Arguing from similar positions: an empirical analysis

Josh Murphy, Elizabeth Black, and Michael Luck

Department of Informatics, King's College London, UK
firstname.surname@kcl.ac.uk

Abstract. Argument-based deliberation dialogues are an important mechanism in the study of agent coordination, allowing agents to exchange formal arguments to reach an agreement for action. Agents participating in a deliberation dialogue may begin the dialogue with very similar sets of arguments to one another, or they may start the dialogue with disjoint sets of arguments, or some middle ground. In this paper, we empirically investigate whether the similarity of agents' arguments affects the dialogue outcome. Our results show that agents that have similar sets of initially known arguments are less likely to reach an agreement through dialogue than those that have dissimilar sets of initially known arguments.

1 Introduction

Autonomous agents must often collaborate with others to achieve their goals, for example when it is impossible or inefficient to achieve them as individuals. One way for a group of agents to coordinate their actions is to participate in argument-based dialogues, which are structured interactions between participants, involving the exchange of formal arguments (e.g., [1]). There are many classes of argument dialogues, one such class being the deliberation dialogue, in which participants attempt to agree on an action. Such dialogues are a rational approach for agents to come to an agreement on how to act, allowing the opportunity for an agent not only to express their action preferences, but also to express the reasons for them. Thus, deliberation dialogues are important as a possible collaboration and coordination mechanism. However, if practical real-world applications for argument-based deliberation dialogues are to be developed, we need to understand the situations in which agents perform successfully in them.

The complexities of agent-based argument dialogues mean that often only a limited number of properties can be studied formally without making overly restrictive simplifications to the problem domain [2]. This can cause formal analysis of agent performance in such dialogues to be difficult. A complementary approach is to use simulation and empirical analysis; for example, Black and Bentley's experiments on simulations with a deliberation dialogue system found that argument-based deliberation dialogues typically outperform a basic consensus forming algorithm [3]. However, while their experiments explore a large and sensitive parameter space, they do not consider the similarity of arguments (they assume that agents have disjoint sets of initial arguments) which could be a contributing factor to the outcome of the dialogue.

In this paper, we also study the behaviour of deliberation dialogues using empirical methods, and investigate the dialogue system studied by Black and Bentley [3],

first presented by Black and Atkinson [4]. We extend Black and Bentley's analysis by considering whether the *similarity* of arguments at the start of the dialogue affects the likelihood of whether agents successfully reach an agreement. This similarity of agent arguments at the start of a dialogue is likely to vary in real-world applications, so it is especially pertinent to understand how this property affects the outcomes of dialogues.

Our results demonstrate that the similarity of the sets of arguments known to each agent has a statistically significant effect on the likelihood of dialogue success. We find that, in contrast to our intuition, the higher the similarity of initial arguments the lower the likelihood of success. We provide a justification for this relationship and analyse the extent of its effect across the parameter space, helping to identify cases where the use of argument-based deliberation dialogues is likely to be useful. The contribution of this paper is thus an analysis of how the similarity of arguments known to agents at the start affects the likelihood that agents will reach agreement through a deliberation dialogue.

The paper is structured as follows. In Section 2 we recapitulate the model of the dialogue system originally presented by Black and Atkinson [4]. In Section 3 we describe our implementation and method of experimentation, including how we varied the similarity of the sets of arguments that agents initially know about. In Section 4 we present the results of our experiments, including an analysis of observed trends and a detailed description of the relationships between variables. We discuss related work in Section 5. Finally, in Section 6, we discuss avenues of future work.

2 Deliberation dialogues

In this section we describe the model that specifies the deliberation dialogues investigated in this paper. This model is the same as that described by Black and Bentley [3], first presented by Black and Atkinson [4], and is based on the popular argument scheme and critical questions approach [5]. First we give details of the argumentation model that agents use to generate and evaluate arguments for and against different actions. We then describe the dialogue system used by agents to exchange these arguments, including the dialogue protocol that defines the structure of a deliberation dialogue, and the strategy that agents use to determine which of their arguments they will exchange.

2.1 Argumentation model

Our key concern is with the performance of the system specified in [4], in which agents have knowledge about the state of the world, about the preconditions and effects of actions they can perform, and about values that are either promoted or demoted by particular changes to the state of the world (these values represent qualitative social interests that an agent wishes to uphold; for example, fairness, health benefit, or personal privacy [6]). An agent can use its knowledge to construct arguments for or against actions by instantiating a *scheme for practical reasoning* [7]: in the current circumstances R, we should/should not perform action A, which will result in new circumstances S, which will achieve goal G, which will promote/demote value V. For example, an agent with the goal to be at the park may be able to construct the following arguments for and against actions to achieve its goal (note that we omit the current and new circumstances from these arguments, assuming the reader can envisage appropriate instantiations).

– **A1:** We should *cycle* (action) because it promotes *personal well-being* (value) in achieving *getting to the park* (goal).
– **A2:** We should not *drive* (action) because it demotes *environmental well-being* (value) in achieving *getting to the park* (goal).
– **A3:** We should *drive* (action) because it promotes *timeliness* (value) in achieving *getting to the park* (goal).

The scheme for practical reasoning is associated with a set of characteristic critical questions (CQs), which can be used to identify challenges to proposals for action that instantiate the scheme. These critical questions each relate to one of three reasoning stages: *problem formulation*, which considers the knowledge agents have about the problem domain (e.g., whether the preconditions and effects of actions are correct, whether state transitions promote or demote particular values); *epistemic reasoning*, where agents determine the current circumstances; and *action selection*, where agents construct and evaluate arguments for and against different action options. The deliberation dialogues we study here consider only action selection, assuming that the other stages have been dealt with previously with other types of dialogue; this action selection stage determines three CQs for consideration (we use the numbering of CQs used in [7]; see [4] for a more detailed justification of the appropriateness of these CQs).

– **CQ 6:** Are there alternate ways of realising the same goal?
– **CQ 9:** Does doing the action have a side effect that demotes some other value?
– **CQ 10:** Does doing the action have a side effect that promotes some other value?

From these CQs we can identify attacks between arguments for and against actions to achieve a particular goal: two arguments *for* different actions attack one another (CQ6); an argument *against* an action a attacks another argument *for* the same action a (CQ9); two arguments *for* the same action that each promote different values attack one another (CQ10). Considering the example arguments given above, A1 attacks A3, A3 attacks A1, and A2 attacks A3.

Each agent has a (total-order) ranking over the values, referred to as its *audience*, which represents the importance it assigns to them. An agent uses its audience to determine the relative strength of arguments according to the values they each promote/demote, and thus whether an attack succeeds as a defeat. In the example above, an agent who finds personal well-being to be a more important value than timeliness will find argument A1 to be stronger than A3 and so will determine that A1 defeats A3, while A3's attack on A1 does not succeed as a defeat.

Given a set of arguments, the attacks between those arguments (determined by the CQs above), and a particular agent's audience, we evaluate the acceptability of an argument with respect to that agent with a Value Based Argumentation Framework (VAF) (introduced in [6]), 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 (if the attack is symmetric) or at least as high as (if the attack is asymmetric) the value of the argument attacked (according to a particular agent's 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 that results from removing all the attacks that are unsuccessful given the audience A. An argument is said to be *acceptable* to the agent if it is part of a maximal admissible set (a *preferred extension*) of the VAF evaluated according to the agent's audience.

An agent considers an action to be *agreeable* if it finds some argument *for* that action to be acceptable. Considering the example arguments given above, if an agent prefers *environmental well-being* to *timeliness*, which it prefers to *personal well-being*, it will find arguments A2 and A1 to be acceptable and conclude that the only agreeable action is to cycle (since this is the only action for which it has an acceptable argument). If, however, the agent prefers *timeliness* to *personal well-being*, which it prefers to *environmental well-being*, it will find arguments A2 and A3 to be acceptable, and so will determine that driving is the only agreeable action to achieve its goal.

Observe that arguments against actions are always acceptable given the instantiation of attacks derived from CQs and these are not considered by the agent in determining which actions it finds agreeable. Intuitively, this is because the CQs are concerned with evaluating presumptive proposals *for* performing some action. It would be possible (and we believe would not affect the results of our experiments) to adapt the VAF generation and evaluation so as to produce the same results in terms of agreeability of actions while avoiding the (perhaps unintuitive) case where both an argument for and an argument against an action are found to be acceptable; we choose here not to adapt the model in order that our results are relatable to previous work [4,3].

We can also see that (as in [4]) if an attack is symmetric, then an attack only succeeds in defeat if the attacked argument's value is more preferred than the value of the argument being attacked; however, if an attack is asymmetric, then an attack succeeds in defeat if the attacking argument's value is at least as preferred as the value of 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' values are equally preferred, then it is undesirable for the argument for the action to withstand the attack. If we have a symmetric attack where the values of the arguments attacking one another are equally preferred, then it must be the case that each argument is for a distinct action but promotes the same value; here, the attack does not succeed as a defeat, since it is reasonable to choose either action.

We have described the mechanism that an agent uses to 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 the acceptability of the arguments and, from this, the agreeability of actions. Next, we describe the dialogue system that agents use to jointly reason about the agreeability of actions.

2.2 Dialogue System

Deliberation dialogues take place between two participating agents (each with an identifier taken from the set $\mathcal{I} = \{x, \overline{x}\}$) and we assume that the dialogue participants have already agreed to participate in a deliberation dialogue in order to agree on an action to perform in order to achieve some mutual goal (this goal is the *topic* of the dialogue). At the start of the dialogue, each agent has available to it a set of arguments for and against actions to achieve the goal, which are those arguments it can construct from its private

knowledge about the state of the world, the different actions that can be performed, and the values promoted or demoted by those actions. Each agent also has an audience (its ranking over the values).

During the course of the dialogue, agents take it in turns to make a single *dialogue move*. There are four types of dialogue move that participants may make:

- `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`).

A dialogue terminates under two conditions: once two consecutive `pass` moves appear (in which case the dialogue is a *failure*, and no agreement has been reached), or two consecutive `agree` moves appear (in which case the dialogue is a *success*).

In order to evaluate which actions it finds agreeable at a point in the dialogue, an agent considers all the arguments it is aware of at this point and evaluates them as described in the previous section; it thus constructs a VAF consisting of the arguments it is initially aware of at the start of the dialogue and those arguments that have been asserted previously in the dialogue by the other agent, and evaluates this according to its audience. An action is *agreeable* to the agent if there is some argument *for* that action that it finds acceptable given this evaluation. Note that the set of actions that are agreeable to an agent may change over the course of the dialogue, due to it becoming aware of new arguments as they are asserted by the other participant.

A dialogue protocol specifies which moves are permissible for an agent x during x's turn in a deliberation dialogue with topic p as follows:

- It is permissible to `assert` an argument A iff the argument is for or against an action to achieve the topic p 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`.

While the dialogue protocol defines a set of moves it is permissible to make, an agent uses a particular *strategy* to decide which of the permissible moves to select. 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; otherwise
- if it is permissible to `assert` a positive argument *for* an action that the agent finds *agreeable*, then assert some such argument; otherwise
- 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; otherwise
- make a `pass` move.

3 Investigating similarity

Previous work has considered whether there is a relationship between the number of unique values and actions being argued over, the number of arguments known by agents, and the likelihood of agents reaching agreement through use of the deliberation dialogue system [3]. However, in those experiments the sets of arguments agents know at the start of the dialogue are always disjoint. It is possible, perhaps even likely, that in real world examples of agent dialogues there will be some overlap in the agents' initial argument sets. Thus, we are interested here in the question of whether the similarity of agents' initial arguments sets has an effect on the resulting dialogue.

To investigate this we perform experiments where we vary not only the number of unique values and actions being argued over and the number of arguments known, but also sim (a measure of the similarity of the sets of arguments known by each agent at the start of the dialogue). We thus require four parameters as follows.

1. acts : The number of unique actions that can be argued about.
2. vals : The number of unique values that can be promoted or demoted by the actions.
3. args : The number of unique arguments in the union of both agents' initial arguments.
4. sim : A measure of the similarity of the agents' sets of initial arguments.

To run experiments across the parameter space, a random scenario generator is required; this initialises the arguments known to each agent at the start of the dialogue (referred to as their *initial arguments*) and their audiences. For each run of the simulation, the scenario generator is given acts actions, and vals values. It then generates all possible arguments that can be constructed from the set of actions and the set of values: for each action and value pair there are two arguments that can be produced; one argument that claims performing the action will promote the value; and the other argument that claims performing the action will demote the value. Therefore, the set of all possible arguments contains $2 \times \text{acts} \times \text{vals}$ many arguments.

Then, random arguments are removed from the set of all possible arguments until it contains args arguments. Note that if $\text{args} = 2 \times \text{acts} \times \text{vals}$ then no arguments need to be removed. Half of the arguments remaining in the set are randomly distributed to one agent, with the other half being distributed to the other agent. The arguments that are distributed to an agent simulate the set of initial arguments that it can generate using its knowledge. The set of initial arguments distributed to an agent x is denoted R^x.

It is clear to see that, at this point, R^{x_i} and R^{x_j} would be disjoint sets. However, this is not always the case in agent dialogues. Two arguing agents are likely to have some overlaps in their knowledge and hence may be able to generate and communicate the same arguments. We introduce the sim parameter to determine how similar the sets R^{x_i} and R^{x_j} should be — the higher the value of sim the more arguments that are shared between agents. So, once R^{x_i} and R^{x_j} have initially been determined, $(\text{args}/2) \times \text{sim}$ random arguments from each set are copied into the other set. It can be seen that after this sharing process, if *sim* $= 1$ then agents will have args many arguments each, and the arguments the agents each have will be identical. Similarly, if $\text{sim} = 0$ then the agents will have $\text{args}/2$ arguments each, and the arguments each agent has will remain disjoint (note, this is equivalent to the situation studied by Black and Bentley [3]).

The *total number of arguments* in a dialogue scenario refers to the sum of the number arguments initially known to one agent plus the number of arguments initially known to the other agent, and is calculated from the experiment parameters according to the following formula $\lceil \mathsf{args} + (\mathsf{args} \times \mathsf{sim}) \rfloor$.

Our experiments investigate whether the similarity of agents' initial arguments has an effect on the simulated deliberation dialogues, across the following different parameter combinations.

- $\mathsf{sim} \in \{0, 0.1, \ldots, 0.9, 1.0\}$,
- $\mathsf{vals} \in \{2, 4, 6, 8, 10\}$,
- $\mathsf{acts} \in \{2, 4, 6, 8, 10\}$,
- $\mathsf{args} \in \{2, 3, \ldots, (\mathsf{vals} \times \mathsf{acts} \times 2)\}$.

The randomised nature of the scenario generator and resulting simulated dialogue means that generated dialogues are not only sensitive to the input parameters, but also an element of chance. As a result, many dialogues must be simulated for each parameter combination: it is not sufficient only to run a single instance of a dialogue because two dialogues generated with the same parameter combination can still differ on the distribution of arguments among the agents, and the randomised aspect of the agents' strategy (agents select a random dialogue move when more than one is determined by the strategy). Thus, for each parameter combination, we simulate 1,000 dialogues and, for each dialogue, we record whether it ended successfully (with both agents having agreed on an action) or unsuccessfully (with agents failing to reach an agreement).

The argumentation model, dialogue system, and scenario generator were implemented in Java (independently from any argumentation libraries), and all simulations and experiments were run on a standard workstation computer. The source code can be found online at `github.com/joshlmurphy`.

4 Results

Black and Bentley [3] also studied the likelihood of success across the parameter space considered here, but only for dialogues in which $\mathsf{sim} = 0$. By limiting our parameter space to dialogues in which $\mathsf{sim} = 0$ we obtain a very close reproduction of results: we again witness that successful dialogues are more likely with higher numbers of actions and values, and we can observe the relationship between the total number of arguments and the likelihood that the dialogue ends successfully (for low numbers of values and actions there is a decrease in the likelihood of dialogue success as the number of arguments increases, while for higher numbers the relationship is more complex, with likelihood initially decreasing as the total number of arguments increases up to a certain point, after which the likelihood of dialogue success begins to increase).

However, by considering the different values of sim, we are able to make a number of empirical observations from which novel conclusions can be drawn. In each of the following subsections, we describe a particular aspect of our results, provide an explanation for what has been observed, and discuss the significance of the result.

A representative subset of our results is shown in Figures 1–4. Figures 1–3 each presents three graphs showing the percentage of dialogues that end in success (y-axis), at different numbers of total arguments (x-axis), for different values of sim (the darker

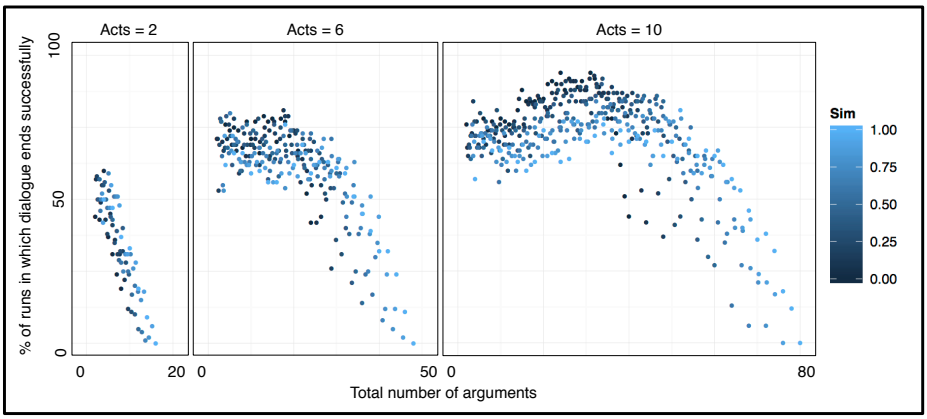

Fig. 1. Graphs to show relationship between total no. of arguments and % of dialogues that ended successfully, for different values of sim when vals=2 (1000 runs for each parameter setting).

the shade of the plot, the lower the value of sim). The figures show the results for dialogues where vals $= 2$ (Figure 1), vals $= 6$ (Figure 2), and vals $= 10$ (Figure 3). The graphs in each figure show the results for dialogues where acts $= 2$ (leftmost), acts $= 6$ (centre), and acts $= 10$ (rightmost). Each point represents the average of 1,000 simulated dialogues with that parameter combination. Similar results were seen across all combinations of vals and acts; we present only a representative sample here.

4.1 Dialogues tend to fail with many arguments

From the results in Figures 1–3 we can see that dialogue success is very unlikely at high levels of total arguments (every graph tails into a 0% rate of dialogue success as the number of total arguments tends towards its maximum value for the parameter combination). The reason for this is that if an agent believes every possible argument over a set of values and actions then it will find no action acceptable: all arguments for doing a particular action because that action promotes some value will be defeated by the negative argument that demotes that action for the same value, and hence the action will not be agreeable to the agent. In the case where agents start the dialogue with every possible argument over a set of values and actions, the agents begin the dialogue finding no actions agreeable and have no possibility of ever finding an action agreeable (since they know all arguments, no asserted argument during the dialogue will change the actions that are acceptable); this corresponds to the plot in the graphs where sim $= 1$, and the total number of arguments is $2 \times$ acts $\times$ vals.

This observation cannot be made without considering dialogues in which sim $\neq 0$ because, with low similarities, higher numbers of total arguments cannot be reached and so at low similarities, dialogues cannot have a large enough number of arguments to reveal this trend. This can be seen in Figures 1–3 where no plots for sim $= 0$ exist beyond 50% of the graphs' maximum of the number of total arguments.

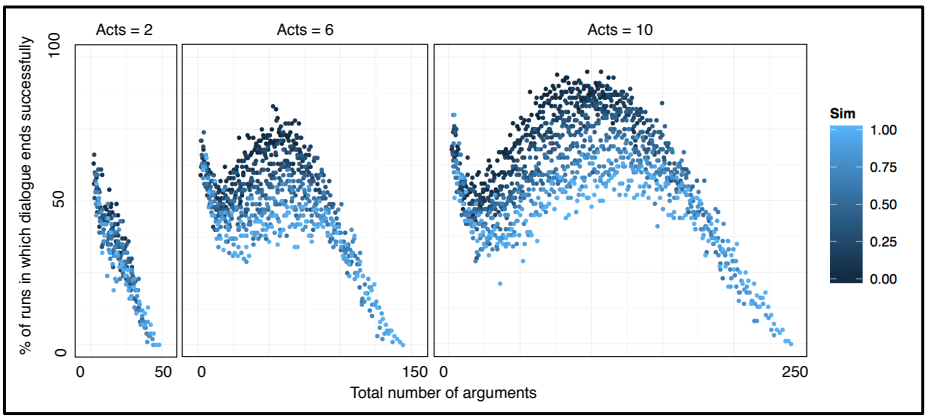

Fig. 2. Graphs to show relationship between total no. of arguments and % of dialogues that ended successfully, for different values of sim when vals=6 (1000 runs for each parameter setting).

Past a certain point, the more arguments an agent knows, the greater the chance that dialogue success is impossible, and this effect becomes severe at high levels of total number of arguments. Thus, it is not the case that the complete failure of dialogues for very high number of total arguments is the fault of the dialogue system but rather is down to the likely impossibility of an agent finding any action agreeable when knowing this many arguments. In real-world scenarios, it is unlikely that an agent will have arguments both for an against an action motivated by the same value, since one would expect this to be resolved in the problem formulation stage of reasoning, and so we consider these types of dialogue to be unrealistic.

Thus, importantly, our results show that when using the deliberation dialogue, agents will not come to an agreement when it would not be rational for them to agree to do any of the possible actions. This result was proven theoretically by Black and Atkinson [4].

4.2 Dialogues are less successful as sim increases

Given these initial results, we investigated whether the likelihood of success of a dialogue (measured by whether the dialogue ends in agreement or not) is affected by the similarity of the two agents' initial arguments (measured by the sim parameter). Looking at Figures 1–3, we can see how the sim parameter affects the rate of dialogue success across different numbers of values, and actions, and total numbers of arguments. Perhaps surprisingly, the general trend is that agents that have similar sets of initial arguments are less likely to reach an agreement compared to agents that have dissimilar sets of initial arguments. The trend violates the intuition that agents with similar knowledge should be able to agree more easily. Indeed, this trend was present across the entire parameter space (except from when both acts $= 2$ and vals $= 2$, which we discuss in Section 4.3), so we present only a representative subset of the results. We observed very similar results for other combinations of vals and acts.

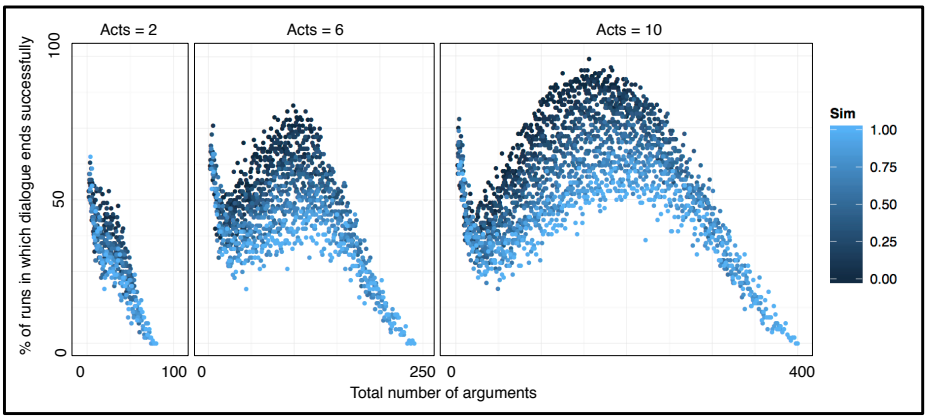

Fig. 3. Graphs to show relationship between total no. of arguments and % of dialogues that ended successfully, for different values of sim when vals=10 (1000 runs for each parameter setting).

We assessed the relationship between the similarity of agents' initial arguments and the rate of success of the dialogue averaged over the total number of arguments in dialogues where the number of actions was 10 and the number of values was 10. This assessment was undertaken by calculating a Pearson product-moment correlation coefficient, which showed that there is a very strong, negative relationship between the two variables (coefficient $r = -0.96$, statistical significance $p < 0.001$), indicating that the more similar the agents' initial arguments the less likely the dialogue will end in success. The scatterplot in Figure 4 displays these results.

We explain this relationship as follows. When a dialogue is initialised with sim $=$ 1 (i.e. agents' initial sets of beliefs are identical) any argument an agent asserts will already be known by the other agent. In these dialogues, the agents' sets of known arguments remain the same throughout the dialogue (since any asserted argument will already be known by both agents) so the actions an agent finds agreeable at the start of the dialogue remain the same at every subsequent turn. If agents do not have any agreeable actions in common at the start, then they never will, so the dialogue will fail. Conversely, when a dialogue is initialised with sim $= 0$ (i.e. agents' initial arguments are entirely disjoint), any argument an agent asserts throughout the dialogue will be novel for the other agent, potentially changing the actions it finds agreeable, and hence the actions that are agreeable to both agents. The more often an assert move changes the actions agreeable to both agents, the more likely it is that throughout the course of the dialogue there will be a point at which there is at least one action agreeable to both agents. In summary, the lower the similarity of the initial arguments, the greater the chance there will be at least one point in the dialogue at which agents mutually find at least one action agreeable, and hence the greater the chance of dialogue success.

Understanding the relationship between the similarity of agents' arguments at the start of a dialogue and the likelihood of dialogue success is important to understand situations in which deliberation dialogues are useful in trying to agree on an action; this can help identify real-world scenarios in which this technique can usefully be applied.

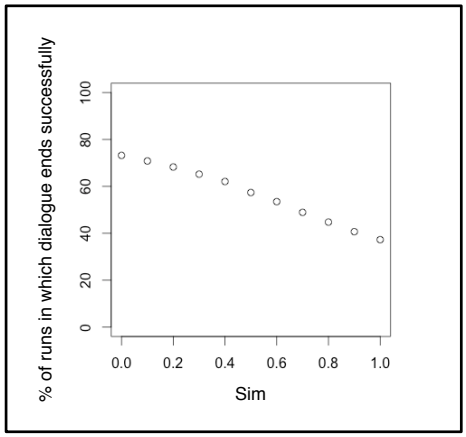

Fig. 4. A scatterplot to show the relationship between the similarity of initial belief sets and the rate of success of the dialogue averaged over the total number of arguments (1000 runs for each parameter setting), in dialogues where vals=10 and acts=10.

4.3 The impact of similarity increases with the number of values

Varying sim for dialogues with a low number of values produces a relatively small effect on the likelihood of the success of the dialogue. For example, dialogues with 2 values are affected only slightly by changing sim; as can be seen in Figure 1, the distances between plots for $\mathsf{sim} = 1$ and $\mathsf{sim} = 0$ are low, within 15%. Looking at Figure 2 where the dialogues have 6 values, the distances between plots for $\mathsf{sim} = 1$ and $\mathsf{sim} = 0$ are wider in general, and this is evidence of an increasing effect of sim at higher values. The distances are greater still for dialogues with 10 values, as seen in Figure 3, where we observe a nearly 50% difference in the likelihood of success of the dialogue between dialogues where $\mathsf{sim} = 1$ and $\mathsf{sim} = 0$.

Generalising these results, we can say that the impact of agents having similar sets of initial arguments on the likelihood of dialogue success increases as the number of values that agents argue over increases. This tells us that similarity has a large effect on the likelihood of dialogue success in such scenarios.

4.4 Success most likely at around 50% of maximum total arguments

For dialogues in which $\mathsf{vals} = 2$ or $\mathsf{acts} = 2$ we observe a general decrease in the likelihood of dialogue success as the total number of arguments increases. Furthermore, for dialogues in which $\mathsf{acts} = 2$ we observe a *linear* decrease in the likelihood of dialogue success as the total number of arguments increases, regardless of the number values. This relationship can be seen in the relevant graphs in Figures 1–3, and was also observed by Black and Bentley [3].

The relationship between the total number of arguments and the likelihood of success is more complex when we consider dialogues in which $\mathsf{vals} > 2$ and $\mathsf{acts} > 2$. The relationship can be described in three stages. First, in the lowest 10% of a graph's

maximum total number of arguments we observe a decrease in the likelihood of dialogue success similar to that in lower numbers of values and actions. However, in the second stage, after the 10% point up to approximately 50% of a graph's maximum total number of arguments, the trend reverses and we observe an increase in the likelihood of dialogue success as the total number of arguments increase. The trend reverses again in the third stage, after 50% of a graph's maximum total number of arguments onward, where we observe a tail off towards a 0% likelihood of dialogue success. This relationship can be seen in the relevant graphs in Figures 1–2. This more complex relationship was not observed by Black and Bentley [3] because very high total numbers of arguments can only be reached by considering $\mathsf{sim} > 0$. Dialogues with a low sim are less affected in the initial stage of the relationship and are more greatly affected in the second stage (the trough is shallower, and the peak is higher), whereas dialogues with a high sim are more affected in the initial stage of the relationship and are less affected in the second stage (the trough is deeper, and the peak is lower).

The shape of the relationship between the total number of arguments and the likelihood of dialogue success as described here would have been extremely difficult to prove using formal methods. However, by using the experimental approach we are able to investigate performance across the entire parameter space. The observation of the shape of the relationship is useful because it allows us to predict accurately the chance a dialogue will succeed for any given parameter combination.

5 Related work

Our experiments are closely related to those of Black and Bentley [3], which are based on the same argumentation model and dialogue system [4] as the work presented in this paper. Their work was perhaps the first to use empirical methods to evaluate the benefit of using deliberation dialogues. In their experiments, they vary the number of values and actions being deliberated over, and the number of arguments available to agents at the start of the dialogue and show that the deliberation dialogue system typically outperforms consensus forming. Here, we expand the parameter space to also vary the similarity of the arguments that the agents have and show that this is an important factor in the success of a deliberation dialogue.

Kok *et al.* similarly take an empirical approach to the investigation of argument-based deliberation dialogues [9]. They focus on the expressive potential of argumentation by using a deliberation dialogue system that allows agents to communicate using elaborate arguments, assuming that agents that are able to express themselves better would be able to perform more efficiently and more effectively. They show that an arguing strategy offers increased effectiveness over a non-arguing strategy. In their work, agents' arguments are generated from their respective knowledge bases, but they do not consider how the performance of the dialogues depends on the similarity of these knowledge bases, or the similarity of arguments that are generated from them.

In considering groups, Toniolo *et al.* investigate how argument-based deliberation dialogues can be used by a team of agents that have their own potentially conflicting goals and norms [10]. Using an empirical evaluation of their model, they find that argument dialogues are a more effective means of agent coordination than collaborative

plans (using the metric of the feasibility of the resulting plan). While their work does consider agents as heterogeneous with their own goals and norms, they do not consider how the similarity of their goals and norms (and hence their arguments) affects the quality of the plans produced.

Finally, Medellin-Gasque *et al.* present a dialogue protocol for deliberation and persuasion dialogues, in which agents argue over cooperative plans [11]. Interestingly, they investigate the impact of the agents' dialogue strategies on the result of the dialogue. Similar to our work, their dialogue system is based on the critical questions approach [5]. They implement 3 different agent strategies (a random strategy, and 2 strategies that place some priority over dialogue moves), which they test over a limited number of cases (20 initial states, generated from 4 different sets of information, and 5 different preference orders over values). Their results show that, for the cases and strategies tested, the quality of the outcome of the dialogue does not vary by altering the agents' strategies, but by using a priority strategy rather than a random strategy, the outcome can be reached more efficiently. Thus, agents' dialogue strategies can be an important consideration for dialogues, in at least some initial circumstances. In the results we present here, we have considered only a single dialogue strategy, however we have run some preliminary experiments that consider the performance of two other strategies, one strategy that selects a random move to make from the entire set of valid moves, and one strategy that is similar to that which is presented in this paper but prioritises assert moves according to the agents preferences (assert moves for positive arguments with high preference actions and assert moves for negative assert moves with low preference actions have priority). Results from these preliminary experiments indicate that the relationship between the similarity of agents' initial arguments and the likelihood of dialogue success is independent of the strategy used.

6 Discussion

Our results show how, in the argument-based deliberation dialogues investigated here, the similarity of agents' initial arguments affects the likelihood that a dialogue ends in success. We found dialogues with high similarities of initial arguments are less likely to end in agreement than dialogues with low similarities of initial arguments, because the higher the similarity of initial arguments the less potential for agents to reach a point in the dialogue at which there exists at least one action that is agreeable to both agents. Using an empirical approach, our investigation allowed a total analysis of the parameter space over a large sample size of dialogues. Our results identify scenarios in which using a deliberation dialogue is likely to lead to an agreement being reached.

In our investigation we explored the entire range of possible similarities of agents' initial arguments: from dialogues where agents started with entirely disjoint sets of initial arguments to dialogues where agents started with identical sets of initial arguments. Across this range we identified a statistically significant effect of similarity on the likelihood of dialogue success, but, it is unclear to what extent this range typically exists in real-world scenarios. The relationship between the sets of initial arguments we randomly generate to those seen in real-world applications is also not understood (for example, dialogues that were generated with a very high number of total arguments are

probably not realistic). The lack of real-world data is an identified problem in research relating to applications of argumentation.

A potential solution for ensuring that our results relate to examples of argument dialogues in the real-world would be to use data from argument corpora for our experiments (e.g. [12]), rather than the randomly generated scenarios we used. Such an argument corpus is an organised collection of examples of real-world argument dialogues presented in a standardised format. However, there are three main limitations of current argument corpora. The first is that they are limited in the number of argument examples that they contain, and this would limit the viability of an experimental approach that uses this data. Basing conclusions on the small samples of data provided by argument corpora could be difficult to justify. In contrast, we were able to run 1,000 dialogues for each parameter combination (around 5.5 million dialogues), ensuring statistical significance of our observations. The second limitation is that the corpora typically focus on only a limited scope of topics: humans engage in deliberation dialogues on a wide range of topics, and only some of these are captured in current argument corpora (they tend to focus on legal or governmental dialogues). Focusing on only a subset of potential dialogue topics limits the conclusions that can be drawn from the data to that specific topic domain. The final limitation is the way in which arguments in the corpora are formatted. Transcription of natural language arguments to a standardised format is a complex process. Work is being done to allow this step to be done computationally [13], but, particularly as it is tied to the problem of natural language processing, it is likely to be a long time before this is an automatic and successful method. As a result, the transcribing and formatting of the arguments are often done by humans. This leads to a potential bias in the corpus, and could come from a number of sources: the human's aptitude for formal argumentation, personal opinions on the argument topic, and the knowledge the transcriber has of the argument topic.

There is a question as to whether measuring the quality of a deliberation dialogue simply on whether agents reach an agreement is the best or only measure. According to Walton and Krabbe [14], while there is a *public* goal to reach an agreement that is ascribed to by both agents in a deliberation dialogue, agents also have a *private* goal to influence the agreed upon action to one that is as favourable as possible to itself. Working out a suitable metric for the success of an agent's private goal is non-trivial as it is unclear how to accurately measure the influence an agent has had on the dialogue, and it is unclear how to measure which action is an agent's most favoured (should it be the agreeable action that promotes the highest value given local beliefs of the agent, or given global beliefs of the system). There are also other factors that could be used to measure the outcome of the dialogue: efficiency/speed of the dialogue (what resources were spent during the dialogue?), soundness of the agreed upon action (is the agreed upon action the best course of action from a global perspective?), and fairness (is the outcome representative of all of the agents' preferences?). For example, Black and Bentley assign scores to dialogue outcomes, depending on whether the agreed upon action is globally agreeable to both, one, or neither agent. However, there are many other possible ways to measure the quality of a deliberation dialogue.

Walton *et al.* [15] question whether models of deliberation dialogues are able to actually capture the richness and depth of human-like deliberation dialogues. Specifically,

they consider dialogues in which information available to participants of the dialogue is dynamic. This is certainly a limitation of our investigations since the knowledge the agents have remains the same throughout the duration of the dialogue. If we extended the dialogue system to simulate changing knowledge of the environment during the course of the dialogue, an interesting investigation would be to see how the similarity of the information/arguments made available to both agents would affect the dialogue (i.e. what happens if the information made available to agents becomes gradually more different or if the information becomes gradually more similar?).

References

1. McBurney, P., Parsons, S.: Dialogue games for agent argumentation. In Simari, G., Rahwan, I., eds.: Argumentation in Artificial Intelligence. Springer (2009) 261–280
2. Rahwan, I.: Argumentation in multi-agent systems. Autonomous Agents and Multi-Agent Systems **11** (2005) 115–125
3. Black, E., Bentley, K.: An empirical study of a deliberation dialogue system. In Modgil, S., Oren, N., Toni, F., eds.: Theory and Applications of Formal Argumentation. Springer (2012) 132–146
4. Black, E., Atkinson, K.: Choosing persuasive arguments for action. In: Proceedings of the Tenth International Conference on Autonomous Agents and Multi-Agent Systems. (2011) 905–912
5. Walton, D.N.: Argumentation Schemes for Presumptive Reasoning. Lawrence Erlbaum Associates (1996)
6. 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
7. 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
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. Kok, E.M., Meyer, J.J.C., Prakken, H., Vreeswijk, G.A.: Testing the benfits of structured argumentation in multi-agent deliberation dialogues. In: Proceedings of the Eleventh International Conference on Autonomous Agents and Multiagent Systems. (2012) 1411–1412
10. Toniolo, A., Norman, T.J., Sycara, K.P.: An empirical study of argumentation schemes for deliberative dialogue. In: Proceedings of the Twentieth European Conference on Artificial Intelligence. (2012) 756–761
11. Medellin-Gasque, R., Atkinson, K., Bench-Capon, T., McBurney, P.: Strategies for question selection in argumentative dialogues about plans. Argument & Computation **4**(2) (2013) 151–179
12. Reed, C.: Argument corpora. Technical report, University of Dundee Technical Report, Available online at `www.arg.dundee.ac.uk/corpora` (2013)
13. Green, N., Ashley, K., Litman, D., Reed, C., Walker, V., eds.: Proceedings of the First Workshop on Argumentation Mining. Association for Computational Linguistics (2014)
14. Walton, D.N., Krabbe, E.C.W.: Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. SUNY Press (1995)
15. Walton, D., Toniolo, A., Norman, T.J.: Missing phases of deliberation dialogue for real applications. In: Proceedings of the Eleventh International Workshop on Argumentation in Multi-Agent Systems, Springer (2014)