Investigating Strategic Considerations in Persuasion Dialogue Games

Christos HADJINIKOLIS a, Sanjay MODGIL a, Elizabeth BLACK a,
Peter MCBURNEY a, and Michael LUCK a

a Department of Informatics, King's College London, UK

Abstract. This paper provides a framework for argumentation-based persuasion dialogues that enables a participant to implement strategies based on its modelling of its interlocutor’s knowledge. The framework is defined on the basis of the recent ASPIC + general model of argumentation, and thus accommodates a range of possible instantiations. We extend existing works on persuasion by accounting for both admissible and grounded semantics, and also by allowing participants to not only move arguments that attack those of their interlocutor, but also preferences which undermine the success of these attacks as defeats. We also state formal results for these dialogues, and then use these dialogues to illustrate that appropriate mechanisms for strategising need to account for the logical content of arguments, rather than just rely on their abstract specification.

Keywords. dialogue, strategies, argumentation

Introduction

This paper deals with the notion of strategising in argument based dialogue systems. In general, such systems formalise how participants in various kinds of dialogue, exchange locutions with respect to a dialogical objective. These locutions may explicitly or implicitly define arguments whose interaction and evaluation bears on the outcome of the dialogical objective. In such systems, dialogues are perceived as games, where at any given stage, the dialogue’s protocol determines a set of possible moves (locutions) that an agent can play (utter) in reply to a move (locution) of its interlocutor. The strategy problem concerns choosing a move out of that set, so as to maximise a participant’s chances of satisfying its self-interested objectives.

Some researchers propose that a participant’s strategic considerations be based on the anticipated outcomes of the various choices, where these outcomes are evaluated based on how the dialogue may proceed given the participant’s modelling of its interlocutor’s replies to each choice. This naturally reflects real world dialogues in which we select utterances based upon what we believe are our interlocutor’s beliefs (and goals). This is known as opponent modelling, and is employed in [2,6,12,17,18]. Other researchers rely on the dialectical obligations of a participant for implementing and employing a strategy [1]. These are, as explained in [14], the expectancies created by commitments in a dialogue, such as supporting a proposition when challenged or else retracting it. However, while the above approaches are successful in providing theoretically sound
methods for dealing with the strategy problem, the fact that some of them (e.g. [12,16]) rely on abstract level argumentation frameworks for defining both the set of the possible response moves, as well as the strategic concepts that characterise a participant’s choice, results in not accounting for the logical content and structure of arguments, and their possible effects on a dialogue game. The latter concerns the possibility of new information being made available as a result of dialogical locutions; information that may be used to construct arguments that are not available from the outset of the game. Even in approaches that do account for the underlying logical instantiation of arguments, this dynamic aspect is largely neglected. There are some exceptions (e.g. [3,8]) that do account for the dynamic construction of arguments. However, as far as we are aware, no existing framework both allows for a logical conception of arguments and considers strategies for highly competitive contexts that rely on the notion of opponent modelling.

In this work we rely on a framework for argumentation-based persuasion dialogues that enables a participant to implement strategies based on its modelling of its interlocutor’s knowledge and goals. So as not to compromise the generality of the framework, we define the dialogical framework (i.e., the locutions, arguments, their relations and evaluation) on the basis of the recent ASPIC$^+$ model of argumentation [11,15]. ASPIC$^+$ has been shown to capture a number of well known logical approaches to argumentation and to satisfy [5]’s rationality postulates. We review ASPIC$^+$ in Section 1 and then we briefly present a general dialogical framework in Section 2, based on which we rely for instantiating a persuasion dialogue instance that we present in section 2.1. In contrast with existing work on persuasion dialogues (e.g., [14]), we allow the participants to not only move arguments that attack those of their interlocutor, but also (possibly contradictory) preferences that undermine the success of these attacks as defeats. Existing works allow only for the moving of defeating arguments, and so need to make the unrealistic assumption that the preferences and so defeat relations of the dialogue participants are the same. We define two persuasion dialogue protocols that conform to the grounded and credulous (admissible/preferred) semantics, and state soundness and fairness (a form of completeness) results for these dialogues (space limitations preclude inclusion of proofs).

The final section concerns an investigation of strategic considerations in persuasion dialogues. We show how participants may strategise based on their beliefs about their interlocutor’s knowledge, and how such considerations need to account for the logical content of arguments. Particularly, we focus on showing how the abstract approach fails to accommodate the dynamics of dialogue, whereby new arguments may be constructed during the course of the dialogue process.

1. Background

Prakken in [15] instantiates Dung’s abstract approach by assuming an unspecified logical language $\mathcal{L}$, and by defining arguments as inference trees formed by applying strict or defeasible inference rules of the form $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$ and $\varphi_1, \ldots, \varphi_n \Rightarrow \varphi$, interpreted as ‘if the antecedents $\varphi_1, \ldots, \varphi_n$ hold, then without exception, respectively presumably, the consequent $\varphi$ holds’.

To define attacks, minimal assumptions on $\mathcal{L}$ are made; namely that certain wff (well formed formulæ) are a contrary or contradictory of certain other wff. Apart from this the framework is still abstract: it applies to any set of strict and defeasible inference rules,
and to any logical language with a defined contrary relation. The basic notion of \(\text{ASPI}^+\) is an argumentation system. Arguments are then constructed w.r.t. a knowledge base that is assumed to contain three kinds of formulæ.

**Definition 1.** Let \(\text{AS} = (\mathcal{L}, -, R, \leq)\) be an argumentation system where:

- \(\mathcal{L}\) is a logical language.
- \(-\) is a contrariness function from \(\mathcal{L}\) to \(2^\mathcal{L}\), such that:
  - \(\varphi\) is a contrary of \(\psi\) if \(\varphi \in \overline{\psi}\) and \(\psi \notin \overline{\varphi}\);
  - \(\varphi\) is a contradictory of \(\psi\) (denoted by \(\varphi = \overline{\psi}\)), if \(\varphi \in \overline{\psi}\) and \(\psi \in \overline{\varphi}\).
- \(R = R_s \cup R_d\) is a set of strict \((R_s)\) and defeasible \((R_d)\) inference rules such that \(R_s \cap R_d = \emptyset\).
- \(\leq\) is a pre-ordering on \(R_d\).

A knowledge base in an argumentation system \((\mathcal{L}, -, R, \leq)\) is a pair \((\mathcal{K}, \leq')\) where \(\mathcal{K} \subseteq \mathcal{L}\) and \(\leq'\) is a pre-ordering on the non-axiom premises \(\mathcal{K} \setminus K_n\). Here, \(\mathcal{K} = K_n \cup K_p \cup K_a\) where these subsets of \(\mathcal{K}\) are disjoint: \(K_n\) is the (necessary) axioms (which cannot be attacked); \(K_p\) is the ordinary premises (on which attacks succeed contingent upon preferences), and; \(K_a\) is the assumptions (on which attacks are always successful, cf. assumptions in [4]).

Arguments are now defined, where for any argument \(A\), \(\text{Prem}\) returns all the formulæ of \(\mathcal{K}\) (premises) used to build \(A\); \(\text{Conc}\) returns \(A\)'s conclusion; \(\text{Sub}\) returns all of \(A\)'s sub-arguments; and \(\text{Rules}\) returns all rules in \(A\).

**Definition 2.** An argument \(A\) on the basis of a knowledge base \((\mathcal{K}, \leq')\) in an argumentation system \((\mathcal{L}, -, R, \leq)\) is:

1. \(\varphi\) if \(\varphi \in \mathcal{K}\) with: \(\text{Prem}(A) = \{\varphi\}\); \(\text{Conc}(A) = \varphi\); \(\text{Sub}(A) = \{\varphi\}\); \(\text{Rules}(A) = \emptyset\).
2. \(A_1, \ldots, A_n \vdash \psi\) if \(A_1, \ldots, A_n\) are arguments such that there exists a strict/defeasible rule \(\text{Conc}(A_1), \ldots, \text{Conc}(A_n) \vdash \psi\) in \(R_s/R_d\).

   \(\text{Prem}(A) = \text{Prem}(A_1) \cup \ldots \cup \text{Prem}(A_n)\); \(\text{Conc}(A) = \psi\);
   \(\text{Sub}(A) = \text{Sub}(A_1) \cup \ldots \cup \text{Sub}(A_n) \cup \{A\}\);
   \(\text{Rules}(A) = \text{Rules}(A_1) \cup \ldots \cup \text{Rules}(A_n) \cup \{\text{Conc}(A_1), \ldots, \text{Conc}(A_n) \vdash \psi\}\)

Three kinds of attack are defined for \(\text{ASPI}^+\) arguments. \(B\) can attack \(A\) by attacking a premise or conclusion of \(A\), or an inference step in \(A\). For the latter undercutting attacks, it is assumed that applications of inference rules can be expressed in the object language; the precise nature of this naming convention will be left implicit. Some kinds of attack succeed as defeats independently of preferences over arguments, whereas others succeed only if the attacked argument is not stronger than the attacking argument. The orderings on defeasible rules and non-axiom premises (we assume their usual strict counterparts, i.e., \(l < l'\) if \(l \leq l'\) and \(l' \nless l\)) are assumed to be used in defining an ordering \(\leq\) on the constructed arguments. Unlike [15] we explicitly define in this paper a function \(\rho\) that takes as input a knowledge base in an argumentation system (and so the defined arguments and orderings on rules and premises) and returns \(\leq^+\). Henceforth, we assume the strict counterpart \(\prec\) of \(\leq\).

\(^1\)See [15] for ways in which such a function would define \(\leq\) according to the weakest or last link principles.
Definition 3. A attacks $B$ iff $A$ undercut, rebuts or undermines $B$, where:

- $A$ undercut a argument $B$ (on $B'$) iff $\text{Conc}(A) \in B'$ for some $B' \in \text{Sub}(B)$ of the form $B'_1, \ldots, B'_n \Rightarrow \psi$.
- $A$ rebuts argument $B$ (on $B'$) iff $\text{Conc}(A) \in \overline{\varphi}$ for some $B' \in \text{Sub}(B)$ of the form $B'_1, \ldots, B'_n \Rightarrow \varphi$. In such a case $A$ contrary-rebuts $B$ iff $\text{Conc}(A)$ is a contrary of $\varphi$.
- Argument $A$ undermines $B$ (on $B'$) iff $\text{Conc}(A) \in \overline{\varphi}$ for some $B' = \varphi, \varphi \in \text{Prem}(B) \setminus \mathcal{K}_n$. In such a case $A$ contrary-undermines $B$ iff $\text{Conc}(A)$ is a contrary of $\varphi$ or if $\varphi \in K_\alpha$.

An undercut, contrary-rebut, or contrary-undermine attack is said to be preference-independent, otherwise an attack is preference-dependent.

Then, $A$ defeats $B$ (denoted $A \rightarrow B$) iff $A$ attacks $B$ (denoted $A \rightarrow B$) on $B'$, and either: $A \rightarrow B$ is preference-independent, or: $A \rightarrow B$ is preference-dependent and $A \not\sim B'$.

Definition 4. An argumentation theory is a triple $\mathcal{AT} = (\mathcal{AS}, KB, p)$ where $\mathcal{AS}$ is an argumentation system, $KB$ is a knowledge base in $\mathcal{AS}$ and $p : \mathcal{AS} \times KB \rightarrow \preceq$ such that $\preceq$ is an ordering on the set of all arguments that can be constructed from $KB$ in $\mathcal{AS}$.

The justified arguments under the full range of Dung semantics [7] can then be defined. To recap, a Dung framework consists of a set of arguments $\mathcal{A}$ and a binary relation $\mathcal{B}$ over $\mathcal{A}$. $S \subseteq \mathcal{A}$ is conflict free iff $\forall X, Y \in S, (X, Y) \notin \mathcal{B}$. $X \in \mathcal{A}$ is acceptable w.r.t. some $S \subseteq \mathcal{A}$ iff $\forall Y$ s.t. $(Y, X) \in \mathcal{B}$ implies $\exists Z \in S$ s.t. $(Z, Y) \in \mathcal{B}$. A conflict free set $S$ is an admissable extension iff $X \in S$ implies $X$ is acceptable w.r.t. $S$; a complete extension iff $X \in S$ iff $X$ is acceptable w.r.t. $S$; a preferred extension iff $X$ is a set inclusion maximal complete extension; the grounded extension iff it is the set inclusion minimal complete extension. For $s \in \{\text{complete, preferred, grounded}\}$, $X$ is sceptically or credulously justified under the $s$ semantics if $X$ belongs to all, respectively at least one, $s$ extension.

Thus, if $\mathcal{A}$ is the set of (c-consistent) arguments on the basis of an ASPIC$^+$ argumentation theory $\mathcal{AT}, \mathcal{C}$ the attack relation over these arguments, and $\mathcal{D}$ the defeat relation obtained from $\mathcal{C}$ and the preference ordering $\preceq$, then letting $\mathcal{D}$ be the binary relation $\mathcal{B}$, the justified arguments of $\mathcal{AT}$ are the justified arguments of the Dung framework $(\mathcal{A}, \mathcal{D})$. In [15] it is shown that under some intuitive assumptions on the strict knowledge and the preference relation $\preceq$, ASPIC$^+$ satisfies all of [5]'s rationality postulates for argumentation.

In summary [15] and subsequently [11]'s ASPIC$^+$ provides a general framework that accommodates a number of possible logical approaches to argumentation$^2$ and satisfies [5]'s rationality postulates.


In this section we provide an abstract general system for dialogue, based on ASPIC$^+$. We provide the basic elements responsible for both regulating a dialogue game, and allowing for strategies to be implemented. We then focus on formalising a persuasion dialogue instance of it.

We begin by assuming an environment of multiple agents $Ag_1, \ldots, Ag_\nu$, where each $Ag_i$ can engage in dialogues in which its strategic selection of locutions may be based on what $Ag_i$ believes its interlocutor (in the set $Ag_{j \neq i}$) knows. Accordingly, and in similar sense to the approach employed in [12], each $Ag_i$ maintains a model of its possible opponent agents, though in contrast with [12], the model consists of the goals and knowledge other agents may use to construct arguments and preferences, rather than just the abstract arguments and their relations. We assume that all agents share the same contrary relation $\neg$, the same language $\mathcal{L}$, and the same way of defining preferences over arguments based on the pre-orderings over non-axiom premises and defeasible rules (i.e., all agents share the same function $p$).

**Definition 5.** Let $\{Ag_1, \ldots, Ag_\nu\}$ be a set of agents. For $i = 1 \ldots \nu$, the agent theory of $Ag_i$ is a tuple $Ag_{Ti} = < S_i, 1, \ldots, S_i, \nu >$ such that for $j = 1 \ldots \nu$, each sub-theory $S_i = \langle AT_{ij}, G_{ij} \rangle$ where $AT_{ij}$ is what $Ag_i$ believes is the argumentation theory $(AS_{ij}, KB_{ij}, p_{ij})$ of $Ag_j$ and $G_{ij}$ is what $Ag_i$ believes are the goals of $Ag_j$, and:

- If $j = i$, $AT_{ij}$ and $G_{ij}$ are respectively $Ag_i$’s own argumentation theory and goals.
- For $i, j, k, m = 1 \ldots n$, let $S_{ij}, S_{km}$ be any two distinct sub-theories of the form $\langle AT_{ij}, G_{ij} \rangle, \langle AT_{km}, G_{km} \rangle$, where $AT_{ij} = (AS_{ij}, KB_{ij}, p_{ij})$, $AT_{km} = (AS_{km}, KB_{km}, p_{km})$. Then $p_{ij} = p_{km}$, $\mathcal{L}_{ij} = \mathcal{L}_{km}$ and $\mathcal{S}_{ij} = \mathcal{S}_{km}$.

We represent the set of all discrete agent-theories of a number of agents equal to $\nu$ operating in a multi-agent environment through a two dimensional matrix $M_{\nu, \nu}$ of the form presented in Figure 1a (note that henceforth we may omit subscripts identifying pre-orderings and rules specific to a given agent). We refer to this matrix as Multi-Agent Omni-Base (MAOB).

Agents may participate in dialogues belonging in $DT = \{\text{Persuasion, Negotiation, Inquiry, Deliberation, Information-seeking}\}$ [19]. We assume a set $SA$ of speech acts that cover the full range of speech acts employed in different dialogue types. Examples of speech acts commonly employed include Accept, Reject, Offer, Argue, Challenge, Inform, Question, etc. Locutions exchanged between participants in dialogues are conjoinings of speech-acts augmented with content. In this paper we express these locutions as dialogue moves ($DM$) the content of which is defined w.r.t. the elements of an AS-

![Figure 1](image_url)
argumentation theory and its defined arguments and preferences. In this sense, a dialogue \( \mathcal{D} \) can then be defined as a sequence of dialogue moves \( \{DM_0, ..., DM_n\} \), each of them being introduced into the game against another move already in \( \mathcal{D} \), and contingent upon satisfying certain conditions defined by a dialogue protocol.

In general, a protocol defines a dialogue in three aspects. These concern, turntaking; backtracking; the legality of a locution, and; the game’s termination rules. In brief, turntaking specifies the participant to move next and the number of dialogue moves she can make; backtracking concerns whether a participant is allowed to return to a previous point in the game to introduce an alternative reply, and; the legality of a move is concerned with explicit rules related with the dialogical objective of a game, a participant’s role in it, and the commitments [14] made by her during the game. The latter we assume to be stored in a commitment store \( CS \) being constantly updated based on the dialogue moves that a participant introduces, based on the following definition:

**Definition 6.** Given a set of agents \( \{Ag_1, ..., Ag_\nu\} \) participating in a dialogue \( \mathcal{D} = \langle DM_0, ..., DM_n \rangle \), for any agent \( Ag_i \) we define the evolution of its commitment store \( CS_i \) such that for \( j = 0 \ldots n \), \( CS_i = \emptyset \) and \( CS_{i+1} \) is obtained by updating \( CS_i \) with the effects of the dialogue move \( DM_j \) through the following update function:

\[
U_{CS}(CS_i, DM_j) \rightarrow CS_{i+1}
\]

We will in the following section show how the definition of a protocol for specific dialogue games implicitly defines turntaking and backtracking aspects of a game, as well as the legality of moves. Termination rules commonly specify that a dialogue terminates when a participant has no possible response. However, one may wish to allow for termination to be decided by the agents themselves (for whatever reasons). In either case, given termination of a dialogue game, the success or failure of a dialogue game must be defined with respect to the role a participant plays in a dialogue. We assume that these roles are defined with respect to an agent’s goals \( G \), i.e.: if the goal of an agent is to persuade its interlocutor of the truth of a claim \( \varphi \), then the agent’s role in the ensuing dialogue is that of ‘proponent’, which happens to comply with the dialogical objective of a persuasion dialogue. Note however that an agent’s goals may not always comply with the dialogical objective. An indicative set of goals could be the following: \( G \in \{prove X, disprove X, delay Ag_j, mislead Ag_j, meet deadline\} \).

Finally, we assume a general strategy function \( Str \) used for selecting a move from amongst the legal moves (defined by the protocol) available to a participant at any stage in the course of a dialogue. Essentially, the function makes use of \( Ag_i \)’s beliefs about the knowledge of its interlocutor \( S_i \) and its commitment store \( CS_i \), so as to simulate the possible ways based on which the current dialogue may be extended –expressed as a tree. The participant can then evaluate which of these dialogues result in success, and so make the choice of move accordingly. A detailed example of the employment of this function is presented in Section 3.

2.1. Formalising Persuasion Dialogues

In this section we formalise persuasion dialogues as instances of the previous section’s general framework. In such dialogues, agents debate the truth of a claim \( \varphi \), where agents adopting the role of proponent \( Pr \), persuade agents who may adopt the role of oppo-
nents (Op) seeking to challenge the truth of \( \varphi \). We build on Prakken’s model of persuasion dialogues [14] that assumes a single proponent and opponent. In [14] Prakken specifies speech-acts claim, why, argue, concede, and retract whose use in locutions, together with the reply structure of a dialogue game, result in a dialogue tree whose root is the initial claim \( \varphi \) proposed by \( Pr \). The locutions in the tree implicitly define a Dung graph of arguments related by defeats. Then, an any-time winning definition is provided, based on an inductive dialectical labelling of the locutions in the dialogue tree, i.e. \( Pr \) is currently winning iff the dialectical status of the root node is labelled in. Soundness and fairness results are then provided with respect to a game for the grounded semantics. Soundness is satisfied if \( \varphi \) is labelled in implies an argument for \( \varphi \) is in the grounded extension of the implicitly defined Dung graph, whereas fairness concerns the reverse.

We take a similar approach to [14] with the following differences and extensions. Firstly, for simplicity of presentation we consider only argue speech acts, and leave for future work the implicit construction of arguments through the use of other speech acts\(^3\). Secondly, persuasion dialogues in our framework are explicitly linked to the ASPIC\(^+\) framework. Thirdly, we not only define a game for grounded semantics, but also a game for credulous (preferred) semantics. We also state soundness and fairness results for both games. Fourthly, current dialogical frameworks (such as [14]) that allow exchange of arguments that defeat each other, need to assume that agents share the same preferences. However this is clearly unrealistic. While two agents may agree that \( X \) attacks \( Y \), one may believe that \( X \prec Y \) and so \( X \) does not defeat \( Y \), while the other may believe that \( Y \prec X \) and so \( X \) defeats \( Y \). We accommodate the possibility of conflicting preferences by allowing agents to move arguments that attack rather than defeat, and then separately move possibly conflicting preferences.

In what follows we assume agents \( Pr \) and \( Op \) with theories as defined in Definition 5, and simply subscripts \( pr \) and \( op \) to identify their argumentation theories and commitment stores as well as the discrete components found in them. Valid dialogue move contents are of the form argue : \( X \) where \( X \) is an ASPIC\(^+\) argument or \( X \) is a tuple \((a, b)\) where \( a \subseteq a' \), \( b \subseteq b' \); that is, a tuple of pre-orderings on non-axiom premises and defeasible rules respectively. Intuitively, a \( DM \) with content \((a, b)\) provides the basis for defining a preference over arguments moved earlier in the dialogue, via the shared function \( p \) described in Section 1. Thus, if \( X \) has been moved (in \( DM_{i+1} \)) as an argument attacking \( Y \) (in \( DM_i \)), then \((a, b)\) may be moved (in \( DM_{i+2} \)) as a reply to \( Y \), where \((a, b)\)’s orderings over \( X \) and \( Y \)’s contained elements determine (via \( p \)) that \( X \prec Y \).

Henceforth, we may as an abuse of notation reference content of the form \((a, b)\) in terms of the argument ordering it defines. The commitment update function can then be defined so that if the content in a move \( DM \) is an argument \( A \) or a tuple \((a, b)\), then the commitment store of the corresponding participant is updated with \( Rule(A) \cup Prem(A) \), respectively \( a, b \); the commitment stores of all other dialogue participants remain the same. Henceforth we will assume that \( Pr \) (\( Op \)) can introduce moves whose content is obtained from their own argumentation theory \( AT_{Pr} \) (\( AT_{Op} \)) and the knowledge in the commitment store \( CS_{Op} \) (\( CS_{Pr} \)) of its interlocutor.

We now define core protocol rules for persuasion protocols conducted according to the credulous and grounded semantics. These rules define what constitutes a legal

\(^3\)In [14], an argument \( X = \{p \text{ since } q \text{ and } q \text{ implies } p, q \text{ since } r \text{ and } r \text{ implies } q' \} \) can be constructed by first arguing \( X' = \{p \text{ since } q \text{ and } q \text{ implies } p' \} \) and then in response to ‘why \( q \)’, moving \( X'' = \{q \text{ since } r \text{ and } r \text{ implies } q' \} \). In this paper we assume \( X \) is moved in a single location.
dialogue. Note that since \( Pr \) and \( Op \) share the same contrary relations, there is agreement as to whether a given argument attacks another. We will henceforth refer to a dialogue move as a tuple \( DM = < I, argue > \), where \( I \in \{ Pr, Op \} \), \( I = Pr \) if \( I = Op \), while we may omit the \( argue \) speech act, referring only to its content.

**Definition 7.** \( D = < DM_0, \ldots, DM_n > \) is a legal persuasion dialogue if:

1. \( DM_0 = < Pr, X > \) (the dialogue begins with \( Pr \) proposing an argument and is said to be ‘a persuasion dialogue for \( X \)’)
2. For \( i = 0 \ldots n - 1 \), if \( DM_i = < I, argue > \) then \( DM_{i+1} = < I, argue > \) (\( Pr \) and \( Op \) take turns)
3. For \( i = 1 \ldots n \), each \( DM_i \) is a reply to some \( DM_j \), \( j < i \) (i.e., backtracking is allowed), where either:
   - \( DM_j = < I, X > \), \( DM_i = < I, Y > \) where \( Y \) attacks \( X \), or;
   - \( DM_j = < I, X > \), \( DM_i = < I, X \prec Y > \) and \( DM_j \) is a reply to some \( DM_k \), \( k < j \) and \( DM_k = < I, Y > \), or;
   - \( DM_j = < I, X \prec Y > \), \( DM_i = < I, X \prec Y > \).

Notice that 2) allows the moving of a preference over arguments to invalidate the success of an attack as a defeat, and 3) allows the moving of a conflicting preference ordering. By licensing backtracking, it is easy to see that a persuasion dialogue \( D \) can be represented as a tree with \( n \) leaf nodes, consisting of \( n \) disputes—paths from the root node to a leaf node—where every child node is a dialogue move in reply to its parent. Each new dispute results from a backtracking move by \( Pr \) or \( Op \). We then distinguish between the grounded and credulous persuasion dialogues by augmenting the above core rules with rules that place restrictions on \( Pr \)’s respectively \( Op \)’s moves. These rules are essentially those used in the corresponding argument game proof theories in [10].

**Definition 8.** Let \( D = < DM_0, \ldots, DM_n > \) be a legal persuasion dialogue, and \( T_D = \{ d_1, \ldots, d_m \} \) the set of disputes defining the dialogue tree. Then:

1. \( D \) is a legal grounded persuasion dialogue iff \( \forall d \in T_D \), no two dialogue moves in \( d \) moved by \( Pr \), have the same content (i.e., \( Pr \) cannot repeat moves in any given dispute)
2. \( D \) is a legal credulous persuasion dialogue iff \( \forall d \in T_D \), no two dialogue moves in \( d \) moved by \( Op \), have the same content (i.e., \( Op \) cannot repeat moves in any given dispute)

Henceforth, \( P_G \) will denote the protocol defined by Definition 7 and 8.1, and \( P_C \) the protocol defined by Definition 7 and 8.2.

We can now define an any-time evaluation of the winner of a dialogue, based on the use of dialectical labellings of dialogue trees as defined in [14], as follows:

- a node is labelled \( \text{in} \) iff all of its children-nodes are labelled \( \text{out} \)
- a node is labelled \( \text{out} \) iff it has at least one child labelled \( \text{in} \)

We then require the notion of a winning-strategy (based on [10]):

**Definition 9.** Given a dialogue tree \( T_D \) with root node \( DM_0 = < Pr, X > \) labelled \( \text{in} \), then \( T' \) is a winning-strategy for \( X \) if \( T' \) is a finite sub-tree of \( T_D \) such that:

- For every \( DM \in T' \) with \( DM = < Op, argue > \), there exists a \( DM' = < Pr, argue > \), where \( DM' \) is a child of \( DM \).
No two arguments $X, Y$ moved by $Pr$ in $T'$ attack each other (i.e., the arguments moved by $Pr$ are conflict free)\(^4\).

Thus at any stage of a dialogue for $X$, $Pr$ winning is contingent on identifying a winning-strategy for $X$, while $Op$ winning is contingent on the absence of such a strategy.

We now come to presenting soundness and fairness results. To recap both $Pr$ and $Op$ can move arguments and preferences constructed from their own argumentation theories and the commitment stores of their interlocutor. The latter contains all the content exchanged in locutions. What we wish to show is soundness and fairness with respect to the Dung framework of arguments and defeats defined by the knowledge in the commitment stores. However, we first need to account for the possibility of conflicting preference information. Essentially, suppose $X \prec Y$ moved by $Pr$ and $Y \prec X$ moved by $Op$ (i.e., the pre-orderings on rules and premises defining these preferences) in the commitment store. It can easily be seen that under the rules for the grounded game, $Op$’s preference will win out over $Pr$’s preference, and under the rules for the credulous game, $Pr$’s preference will win out over $Op$’s preference\(^5\).

So, let us define an $AF_\varnothing = (A, D)$, where the arguments $A$ are defined by the rules and premises in the commitment stores of $\varnothing$, and the defeats are defined based on the attacks between these arguments, and the preferences elicited from: (a) the pre-orderings moved by $Pr$, maximally consistently extended with the pre-orderings moved by $Op$ in the case of the credulous game; (b) the pre-orderings moved by $Op$, maximally consistently extended with the pre-orderings moved by $Pr$, in the case of the grounded game. Soundness and completeness can then be shown under the following completeness condition which essentially states that agents make all the moves they can make given the knowledge committed to in the commitment stores:

**Definition 10.** Let for any $i = 1, \ldots, n - 1$, $DM = DM_i$ in the dialogue $\varnothing = <DM_0, \ldots, DM_n, >$. $\varnothing$ is complete iff one can construct a reply $DM'$ to $DM$ where the content of $DM'$ is obtained based on the commitment stores $CSPr_{i+1}$ and $CSOp_{i+1}$, and $DM'$ is a legal reply under $\mathcal{P}_G$ ($\mathcal{P}_C$), then $DM'$ is a reply to $DM$ in $\varnothing$.

**Theorem 1** (Soundness). Given a finite complete grounded (credulous) persuasion dialogue $\varnothing$ for $X$, then if $Pr$ is winning, $X$ is in the grounded (an admissible and so preferred) extension of $AF_\varnothing$.

**Theorem 2** (Fairness). Given a finite $AF_\varnothing = (A, D)$ defined by a dialogue $\varnothing$, then for any $X \in A$, if $X$ is in the grounded (an admissible and so preferred) extension of $AF_\varnothing$, then there exists a finite complete grounded (credulous) persuasion dialogue for $X$, such that $Pr$ is winning.

---

\(^4\)Note that arguably, the notion of a winning-strategy could be enforced via the legal moves licensed by the protocol; for example preventing $Pr$ from moving conflicting arguments in a dialogue. Space limitations preclude presentation of a more complex protocol definition.

\(^5\)Intuitively, consider these conflicting preferences as mutually attacking arguments with no other incoming attacks. Then $Pr$’s argument (preference) will be justified under the credulous, but not grounded semantics.
3. Strategic Considerations

We now investigate through an example how agents engaging in Section’s 2.1 grounded game dialogues may make use of their beliefs about their opponent’s knowledge to strategise, through the employment of the general strategy function described in Section 2.

Example 1. Suppose an argumentation system \((\mathcal{L}, \neg, \mathcal{R}, \leq)\) where:

- \(\mathcal{L}\) is a language of propositional literals, composed from a set of propositional atoms \(\{a, b, c, \ldots\}\) and the symbols \(\neg\) and \(\leq\) respectively denoting strong and weak negation (i.e., negation as failure). \(\alpha\) is a strong literal if \(\alpha\) is a propositional atom or of the form \(\neg\beta\) where \(\beta\) is a propositional atom. \(\alpha\) is a wff of \(\mathcal{L}\) if \(\alpha\) is a strong literal or of the form \(\beta\) where \(\beta\) is a strong literal.
- For a wff \(\alpha, \beta\) and \(\neg\alpha\) are contradictories and \(\alpha\) is a contrary of \(\neg\alpha\).

Consider a grounded persuasion dialogue \(\mathcal{D} =< \mathcal{DM}_0, \ldots, \mathcal{DM}_k>\) is to be extended with a \(Pr\) move \(\mathcal{DM}_{k+1}\), and let us assume that for \(Pr\) to win, the dialectical labelling of \(\mathcal{DM}_k\) must be made out. Figure 2a (resp. 2b) shows \(Pr\)’s own (resp. what \(Pr\) believes is \(Op\)’s) premises, rules, pre-orderings and goals, relevant to extending \(\mathcal{D}\). Accordingly, Figures 2c and 2d, illustrate the set of arguments we assume that \(Pr\) can construct based on \(S'_{Pr,Pr}\) and \(Pr\)’s beliefs about the arguments that \(Op\) can construct based on \(S'_{Pr,Op}\). Notice that because of the absence of premise \(p \notin K_{Pr,Op}\), \(Pr\) believes that \(Op\) is unable to instantiate argument \(E\) (Figure 2e). \(Pr\) has a choice of replying to \(\mathcal{DM}_k\) with arguments \(A\) or \(A'\) and thus simulates the following two dialogue trees \((T1, T2)\) based on them, of which the second is depicted in Figure 2f:

- \(T1\): \(Pr\) moves \(A\) leading to an immediate victory for \(Pr\) since it cannot be countered.
- \(T2\): \(Pr\) moves \(A'. \ Op\) then replies with \(B\), giving \(Pr\) a choice between \(C\) and \(D\). In its simulation \(Pr\) opts for \(D\), which leads to a repetition of \(B\) by \(Op\) as licensed by protocol \(\mathcal{DG}\), followed by \(Pr\) replying with \(C\). \(Pr\) wins this dispute, making \(\mathcal{DM}_k\) out (this would also have been the case if \(Pr\) had chosen \(C\) rather than \(D\) at the earlier choice point). Since \(p\) is now in \(Pr\)’s commitment store, then given \(Pr\)’s beliefs about \(Op\)’s knowledge, \(Pr\) simulates \(Op\)’s use of this commitment to construct \(E\) and simulates \(Op\)’s backtracking moving \(E\) in reply to \(A'\), thus making \(\mathcal{DM}_k\) in. Hence \(Pr\) backtracks to move \(A\) against \(\mathcal{DM}_k\), followed by \(Op\) reusing \(E\) against \(A\) which again results in \(\mathcal{DM}_k\) in. Based on \(\mathcal{DG}\), \(Pr\) cannot repeat \(A\) against \(E\) and thus loses the game.

The above example illustrates that if \(Pr\), in its strategising, accounts for the logical content of arguments updating the commitment store, the choice of content for \(\mathcal{DM}_{k+1}\) makes a difference to the outcome of the actual dialogue, under the assumption that \(Pr\)’s beliefs about \(Op\)’s knowledge is indeed accurate. \(Pr\) prefers to move \(A\) rather than \(A'\), as the latter would result in there being no winning-strategy for \(Pr\). Notice that if one were to rely on an abstract representation of the employed arguments, disregarding their logical contents, the simulated dialogues (we show only the arguments) would have been <A>, and <A',B,C> or <A',B,D,B,C>, all of which would make \(\mathcal{DM}_k\) out and \(Pr\) winning. In other words, \(Pr\) would be indifferent to choosing between \(A\) and \(A'\) since the construction and use of argument \(E\) would not have been simulated. In this

\[\text{\textsuperscript{6}}\text{In and out labelled nodes are expressed with double respectively single lines. Dashed arrows concern the possible replies that may follow after a dialogue move.}\]
Figure 2. (a) & (b) illustrate a subset of Pr’s own knowledge ($S_{PrPr}^t \subseteq S_{PrPr}^s$) respectively beliefs about Op’s knowledge ($S_{PrOp}^t \subseteq S_{PrOp}^s$). (c) & (d) concern respectively the set of arguments $A_{Pr}^t \subseteq A_{Pr}$ that Pr can construct based on $S_{PrPr}^t$, and the set of arguments $A_{Op}^t \subseteq A_{Op}$ that Pr assumes Op can construct based on $S_{PrOp}^t$. (e) Argument $E$ (f) the simulated dialogue tree ($T_2$) instantiated if $DM_{k+1} = < Pr, A^t >$

respect, we argue that a purely abstract approach is characterised by severe limitations, as it fails to accommodate the fact that new arguments can be made available during the course of a dialogue, due to the dynamic evolution of knowledge available for argument construction (as shown by the use of the commitments of one agent in the arguments constructed by another).

Further in relation to the provided example and to the use of preferences, assume that after the deployment of $E$ by Op against $A$, Pr updates its preference-orderings such that not only $s > w$, but also $p \Rightarrow a > w, p \Rightarrow \neg a$. The latter would, under the weakest link principle, give the argument ordering $A \succ E$, which Pr can simulate moving in as a reply to $E$, thus making $DM_k$ out. In a similar sense, suppose Pr assumes that Op will also update its preference-orderings such that in addition to $w > s$, Op also believes $p > s$ and $w, p \Rightarrow \neg a > p \Rightarrow a$. Then, again under the weakest link principle, this would produce a counter argument ordering $E \succ A$, which if included in Pr’s simulation as Op’s reply to $A \succ E$, and based on $\mathcal{R}_G$ which dictates that Pr cannot repeat $A \succ E$ in the same dispute, will result in the dialogue ending with $DM_k$ in.

4. Conclusions

In this paper we have provided an argumentation-based framework for persuasion dialogues that enables a formal off-line analysis, based on a logical conception of arguments, that an agent may undertake in order to strategise over the choice of moves to make in a dialogue game, based on its model of its opponents. Though our approach concerns just a single dialogue type, we believe that it can be easily generalised so as to include other types of dialogues. Our aim was to make two main contributions to the study of dialogue games. Firstly, we defined persuasion dialogues related to those described

---

In this paper we do not formally model the mechanism an agent uses to update its priority ordering over rules and premises. We will assume agents use generic principles to do so, e.g. the well known specificity principle, and the temporal principle (which orders newly acquired knowledge over older knowledge).
in [13], but extended to account for admissible semantics, while we also allow agents
to move preferences that undermine the success of attacks as defeats. These preferences
may be contradictory and are effectively treated as mutually attacking arguments. The
latter we consider to be a novel property of our system, while it suggests future work,
building on [9], to enable agents to argue about their preferences. The second concerns
the fact that the provided dialogue framework is \textit{ASPIC}^+\textit{-based, and thus allows for a
more concrete logical analysis w.r.t. the underlying logic. Indeed we contrasted such
an analysis with abstract opponent modelling (such as that deployed in [12]), showing
that appropriate mechanisms for strategising need to account for the logical content of
arguments. Finally we note that because \textit{ASPIC}^+ explicitly models the logical content
and structure of arguments, while accommodating many existing logical approaches to
argumentation, we can claim a similar level of generality for our dialogical framework.

References

[7] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning,
[9] S. Modgil. Reasoning about preferences in argumentation frameworks. Artificial Intelligence, 173(9-
1021–1026, 2011.
2006.
international joint conference on Autonomous Agents and Multi-Agent Systems - Volume 2, pages 1031–
1038, 2008.
abilistic approach to legal disputes. In Proceedings of the 20th annual conference on Legal Knowledge