REC-02-19347, NSF IIS-0329037, and EU PF6-IST 002307 (ASPIC), and has benefited from conversations with Leila Amgoud, Eva Cogan, Peter McBurney and Mike Wooldridge. We are grateful to anonymous reviewers of a previous version of this paper for helping us to clarify our thoughts on this subject.

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. B. Chaib-Draa and F. Dignum. Trends in agent communication language. *Computational Intelligence*, 18(2):89–101, 2002.
3. 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.
4. 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.
5. R. A. Flores and R. C. Kremer. To commit or not to commit. *Computational Intelligence*, 18(2):120–173, 2002.
6. D. M. Gabbay and J. Woods. More on non-cooperation in Dialogue Logic. *Logic Journal of the IGPL*, 9(2):321–339, 2001.
7. D. M. Gabbay and J. Woods. Non-cooperation in Dialogue Logic. *Synthese*, 127(1-2): 161–186, 2001.
8. P. Gärdenfors. *Knowledge in Flux*. MIT Press, 1988.
9. C. L. Hamblin. Mathematical models of dialogue. *Theoria*, 37:130–155, 1971.
10. S. Kraus, K. Sycara, and A. Evenchik. Reaching agreements through argumentation: a logical model and implementation. *Artificial Intelligence*, 104(1–2):1–69, 1998.
11. P. McBurney and S. Parsons. Representing epistemic uncertainty by means of dialectical argumentation. *Annals of Mathematics and Artificial Intelligence*, 32(1–4):125–169, 2001.
12. S. Parsons, P. McBurney, and M. Wooldridge. Some preliminary steps towards a meta-theory for formal inter-agent dialogues. In Iyad Rahwan, editor, *Proceedings of the 1st International Workshop on Argumentation in Multiagent Systems*, New York, 2004.
13. 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.
14. 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.
15. 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.
16. 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.
17. E. Sklar and S. Parsons. Towards the application of argumentation-based dialogues for education. In N. R. Jennings, C. Sierra, E. Sonenberg, and M. Tambe, editors, *Proceedings of the 3rd International Conference on Autonomous Agents and Multi-Agent Systems*. IEEE Press, 2004.
18. E. Sklar, S. Parsons, and M. Davies. When is it okay to lie? a simple model of contraditcion in agent-based dialogues. In *Proceedings of the First Workshop on Argumentation in Muliagent Systems*, 2004.
19. K. Sycara. Argumentation: Planning other agents' plans. In *Proceedings of the Eleventh Joint Conference on Artificial Intelligence*, pages 517–523, 1989.
20. R. D. Tennent. *Semantics of Programming Languages*. International Series in Computer Science. Prentice Hall, Hemel Hempstead, UK, 1991.
21. D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press, Albany, NY, USA, 1995.