# Mediation of Debates with Dynamic Argumentative Behaviors

Emmanuel Hadoux, Aurélie Beynier, Nicolas Maudet, Paul Weng

▶ **To cite this version:**

Emmanuel Hadoux, Aurélie Beynier, Nicolas Maudet, Paul Weng. Mediation of Debates with Dynamic Argumentative Behaviors. 7th International Conference on Computational Models of Argument, Sep 2018, Warsaw, Poland. pp.249-256, 10.3233/978-1-61499-906-5-249 . hal-01882384

**HAL Id: hal-01882384**
**https://hal.science/hal-01882384**

Submitted on 26 Sep 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

*Computational Models of Argument*
*S. Modgil et al. (Eds.)*

*© 2018 The authors and IOS Press.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/978-1-61499-906-5-249*

# Mediation of Debates with Dynamic Argumentative Behaviors

Emmanuel HADOUX [a] Aurélie BEYNIER [b] Nicolas MAUDET [b] and Paul WENG [c]
[a] *Department of Computer Science, University College London, London, UK*
[b] *Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6, Paris, France*
[c] *Shanghai Jiao Tong University, UM-SJTU Joint Institute, Shanghai, China*

**Abstract.** Mediation is a process for resolving conflicts among several entities. In argumentation debates, conflicting agents that may be organized as teams exchange arguments to persuade each other. In this paper, we consider an automated mediator, which assigns the speaking slots to agents so as to optimize some objectives and ensure the fairness of the debate. We propose a general setting where the argumentation strategies of the agents are probabilistically known and may evolve over time. We show that the problem can be solved as a semi-Markov decision problem with hidden modes.

**Keywords.** Strategies in argumentation, Argumentation and probability

## 1. Introduction

Mediating debates is a longstanding issue in democracies. As early as 1876, Henry Martin Robert designed a set of rules, Robert's Rule of Order [19], which prescribe how assembly discussions should be conducted. For instance, a general guideline of the rules is that "no member can speak twice on the same issue until everyone else wishing to speak has spoken once". But it also goes in much deeper details regarding the agenda of a meeting, the amendments and the motions, the votes, and how the "floor" (the right to speak) can be allocated in assembly discussions. Prakken and Gordon formalized (some of) those rules and argue that they may be used in electronic debates [17]. But they also discuss reasons why this may not be an obvious choice. In particular, they mention that electronic debates are not synchronous discussions. Perhaps one further difference is that rules of order are assumed to take place in settings where participants can be assumed (to some extent) to have equal power and access to the floor.

We emphasize this point because it has some important consequences as to how the fairness of mediation should be interpreted. Often, it is understood simply as *strict neutrality*. Every party or speaker should be treated the same. In many debate contexts, however, neutrality is not appropriate. For instance, by simply looking at the number of speak turns, a mediator may be neutral. However, if a party always had the floor in a situation where it could not utter any sensible argument, this may not be seen as fair. In this paper we claim that some recent advances in the modeling of debates can provide formal tools for an optimal mediation.

In our setting, we suppose that agents are split into several teams, exchanging arguments to persuade each other. The mediator should decide which agent of which team will speak next. To solve this decision problem, our mediator exploits her knowledge about the debating agents, *i.e.*, about their argumentative strategies. One of the main issues raised is the amount of information that the mediator has at her disposal. While it is conceivable that the mediator knows which team each agent belongs to, it is difficult to assume that she could assign a deterministic strategy to each agent, or that agents play optimally. Instead, agents will be viewed as reasoning with probabilistic strategies [8]. This models both the fact that an agent can act non-deterministically and that the mediator does not know perfectly the strategy of each agent. However, assuming that those strategies are stationary (*i.e.*, do not evolve during the debate) may also be too strong. Indeed, under time pressure in particular, realizing that she could not satisfy her own goal, an agent may use a more aggressive strategy in the hope of, at least, avoiding the other party attaining its own. As the debate progresses, each agent may change her current topic or her argumentative behavior representing her state-of-mind.

Following previous works on decision-making in non-stationary environments [6], each possible stationary strategy of an agent will be referred to as an argumentative *behavior*. Transitions between behaviors will formalize the non-stationarity of the agents' argumentative behaviors (each time an agent changes her strategy, she will move to another argumentative behavior). Of course, the agents' current behavior cannot be directly observed by the mediator. Our work relates to a series of papers investigating how reasoning techniques can be exploited in the context of argumentation-based interaction. Depending on the assumption made (see [21] for a survey), these techniques range from heuristic-based approaches [11], game-theoretical analysis, or opponent modeling [4,18]. Frameworks optimizing the sequence of moves of a debating agent facing a stochastic opponent have been developed in a setting with probabilistic strategies [9,5] or in a more general setting [7]. This differs significantly from our perspective in the sense that a mediator has to allocate turns to agents involved in a debate.

The problem of mediation has recently emerged as an important challenge for formal argumentation. In [16] a persuasion dialogue game for two players is extended to consider a neutral *adjudicator*. In [10], a dialectical system designed for mediation is proposed. Such dialogue games look in detail at the types of moves that can be played in this context, and prescribe what agents can play, but not *how* the mediator should play. None of the work above takes the perspective of a mediating agent. In this paper, we propose a general formalization of the mediator's decision problem (Section 2) and we argue that the problem can be modeled as a Markov Decision Problem with *hidden modes* (Section 3). Finally, experiments demonstrate the relevance of our approach (Section 4).

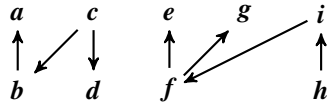
## 2. Dynamic Mediation Problems

In order to elaborate a mediation strategy, there is a need to represent possible argumentation strategies of debating agents. When it is not possible to assume full knowledge of all agents' strategies or when they act non-deterministically, one can use probabilities to reflect the likelihood that an agent plays a given argument or attack. Following this idea, the *Argumentation problems with Probabilistic Strategies* (APS) framework has been proposed to model argumentation problems using probabilistic executable logic [9].

Inspired by the APS formalization, we propose the *Dynamic Mediation Problem* (DMP) framework to consider a strategic mediator managing turn-taking between agents with non-stationary strategies. Our objective is to allow a mediator to decide for the best turn-taking sequence by adapting to the changes of argumentative behaviors (*i.e.*, due to non-stationarity of the agents). A DMP is defined by a tuple $\langle \mathcal{D}, \mathcal{T}, \mathcal{A}, \mathcal{E}, \mathcal{P}, (\mathcal{M}_i)_{i\in\mathcal{D}}, ((\mathcal{R}_i^\mu)_{\mu\in\mathcal{M}_i})_{i\in\mathcal{D}}, (\mathcal{B}_i)_{i\in\mathcal{D}}, (g_j)_{j=0\ldots|\mathcal{T}|}, (\mathcal{F}_i)_{i\in\mathcal{D}}\rangle$ with $\mathcal{D}$, a set of agents; $\mathcal{T}$, a set of teams (*i.e.*, subset of agents) where $\mathcal{T}$ forms a partition of $\mathcal{D}$; $\mathcal{A}$, a set of arguments; $\mathcal{E}$, a set of attacks $e(x, y)$, meaning argument $x$ attacks argument $y$; $\mathcal{P}$ a set of all possible *public* states; $\mathcal{M}_i$, the set of argumentative behaviors for agent $i$; $\mathcal{R}_i^\mu$, the set of rules of agent $i$ in mode $\mu \in \mathcal{M}_i$; $\mathcal{B}_i : \mathcal{M}_i \times \mathcal{M}_i \to [0, 1]$ models the probability of agent $i$ to change from one behavior to another; $g_j$, the goal of team $\mathcal{T}_j$ and $g_0$, the mediator's goal and $\mathcal{F}_i : \mathcal{M}_i \times \mathcal{M}_i \times \mathbb{N} \to [0, 1]$ models the probability of agent $i$ to move from one behavior to another after a given number of steps in the first behavior.

A *rule* $r \in \mathcal{R}_i$ is of the form $r$ : *prem* $\Rightarrow Pr(Acts_i)$ where premise *prem* is a conjunction of predicates $a(\cdot), h_i(\cdot)$ and $e(\cdot, \cdot)$ (or their negations) applied to one or more arguments. $Pr(Acts_i)$ is a probability distribution over a set $Acts_i$ of possible acts. An act $\alpha \in Acts_i$ is a set of modifications on predicates of the public space and private state of agent $i$: $\boxplus(p)$ (resp. $\boxminus(p)$) stands for adding (resp. removing) $p$ to (resp. from) the public space, where $p$ is either $a(x)$ or $e(x, y)$, and $\oplus(p)$ (resp. $\ominus(p)$) corresponds to adding (resp. removing) predicate $p$ to (resp. from) the private state of agent $i$. Let $Pr(Acts_i) = [p_1/act_1 \vee p_2/act_2 \vee \ldots \vee p_k/act_k]$ denote a distribution yielding $act_j$ with probability $p_j$. A rule can only be fired by agent $i$ if its premise is fulfilled. Premises formalize the conditions (*i.e.*, arguments or attacks) that must hold in order to play a rule.

A goal (for a team or for the mediator) consists in having some arguments present or absent from the grounded extension [3] of the arguments played in the public debate space. We say that an agent plays a *vacuous act* when it does not change the state of the debate. It is important that agents are able to skip the turn so that the debate can continue, to avoid potential deadlock. When each team has played the *skip* act, the debate stops. It then triggers the checking of acceptability of the goals of each team and of the mediator.




**Example 1.** *A government is discussing a bill to legalize communication surveillance. Two teams debate at the legislative assembly: the pro- and the anti-bill. The modeling contains 9 arguments (4 pros and 5 cons),* e.g.*, (**a**) anonymization software should not be seen as suspicious, (**e**) no judge is required to monitor a user, (**g**) the government can possibly abuse control. We picture the attack graph between arguments. Assume the goal of the second team is* $g_2 = \{in(\boldsymbol{d}), in(\boldsymbol{h})\}$*. Under the grounded semantics, **a**, **c**, **h**, and **f** are acceptable, thus* $g_2$ *is not fully satisfied.*

*We only give below examples of rules in one of the modes for two agents* $i$ *and* $j$ *from two opposite teams:* $i$ *(resp.* $j$*) belongs to team* $\mathcal{T}_1$ *(resp.* $\mathcal{T}_2$*). For conciseness, we remove the attacks from the rules, though they are still used to determine which arguments are attacked or defended. The goals are* $g_1 = \{in(\boldsymbol{c}), in(\boldsymbol{i})\}$, $g_2 = \{in(\boldsymbol{d}), in(\boldsymbol{h})\}$.

$\mathcal{R}_i : \{\emptyset \Rightarrow [0.7/\boxplus a(\boldsymbol{a}) \vee 0.3/\boxplus a(\boldsymbol{e})]$
$a(\boldsymbol{b}) \Rightarrow [0.55/\boxplus a(\boldsymbol{g}) \vee 0.45/\boxplus a(\boldsymbol{c})]$,
$a(\boldsymbol{d}) \Rightarrow [0.5/\boxplus a(\boldsymbol{i}) \vee 0.5/\boxplus a(\boldsymbol{c})]$,
$a(\boldsymbol{f}) \Rightarrow [0.9/\boxplus a(\boldsymbol{c}) \vee 0.1/\boxplus a(\boldsymbol{i})]$,
$a(\boldsymbol{e}) \wedge a(\boldsymbol{f}) \Rightarrow [1.0/\boxplus a(\boldsymbol{i})]\}$

$\mathcal{R}_j : \{\emptyset \Rightarrow [0.6/\boxplus a(\boldsymbol{d}) \vee 0.4/\boxplus a(\boldsymbol{h})]$,
$a(\boldsymbol{a}) \Rightarrow [0.7/\boxplus a(\boldsymbol{d}) \vee 0.3/\boxplus a(\boldsymbol{b})]$,
$a(\boldsymbol{e}) \Rightarrow [0.8/\boxplus a(\boldsymbol{f}) \vee 0.2/\boxplus a(\boldsymbol{f})]$,
$a(\boldsymbol{g}) \Rightarrow [0.5/\boxplus a(\boldsymbol{f}) \vee 0.5/\boxplus a(\boldsymbol{b})]$,
$a(\boldsymbol{i}) \Rightarrow [1.0/\boxplus a(\boldsymbol{h})]\}$

As the debate progresses, each agent may change her argumentative behavior. We shall exploit the typology of constructive versus destructive behavior [13]. This constitutes a basic case that has the advantage of being grounded in argumentation theory, and which can be easily extended with mixtures of these extreme behaviors. In the *constructive behavior*, the agent will favor acts that build her goals, while in the *destructive behavior* she seeks to destroy the arguments that are potential goals of the other team. Probabilities over acts in the rules thus vary from one mode to another in order to reflect this argumentative behavior. We also define the *mean behavior*, which takes the mean probability per act. It will be used in the sequel as an evaluation basis. In our examples, we consider only two behaviors but our approach allows for any number.

**Example 2.** ***Example 1 cont'd.*** *Consider the following modes, capturing different attitudes for an agent of team $\mathcal{T}_2$ when argument **a** holds on the public state (either to put forward argument **d** or instead attack argument **a** by playing argument **b**).*

| | |
|---|---|
| *constructive:* | $a(\boldsymbol{a}) \Rightarrow [0.7/\boxplus a(\boldsymbol{d}) \vee 0.3/\boxplus a(\boldsymbol{b})]$ |
| *destructive:* | $a(\boldsymbol{a}) \Rightarrow [0.2/\boxplus a(\boldsymbol{d}) \vee 0.8/\boxplus a(\boldsymbol{b})]$ |
| *mean:* | $a(\boldsymbol{a}) \Rightarrow [0.45/\boxplus a(\boldsymbol{d}) \vee 0.55/\boxplus a(\boldsymbol{b})]$ |

*Mediator's Objectives.* We distinguish different types of objectives of the mediator, and classify these principles as belonging either to the efficiency or the fairness of the debate.

*Debate Efficiency.* An important feature to define the debate efficiency is the *goal of the debate*. Strict neutrality would imply that the mediator holds an empty goal. However, it is legitimate for the mediator to have an impartial goal i.e. not to favor any team a priori. This could typically correspond to the goal of the interaction itself, which depends on the type of interaction considered [22,15]. Indeed, the main objective of the mediator is to lead the debate to its expected outcome. This is expressed by a goal ($g_0$) that the mediator pursues. Then the following principles can be considered. (1) At each turn, the debate yields a public state where the goal of the mediator can be evaluated (*Impact on audience (Imp)*). Sometimes it makes sense to do so *at each step* of the debate, *e.g.*, for debates broadcast on radio, where an audience might be convinced depending on how long they were exposed to arguments. Sometimes only *end state* of the debate is relevant, as in a trial, where only the ultimate state is considered by the judges. (2) Making regular progress in the debate should be favored, *i.e.*, circular arguments [12] or empty moves should be discouraged, and the mediator is legitimate to intervene to avoid this (*Progress of the debate (Prog)*). Finally, (3) short debates are preferred (*Length of the debate (Len)*).

**Example 3.** ***Example 1 cont'd.*** *A possible impartial goal could be that both teams manage to reach a consensus at least on argument **g**, i.e., $g_0 = \{in(\boldsymbol{g})\}$. Another type of impartial goal would be $g_0 = \{in(\boldsymbol{c}) \vee in(\boldsymbol{h})\}$. These goals are impartial in the sense that this is a disjunction which does not discriminate between the goals of the teams.*

*Debate Fairness.* The following properties are crucial to ensure that the debate is conducted in a way that is fair to all the participating agents. (1) One of the main guidelines of Robert's Rules of Order is *Alternation between teams (Alt)*. [10] also note that fairness is achieved in their system by balancing the agents' positions. We reformulate this rule in the context of several teams: "the turn should not be given to the same team again, as long as all the other teams did not have the opportunity to speak". (2) Priority should be given to agents who have (supposedly) a move directly connected to the most recent argument made in the debate (*Fair opportunity to respond (Resp)*). This captures a notion of relevance, but not as stringent as the one used in [14,1], which asks the status of the issue to be impacted by the move. Finally, (3) *Full participation (Part)*states that as long as agents have something relevant to say, they must in principle be allowed to do so.

Clearly, all of these principles may not be satisfied simultaneously: they are even sometimes contradictory. In the next section, we are going to see how all these principles can be captured through some appropriate setting of states, goals and reward functions.

## 3. DMP as Sequential Decision-making

*Dealing with Non-stationary Behaviors.* The decision problem of the mediator can be viewed as a sequential decision-making problem under uncertainty. As debating agents may change their behaviors over time, traditional Markov decision models fail to represent such non-stationary problems. To remedy this, *Hidden-Semi-Markov-Mode MDP* (HS3MDPs) [6] have been proposed to model problems where the non-stationarity of the environment is modeled as a finite non-controlled Semi-Markov chain, whose states called modes correspond to stationary MDPs. Formally, an HS3MDP is defined by a tuple $\langle M, C, H\rangle$ with $M$ a set of modes; $C : M \times M \to [0, 1]$ a transition function between modes and $H : M \times M \times \mathbb{N} \to [0, 1]$ a duration function. Each mode $m_k \in M$ corresponds to a stationary MDP, characterized by a tuple $\langle S, A, T_k, R_k\rangle$ where $S$ is a set of states; $A$ a set of actions; $T_k : S \times A \times S \to [0, 1]$ a transition function over states and $R_k : S \times A \to \mathbb{R}$ a reward function.

The duration function $H$ sets the stochastic number of decision steps the environment will stay in the current mode before being allowed to change to another one. When the number of steps remaining in the current mode reaches 0, the next mode is drawn using the function $C$. In this model, the current state $s \in S$ of the problem is directly observable, while the current mode $m \in M$ and duration $h$ are not.

The following two observations suggest that HS3MDPs can handle DMP problems: (1) there is a fixed and known number of possible environment dynamics, which correspond to all combinations of the debating agents' behaviors; (2) an environment dynamics mode prevails for several time steps since agents engage in a consistent behavior and keep the same behavior over several time steps of the debate. Formally, the decision problem of the mediator in a DMP can be modeled as an HS3MDP:

• $M = \prod_{i\in\mathcal{D}} \mathcal{M}_i$ the set of all possible combinations $(\mu_1, \cdots, \mu_{|\mathcal{D}|})$ of modes such as $\mu_i$ is a behavior of agent $i$ in the DMP. The elements of $M$ are numbered and denoted $m_k$. Each mode $m_k$ corresponds to an MDP $\langle S, A, T_k, R_k\rangle$ with $S = \mathcal{P} \times \{1, \ldots, |\mathcal{T}|\}$, all possible combinations of public states, plus the team of the agent who has just spoken, $A = \mathcal{D}$, as an action consists of allowing one agent to fire one rule and $T_k$ and $R_k$ (for each mode $m_k \in M$), as specified below.

• $C : M \times M \rightarrow [0, 1]$ the transition function over modes induced by $\mathcal{B}_i$ of the DMP, assuming independence between the changes of the agents' behavior,

• $H : M \times M \times \mathbb{N} \rightarrow [0, 1]$ the mode duration function derived from $\mathcal{F}_i$ of the DMP with the independence assumption. There is a (HS3MDP) mode change if the duration of at least one agent's behavior is equal to zero.

Let *Acts*$(r)$ be the set of acts for rule $r$, $\mathcal{R}_i^{\mu_i}$ be the set of rules of agent $i$ in her current mode $\mu_i$ and $\mathcal{R}_i^{\mu_i}[s]$ be the subset of rules of $\mathcal{R}_i^{\mu_i}$ compatible with state $s$. The transition function $T_k$ in mode $m_k = (\mu_1, \ldots, \mu_{|\mathcal{D}|})$ is defined as follows: $\forall s \in S, \forall i \in A = \mathcal{D}, \forall r \in \mathcal{R}_i^{\mu_i}[s], \forall l \in Acts(r)$, $T_k(s, i, s_{r,l}) = 1/|\mathcal{R}_i^{\mu_i}[s]| \cdot p_{r,l}^{\mu_i}$ where $s_{r,l}$ is the state resulting from the application of act $l$ of rule $r$ on state $s$ and $p_{r,l}^{\mu_i}$ is the probability of act $l$ of rule $r$ when in mode $\mu_i$ (given by the DMP specification).

*Capturing Efficiency and Fairness Principles.* The reward function $R_k$ formalizes the objectives of the mediator and has to be defined in compliance with the problem. These different objectives can generally be captured with a specifically designed reward function, and simultaneous objectives can be handled by combining several reward signals. Some of them would require to augment the state space.

• for *efficiency*: *(Prog)* is captured by assigning negative rewards to *vacuous act*. A less obvious situation to avoid are *circular arguments*: it would also be possible to penalize with a negative reward, but one would need to augment the state space to represent the *history* of the moves. For *(Imp)*, recall that the mediator may have some impartial goals specific to the debate. This is simply handled by giving a positive reward when (parts of) those goals hold. We distinguish *final* v. *step-wise* reward: in the latter case, the reward for a fulfilled goal is given at each step as the goal holds. In the former approach, a reward is only given at the end of the debate if the goal of the mediator holds. For *(Len)*, it suffices to penalize every time-step with a small negative reward value, or alternatively to use an adequate discount factor.

• for *fairness*: to favor alternation *(Alt)* between two teams, a negative reward can be given to the mediator if she lets the same team speak twice consecutively. When considering more than two teams, the state has to be augmented with the history of the $|\mathcal{D} - 1|$ last teams' turns. Regarding *(Resp)*, the mediator may be rewarded if a move is relevant, *i.e.*, related to the last (or some recent) moves. To that aim, the state space would need to be augmented to keep track of a few last moves. For simplicity, we do not take it into account in our model. Finally, *(Part)* is guaranteed in our model since the debate only ends when each team has triggered the skip act.

In short, the designer only has to specify *goal* and *relevance rewards*, along with *progress*, *length* and *alternation penalties* to model the mediator of her choice. It is possible to play on the parameters and combine these dimensions to obtain various types of mediator. In the end, the mediator aims at optimizing the expected sum of rewards discounted by a discount factor $\gamma \in [0, 1)$.

## 4. Solving a DMP and experiments

We ran experiments to test the relevance of formalizing the possible behaviors of the agents in the decision process. We compared the performance of the mediator while making decisions using an HS3MDP policy against a policy issued from a mean model over all behaviors. Indeed, this common method (see, *e.g.*, [2]) approximates the non-

stationarity to solve the problem with standard algorithms. Given an instance of debate mediation, the mean model is defined by averaging over the behaviors, rule by rule, the probability distributions over possible acts. We obtain a "mean" MDP with stationary state transition and reward functions. The HS3MDP and the "mean" MDP are then solved using POMCP [20] that we modified and improved using the structure of HS3MDPs [6].

Each part of Table 1 corresponds to respectively 3 agents in one team v. 4 in the other, 12 v. 12, 25 v. 25 and finally 50 v. 50. For each team size, we generate 100 instances of the problem described in Example 1 with different probabilities on acts in the rules. They are defined randomly for each agent with respect to the behaviors, *i.e.*, in the constructive behavior, the probability of the act moving the debate towards the goal is higher than the probability of trying to defeat the opponent. The mediator's goal is randomized for each instance. We record the mediator's performance (*i.e.*, discounted sum of rewards) for each instance and average over the 100 instances. We also increase the numbers of simulations done by POMCP while averaging on 1000 runs with the given number of simulations. The number of simulations is the number of Monte-Carlo executions done in the simulator before executing in the real environment the best action found. It starts with eight simulations and doubles the number of simulations until it takes more than one hour for 1000 runs (for at least one of the 100 averaging instances). POMCP results tend towards the optimal results when increasing the number of simulations [20].

We use a goal reward of 10 for each part of the goal accepted, -100 for the alternation penalty and a discount factor of 0.9 accounting for both the length and the progress penalties. Reported performances correspond to the results obtained by the mediator using *step-wise* and *final* reward functions. The left (resp. right) value of each column is obtained using the "mean" model (resp. the HS3MDP model). Bold face values mean that the relative improvement is at least 1% and that the difference is statistically significant. Both values are in bold face in the opposite case. For all sizes of instances, with a sufficient number of simulations, one can see that HS3MDP always outperforms the "mean" model. However, as the size of the instances increases, more simulations are needed to outperform the "mean" model. In fact, without enough simulations, the additional information brought by the HS3MDP model is not used and leads to wrong choices of actions when the model believes to be in a wrong mode. Nonetheless, it has to be noticed that, even for large-sized instances, the number of simulations required to outperform the "mean" model remains small. HS3MDPs can lead to significant improvements since the relative improvements are up to 79%. Apart from the results for Teams 12-12 and Teams 50-50 at 256 simulations for the "final" reward function, all results are statistically significant with $p < 0.05$ and most of them with $p < 0.001$ under a Student t-test.

The high alternation penalty has been defined as shown previously to verify that the objective (Alt) can be learned with an appropriate setting of the reward function.

Finally, to further validate our method, we performed a set of experiments with an impartial mediator who has disjunctive goals, *i.e.* disjunction of random parts of the goals of each team. Again, we can observe in Table 2 that our method consistently performs at least as good as the "mean" model, with relative improvements of up to 43%.

## References

[1] E. Bonzon and N. Maudet. On the outcomes of multiparty persuasion. In *AAMAS*, pages 47–54, 2011.
[2] B. C. da Silva, D. W. Basso, A. L. Bazzan, and P. M. Engel. Dealing with non-stationary environments using context detection. In *ICML*, 2006.

**Table 1.** Perf. for Teams 3-4, 12-12, 25-25 and 50-50

| Teams | # Sim. | Step-wise | Final |
|---|---|---|---|
| 3-4 | 8 | -93.66 / **-86.28** | -116.97 / **-108.90** |
| | 16 | -52.26 / **-39.40** | -79.27 / **-64.99** |
| | 32 | -10.29 / **-4.99** | -35.49 / **-30.40** |
| | 64 | 3.12 / **4.46** | -21.27 / **-19.73** |
| | 128 | 4.57 / **5.73** | -19.89 / **-18.82** |
| | 256 | 4.36 / **5.97** | -19.90 / **-18.51** |
| 12-12 | 64 | -8.70 / **-5.82** | -36.20 / **-32.87** |
| | 128 | 16.15 / **16.75** | -9.81 / **-9.46** |
| | 256 | 20.58 / **20.87** | **-4.94** / **-4.97** |
| 25-25 | 128 | -5.08 / **-3.56** | -31.93 / **-30.68** |
| | 256 | -15.59 / **16.74** | -10.56 / **-10.28** |
| 50-50 | 256 | -1.50 / **-0.31** | -27.84 / **-27.32** |

**Table 2.** Results for Teams 3-4 with disjunctive goals

| # Sim. | Step-wise |
|---|---|
| 8 | -73.71/**-64.58** |
| 16 | -29.62/**-16.91** |
| 32 | 13.94/**16.41** |


[3] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[4] C. Hadjinikolis, Y. Siantos, S. Modgil, E. Black, and P. McBurney. Opponent modelling in persuasion dialogues. In *IJCAI*, 2013.

[5] E. Hadoux, A. Beynier, N. Maudet, P. Weng, and A. Hunter. Optimization of probabilistic argumentation with Markov decision models. In *IJCAI*, pages 2004–2010, 2015.

[6] E. Hadoux, A. Beynier, and P. Weng. Solving Hidden-Semi-Markov-Mode Markov Decision Problems. In *Scalable Uncertainty Management*, pages 176–189. Springer, 2014.

[7] E. Hadoux and A. Hunter. Strategic sequences of arguments for persuasion using decision trees. In *AAAI*, 2017.

[8] A. Hunter. A probabilistic approach to modelling uncertain logical arguments. *International Journal of Approximate Reasoning*, 54(1):47–81, 2013.

[9] A. Hunter. Probabilistic strategies in dialogical argumentation. In *SUM, LNCS vol. 8720*, 2014.

[10] M. Janier, M. Snaith, K. Budzynska, J. Lawrence, and C. Reed. A system for dispute mediation: The mediation dialogue game. In *Computational Models of Argument*, 2016.

[11] D. Kontarinis, E. Bonzon, N. Maudet, and P. Moraitis. Empirical evaluation of strategies for multiparty argumentative debates. In *CLIMA*, pages 105–122, 2014.

[12] J. D. Mackenzie. Question-begging in non-cumulative systems. *Journal of philosophical logic*, 8(1):117–133, 1979.

[13] D. J. Moore. *Dialogue game theory for intelligent tutoring systems, PhD Thesis*. 1993.

[14] H. Prakken. Formalizing robert's rules of order. an experiment in automating mediation of group decision making. Technical Report GMD Report 12, 1998.

[15] H. Prakken. Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21:163–188, 2006.

[16] H. Prakken. A formal model of adjudication dialogues. *Artificial Intelligence and Law*, 16:305–328, 2008.

[17] H. Prakken and T. F. Gordon. Rules of order for electronic group decision making - A formalization methodology. In *Collaboration between Human and Artificial Societies, Coordination and Agent-Based Distributed Computing*, 1999.

[18] T. Rienstra, M. Thimm, and N. Oren. Opponent models with uncertainty for strategic argumentation. In *IJCAI*, 2013.

[19] H. M. Robert. *Robert's Rules of Order Newly Revised, 11th ed.* Da Capo Press, 2011.

[20] D. Silver and J. Veness. Monte-Carlo planning in large POMDPs. In *NIPS*, pages 2164–2172, 2010.

[21] M. Thimm. Strategic argumentation in multi-agent systems. *Künstliche Intelligenz, Special Issue on Multi-Agent Decision Making*, 28(3):159–168, 2014.

[22] D. Walton and E. Krabbe. *Commitment in dialogue: basic concepts of interpersonal reasoning*. State University of New York Press, 1995.