Choosing persuasive arguments for action: a technical report*

Elizabeth Black
University of Utrecht
lizblack@cs.uu.nl

Katie Atkinson
University of Liverpool
katie@liverpool.ac.uk

February 4, 2011

Abstract

We present a dialogue system that allows agents to exchange arguments in order to come to an agreement on how to act. When selecting arguments to assert, an agent uses a model of what is important to the recipient agent. The system lets the agents agree to an action that each finds acceptable, but does not necessarily demand that they resolve their differing preferences. We present an analysis of the behaviour of our system and develop a mechanism with which an agent can develop a model of another's preferences.

1 Introduction

Agents engaged in a deliberation dialogue share the aim to reach an agreement about how to act in order to achieve a particular goal [19]. Deliberating agents are co-operative in that they each aim for agreement; however, individually they may each wish to influence the outcome in their own favour. We assume that agents do not mislead one another and will come to an agreement wherever possible; however, each agent aims to satisfy its own preferences.

We build on an existing system for deliberation that provides a dialogue strategy which allows agents to come to an agreement about how to act, despite the fact that they may have different preferences and thus may each be agreeing for different reasons [6]; this system couples a dialectical setting with formal methods for argument evaluation and allows strategic manoeuvring in order to influence the dialogue outcome. The analysis of the simple strategy defined in [6] provides a foundation upon which we build here in order to investigate a more sophisticated strategy that takes into account the *proponent's* (that is, the agent who asserts the argument) perception of the *recipient* (the agent who receives the argument).

We present a novel deliberation strategy, which allows a proponent to use its perception of the recipient to guide its dialogue behaviour, and we perform a detailed

*This paper is the same as [7] but with proofs included.

analysis of the behaviour of our system. Such an analysis is crucial as it allows one to determine which applications our system is suitable for; it can also guide the development of new deliberation strategies with properties that do not hold for the strategy presented here.

The type of investigation presented here is commonly missing from comparable dialogue systems (in part because historically such work has focussed on defining rules to constrain dialogue interaction, rather than on strategies for manoeuvring within the constraints); our analysis gives us a better understanding of how the strategy design affects dialogue outcome, which is crucial if we are to deploy dialogue systems effectively.

We also present a mechanism that enables agents to model preference information about others. When presenting proposals to others, a key consideration is how the proposal appears to the recipient; if an option presented does not meet the preferences of other dialogue participants, then it will be rejected. We present a mechanism with which an agent can develop a model of what is important to another agent and show how it can be used to help agents make proposals that are more likely to be agreeable.

Our paper is structured thus: in Sect. 2 we present the reasoning mechanism (recapitulated from [6]) through which agents can construct and evaluate arguments about action; in Sect. 3 we define the dialogue system, which is adapted from that presented in [6] in order to allow a proponent to take into account its model of the recipient when selecting an utterance to make; a detailed analysis of the behaviour of the dialogue system is given in Sect. 4 and Sect. 5 presents our mechanism for modelling another agent; we consider related work in Sect. 6; Sect. 7 concludes the paper.

2 Practical arguments

Our account is based upon a popular approach to argument characterisation, whereby argumentation schemes and critical questions are used as presumptive justification for generating arguments and attacks between them [18]. Arguments are generated by an agent instantiating a *scheme for practical reasoning* which makes explicit the following elements: the initial circumstances where action is required; the action to be taken; the new circumstances that arise through acting; the goal to be achieved; the social value promoted by realising the goal in this way. The scheme is associated with a set of characteristic critical questions (CQs) that can be used to identify challenges to proposals for action that instantiate the scheme. An unfavourable answer to a CQ will identify a potential flaw in the argument. Since the scheme makes use of what are termed as 'values', this caters for arguments based on subjective preferences as well as more objective facts. Such values represent qualitative social interests that an agent wishes (or does not wish) to uphold by realising the goal stated [3].

To enable the practical argument scheme and critical questions approach to be precisely formalised for use in automated systems, in [2] it was defined in terms of an Action-based Alternating Transition System (AATS) [20], which is a structure for modelling game-like multi-agent systems where the agents can perform actions in order to attempt to control the system in some way. Hence, we use an adaptation of the formalisms (first presented in [5]) to define a *Value-based Transition System* (VATS) as

follows.

Definition 1: *A* **value-based transition system** *(***VATS***) for an agent x, denoted S^x, is $\langle Q^x, q_0^x, Ac^x, Av^x, \rho^x, \tau^x, \Phi^x, \pi^x, \delta^x \rangle$ s.t.:*
Q^x is a finite set of states*;*
$q_0^x \in Q^x$ is the designated initial state*;*
Ac^x is a finite set of actions*;*
Av^x is a finite set of values*;*
$\rho^x : Ac^x \mapsto 2^{Q^x}$ is an action precondition function*, which for each action $a \in Ac^x$ defines the set of states $\rho(a)$ from which a may be executed;*
$\tau^x : Q^x \times Ac^x \mapsto Q^x$ is a partial system transition function*, which defines the state $\tau^x(q, a)$ that would result by the performance of a from state q—n.b. as this function is partial, not all actions are possible in all states (cf. the precondition function above);*
Φ^x is a finite set of atomic propositions*;*
$\pi^x : Q^x \mapsto 2^{\Phi^x}$ is an interpretation function*, which gives the set of primitive propositions satisfied in each state: if $p \in \pi^x(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q; and*
$\delta^x : Q^x \times Q^x \times Av^x \mapsto \{+, -, =\}$ is a valuation function*, which defines the* status *(promoted ($+$), demoted ($-$), or neutral ($=$)) of a value $v \in Av^x$ ascribed by the agent to the transition between two states: $\delta^x(q, q', v)$ labels the transition between q and q' with respect to the value $v \in Av^x$.*
Note, $Q^x = \emptyset \leftrightarrow Ac^x = \emptyset \leftrightarrow Av^x = \emptyset \leftrightarrow \Phi^x = \emptyset$.

An agent has its own individual VATS; any two agents' VATSs are *not necessarily* the same. Given its VATS, an agent can now instantiate the practical reasoning argument scheme in order to construct arguments for (or against) actions to achieve a particular goal because they promote (or demote) a particular value.

Definition 2: *An* **argument** *constructed by an agent x from its VATS S^x is a 4-tuple $A = \langle a, p, v, s \rangle$ s.t.: $q_x = q_0^x$; $a \in Ac^x$; $\tau^x(q_x, a) = q_y$; $p \in \pi^x(q_y)$; $v \in Av^x$; $\delta^x(q_x, q_y, v) = s$ where $s \in \{+, -\}$. We define the functions: $\mathsf{Act}(A) = a$; $\mathsf{Goal}(A) = p$; $\mathsf{Val}(A) = v$; $\mathsf{Sign}(A) = s$. If $\mathsf{Sign}(A) = +(-resp.)$, then we say A is a* **positive** *(***negative** *resp.) argument* **for** *(***against** *resp.) action a. We denote the* **set of all arguments an agent** *x* **can construct from** *S^x as $Args^x$; we let $Args_p^x = \{A \in Args^x \mid \mathsf{Goal}(A) = p\}$. The set of* **values** *for a set of arguments $\mathcal{X}$ is defined as $\mathsf{Vals}(\mathcal{X}) = \{v \mid A \in \mathcal{X}$ and $\mathsf{Val}(A) = v\}$.*

If we take a particular argument for an action, it is possible to generate attacks on that argument by posing the various CQs related to the practical reasoning argument scheme. In [2], details are given of how the reasoning with the argument scheme and posing CQs is split into three stages: *problem formulation*, where the agents decide on the facts and values relevant to the particular situation under consideration for constructing and, if necessary, aligning their VATSs; *epistemic reasoning*, where the agents determine the current situation with respect to the structure formed at the previous stage; and *action selection*, where the agents develop, and evaluate, arguments and counter arguments about what to do. Here, we assume that the agents' problem formulation and epistemic reasoning are sound and that any dispute between them relating to these stages has been resolved; hence, we do not consider the CQs that arise in these stages. That leaves CQ5-CQ11 for consideration (as numbered in [2]):

CQ5: Are there alternative ways of realising the same consequences?
CQ6: Are there alternative ways of realising the same goal?
CQ7: Are there alternative ways of promoting the same value?
CQ8: Does the action have a side effect that demotes the value?
CQ9: Does the action have a side effect that demotes another value?
CQ10: Does doing the action promote some other value?
CQ11: Does doing the action preclude some other action that would promote some other value?

We do not consider CQ5 or CQ11 further, as the focus here is to agree to an action that achieves the *goal*; hence, incidental consequences (CQ5) and other potentially precluded actions (CQ11) are of no interest. We focus instead on CQ6-CQ10; agents participating in a deliberation dialogue use these CQs to identify attacks on proposed arguments for action. These CQs generate a set of arguments for and against different actions to achieve a particular goal, where each argument is associated with a motivating value. To evaluate the status of these arguments we use a Value Based Argumentation Framework (VAF), an extension of the argumentation frameworks (AF) of Dung [10] (introduced in [3]). In an AF an argument is admissible with respect to a set of arguments S if all of its attackers are attacked by some argument in S, and no argument in S attacks an argument in S. In a VAF an argument succeeds in defeating an argument it attacks if its value is ranked higher than the value of the argument attacked; a particular ordering of the values is characterised as an *audience*. Arguments in a VAF are admissible with respect to an audience A and a set of arguments S if they are admissible with respect to S in the AF which results from removing all the attacks which are unsuccessful given the audience A. A maximal admissible set of a VAF is known as a *preferred extension*.

Although VAFs are often considered abstractly, here we give an instantiation in which we define the attack relation between the arguments. Condition 1 of the following attack relation allows for CQ8 and CQ9; condition 2 allows for CQ10; condition 3 allows for CQ6 and CQ7. Note that attacks generated by condition 1 are not symmetrical, whilst those generated by conditions 2 and 3 are.

Definition 3: *An* **instantiated value-based argumentation framework (iVAF)** *is defined by a tuple $\langle \mathcal{X}, \mathcal{A} \rangle$ s.t. $\mathcal{X}$ is a finite set of arguments and $\mathcal{A} \subset \mathcal{X} \times \mathcal{X}$ is the* **attack relation**. *A pair $(A_i, A_j) \in \mathcal{A}$ is referred to as "A_i attacks A_j" or "A_j is attacked by A_i". For two arguments $A_i = \langle a, p, v, s \rangle$, $A_j = \langle a', p', v', s' \rangle \in \mathcal{X}$, $(A_i, A_j) \in \mathcal{A}$ iff $p = p'$ and either: (1) $a = a'$, $s = -$ and $s' = +$; or (2) $a = a'$, $v \neq v'$ and $s = s' = +$; or (3) $a \neq a'$ and $s = s' = +$.*

An **audience** *for an agent x over the values V is a binary relation $\mathcal{R}^x \subset V \times V$ that defines a* total order *over V where exactly one of (v, v'), (v', v) is a member of $\mathcal{R}^x$ for any distinct v, $v' \in V$. If $(v, v') \in \mathcal{R}^x$ we say that v is* **preferred to** *v', denoted $v \succ_x v'$. We say that an argument A_i is* **preferred to** *the argument A_j in the audience $\mathcal{R}^x$, denoted $A_i \succ_x A_j$, iff $\mathsf{Val}(A_i) \succ_x \mathsf{Val}(A_j)$. If R^x is an audience over the values V for the iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$, then $\mathsf{Vals}(\mathcal{X}) \subseteq V$.*

We use the term 'audience' to be consistent with the literature. Note, however, audience does not refer to the preference of a *set* of agents; rather, it represents a particular agent's preferences.

Given an iVAF and a particular agent's audience, we can determine acceptability of an argument as follows. Note that if an attack is symmetric, then an attack only succeeds in defeat if the attacker is *more preferred* than the argument being attacked; however, as in [3], if an attack is asymmetric, then an attack succeeds in defeat if the attacker is *at least as preferred* as the argument being attacked.

Definition 4: *Let $\mathcal{R}^x$ be an audience and let $\langle \mathcal{X}, \mathcal{A} \rangle$ be an iVAF.*
For $(A_i, A_j) \in \mathcal{A}$ s.t. $(A_j, A_i) \notin \mathcal{A}$, A_i **defeats** *A_j under $\mathcal{R}^x$ if $A_j \not\succ_x A_i$.*
For $(A_i, A_j) \in \mathcal{A}$ s.t. $(A_j, A_i) \in \mathcal{A}$, A_i **defeats** *A_j under $\mathcal{R}^x$ if $A_i \succ_x A_j$.*
An argument $A_i \in \mathcal{X}$ is **acceptable w.r.t** *S under $\mathcal{R}^x$ ($S \subseteq \mathcal{X}$) if: for every $A_j \in \mathcal{X}$ that defeats A_i under $\mathcal{R}^x$, there is some $A_k \in S$ that defeats A_j under $\mathcal{R}^x$.*
A subset S of $\mathcal{X}$ is **conflict-free** *under $\mathcal{R}^x$ if no argument $A_i \in S$ defeats another argument $A_j \in S$ under $\mathcal{R}^x$.*
A subset S of $\mathcal{X}$ is **admissible** *under $\mathcal{R}^x$ if: S is conflict-free in $\mathcal{R}^x$ and every $A \in S$ is acceptable w.r.t S under $\mathcal{R}^x$.*
A subset S of $\mathcal{X}$ is a **preferred extension** *under $\mathcal{R}^x$ if it is a maximal admissible set under $\mathcal{R}^x$.*
An argument A is **acceptable** *in the iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$ under audience $\mathcal{R}^x$ if there is* some *preferred extension containing it.*

We can define a *winning value* for an iVAF and a particular agent's audience: a value is a winning value for an agent if there is an argument that promotes that value and is acceptable under the agent's audience. Note that the winning value is not necessarily the most preferred, rather the one that motivates some undefeated argument *for* an action.

Definition 5: *Let $\mathcal{R}^x$ be an audience and $\langle \mathcal{X}, \mathcal{A} \rangle$ be an iVAF. The value v is a* **winning value** *in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$ iff $\exists A \in \mathcal{X}$ s.t. A is acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$, $\mathsf{Sign}(A) = +$ and $\mathsf{Val}(A) = v$.*

It is clear (from the definition of an iVAF) that if all the arguments that appear in an iVAF relate to the same goal, then there is at most one winning value for a given audience.

Proposition 1: *Let $\mathcal{R}^x$ be an audience and let $\langle \mathcal{X}, \mathcal{A} \rangle$ be an iVAF. If $\forall A, A' \in \mathcal{X}$, $\mathsf{Goal}(A) = \mathsf{Goal}(A')$ and v and v' are both winning values in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$, then $v = v'$.*

Proof: *Assume $v1$ is a winning value in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$, therefore $\exists A1 = (a1, p, v1, +) \in \mathcal{X}$ s.t. $A1$ is acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$. Assume $v2$ is a winning value in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$, therefore $\exists A2 = (a2, p, v2, +) \in \mathcal{X}$ s.t. $A2$ is acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$. Assume that $v1 \neq v2$ are distinct. From Def 3., $A1$ and $A2$ attack one another. As $A1$ is acceptable, it must be the case that $v1$ is at least as preferred as $v2$. As $A2$ is acceptable, it must be the case that $v2$ is at least as preferred as $v1$. Exactly one of $(v1, v2)$, $(v2, v1)$ must be in $\mathcal{R}^x$ (Def. 3) — contradiction.* □

We have defined a mechanism with which an agent can determine attacks between arguments for and against actions; it can then use an ordering over the values that motivate such arguments (its audience) in order to determine their acceptability. Next, we define our dialogue system, which significantly enhances that presented in [6] in order to allow a proponent to take into account its perception of the recipient's audience.

Move	Format
$open$	$\langle x, \mathsf{open}, \gamma \rangle$
$assert$	$\langle x, \mathsf{assert}, A \rangle$
$agree$	$\langle x, \mathsf{agree}, a \rangle$
$close$	$\langle x, \mathsf{close}, \gamma \rangle$

Table 1: Format for moves used in deliberation dialogues: γ is a goal; a is an action; A is an argument; x is an agent identifier.

3 Dialogue system

The communicative acts in a dialogue are called *moves*. We assume that there are always exactly two agents (*participants*) taking part in a dialogue, each with its own identifier taken from the set $\mathcal{I} = \{Ag1, Ag2\}$. Each participant takes it in turn to make a move to the other. We refer to participants using the variables x and $\overline{x}$ such that: x is 1 if and only if $\overline{x}$ is 2; x is 2 if and only if $\overline{x}$ is 1.

A move in our system is of the form $\langle Agent, Act, Content \rangle$. $Agent$ is the identifier of the agent generating the move, Act is the type of move, and the $Content$ gives the details of the move. The format for moves used in deliberation dialogues is shown in Table 1, and the set of all moves meeting the format defined in Table 1 is denoted $\mathcal{M}$. Note, $\mathsf{Sender} : \mathcal{M} \mapsto \mathcal{I}$ is a function such that $\mathsf{Sender}(\langle Agent, Act, Content \rangle) = Agent$.

We now informally explain the different types of move: an *open* move $\langle x, \mathsf{open}, \gamma \rangle$ opens a dialogue to agree on an action to achieve the goal γ; an *assert* move $\langle x, \mathsf{assert}, A \rangle$ asserts an argument A for or against an action to achieve a goal that is the topic of the dialogue; an *agree* move $\langle x, \mathsf{agree}, a \rangle$ indicates that x agrees to performing action a to achieve the topic; a *close* move $\langle x, \mathsf{close}, \gamma \rangle$ indicates that x wishes to end the dialogue.

A dialogue is simply a sequence of moves, each of which is indexed by the timepoint when the move was made. Exactly one move is made at each timepoint.

Definition 6: *A* **dialogue**, *denoted D^t, is a sequence of moves $[m_1, \ldots, m_t]$ involving two participants in $\mathcal{I} = \{Ag1, Ag2\}$, where $t \in \mathbb{N}$ and the following conditions hold: (1) m_1 is a move of the form $\langle x, \mathsf{open}, \gamma \rangle$ where $x \in \mathcal{I}$; (2) $\mathsf{Sender}(m_s) \in \mathcal{I}$ for $1 \leq s \leq t$; (3) $\mathsf{Sender}(m_s) \neq \mathsf{Sender}(m_{s+1})$ for $1 \leq s < t$. The* **topic** *of the dialogue D^t is returned by $\mathsf{Topic}(D^t) = \gamma$. The set of all dialogues is denoted $\mathcal{D}$.*

The first move of a dialogue D^t must always be an open move (condition 1 of the previous definition), every move of the dialogue must be made by a participant (condition 2), and the agents take it in turns to send a move (condition 3). In order to terminate a dialogue, either: two close moves must appear one immediately after the other in the sequence (a *matched-close*); or two moves agreeing to the same action must appear one immediately after the other in the sequence (an *agreed-close*).

Definition 7: *Let D^t be a dialogue s.t. $\mathsf{Topic}(D^t) = \gamma$. We say that either : m_s ($1 < s \leq t$) is a* **matched-close for** *D^t iff $m_{s-1} = \langle x, \mathsf{close}, \gamma \rangle$ and $m_s = \langle \overline{x}, \mathsf{close}, \gamma \rangle$; else m_s ($1 < s \leq t$) is an* **agreed-close for** *D^t iff $m_{s-1} = \langle x, \mathsf{agree}, a \rangle$ and $m_s = \langle \overline{x}, \mathsf{agree}, a \rangle$. We say D^t has a* **failed outcome** *iff m_t is a matched-close, whereas we*

say D^t has a **successful outcome** *of a iff $m_t = \langle x, \mathsf{agree}, a\rangle$ is an agreed-close.*

So a matched-close or an agreed-close will terminate a dialogue D^t but only if D^t has not already terminated.

Definition 8: *Let D^t be a dialogue. D^t* **terminates at** *t iff m_t is a matched-close or an agreed-close for D^t and $\neg\exists s$ s.t. $s < t$, D^t* **extends** *D^s (i.e. the first s moves of D^t are the same as the sequence D^s) and D^s terminates at s.*

We shortly give the particular protocol and strategy functions that allow agents to generate deliberation dialogues. First, we introduce some subsidiary definitions. At any point in a dialogue, an agent x can construct an iVAF from the union of the arguments it can construct from its VATS and the arguments that have been asserted by the other agent; we call this x's *dialogue iVAF*.

Definition 9: *A* **dialogue iVAF** *for an agent x participating in a dialogue D^t is denoted* $\mathrm{dVAF}(x, D^t)$. *If D^t is the sequence of moves $= [m_1, \ldots, m_t]$, then* $\mathrm{dVAF}(x, D^t)$ *is the iVAF $\langle \mathcal{X}, \mathcal{A}\rangle$ where $\mathcal{X} = Args^x_{\mathsf{Topic}(D^t)} \cup \{A \mid \exists m_k = \langle \overline{x}, assert, A\rangle (1 \leq k \leq t)\}$.*

An action is *agreeable* to an agent x if and only if there is some argument *for* that action that is acceptable in x's dialogue iVAF under the audience that represents x's preference over values. Note that the set of actions that are agreeable to an agent may change over the course of the dialogue.

Definition 10: *An action a is* **agreeable** *in the iVAF $\langle \mathcal{X}, \mathcal{A}\rangle$ under the audience $\mathcal{R}^x$ iff $\exists A = \langle a, \gamma, v, +\rangle \in \mathcal{X}$ s.t. A is acceptable in $\langle \mathcal{X}, \mathcal{A}\rangle$ under $\mathcal{R}^x$. We denote the* **set of all actions that are agreeable to an agent** *x* **participating in a dialogue** *D^t as* $\mathsf{AgActs}(x, D^t)$, *s.t. $a \in \mathsf{AgActs}(x, D^t)$ iff a is agreeable in* $\mathrm{dVAF}(x, D^t)$ *under $\mathcal{R}^x$.*

A protocol is a function that returns the set of moves that are permissible for an agent to make at each point in a particular type of dialogue. Here we give a protocol for deliberation. It takes the dialogue that the agents are participating in and the identifier of the agent whose turn it is and returns the set of permissible moves.

Definition 11: *The* **deliberation protocol** *for agent x is a function $\mathsf{Protocol}_x : \mathcal{D} \mapsto \wp(\mathcal{M})$. Let D^t be a dialogue ($1 \leq t$) s.t. $\mathsf{Sender}(m_t) = \overline{x}$ and $\mathsf{Topic}(D^t) = \gamma$.*

$$\mathsf{Protocol}_x(D^t) = P^{\mathsf{ass}}_x(D^t) \cup P^{\mathsf{ag}}_x(D^t) \cup \{\langle x, \mathsf{close}, \gamma\rangle\}$$

where the following are sets of moves and $x' \in \mathcal{I}$:

$$\begin{aligned}
P^{\mathsf{ass}}_x(D^t) \quad &= \{\langle x, \mathsf{assert}, A\rangle \mid \mathsf{Goal}(A) = \gamma \\
&\quad \textbf{and} \\
&\quad \neg\exists m_{t'} = \langle x', assert, A\rangle (1 < t' \leq t) \\
P^{\mathsf{ag}}_x(D^t) \quad &= \{\langle x, \mathsf{agree}, a\rangle \mid \ \textbf{either} \\
&\quad (1) m_t = \langle \overline{x}, agree, a\rangle \\
&\quad \textbf{else} \\
&\quad (2)(\exists m_{t'} = \langle \overline{x}, assert, \langle a, \gamma, v, +\rangle\rangle (1 < t' \leq t)) \\
&\qquad \textbf{and} \\
&\qquad (\ \textbf{if } \exists m_{t''} = \langle x, agree, a\rangle \\
&\qquad \textbf{then } \exists A, m_{t'''} = \langle x, assert, A\rangle \\
&\qquad\quad (t'' < t''' \leq t))\}
\end{aligned}$$

The protocol states that it is permissible to assert an argument for or against an action to achieve the topic of the dialogue as long as that argument has not previously been asserted in the dialogue. An agent can agree to an action that has been agreed to by the other agent in the preceding move (condition 1 of P_x^{ag}); otherwise an agent x can agree to an action that has been proposed by the other participant (condition 2 of P_x^{ag}) as long as if x has previously agreed to that action, then x has since then asserted some new argument. This is because we want to avoid the situation where an agent keeps repeatedly agreeing to an action that the other agent will not agree to: if an agent makes a move agreeing to an action and the other agent does not wish to also agree to that action, then the first agent must before being able to repeat its agree move introduce some new argument that may convince the second agent to agree. Agents may always make a close move.

We have thus defined a protocol that determines which moves it is permissible to make during a dialogue; however, an agent still has considerable choice when selecting which of these permissible moves to make. In order to select one of the permissible moves, an agent uses a particular strategy. Informally, the strategy that we will shortly define selects a move as follows: if it is permissible to make a move agreeing to an *agreeable* action, then make such an agree move; else, if it is permissible to assert an argument *for* an *agreeable* action, then assert some such argument; else, if it is permissible to assert an argument *against* an action that *is not agreeable*, then assert some such argument; else make a close move. When the strategy results in a choice of more than one agree or assert move, an agent must rely on two further functions for selecting from a set of either permissible assert or permissible agree moves.

When selecting a particular assert move, a proponent makes use of its model of the recipient. In particular, when faced with a choice of arguments to assert, an agent will choose one with a motivating value that it believes is highly ranked by the recipient. Thus, a proponent needs to model what it believes could be the recipient's winning value. We define a function that takes a value and, for a given dialogue and recipient, maps to the interval between 0 and 1; the higher the output of this function, the more the proponent believes that the value is the recipient's winning value.

Definition 12: *A* **recipient value model** *is given by the function* $\mathsf{Models}_x^{\overline{x}} : \mathcal{D} \times Av^{\overline{x}} \mapsto [0,1]$ *(*$x, \overline{x} \in \mathcal{I}$*).*

Note, there are many ways this function could be initialised at the beginning of a dialogue. For example: we could initialise all values to 0.5; information from past interactions could be used to guide the initial values; or in highly co-operative settings it may make sense to assume that the agents share similar views, so the values could be initialised to mirror the proponent's value preference.

A proponent selects an argument to assert as follows: if there is a choice of more than one argument to be asserted, then the agent will choose to assert one such argument such that of all the other arguments it could assert, it does not believe that the values that motivate them are more likely to be the recipient's winning value than that which motivates the selected argument.

Definition 13: *Let* $\Psi = \{\langle x, \mathsf{assert}, A_1\rangle, \ldots, \langle x, \mathsf{assert}, A_k\rangle\}$.
The function $\mathsf{Pick}_{\mathsf{ass}}$ *returns a* **chosen assert move** *s.t.*
if $\mathsf{Pick}_{\mathsf{ass}}(\Psi) = \langle x, \mathsf{assert}, A_i\rangle$ *(*$1 \leq i \leq k$*), then* $\neg\exists j$ *(*$1 \leq j \leq k$*) s.t.* $\mathsf{Models}_x^{\overline{x}}(\mathsf{Val}(A_j)) >$

$\mathsf{Strat}_x(D^t) = \mathsf{Pick}_{\mathsf{ag}}(S_x^{\mathsf{ag}})(D^t)$ iff $S_x^{\mathsf{ag}}(D^t) \neq \emptyset$
$\mathsf{Strat}_x(D^t) = \mathsf{Pick}_{\mathsf{ass}}(S_x^{\mathsf{prop}})(D^t)$ iff $S_x^{\mathsf{ag}}(D^t) = \emptyset$ and $S_x^{\mathsf{prop}}(D^t) \neq \emptyset$
$\mathsf{Strat}_x(D^t) = \mathsf{Pick}_{\mathsf{ass}}(S_x^{\mathsf{att}})(D^t)$ iff $S_x^{\mathsf{ag}}(D^t) = S_x^{\mathsf{prop}}(D^t) = \emptyset$ and $S_x^{\mathsf{att}}(D^t) \neq \emptyset$
$\mathsf{Strat}_x(D^t) = \langle x, \mathsf{close}, \mathsf{Topic}(D^t)\rangle$ iff $S_x^{\mathsf{ag}}(D^t) = S_x^{\mathsf{prop}}(D^t) = S_x^{\mathsf{att}}(D^t) = \emptyset$

where the choices for the moves are given by the following subsidiary functions with $x' \in \{x, \overline{x}\}$ and $\mathsf{Topic}(D^t) = \gamma$

$S_x^{\mathsf{ag}}(D^t) = \{\langle x, \mathsf{agree}, a\rangle \in P_x^{\mathsf{ag}}(D^t) \mid a \in \mathsf{AgActs}(x, D^t)\}$

$S_x^{\mathsf{prop}}(D^t) = \{\langle x, \mathsf{assert}, A\rangle \in P_x^{\mathsf{ass}}(D^t) \mid A \in Args_\gamma^x, \mathsf{Act}(A) = a, \mathsf{Sign}(A) = + \text{ and } a \in \mathsf{AgActs}(x, D^t)\}$

$S_x^{\mathsf{att}}(D^t) = \{\langle x, \mathsf{assert}, A\rangle \in P_x^{\mathsf{ass}}(D^t) \mid A \in Args_\gamma^x, \mathsf{Act}(A) = a, \mathsf{Sign}(A) = -, a \notin \mathsf{AgActs}(x, D^t)$ and $\exists m_{t'} = \langle x', \mathsf{assert}, A'\rangle (1 \leq t' \leq t)$ s.t. $\mathsf{Act}(A') = a$ and $\mathsf{Sign}(A') = +\}$

Figure 1: The **strategy** function selects a move according to the following preference ordering (starting with the most preferred): an agree (ag), a proposing assert (prop), an attacking assert (att), a close (close).

$\mathsf{Models}_x^{\overline{x}}(\mathsf{Val}(A_i))$

We also require a function that allows an agent to select a particular permissible move to make from a set of agree moves (denoted $\mathsf{Pick}_{\mathsf{ag}}$). Our analysis in the next section does not depend on the definition of $\mathsf{Pick}_{\mathsf{ag}}$, hence we do not define $\mathsf{Pick}_{\mathsf{ag}}$ here but leave it as a parameter of our system (in its simplest form, $\mathsf{Pick}_{\mathsf{ag}}$ may return an arbitrary agree move from the input set).

We are now able to define a *deliberation strategy*. It takes the dialogue D^t and returns exactly one of the legal moves.

Definition 14: *The* **strategy** *for an agent x is a function $\mathsf{Strat}_x : \mathcal{D} \mapsto \mathcal{M}$ given in Figure 1.*

A *well-formed dialogue* is a dialogue that has been generated by two agents each following this strategy.

Definition 15: *A* **well-formed dialogue** *is a dialogue D^t s.t. $\forall t'$ $(1 \leq t' \leq t)$, $\mathsf{Sender}(m^{t'}) = x$ iff $\mathsf{Strat}_x(D^{t'-1}) = m_{t'}$*

We now give a short example. There are two participants, $Ag1$ and $Ag2$, who have the shared goal of doing something together on Saturday ($ActivityForSat$). The relevant values for this scenario are *company* (C), promoted by spending time with the other agent, *variety* (V) promoted by doing an activity the agent has not done recently, *distance* (D), promoted by doing a nearby activity, and *money* (M), promoted by cheap activities. The participants have the following audiences.

$C \succ_{Ag1} D \succ_{Ag1} V \succ_{Ag1} M$
$M \succ_{Ag2} V \succ_{Ag2} D \succ_{Ag2} C$

To save space, we only consider $Ag1$'s recipient value model of $Ag2$, which is initialised as follows (presumably based on some background knowledge that $Ag1$ has). $\mathsf{Models}_{Ag1}^{Ag2}(D^1, val) = 1$ iff $val = C$; 0.9 iff $val = D$; 0.8 iff $val = V$ or $val = M$.

The agents' initial dialogue iVAFs can be seen in Figs. 2 and 3, where the nodes

represent arguments and are labelled with the action that they are for (or the negation of the action that they are against) and the value they are motivated by. The arcs represent the attack relation and a double circle round a node means that the argument that it represents is acceptable to that agent.

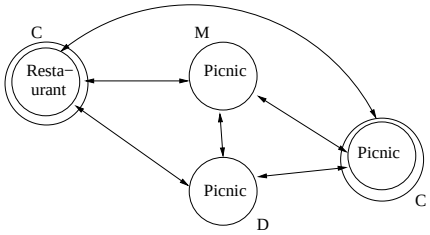

Figure 2: Agent $Ag1$'s dialogue iVAF at $t = 1$, *dVAF*$(Ag1, D^1)$.

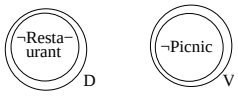

Figure 3: Agent $Ag2$'s dialogue iVAF at $t = 1$, *dVAF*$(Ag2, D^1)$.

Agent $Ag2$ starts the dialogue with the move m_1. At this point there are two arguments that are acceptable to $Ag1$:
$\langle Restaurant, ActivityForSat, C, + \rangle$;
$\langle Picnic, ActivityForSat, C, + \rangle$.
Agent $Ag1$ currently believes that C is most likely the winning value for $Ag2$ (as $\mathsf{Models}^{Ag2}_{Ag1}(D^1, C) = 1$) and so it selects an argument motivated by C to assert.

$m_1 = \langle Ag2, \mathsf{open}, ActivityForSat \rangle$

$m_2 = \langle Ag1, \mathsf{assert}, \langle Restaurant, ActivityForSat, C, + \rangle \rangle$

This new argument is added to $Ag2$'s dialogue iVAF, to give *dVAF*$(Ag2, D^2)$ (Fig. 4).

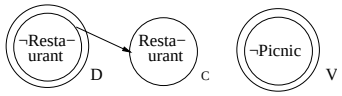

Figure 4: Agent $Ag2$'s dialogue iVAF at $t = 2$, *dVAF*$(Ag2, D^2)$.

As $Ag2$ actually prefers value D to value C, this new argument is not acceptable to it. In fact, there are no actions currently agreeable to $Ag2$ (as there are no acceptable arguments for an action in its dialogue iVAF) and so $Ag2$ makes an attacking move by asserting its argument against going to the restaurant (as it is far away).

$m_3 = \langle Ag2, \mathsf{assert}, \langle Restaurant, ActivityForSat, D, - \rangle \rangle$

This new argument is added to $Ag1$'s dialogue iVAF, to give *dVAF*$(Ag1, D^3)$ (Fig. 5). As $Ag2$ did not agree to $Ag1$'s suggestion to go to a restaurant for good

company, $Ag1$ now has reason to believe that in fact C is unlikely to be the winning value for $Ag2$ and so it decrements its recipient value model for this value from 1 to 0.8: $\mathsf{Models}^{Ag2}_{Ag1}(D^3, C) = 0.8$.

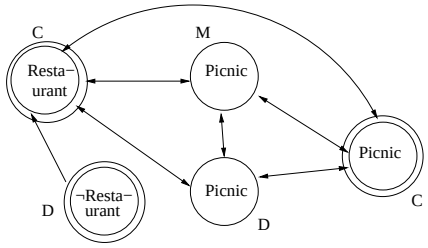

Figure 5: Agent $Ag1$'s dialogue iVAF at $t = 3$, *dVAF*$(Ag1, D^3)$.

Agent $Ag1$ still finds both picnic and restaurant agreeable actions. As it has already asserted its argument for going to the restaurant, it must now choose one of its arguments for going for a picnic to assert. It currently believes that D is likely the winning value for $Ag2$ and so chooses an argument motivated by this value.

$m_4 = \langle Ag1, \mathsf{assert}, \langle Picnic, ActivityForSat, D, + \rangle \rangle$

This new argument is added to $Ag2$'s dialogue iVAF, to give *dVAF*$(Ag2, D^4)$ (Fig. 6). As $Ag2$ in fact prefers value V to value D, the proposed action of going for a picnic is not agreeable to $Ag2$, and so it asserts its argument against this action.

$m_5 = \langle Ag2, \mathsf{assert}, \langle Picnic, ActivityForSat, V, - \rangle \rangle$

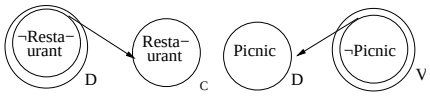

Figure 6: Agent $Ag2$'s dialogue iVAF at $t = 4$, *dVAF*$(Ag2, D^4)$.

This new argument is added to $Ag1$'s dialogue iVAF, to give *dVAF*$(Ag1, D^5)$ (Fig. 7). As $Ag2$ did not agree to $Ag1$'s suggestion to go for a picnic as it is nearby, $Ag1$ now has reason to believe that in fact D is unlikely to be the winning value for $Ag2$ and so it decrements its recipient value model for this value from 0.9 to 0.7: $\mathsf{Models}^{Ag2}_{Ag1}(D^5, D) = 0.7$.

Agent $Ag1$ still finds going for a picnic agreeable, but it now believes that either M or V is likely to be the winning value for $Ag2$. Hence, it asserts its argument for going for a picnic that is motivated by the value M.

$m_6 = \langle Ag1, \mathsf{assert}, \langle Picnic, ActivityForSat, M, + \rangle \rangle$

This new argument is added to $Ag2$'s dialogue iVAF, to give *dVAF*$(Ag2, D^6)$ (Fig. 8). As $Ag1$ is now right in believing that M is the winning value for $Ag1$, $Ag1$ finds this new argument acceptable and so agrees to going for a picnic. Agent $Ag2$ also agrees to this action and the dialogue terminates successfully.

$m_8 = \langle Ag1, \mathsf{agree}, Picnic \rangle$

$m_9 = \langle Ag2, \mathsf{agree}, Picnic \rangle$

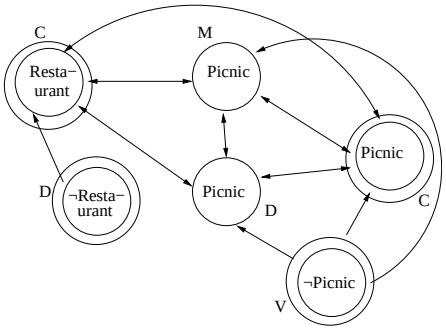

Figure 7: Agent $Ag1$'s dialogue iVAF at $t = 5$, $dVAF(Ag1, D^5)$.

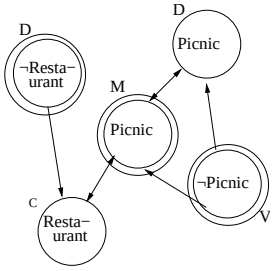

Figure 8: Agent $Ag2$'s dialogue iVAF at $t = 6$, $dVAF(Ag2, D^6)$.

This example illustrates how agents can reach an agreement on an action to achieve a joint goal despite their differing preferences over values; it also shows how an agent may update is model of another's winning value based on their dialogue behaviour.

4 Analysis of the system

In [6], an analysis is given of a more abstract version of the dialogue system discussed here in which neither Pick function is specified, hence the results of [6] all hold for the specialised version of the dialogue system that we present here. In particular: all dialogues generated by our system terminate; if the dialogue terminates with a successful outcome of action a, then a is agreeable to both agents at the end of the dialogue; if there is an action a that is agreeable to both agents when the dialogue terminates, then the dialogue terminates with a successful outcome.

However, for the dialogue system defined in [6], it is sometimes the case that even when there is an action that is agreeable to each agent given the union of their arguments (i.e. agreeable in the **joint iVAF** $\langle \mathcal{X}, \mathcal{A} \rangle$ under each agent's audience, where $\mathcal{X} = Args^x_\gamma \cup Args^{\overline{x}}_\gamma$), the dialogue may still terminate unsuccessfully. As we have now instantiated the $\mathsf{Pick}_{\mathsf{ass}}$ function we are able to present a more detailed analysis of when a dialogue generated by the system will terminate successfully.

First we need to show that if there is an action that is agreeable to both agents in the

joint iVAF and that action is agreeable to one of the agents *at the end of the dialogue*, then the dialogue will terminate with a successful outcome. (This following lemma holds for any instantiation of the Pick functions.)

Lemma 1: *Let D^t be a well-formed deliberation dialogue that terminates at t where $\langle \mathcal{X}, \mathcal{A} \rangle$ is the joint iVAF ($\mathcal{X} = Args^x_\gamma \cup Args^{\overline{x}}_\gamma$).*
If there exists an action a s.t. a is agreeable in the joint iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$ under both $\mathcal{R}^x$ and $\mathcal{R}^{\overline{x}}$ and a is agreeable in dVAF*(x, D^t) under $\mathcal{R}^x$, then the dialogue terminates with a successful outcome.*

Proof: *Assume dialogue terminates unsuccessfully, hence a not agreeable to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$ *under $\mathcal{R}^{\overline{x}}$ (Prop. 3 of [6]). From Lem. 1 of [6], $\overline{x}$ knows all arguments for a. Therefore $\overline{x}$ must know of some argument A' for a' (distinct to a) such that A' defeats all arguments for a under $\mathcal{R}^{\overline{x}}$ and A' is acceptable to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$*, and x must know some argument A'' against a' that defeats A' under $\mathcal{R}^{\overline{x}}$. As a' is not agreeable to x in* dVAF(x, D^t) *(Prop. 3 of [6]), x must have asserted its arguments against a' (because an agent cannot make a close move unless all assert moves exhausted). therefore A' cannot be acceptable to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$ *– contradiction.* □

We can now show that if there is an action agreeable to both agents in the joint iVAF such that *at any point in the dialogue* that action is agreeable to x who knows correctly what $\overline{x}$'s winning value is, then the dialogue will terminate successfully.

Proposition 2: *Let D^t be a well-formed deliberation dialogue that terminates at t where $\langle \mathcal{X}, \mathcal{A} \rangle$ is the joint iVAF ($\mathcal{X} = Args^x_\gamma \cup Args^{\overline{x}}_\gamma$), the value v is the winning value in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^{\overline{x}}$, and the action a is agreeable in the joint iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$ under both $\mathcal{R}^x$ and $\mathcal{R}^{\overline{x}}$. If there exists t' s.t. D^t extends $D^{t'}$ and there exists an argument A for a s.t. A is acceptable in* dVAF$(x, D^{t'})$ *under $\mathcal{R}^x$ and* $\mathsf{Models}^{\overline{x}}_x(D^i, Val) = 1$ *iff $Val = v$, then D^t terminates with a successful outcome.*

Proof: *Assume D^t terminates unsuccessfully, hence a not agreeable to x in* dVAF(x, D^t) *or to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$ *(Lem. 1). Therefore $\overline{x}$ asserts A' at timepoint t'', $D^{t''}$ extends $D^{t'}$ and A' defeats A under $\mathcal{R}^x$. Either: (1) A' is an argument against a. $\overline{x}$ must know an argument A'' for a' that it finds acceptable and that defeats all arguments for a in the joint iVAF under $\mathcal{R}^{\overline{x}}$ (because $\overline{x}$ knows an argument for a motivated by its winning value, either at start of dialogue else x, having perfect knowledge of what $\overline{x}$'s winning value is, would have asserted such an argument; and there cannot be any arguments against a that defeat such an argument for a under $\mathcal{R}^{\overline{x}}$'s audience else a would not be agreeable to $\overline{x}$ in the joint iVAF and so $\overline{x}$ must rather know A'' in order to defeat such an argument). $\overline{x}$ must have previously asserted A'' (as agent will only assert an argument against an action if it has no arguments for an action it can assert). As dialogue terminates unsuccessfully, it must be the case that A'' is unacceptable to x in* dVAF(x, D^t) *and yet acceptable to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$ *(as A'' must be acceptable to $\overline{x}$ in order to ensure that a it not agreeable to $\overline{x}$ but, from Prop. 3 of [6], cannot also be acceptable to x). For any such A'', it must be the case that x knows an argument A''' against a' that defeats A'' under $\overline{x}$'s audience (because a agreeable to $\overline{x}$ in the joint iVAF). x has asserted all such A''' (because close move only made when all assert moves have been exhausted), hence it cannot be that A'' is acceptable to $\overline{x}$ in* dVAF$(\overline{x}, D^t)$ *– contradiction. (2) A' is an argument for an action a' (distinct to a) that is agreeable to $\overline{x}$ in* dVAF$(\overline{x}, D^{t''})$*. Either: (a) A' is acceptable to x in* dVAF$(x, D^{t''})$

(and so each agent agrees to a' and the dialogue terminates successfully); or (b) A' is not acceptable to x in dVAF$(x, D^{t''})$ *Therefore A defeats A' under x's audience – contradiction.* □.

It is interesting to note that it is not always the case that if there is an action that is agreeable to both agents in the joint iVAF and that is agreeable to one of the agents at some point in the dialogue, then the successful outcome of the dialogue will be an action that is agreeable to both agents in the joint iVAF. For example, consider the situation where: $Args_{\gamma}^{Ag1} = \{\langle a1, \gamma, v2, + \rangle, \langle a2, \gamma, v1, - \rangle\}$; $Args_{\gamma}^{Ag2} = \{\langle a2, \gamma, v3, + \rangle, \langle a2, \gamma, v4, - \rangle\}$; $v4 \succ_{Ag1} v3 \succ_{Ag1} v2 \succ_{Ag1} v1$; $v1 \succ_{Ag2} v3 \succ_{Ag2} v2 \succ_{Ag2} v4$. If we construct the joint iVAF for this example, then we see that the action $a1$ is agreeable to both agents and the action $a2$ is agreeable to neither (given the union of their arguments); however, the dialogue generated will terminate successfully with $a2$ as the outcome. This observation is important as it helps to determine the suitability of the strategy defined here for particular applications: if it is imperative that the outcome arrived at is the 'best' possible (in the sense that it is agreeable to each participant given the union of their knowledge), then the strategy we give here is not suitable; whilst if we simply desire that agents reach some agreement, then our strategy may suffice.

There are situations where there is an action agreeable to each agent in the joint iVAF and yet the dialogue still does not terminate successfully (for example, if there is no action agreeable to at least one of the agents at the start of the dialogue). The detailed analysis that we give here of when and why a dialogue terminates successfully is invaluable for the future design of deliberation systems that aim to avoid this situation. Our investigation takes steps towards an understanding of how the design of a deliberation strategy and the subjective preferences of agents affect dialogue behaviour.

5 Modelling agent preferences

We have shown that if a proponent can correctly model the recipient's winning value for the joint iVAF and there is an action agreeable to each given the joint iVAF, then if that action is at any point agreeable to the proponent, the dialogue will terminate successfully. We now consider how a proponent may aim to correctly model the recipient's winning value. Whilst there is much existing work on reasoning about another agent's beliefs, we are not aware of any work that aims at modelling another agent's values.

In order to design a modelling mechanism, we consider what it means to be a winning value. Recall: (Def. 5) a value is a winning value for an agent in an iVAF if there is a *positive* argument that *promotes* that value and that is acceptable under the agent's audience (and so it is not necessarily the most preferred value); (Prop. 1) an agent has at most one winning value for a particular iVAF where all arguments relate to the same goal (since we are dealing with deliberation dialogues with a particular topic, we assume henceforth that all the arguments in an iVAF relate to the same goal).

We can show that if there is no winning value for an iVAF under a particular audience, then it must be the case that for every *positive* argument *for* an action, there is another *negative* argument *against* that action whose value is at least as preferred.

Thus there is only one special case in which there is no winning value for an agent in an iVAF, justifying our approach of modelling what is likely to be an agent's winning value.

Proposition 3: *Let $\langle \mathcal{X}, \mathcal{A} \rangle$ be an iVAF s.t. there is no winning value under audience $\mathcal{R}^x$. If $\exists \langle a, p, v, + \rangle \in \mathcal{X}$, then $\exists \langle a, p, v', - \rangle$ s.t. $(v, v') \notin \mathcal{R}^x$.*

Proof: *There is no winning value therefore (from Def. 5) $\nexists A \in \mathcal{X}$ s.t. A is acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$ and $\mathsf{Sign}(A) = +$. Assume $\exists A = \langle a, p, v, + \rangle \in \mathcal{X}$, A must be defeated under $\mathcal{R}^x$. Either A is defeated by either an argument A' for (i) another action or (ii) against a. If (i), then A' must be acceptable — contradiction.* □

Now we consider what it means if there is an argument motivated by the winning value that is not acceptable. We can show that if there is an argument *for* an action that is motivated by the winning value but that is not acceptable, then there must be an argument *against* that action that is at least as preferred.

Proposition 4: *Let $\langle \mathcal{X}, \mathcal{A} \rangle$ be an iVAF s.t. v is the winning value under $\mathcal{R}^x$. If $\exists A = \langle a, p, v, + \rangle \in \mathcal{X}$ s.t. A not acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$, then $\exists A' = \langle a, p, v', - \rangle$ s.t. $(v, v') \notin \mathcal{X}$.*

Proof: *Assume A is not acceptable therefore A is defeated by either an argument A' (i) for another action or (ii) against a. If (i), then A' must be acceptable and $(\mathsf{Val}(A'), v) \in \mathcal{R}^x$ (so $\mathsf{Val}(A')$ a winning value and distinct from v), but there is at most one winning value (Prop. 1) — contradiction.* □

The previous result considers an iVAF in which v is an agent's winning value. However, we are concerned with modelling the recipient's winning value **in the joint iVAF**, which the agents do not have access to (since this is built from the agents' private knowledge). Thus we must also consider the relationship between an iVAF and its subgraphs. We show that if v is a winning value in an iVAF, but there is an argument for an action a motivated by v that is not acceptable in a subgraph, then either: there must be an argument against that action in the subgraph that is at least as preferred; else there must be an argument in the subgraph for some other action a' that is motivated by a more preferred value than v and there must be an argument that is in the iVAF but not in the subgraph against action a' that defeats this argument.

Proposition 5: *Let $\langle \mathcal{X}, \mathcal{A} \rangle$, $\langle \mathcal{X}', \mathcal{A}' \rangle$ be iVAFs s.t. $\mathcal{X}' \subseteq \mathcal{X}$. If v is the winning value in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$ but $A = \langle a, p, v, + \rangle$ is not acceptable in $\langle \mathcal{X}', \mathcal{A}' \rangle$ under $\mathcal{R}^x$, then either: (1) $\exists \langle a, p, v', - \rangle \in \mathcal{X}'$ s.t. $(v, v') \notin \mathcal{R}^x$; else (2) $\exists \langle a', p, v', + \rangle \in \mathcal{X}'$ s.t. $(v', v) \in \mathcal{R}^x$ and $\exists \langle a', p, v'', - \rangle \in \mathcal{X} \setminus \mathcal{X}'$ s.t. $(v', v'') \notin \mathcal{R}^x$.*

Proof: *Assume A not acceptable in $\langle \mathcal{X}', \mathcal{A}' \rangle$ under $\mathcal{R}^x$, therefore A is defeated by either an argument A' (i) for another action or (ii) against a. If (ii), then we get case (1). If (i), then $A' \langle a', p, v', + \rangle \in \mathcal{X}'$ s.t. $(v', v) \in \mathcal{R}^x$. However, v is the only (Prop. 1) winning value in $\langle \mathcal{X}, \mathcal{A} \rangle$ therefore A' is not acceptable in $\langle \mathcal{X}, \mathcal{A} \rangle$ under $\mathcal{R}^x$. As we know $(v', v) \in \mathcal{R}^x$, it must be the case that $\exists \langle a', p, v'', - \rangle \in \mathcal{X} \setminus \mathcal{X}'$ s.t. $(v', v'') \notin \mathcal{R}^x$.* □

Let us now consider the case where a proponent asserts a positive argument for an action a motivated by the value v, where v is the recipient's winning value in the joint iVAF, and the recipient does not respond with an agree move. From Prop. 5 we see that there are two possible cases.

Case 1: The recipient has a negative argument *against* a that is motivated by a

value that the recipient prefers at least as much as v. In this case, a cannot be agreeable to the recipient in the joint iVAF (since v is the recipient's winning value, therefore all acceptable positive arguments must be motivated by v, and any such argument for a will be defeated by the recipient's argument against a).

Case 2: The recipient has a positive argument for some other action a' that is motivated by a value v' that it prefers more to v and the proponent has an unasserted negative argument against a' that is motivated by a value v'' that the recipient prefers at least as much as v'.

As v is the recipient's winning value, there must be a positive argument in the joint iVAF that is motivated by v and acceptable under the recipient's audience, thus there must be at least one positive argument motivated by v and known to the proponent that falls under Case 2 (since a negative argument that defeats an argument in the recipient's dialogue iVAF will also defeat that argument under the recipient's audience in the joint iVAF). Therefore, if a proponent has asserted all of its positive arguments motivated by v and not elicited an agree, the only way that v can be the recipient's winning value is if the proponent has an unasserted argument against every action agreeable to the recipient that succeeds in defeat under the recipient's audience.

If the proponent knows no unasserted negative arguments, then Case 2 above cannot hold, therefore further limiting the chance of v being the winning value.

We can use these insights to define a simple mechanism for updating an agent's Models function. This function maps each value to the interval between 0 and 1; the higher the output of the function the more the proponent believes that the value is the winning value for the recipient (Def. 12). For reasons of space, here we only consider the case where the proponent has asserted an argument for an action motivated by v and the recipient does not then agree to that action. As we have seen, if the following conditions also hold, the proponent has extra reason to believe that v is not the recipient's winning value:

- the proponent knows no unasserted negative arguments;
- the proponent knows no unasserted positive arguments motivated by v;
- the proponent knows no unasserted positive arguments motivated by v and knows no unasserted negative arguments (in this case it is not possible that v is the recipient's winning value).

We use an update function $\mathsf{Sub}(\mathsf{Models}_{\overline{x}}^{x}(D^t, v), N)$ that decrements $\mathsf{Models}_{\overline{x}}^{x}(D^t, v)$ by N (whilst respecting the function's range boundaries) and captures these situations as follows:

Definition 16: *Let D^t be a dialogue s.t.* $\text{dVAF}(x, D^t) = \langle \mathcal{X}, \mathcal{A} \rangle$, $AssArgs = \{A \mid \exists i (1 \leq i \leq t)$ *s.t.* $m_i = \langle \_, \mathsf{assert}, A \rangle\}$, $m_{t-1} = \langle x, \mathsf{assert}, \langle a, p, v, + \rangle \rangle$, *and* $m_t \neq \langle \overline{x}, \mathsf{agree}, a \rangle$.

Agent x **updates its recipient value model** $\mathsf{Models}_x^{\overline{x}}$ *as follows.*

If $\nexists A \in \mathcal{X}$ *s.t.*
$(\mathsf{Sign}(A) = +, \mathsf{Val}(A) = v$ *and* $A \notin AssArgs$
then if $\nexists A' \in \mathcal{X}$ *s.t.* $\mathsf{Sign}(A') = -$ *and* $A' \notin AssArgs$,
$\mathsf{Models}_x^{\overline{x}}(D^t, v) := 0$,
else $\mathsf{Models}_x^{\overline{x}}(D^t, v) := \mathsf{Sub}(\mathsf{Models}_x^{\overline{x}}(D^{t-1}, v), 0.4)$.
Otherwise
if $\nexists A \in \mathcal{X}$ *s.t.* $\mathsf{Sign}(A) = -$ *and* $A \notin AssArgs$,
then $\mathsf{Models}_x^{\overline{x}}(D^t, v) := \mathsf{Sub}(\mathsf{Models}_x^{\overline{x}}(D^{t-1}, v), 0.2)$.
Otherwise
$\mathsf{Models}_x^{\overline{x}}(D^t, v) := \mathsf{Sub}(\mathsf{Models}_x^{\overline{x}}(D^{t-1}, v), 0.1)$.

In the example in Sect. 3, agent $Ag1$ updates its recipient value model in this manner.

We have thus given a principled mechanism with which an agent can model another agent's winning value, based on their dialogue behaviour. Our mechanism is not intended to be complete, it needs also to consider situations in which it is appropriate to increment the function output for a particular value. Also, the figures that our update mechanism uses for the decrements (which reflect the strength of the reason that the proponent has to believe that v is not the recipient's winning value) could be further refined (particularly with empirical analysis). However, our simple mechanism illustrates how detailed theoretical analysis of system behaviour can be useful in designing dialogue strategies.

6 Related work

Our proposal uses the same underlying dialogue framework as in [5]; however, that work is only similar in that it uses the same dialogue representation. The system defined in [5] is concerned not with deliberation but with a type of inquiry dialogue; it ensures that all relevant arguments are asserted, after which a shared value ordering is applied to determine the outcome.

The system here builds directly on that presented in [6]. We have extended that work by defining a function that allows a proponent to select arguments to assert based on its perception of what is important to the recipient. By specifying the strategy thus, we have been able to perform a more detailed analysis of the behaviour of the system than was previously possible; this fundamental analysis moves us towards a better understanding of the design of dialogue strategies that are suitable for particular applications. We have also provided a mechanism with which an agent can model what is important to the other participant.

Other works allow a proponent to select arguments suited to a particular recipient. In [11] a proponent selects sets of arguments likely to resonate with the recipient by considering the recipient's desires, whilst [17] investigates how a proponent may use the recipient's personality to guide argument selection; however, both of these works

deal with monological rather than dialogical argumentation. The dialogue system proposed in [13] allows an agent to use a model of its opponent's goals and beliefs to select arguments; however, [13] does not consider value based arguments, and the behaviour of the system is not analysed as we have done here.

Deliberation dialogues are considered by [12, 16]. In [12] argument evaluation is not done in terms of AFs, and strategies for reaching agreement are not considered; [16] focusses on goal selection and planning. Practical reasoning using argumentation in agent systems has been addressed by Amgoud and colleagues (see e.g. [1]), but in this work the focus is not on the dialogical aspects nor is there an element to model other participants' preferences.

The proposal of [4] considers how to find particular audiences for which only certain arguments are acceptable and how preferences over values emerge through a dialogue; however, it assumes a static argument graph within which agents are playing moves, whilst agents in our system construct argument graphs dynamically.

The work of [8] allows AFs of individual agents to be merged; it aims to characterise the sets of arguments acceptable by the whole group of agents using notions of joint acceptability. In our work, an agent develops its own individual graph and uses this to determine if it finds an action agreeable, thus maintaining its subjective view.

Prakken [14] considers how agents can come to a public agreement despite their internal views of argument acceptability conflicting, allowing them to make explicit attack and surrender moves. However, Prakken does not explicitly consider value-based arguments, nor does he discuss particular strategies.

Strategic argumentation has been considered in other work. In [9] a dialogue game for persuasion is presented that is based on one originally proposed in [19] but makes use of Dungian AFs. Strategies in [9] concern reasoning about an opponent's beliefs, as opposed to about action proposals with subjective preferences. Strategies for reasoning with value-based arguments are considered in [3], where the objective is to create obligations on the opponent to accept some argument based on his previously expressed preferences. In [3], a fixed joint VAF is assumed, whilst our agents dynamically construct individual dialogue iVAFs. Neither [9] or [3] gives an analysis of how strategy affects dialogue behaviour.

A related emerging area is the application of game theory to argumentation (e.g. [15]). This work has investigated situations under which rational agents will not have any incentive to lie about or hide arguments; although concerned mainly with protocol design, it is likely such work will have implications for strategy design.

7 Concluding remarks

We have presented a dialogue system for joint deliberation, where the agents involved may each have different preferences yet all want an agreement to be reached. The novel strategy that we have defined allows a proponent to take account of the recipient's preferences. The initial analysis that we presented gives us a better understanding of how strategy design affects dialogue behaviour. Furthermore, we have also provided a mechanism to enable a dialogue participant to model what is likely to be the winning value for the other participant; it can then use this model to select arguments for action

that are likely to be persuasive to the other agent. The design of this mechanism was guided by our investigation into the behaviour of iVAFs; however, it is only a first step towards modelling agents' values. Many interesting questions remain, for example: why might a proponent *increase* its belief that a particular value is the winning one for the recipient; how should a proponent initialise its recipient model function at the start of a dialogue?

Another very interesting line of future work is to extend the system so that argumentation theory is also used by the proponent to determine which is the recipient's winning value. We have seen that there can be reasons to believe that v is not the recipient's winning value, these reasons and their different strengths could themselves be modelled as an argumentation framework.

In the dialogue system we have presented here, we have assumed that there are only two participants and that each is following the same strategy. It will be necessary to relax these assumptions in the future if our system is to be applicable in all but the simplest of situations. If we are to meet the ultimate goal of a robust theory for deliberation strategy design, analyses such as the one presented here are a key requirement, providing the foundations for developing and analysing more complex deliberation dialogue systems.

8 Acknowledgements

E. Black funded by the European Union Seventh Framework Programme (FP7/2007-2011) under grant agreement 253911.

References

[1] L. Amgoud, C. Devred, and M.-C. Lagasquie-Schiex. A constrained argumentation system for practical reasoning. In *7th Int. Joint Conf. on Autonomous Agents and Multiagent Systems*, pages 429–436, 2008.

[2] K. Atkinson and T. J. M. Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10–15):855–874, 2007.

[3] T. J. M. Bench-Capon. Agreeing to differ: Modelling persuasive dialogue between parties without a consensus about values. *Informal Logic*, 22(3):231–245, 2002.

[4] T. J. M. Bench-Capon, S. Doutre, and P. E. Dunne. Audiences in argumentation frameworks. *Artificial Intelligence*, 171(1):42–71, 2007.

[5] E. Black and K. Atkinson. Dialogues that account for different perspectives in collaborative argumentation. In *8th Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, pages 867–874, 2009.

[6] E. Black and K. Atkinson. Agreeing what to do. In *7th Int. Workshop on Argumentation in Multi-Agent Systems*, 2010.

[7] E. Black and K. Atkinson. Choosing persuasive arguments for action. In *10th Int. Conf. on Autonomous Agents and Multi-Agent Systems*, 2011.

[8] S. Coste-Marquis, C. Devred, S. Konieczny, M.-C. Lagasquie-Schiex, and P. Marquis. On the merging of Dung's argumentation systems. *Artificial Intelligence*, 171(10–15):730–753, 2007.

[9] J. Devereux and C. Reed. Strategic argumentation in rigorous persuasion dialogue. In *6th Int. Workshop on Argumentation in Multi-Agent Systems*, pages 37–54, 2009.

[10] 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.

[11] A. Hunter. Towards higher impact argumentation. In *Proc. of the 19th American National Conf. on Artificial Intelligence*, pages 275–280, 2004.

[12] P. McBurney, D. Hitchcock, and S. Parsons. The eightfold way of deliberation dialogue. *International Journal of Intelligent Systems*, 22(1):95–132, 2007.

[13] N. Oren and T. J. Norman. Arguing using opponent models. In *6th Int. Workshop on Argumentation in Multi-Agent Systems*, 2009.

[14] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *J. of Logic and Computation*, 15:1009–1040, 2005.

[15] I. Rahwan and K. Larson. Mechanism design for abstract argumentation. In *5th Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, pages 1031–1038, 2008.

[16] Y. Tang and S. Parsons. Argumentation-based dialogues for deliberation. In *4th Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, pages 552–559, 2005.

[17] T. van der Weide, F. Dignum, J.-J, Meyer, H. Prakken, and G. Vreeswijk. Personality-based practical reasoning. In *5th Int. Workshop on Argumentation in Multi-Agent Systems*, pages 3–18, 2008.

[18] D. N. Walton. *Argumentation Schemes for Presumptive Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, USA, 1996.

[19] D. N. Walton and E. C. W. Krabbe. *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press, Albany, NY, USA, 1995.

[20] M. Wooldridge and W. van der Hoek. On obligations and normative ability: Towards a logical analysis of the social contract. *J. of Applied Logic*, 3:396–420, 2005.