An empirical study of a deliberation dialogue system

Elizabeth Black[1] and Katie Bentley[2]

[1] Intelligent Systems Group, Universiteit Utrecht, De Uithof, 3584 CC Utrecht, NL
lizblack@cs.uu.nl
[2] Vascular Biology Lab, London Research Institute, Cancer Research UK, Lincoln's Inn Fields, WC2A 3LY, UK
katie.bentley@cancer.org.uk

Abstract. We present an empirical simulation-based study of the use of value-based argumentation in two-party deliberation dialogues, investigating the impact that argumentation can have on the quality of the outcome reached. Our simulation allows us to vary the number of values, actions and arguments that appear in the system; we investigate how the behaviour of the system changes as these parameters vary. This parameter sensitivity analysis tells us whether a value-based deliberation dialogue system may be useful for a particular real-world application. We measure the quality of the dialogue outcome (i.e. the action that the agents agree to) against a global view of whether that action would be agreeable to each agent if all of the agents' knowledge were taken into account. We compare the deliberation outcome with a simple consensus forming procedure (where no arguments are exchanged). Our results show that the deliberation dialogue system we present outperforms consensus forming.

ACM Category: I.2.11 Multiagent systems. **General terms:** performance, experimentation

Keywords: dialogue, value-based argumentation, simulation, agreement, deliberation.

1 Introduction

There is little work on evaluating whether an argumentation-based approach to a problem is a good approach to take. Most works assume that the decision to use argumentation has already been made and disregard the question of whether there is a better approach to take. We present what we believe to be the first simulation-based study of an argumentation-based deliberation dialogue system, which allows us to start addressing this question and allows us to investigate the effect of varying the parameters of the system.

Simulation is an imperative next step for bridging the gap between argumentation theory and real-world agent applications. Given the complexity of argumentation-based dialogue systems, it is very hard to theoretically investigate their properties without making many restrictive assumptions. In order to gain a full understanding of the behaviour of such systems, theoretical investigations need to be complemented with empirical simulation-based studies. Simulation provides a unique opportunity to generate

large, complex scenarios and analyse their results across thousands of iterations and permutations.

There are few existing works that take a simulation approach to investigating the performance of argumentation-based dialogue systems. Two notable examples are [1, 2]. Each of these focus on a form of argumentation-based negotiation (ABN), where arguments providing reasons for an agent's position are shared; this exchange of information allows the negotiation space to change. In [1], the information exchanged relates to the influence of social commitments between roles, whilst [2] focusses on interest-based negotiation where agents exchange information about their underlying goals and different ways to achieve these.

In [3], ABN is used to address the distributed constraint satisfaction problem. Importantly, the authors have performed experiments with their model to investigate the performance of their argument-based approach. Agents in the system use arguments in the sense that they put forward a proposal and provide a justification for this by giving their local constraints, these constraints are propagated by the receiving agent.

Our deliberation context differs from ABN (which is generally concerned with the allocation of scarce resources), as agents in our system have a shared goal and wish to come to an agreement on how to act in order to achieve that goal. Similarly to the systems discussed above, our agents also share arguments regarding actions to achieve the goal and this allows the set of actions that an agent finds agreeable to change. In our system, however, these arguments are value-based (relating to various social values that may be promoted or demoted by performing an action) and, unlike in [1–3], our agents also use argumentation as the reasoning mechanism with which they determine which actions they find agreeable.

We specifically investigate two questions:

- Do our deliberation dialogues perform better across the entire parameter space than a simple consensus forming approach, where agents try to find an action they each find agreeable without sharing any arguments?
- How does the behaviour of both the dialogue system and the consensus forming mechanism change as the number of arguments, actions and values present in the system varies?

Our results clearly show that the deliberation dialogue system outperforms consensus forming across all parameter combinations. Further, we have identified particular parameter settings that optimise dialogue performance in terms of quality of outcome and length of dialogue. This detailed parameter sensitivity analysis allows a designer of an agent system to evaluate whether value-based deliberation dialogues are useful for their particular application domain.

2 Model

In this section we describe the model that we are simulating. We give details of the value-based argumentation model, the dialogue system, the consensus forming mechanism, the evaluation metric that we use and our experimental set up. The model was written in c++ on a standard workstation. A complete parameter sensitivity analysis of 1.8 million runs took less than an hour to complete.

2.1 Argumentation model

We are investigating the performance of the system formally specified in [4], which is based on the popular argument scheme and critical question approach [5]. Arguments are generated by an agent instantiating a **scheme for practical reasoning** [6]: In the current circumstances R, we should perform action A, which will result in new circumstances S, which will achieve goal G, which will promote value V.

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 to uphold by realising the goal stated [7].

An agent has a **Value-based Transition System** (VATS), that it uses to instantiate the scheme for practical reasoning. This transition system represents the agent's knowledge about the effect of actions and the values that are promoted or demoted. (For brevity, we omit the definition here; the reader is referred to [4].) Given its VATS, an agent can 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. Note that here we are focussing on the **choice of action** stage (as defined in [6]), we assume that any discrepancies between the agents in either the problem formulation or epistemic reasoning stages have been resolved (perhaps with some other type of dialogue); thus, for example, agents do not need to question here whether an action in question does achieve the desired goal or whether a certain set of circumstances hold.

Definition 1: *An* **argument** *constructed by an agent x from its VATS is a 4-tuple $A = \langle a, p, v, s\rangle$ where:*
$s = +$ iff a is an **action** *that will achieve* **goal** *p and will* **promote** *value v;*
$s = -$ iff a is an **action** *that will achieve* **goal** *p but will* **demote** *value v.*
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 its VATS** *as $Args^x$; we let $Args^x_p = \{A \in Args^x \mid \mathsf{Goal}(A) = p\}$. 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. The relevant CQs are used to 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) (introduced in [7]), an extension of the argumentation frameworks (AF) of Dung [8]. 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 or at least as high as 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. This attack relation is derived from the CQs, for details the reader is referred to [4].

Definition 2: *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) are members 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 $v \succ_{\mathcal{R}^x} v'$.We say that an argument A_i is* **preferred to** *the argument A_j in the audience $\mathcal{R}^x$, denoted $A_i \succ_{\mathcal{R}^x} A_j$, iff $\texttt{Val}(A_i) \succ_{\mathcal{R}^x} \texttt{Val}(A_j)$. If R^x is an audience over the values V for the iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$, then $\texttt{Vals}(\mathcal{X}) \subseteq V$.*

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 preference over values.

Given an iVAF and a particular agent's audience, we can determine acceptability of an argument as follows. Note that (as in [4]) if an attack is symmetric, then an attack only succeeds in defeat if the attacker is more preferred than the argument being attacked; however, if an attack is asymmetric, then an attack succeeds in defeat if the attacker is at least as preferred as the argument being attacked. Asymmetric attacks occur only when an argument against an action attacks another argument for that action; in this case, if both arguments are equally preferred then we do not wish the argument for the action to withstand the attack. If we have a symmetric attack where the arguments attacking one another are equally preferred, then we must have arguments for two different actions that promote the same value; here, the defeat is not successful, since it is reasonable to choose either action.

Definition 3: *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 $A_j \not\succ_{\mathcal{R}^x} A_i$.*

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 $A_i \succ_{\mathcal{R}^x} A_j$.*

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 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.

2.2 Dialogue system

The dialogue system investigated here is formally defined in [4]. For readability and brevity, we omit the formal definitions here but informally describe the 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 $\mathcal{I} = \{Ag1, Ag2\}$ and each with a knowledge base of arguments that it knows about (those it can construct from its VATS). Each participant takes it in turn to make a move to the other participant. We refer to participants using the variables x and $\overline{x}$ such that: x is $Ag1$ if and only if $\overline{x}$ is $Ag2$; x is $Ag2$ if and only if $\overline{x}$ is $Ag1$.

We assume that the participants have agreed to partake in a deliberation dialogue whose **topic** is the joint goal in question. During the dialogue, agents can either:

- `assert` a positive argument (an argument *for* an action);
- `assert` a negative argument (an argument *against* an action);
- `agree` to an action;
- indicate that they have no arguments that they wish to assert (with a `pass`).

The agents take it in turn to make a single move. A dialogue terminates under one of two conditions: **failure**, when two `pass` moves appear one immediately followed by the other in the dialogue; **success** with **outcome** a, when two moves each agreeing to the action a appear one immediately followed by the other in the dialogue.

In order to evaluate which actions it finds agreeable at a point in a dialogue with topic p, an agent x considers the iVAF that it constructs from all the arguments that it currently has available to it relating to p; this consists of the arguments from its own VATS, as well as the arguments that the other agent has asserted thus far. We call this agent x's **dialogue iVAF**, which is the iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$ where $\mathcal{X} = Args_p^x \cup \{A \mid \overline{x}$ has previously asserted A during the dialogue$\}$. 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, due to its dialogue iVAF changing as arguments asserted by $\overline{x}$ are added to it.

The **protocol** defines which moves an agent x (whose turn it is) is allowed to make at any point in a deliberation dialogue with topic p as follows:

- It is permissible to `assert` an argument A iff $\texttt{Goal}(A) = p$ (i.e. the argument is for or against an action to achieve the topic of the dialogue) and A has not been asserted previously during the dialogue.
- It is permissible to `agree` to an action a iff either:
  - the immediately preceding move was an `agree` to the action a, or

    - the other participant $\overline{x}$ has at some point previously in the dialogue asserted a positive argument A for the action a.
- It is always permissible to `pass`.

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. The **strategy** that our agents use is as follows:

- If it is permissible to `agree` to an action that the agent finds *agreeable*, then make such an `agree` move; else
- if it is permissible to `assert` a positive argument *for* an action that the agent finds *agreeable*, then assert some such argument; else
- if it is permissible to `assert` a negative argument *against* an action and the agent finds that action *not agreeable* then assert some such argument; else
- make a `pass` move.

We have now defined how our dialogue system regulates the moves that agents may make, and the strategy that the agents use to select one of the permissible moves to make. (For an example of a dialogue produced by this system, please refer to [4].) Next, we define a method with which two agents may form a consensus without exchanging any arguments.

2.3 Consensus forming

In order to start investigating the question of whether it is worth using argumentation-based deliberation dialogues to decide how to act to achieve a shared goal, we compare outcomes produced by our dialogue system with those produced by a simple consensus forming method. For two agents $x, \overline{x}$ who are about to enter into a deliberation dialogue with topic p, the outcome produced by **consensus forming** is simply the *intersection* of the following two sets:

- the set of actions to achieve p that agent x finds agreeable at the start of the dialogue;
- the set of actions to achieve p that agent $\overline{x}$ finds agreeable at the start of the dialogue.

That is to say, the consensus set contains all the actions that each agent finds agreeable, given the arguments they can construct from their VATS and without any exchange of arguments. If consensus forming returns a non-empty set, then we say that a **consensus was found** and that the consensus forming was **successful**.

This gives us a non-argumentative approach to which we can compare our dialogue system. We next discuss how we compare these systems, namely on the quality of outcome.

2.4 Measuring quality of outcome

Unless they exchange all arguments, agents in our system only ever have a partial view of all of the available knowledge. We can, however, take a global view of which potential outcomes are best for each of the agents. For this purpose, we define for a dialogue the **omniscient argumentation framework** (OAF), which is the iVAF constructed from the union of the arguments that each participant can construct from its

VATS that relate to the topic of the dialogue. For a dialogue with participants $x, \overline{x}$ and topic p, the associated OAF is thus the iVAF $\langle \mathcal{X}, \mathcal{A} \rangle$ where $\mathcal{X} = Args_p^x \cup Args_p^{\overline{x}}$. We say that an action is **globally agreeable** to an agent x if and only if there is some *positive* argument *for* that action that is acceptable in the OAF under the audience that represents x's value preference.

We can now measure the quality of a particular outcome (i.e. an action to achieve the goal p) by considering whether it is globally agreeable to each agent. Such a quality measure can be applied to both the outcome produced by a dialogue and the outcome produced by consensus forming.

For a particular outcome a, we assign an **outcome quality score** as follows:

- if a is **globally agreeable to both** x and $\overline{x}$, **score 3**;
- if a is **globally agreeable to only one** of x or $\overline{x}$, **score 2**;
- if a is **not globally agreeable to either** x or $\overline{x}$, **score 1**.

If there is **no successful outcome** (i.e. dialogue terminates in failure or consensus forming returns an empty set) then the outcome quality **score is 0**. Where the consensus forming returns a set of more than one action, we assign the outcome quality score to be that of the action from the set which receives the lowest score (since this is the best that the consensus forming method can guarantee to do, given that only one action can be selected).

Our simple scoring metric reflects the intuition that any outcome is better than no outcome, but an outcome that is globally agreeable to an agent is better than one that is not. We plan to study more sophisticated scoring metrics in future work.

2.5 Experimental set up

The dialogue system and consensus forming mechanism were implemented as described in the previous sections. We also implemented a **random scenario generator**; this generates **scenarios** that initialise the agents' knowledge bases (i.e. the arguments known to each agent at the start of the dialogue, which all relate to the joint goal which the agents wish to achieve) and their audiences. The generator takes three parameters $(\mathsf{Args}, \mathsf{Vals}, \mathsf{Acts})$, where

- Args is the number of distinct arguments to appear in the union of the agents' knowledge bases;
- Vals is the number of distinct values that may motivate those arguments;
- Acts is the number of distinct actions that the arguments may relate to.

The generator randomly constructs without replacement (i.e. does not allow duplicate arguments) the required number of arguments from the allowed values and actions and the symbols $\{+, -\}$ (where each combination is equally likely). For example, when given parameters $(8, 2, 2)$, the generator will construct the following set of arguments:
$\{\langle a1, p, v1, +\rangle, \langle a1, p, v1, -\rangle, \langle a1, p, v2, +\rangle, \langle a1, p, v2, -\rangle,$
$\langle a2, p, v1, +\rangle, \langle a2, p, v1, -\rangle, \langle a2, p, v2, +\rangle, \langle a2, p, v2, -\rangle\}$.

(Note, it is not possible for the generator to construct a set of arguments from parameters $(\mathsf{Args}, \mathsf{Vals}, \mathsf{Acts})$ if $\mathsf{Args} > \mathsf{Vals} \times \mathsf{Acts} \times 2$. For a particular number of values and a particular number of actions, the total **possible arguments** is $\mathsf{Vals} \times \mathsf{Acts} \times 2$.)

The generator randomly assigns each agent an audience over the allowed values and it randomly allocates exactly half of the constructed arguments to one agent, and the other half to the other agent. Our generator is therefore simulating the construction of arguments from the agents' VATS. It allows us to run experiments over all possible combinations of the parameters $(\mathsf{Args}, \mathsf{Vals}, \mathsf{Acts})$. In the experiments reported here we consider all possible parameter combinations where:

- $\mathsf{Vals} \in \{2, 4, 6, 8, 10\}$,
- $\mathsf{Acts} \in \{2, 4, 6, 8, 10\}$,
- $2 \leq \mathsf{Args} \leq \mathsf{Vals} \times \mathsf{Acts} \times 2$.

Our experiments investigate how the outcome quality scores of the dialogue system and the consensus forming mechanism compare across the space of possible parameter combinations. We performed 1000 runs of our simulation for each possible parameter combination. In each run, a random scenario is generated. We first calculate the consensus set of the scenario and then simulate a dialogue from the same scenario; we compare the quality scores assigned to the outcomes produced by these two approaches.

3 Results

3.1 Dialogue is significantly more likely to be successful than consensus forming

Figure 1 shows strikingly across all parameter combinations that the frequency of successful consensuses is never as great as the frequency of successful dialogues. There is a significant difference between these two frequencies: across all parameters, consensus forming fails more than 50% of the time, whilst up to 90% of dialogues are successful.

We also found that, across all runs for each possible parameter combination (a total of 1.8 million runs), for every run in which a consensus was found the dialogue produced was also successful. It is not immediately clear whether the converse situation (i.e. a consensus is found but the dialogue produced is not successful) is theoretically impossible, but this result strongly suggests that this may be the case and so identifies a property worthy of theoretical investigation.

Consensus forming is relatively robust to the number of values present in the system; however there is a marked difference when $\mathsf{Vals} = 2$, in which case the frequency of successful consensuses is approximately half that of when $\mathsf{Vals} \in \{4, 6, 8, 10\}$.

When $\mathsf{Acts} = 2$, the highest frequency of consensuses found is seen when Args is equal to approximately 50% of the total arguments possible. A higher number of arguments present in the system leads to a higher frequency of successful consensuses; in contrast, the frequency of successful dialogues drops as the number of arguments present in the system increases (although the number of successful dialogues is still greater than the number of consensuses found).

3.2 Successful dialogues are more likely with higher numbers of actions and values

Looking at the top of Figure 1 in depth, we can see how sensitive the dialogue system is to the parameters. The dialogue system appears to be most sensitive to the parameter settings $\mathsf{Acts} = 2$ and $\mathsf{Vals} = 2$.

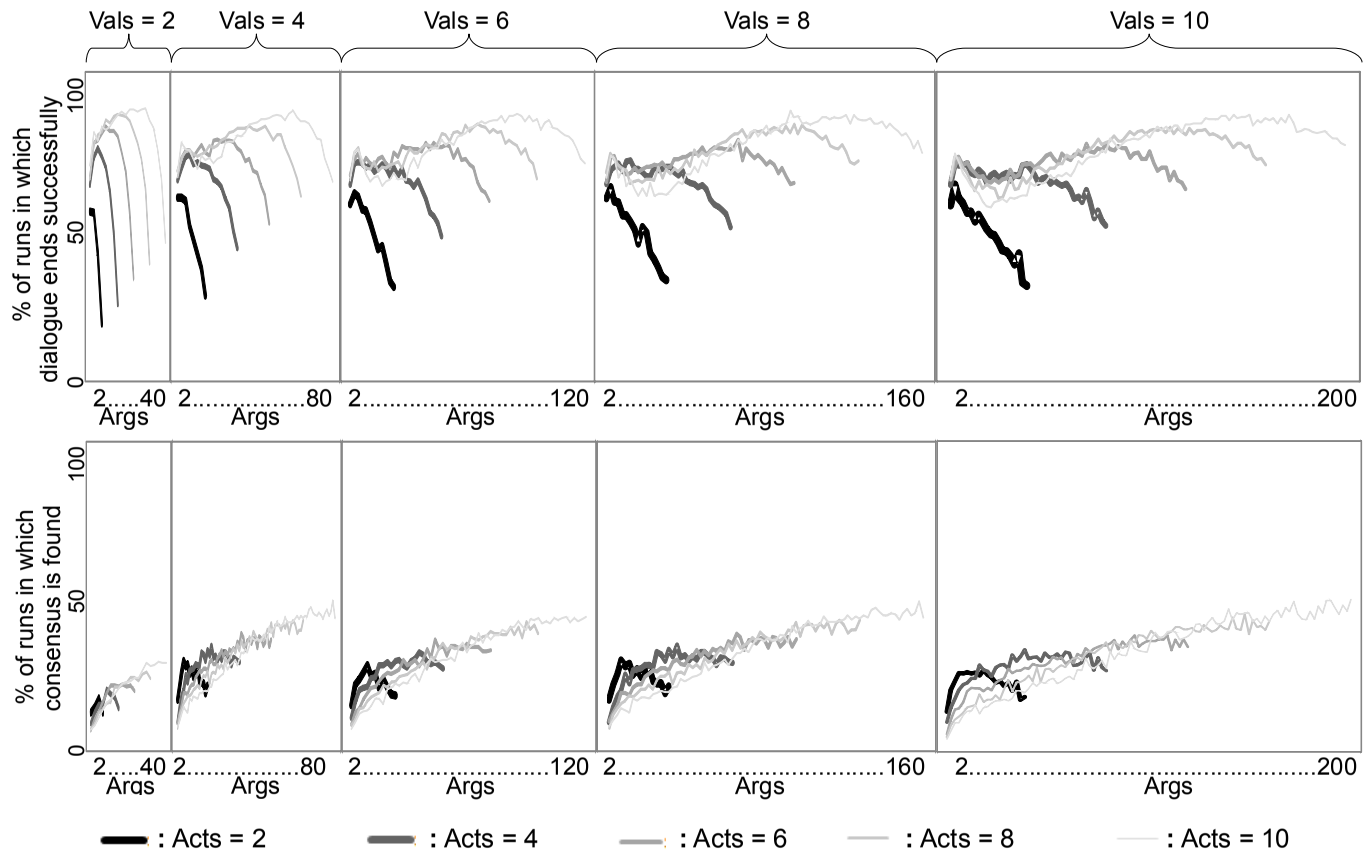

Fig. 1. Top: Percentage of dialogues that ended successfully out of 1000 runs across each possible parameter combination. Bottom: Percentage of 1000 runs across each possible parameter combination in which a consensus was found.

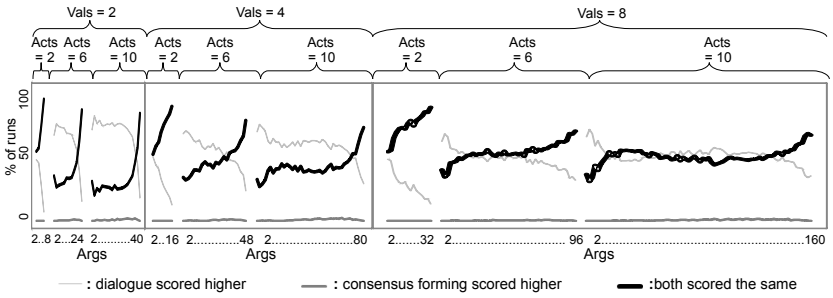

Fig. 2. Percentage of 1000 runs across each possible parameter combination where $\mathsf{Acts} \in \{2, 6, 10\}$ and $\mathsf{Vals} \in \{2, 4, 8\}$ in which: dialogue outcome quality score was higher than consensus outcome quality score; consensus outcome quality score was higher than dialogue outcome quality score; dialogue outcome quality score was the same as consensus outcome quality score.

Across all parameter settings, the frequency of successful dialogues is closely related to the percentage of the total possible arguments present in the system: if $\mathsf{Acts} \in \{4, 6, 8, 10\}$ and $\mathsf{Vals} \in \{4, 6, 8, 10\}$, this frequency peaks when Args is around 75% of the total possible; if $\mathsf{Acts} = 2$, this frequency peaks when $\mathsf{Args} \approx 4$; if $\mathsf{Acts} \in \{4, 6, 8, 10\}$ and $\mathsf{Vals} = 2$, this frequency peaks when Args is around 50% of the total possible.

When $\mathsf{Acts} = 2$, the highest frequency of successful dialogues seen is lower than the highest frequencies seen for the other settings of Acts. Both the maximum and the minimum frequency of successful dialogues recorded is greater when more actions are under consideration, and the minimum frequency of successful dialogues is greater when more values are present in the system.

Generalising these results, we can say that the dialogue system performs better (i.e. reaches agreement more often) when $\mathsf{Acts} \neq 2$. The more values and the more actions present in the system the better the system performs, with the frequency of successful dialogues dependent on the percentage of the total possible arguments present in the system.

3.3 Quality of dialogue outcome is very rarely worse than quality of consensus outcome

We next consider for each run whether the dialogue system or consensus forming resulted in a higher outcome quality score. We investigated this across all possible parameter combinations; since we found a trend that repeats across the whole parameter space, we present in Figure 2 only the results for when $\mathsf{Acts} \in \{2, 6, 10\}$ and $\mathsf{Vals} \in \{2, 4, 8\}$.

Figure 2 shows clearly that only very rarely (in less than 3% of the runs across all possible parameter settings) does consensus forming produce a higher quality outcome than the dialogue system. However, if there are only two actions, then the two methods produce the same quality outcome more often than the dialogue system produces a higher quality outcome. This is a useful observation, particularly considering the higher computational overheads associated with the dialogue system.

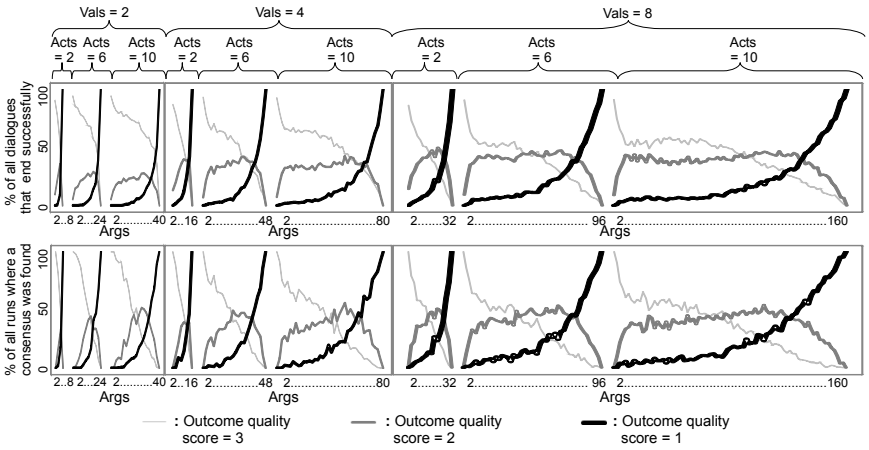

Fig. 3. Top: across all possible parameter settings where Acts $\in \{2, 6, 10\}$ and Vals $\in \{2, 4, 8\}$, percentage *of the dialogues that ended successfully* that received each outcome quality score. Bottom: across all possible parameter settings where Acts $\in \{2, 6, 10\}$ and Vals $\in \{2, 4, 8\}$, percentage of the runs *in which a consensus was found* that received each outcome quality score.

3.4 Successful dialogue outcomes are more likely to be globally agreeable to both agents than successful consensus outcomes

We now consider how the outcome quality score varies for successful outcomes produced by both the dialogue system and consensus forming across the parameter space. We performed this analysis across all possible parameter settings and found a trend that occurs across the entire parameter space; hence we present in Figure 3 only those results where where Acts $\in \{2, 6, 10\}$ and Vals $\in \{2, 4, 8\}$. The top of this figure shows what percentage *of the dialogues that ended in agree* received which outcome quality score. The bottom of this figure shows what percentage *of the runs in which a consensus was found* received which outcome quality score. (Recall the outcome quality score metric: 3 - outcome is globally agreeable to both agents; 2 - outcome is globally agreeable to only of the agents; 1 - outcome is not globally agreeable to either of the agents.)

As discussed earlier, Figure 1 shows that the frequency of dialogues that end successfully is considerably higher than the frequency of consensuses found, and that each of these frequencies vary as the parameters change; thus, it is important to bear in mind here that the percentages denoted on the y-axes of the graphs in Figure 3 relate to different sized sets depending on the particular parameter settings and on whether dialogue outcome or consensus outcome is being considered. Considering only the proportion of *successful* dialogues and consensuses that receive the different outcome quality scores (as seen in Figure 3) allows us to clearly see the following points.

Of the successful outcomes produced by both methods (consensus forming and the dialogue system), a higher proportion of those produced by the dialogue system are globally agreeable to each agent (i.e. outcome quality score = 3). The difference between the proportion of successful dialogues that receive outcome quality score 3 and

the proportion of consensuses that receive outcome quality score 3 is bigger the more actions and the fewer values that are present in the system.

It is interesting to note that the points on the graphs in Figure 3 where the green line (i.e. outcome quality score = 1) and the red line (i.e. outcome quality score = 2) intersect occur at the same position on the x-axis for both the dialogue outcome and the consensus outcome. If $\mathsf{Vals} = 2$, this occurs when Args is equal to approximately 95% of the total possible arguments, otherwise this occurs when Args is equal to approximately 80% of the total possible arguments. Thus, if a successful outcome is produced either by consensus forming or by the dialogue system and there are more than 80% of the total possible arguments present in the system (95% if $\mathsf{Vals} = 2$), it is likely that this outcome is not globally agreeable to either agent.

The quality of successful outcomes produced by both the dialogue system and consensus forming is most sensitive to the number of arguments present in the system, and is little affected by changes to the number of values or actions under consideration. Consensus forming is more sensitive than the dialogue system to the number of arguments.

3.5 Average dialogue outcome quality score is higher than average consensus outcome quality score

Figure 4 shows the average outcome quality score produced by both the dialogue system and the consensus forming mechanism across all parameter settings where $\mathsf{Args} = 25\%$, 50% and 75% of the total possible arguments. It is very clear from these results that, on average, the dialogue system outperforms consensus forming.

Looking at Figure 4 in more depth, we see that the highest outcome quality score averages for the dialogue system are seen when $\mathsf{Vals} = 2$, whilst this parameter setting produces the lowest outcome quality score averages for consensus forming. For all settings of Acts and Vals, the smallest difference between the outcome quality score averages of the two methods is seen when $\mathsf{Args} = 75\%$ of the total possible arguments. For all settings of Vals and Args, the smallest difference between the two outcome quality score averages is seen when $\mathsf{Acts} = 2$. We can conclude that if $\mathsf{Vals} = 2$ and $\mathsf{Acts} \neq 2$, it is likely that the outcome produced by the dialogue system will be higher quality than that produced by consensus forming.

3.6 Dialogue length grows exponentially with increasing arguments

Figure 5 shows that the time it takes to complete dialogues increases exponentially with the number of arguments. However as the number of values increases this trend flattens and increases are more linear. Indeed as values and actions increase the curve becomes almost sigmoidal. This indicates that if speed is a key factor for an applied dialogue system, deliberation dialogues are most useful when either the number of arguments is low or the number of values and actions is high.

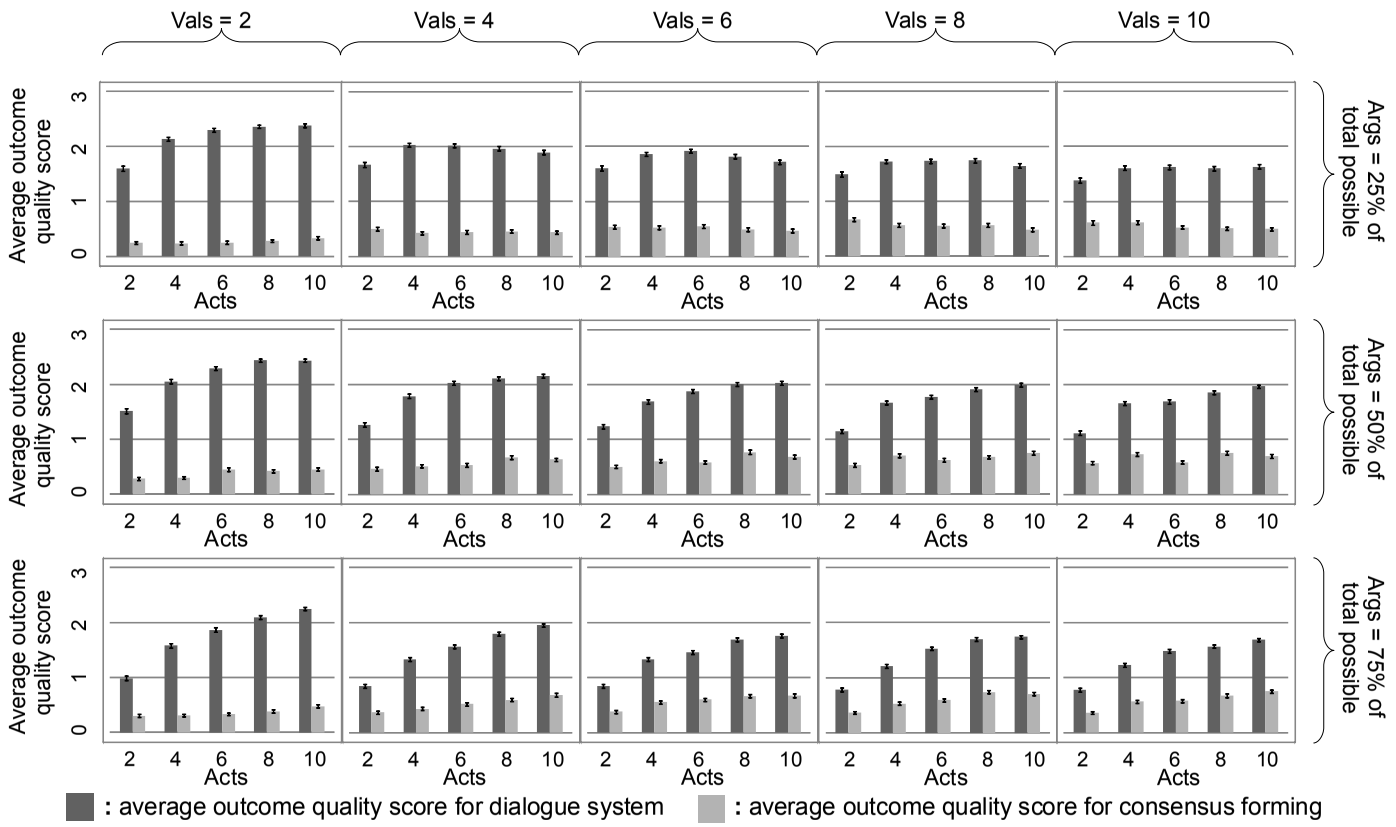

Fig. 4. Average quality outcome score over 1000 runs for both the dialogue system and consensus forming, across every parameter combination where Args = 25%, 50% or 75% of the total possible arguments. The error bars show the standard errors of the means.

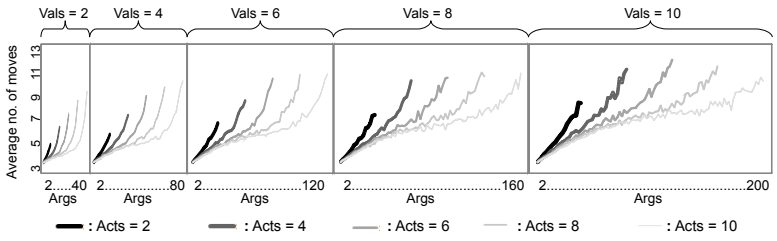

Fig. 5. Average number of moves in a terminated dialogue.

4 Discussion

We have presented empirical results from what we believe is the first simulation-based study of a deliberation dialogue system, where the agents involved used value-based argumentation to determine agreeable actions. Our results show that the dialogue system we present outperforms a simple consensus forming mechanism. We provide an in-depth analysis of the behaviour that can be expected from the system based on the number of actions, values and arguments that are present. For instance, the dialogue system reaches agreement more frequently when there is a higher number of actions and values under consideration; the quality of a successful dialogue outcome is more likely to be higher when there are less than 80% of the total possible arguments present.

These results take a significant step towards demonstrating the applicability of value-based deliberation dialogue systems, as well as demonstrating the importance of complementing theoretical evaluations with simulation-based studies. Our specific quantitative results can be compared against the parameters derived from a particular domain in order to determine the suitability of value-based deliberation dialogues.

Our simulation facilitates many avenues of future work, for example it is simple to adapt it to allow multiple agents and we are particularly interested in investigating different strategies that the agents might use and seeing how these compare with one another. Our next step is to analyse why the system behaves as it does. We have already begun to investigate how the topology of the OAF (which is itself determined by the combination of parameters) affects the dialogue behaviour, and it is clear that they are closely linked. Here, we have restricted the system so that the agents each get exactly half of the arguments present in the system; certainly altering this split will have a marked effect of the behaviour of the system and this is something we are keen to investigate. We also intend to extend our dialogue model to take into account the other stages of practical reasoning (problem formulation and epistemic reasoning [6]).

It would be very interesting to see how an argumentative agent would perform against a non-argumentative agent, such as one that uses classical decision theory to determine the actions it finds agreeable. There is a large body of work on computational social choice (see e.g. [9]), which considers mechanisms with which group decisions can be made. Although beyond the scope of this paper, we plan to compare deliberation dialogues with social choice mechanisms (more sophisticated that the simple consensus forming method presented here). Such comparisons of an argumentation-based

approach with approaches from other fields are of vital importance if we are to demonstrate the value of argumentation theory to the wider field of Artificial Intelligence.

Our investigation here takes a fundamental first step towards evaluating the potential benefit of a value-based deliberation dialogue system; however, it is not clear whether the scenarios that our simulation randomly generates are reflected in any real world setting. For example: Are there any real applications where more than 75% of all possible arguments are present in the system? Is it realistic that negative arguments are as likely to appear within the system as positive arguments? In order to be sure that the results are useful beyond a randomised setting, it is important to test argumentation-based approaches using real world data. This presents a challenge for the community, since it is hard to get access to such data that can be represented as arguments. We plan to collaborate with researchers working on real applications in order to validate our approach.

This simulation has been invaluable in identifying areas of future work that have the potential to be of benefit to real world applications, and in providing us with an implemented framework that we can adapt to investigate these areas.

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

References

1. Karunatillake, N.C., Jennings, N.R., Rahwan, I., McBurney, P.: Dialogue games that agents play within a society. Artificial Intelligence **173**(9-10) (2009) 935–981
2. Pasquier, P., Hollands, R., Rahwan, I., Dignum, F., Sonenberg, L.: An empirical study of interest-based negotiation. Autonomous Agents and Multi-Agent Systems **22**(2) (2011) 249–288
3. Jung, H., Tambe, M.: Towards argumentation as distributed constraint satisfaction. In: Proc. of AAAI Fall Symposium on Negotiation Methods for Autonomous Cooperative Systems. (2001)
4. Black, E., Atkinson, K.: Choosing persuasive arguments for action. In: Proc. of the 10th Int. Conf. on Autonomous Agents and Multi-Agent Systems. (2011) 905–912
5. Walton, D.N.: Argumentation Schemes for Presumptive Reasoning. Lawrence Erlbaum Associates, Mahwah, NJ, USA (1996)
6. Atkinson, K., Bench-Capon, T.J.M.: Practical reasoning as presumptive argumentation using action based alternating transition systems. Artificial Intelligence **171**(10–15) (2007) 855–874
7. Bench-Capon, T.J.M.: Agreeing to differ: Modelling persuasive dialogue between parties without a consensus about values. Informal Logic **22**(3) (2002) 231–245
8. Dung, P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence **77** (1995) 321–357
9. Chevaleyre, Y., Endriss, U., Lang, J., Maudet., N.: A short introduction to computational social choice. In: Proc. of the 33rd Conf. on Current Trends in Theory and Practice of Computer Science, LNCS 4362. (2007) 51–69