Evaluation of Dishonest Argumentation based on an Opponent Model: A Preliminary Report

Kazuyuki Kokusho and Kazuko Takahashi
School of Science&Technology, Kwansei Gakuin University, 2-1, Gakuen, Sanda, 669-1337, Japan
dbq68680@kwansei.ac.jp, ktaka@kwansei.ac.jp

Keywords: Argumentation, Strategy, Persuasion, Dishonesty, Opponent Model.

Abstract: This paper discusses persuasive dialogue in a case where dishonesty is permitted. We have previously proposed a dialogue model based on a predicted opponent model using an abstract argumentation framework, and discussed the conditions under which a dishonest argument could be accepted without being detected. However, it is hard to estimate the outcome of a dialogue, or identify causality between agents' knowledge and the result. In this paper, we implement our dialogue model and execute argumentations between agents under different conditions. We analyze the results of these experiments and discuss about them. In brief, our results show that the use of dishonest arguments affects the likelihood of successfully persuading the opponent, or winning a debate game, but we could not identify a relationship between the results of a dialogue and the initial argumentation frameworks of the agents.

1 INTRODUCTION

The aim of persuasion is to change an opponent's mind. An argumentation framework is a useful mechanism to manage a persuasive dialogue as a computational model, and there have been many studies on argumentation frameworks (Amgoud and de Saint-Cyr, 2013; Bench-Capon, 2003; Black and Hunter, 2015; Prakken, 2006; Rahwan and Simari, 2009). In the persuasion dialogue model, agents generally have independent knowledge or beliefs, which change every time they get their opponent's argument. Therefore, when there are multiple possible counter-arguments for the same argument, persuasion sometimes succeeds, but sometimes fails, depending on which argument is selected. Strategic selection of an argument can be done considering what an opponent knows (Hunter, 2015; Rienstra et al., 2013).

Agents sometimes try to persuade their opponents by presenting dishonest arguments and moreover, they may be revealed. In this case, agents need their prediction on their opponent's knowledge or belief. Therefore, it is essential to make a dialogue model with on an opponent model, if we formalize a dishonest argument and suspect of the truth of an argument. However, the possibility that an agent could present a dishonest argument has not been included in most of strategic argumentation models proposed so far. Takahashi et al. formalized dishonest argumentation using an opponent model (Takahashi and Yokohama, 2017). Of the several types of dishonesty, they focused on deception. That is, an agent intentionally hiding something they know.

Consider the following situation in which students are selecting a research laboratory[1]. This example shows how opponent models are used in giving a dishonest argument and pointing out the dishonesty.

> [Labo Selection Example]
> Alice tries to persuade Bob to apply to the same laboratory. Both know that Professor Charlie is strict and not generous. Alice, who prefers strict professors, wants to apply to Charlie's laboratory. However, Bob wants to work for a generous professor, but not for a strict professor. Alice knows Bob's preference. In addition, Alice thinks that Bob only knows Charlie is strict as Charlie's reputation.

Alice considers that if she said "Let's apply to Charlie's laboratory, because he is strict," then Bob might reject her proposal. Therefore, she says "Let's apply to Charlie's laboratory, because he is generous," hiding the fact that Charlie is not generous, to persuade Bob. However, Bob, who knows Charlie's reputation, suspects its truth, and may say, "No, I don't want to, because Charlie is not generous. Don't try

[1]This is another version of the example shown in (Takahashi and Yokohama, 2017).

to persuade me by hiding that fact." In this example, Alice deceives Bob by predicting his response. But as Bob knows what Alice knows, he suspects her of dishonesty and challenges her deception. A reason for a failure of the persuasion is that Alice does not know that Bob knows Charlie is not generous.

The success of persuasion depends on what an agent knows and what strategy is taken. Parsons et al. investigated the relationships between agents' initial knowledge and the outcome of the dialogue for several agents' tactics (Parsons et al., 2003). Yokohama et al. developed a strategy using an oppoment model for honest argumentation that will never fail to persuade, and proved its correctness (Yokohama and Takahashi, 2016). It is interesting to consider whether such a strategy exists in the case of a dishonest dialogue, and it is particularly interesting to investigate the effect of the deception and being silent, on the outcome of a dialogue.

There have been many studies on computational argumentation, but few of these studies evaluated the agents' strategies, where probabilistic approach is applied to learn an opponent model. They proposed metrics for evaluating a dialogue, such as a length of a dialogue and the number of arguments that agents have agreed on (Hadjinikolis et al., 2013; Thimm, 2014; Rahwan et al., 2009). And dishonesty has not been considered there.

In this paper, we implement a dialogue based on the model proposed in (Takahashi and Yokohama, 2017), where the agent predicts their opponent's argumentation framework. We define the concepts of dishonest argument and suspicious argument, by means of the acceptance of arguments in this model. We execute argumentations under different conditions and show the experimental results and their analysis. The main purpose of the experiment is to investigate the effect of deception or being silent, and to identify the relationship between these strategies and particular protocols or argumentation frameworks.

The results show that the use of dishonest arguments affects the likelihood of successfully persuading the opponent, or winning a debate game. But we could not identify a relationship between the results of a dialogue and the argumentation framework of the agents.

The rest of the paper is organized as follows. Section 2 describes the argumentation framework on which our model is based. In Section 3 we formalize our dialogue protocol and the concepts related to dishonesty. In Section 4, we present and evaluate the results of our simulations. In Section 5, we compare our approach to other approaches. Finally, in Section 6 we present our conclusions.

2 ARGUMENTATION FRAMEWORK

Dung's abstract argumentation framework is defined as the pair of a set and a binary relationship on the set (Dung, 1995).

Definition 2.1 (argumentation framework)**.** An argumentation framework *is defined as a pair* $\langle AR, AT\rangle$ *where AR is the set of arguments and AT is a binary relationship on AR, called* an attack*. If* $(A, A') \in AT$*, we say that A* attacks A'*.*

Definition 2.2 (sub-AF)**.** *Let* $\mathcal{AF}_1 = \langle AR_1, AT_1\rangle$ *and* $\mathcal{AF}_2 = \langle AR_2, AT_2\rangle$ *be argumentation frameworks. If* $AR_1 \subseteq AR_2$ *and* $AT_1 = AT_2 \cap (AR_1 \times AR_1)$*, then it is said that* $\mathcal{AF}_1$ *is* a sub-argumentation framework (sub-AF, in short) *of* $\mathcal{AF}_2$ *and denoted by* $\mathcal{AF}_1 \subseteq \mathcal{AF}_2$*.*

We define the semantics of a given argumentation framework based on labelling (Baroni et al., 2011).

Definition 2.3 (labelling)**.** *Let* $\mathcal{AF} = \langle AR, AT\rangle$ *be an argumentation framework. A labelling is a total function* $\mathcal{L}^{\mathcal{AF}}$*: from AR to* $\{in, out, undec\}$*.*

The idea underlying the labelling is to give each argument a label. Specifically, the label *in* means that the argument is accepted in the argumentation framework, the label *out* means that the argument is rejected, and the label *undec* means that the argument is neither accepted nor rejected.

Definition 2.4 (complete labelling)**.** *Let* $\mathcal{AF} = \langle AR, AT\rangle$ *be an argumentation framework and* $\mathcal{L}^{\mathcal{AF}}$ *be its labelling. If the following condition holds for each* $A \in AR$*, then* $\mathcal{L}^{\mathcal{AF}}$ *is* a complete labelling of $\mathcal{AF}$*.*

1. $\mathcal{L}^{\mathcal{AF}}(A) = in$ *iff* $\forall A' \in AR$ ($(A', A) \in AT \Rightarrow \mathcal{L}^{\mathcal{AF}}(A') = out$).
2. $\mathcal{L}^{\mathcal{AF}}(A) = out$ *iff* $\exists A' \in AR$ ($(A', A) \in AT \wedge \mathcal{L}^{\mathcal{AF}}(A') = in$).
3. $\mathcal{L}^{\mathcal{AF}}(A) = undec$ *iff* $\mathcal{L}^{\mathcal{AF}}(A) \neq in \wedge \mathcal{L}^{\mathcal{AF}}(A) \neq out$*.*

Note that if an argument A is attacked by no arguments, then $\mathcal{L}^{\mathcal{AF}}(A) = in$.

There are various semantics based on labelling, but here, we use the term "labelling" to mean grounded labelling. Every argumentation framework has a unique grounded labelling.

Definition 2.5 (grounded labelling)**.** *Let* $\mathcal{AF}$ *be an argumentation framework. The* grounded labelling *of* $\mathcal{AF}$ *is a complete labelling* $\mathcal{L}^{\mathcal{AF}}$ *where a set of arguments that are labelled 'in' is minimal with respect to set inclusion.*

For example, Figure 1 shows an argumentation framework $\langle\{A,B,C,D,E\},\{(A,B),(B,C),(C,D),(D,E),(E,D)\}\rangle$ with its grounded labelling.

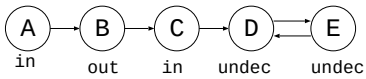


Figure 1: Labelled argumentation framework.

3 ARGUMENTATIVE DIALOGUE MODEL

We describe the argumentative dialogue model presented in (Takahashi and Yokohama, 2017).

3.1 Argumentation Frameworks

An argumentative dialogue is a sequence of arguments provided by agents following the protocol. Each agent has her own argumentation framework, as well as her prediction of the opponent's argumentation framework, and makes a move in a dialogue using them. When an argument is given, then these argumentation frameworks are updated.

Consider a dialogue between agents X and Y. We assume *a universal argumentation framework (UAF)* $\mathcal{UAF}$ as usual that contains every argument that can be constructed from all the available information in the universe. We naturally assume that $\mathcal{UAF}$ does not contain an argument that attacks itself. Let $\mathcal{AF}_X$ and $\mathcal{AF}_Y$ be argumentation frameworks of X and Y, respectively, where $\mathcal{AF}_X$, $\mathcal{AF}_Y \subseteq \mathcal{UAF}$; let $\mathcal{PAF}_Y$ and $\mathcal{PAF}_X$ be X's prediction of Y's argumentation framework and Y's prediction of X's argumentation framework respectively. That is, X has two argumentation frameworks, $\mathcal{AF}_X$ and $\mathcal{PAF}_Y$, and Y has $\mathcal{AF}_Y$ and $\mathcal{PAF}_X$. We assume several inclusion relationships among these argumentation frameworks. First, we assume $\mathcal{PAF}_X \subseteq \mathcal{AF}_X$ and $\mathcal{PAF}_Y \subseteq \mathcal{AF}_Y$, because common sense or widely prevalent facts are known to all agents, while there may be some facts that only the opponent knows and other facts that the agent is not sure whether the opponent knows. Second, we assume that $\mathcal{PAF}_Y \subseteq \mathcal{AF}_X$, $\mathcal{PAF}_X \subseteq \mathcal{AF}_Y$, because a prediction is made using an agent's own knowledge.

3.2 A Dialogue Protocol

We introduce three types of acts in a persuasion dialogue to focus on clarifying the effect of deception, although other acts can be considered.

Definition 3.1 (act)**.** *An act is assert, suspect, or excuse.*

Definition 3.2 (move)**.** A move *is a triple* (X,R,T)*, where* X *is an agent,* R *is an argument, and* T *is an act.*

Definition 3.3 (dialogue)**.** A dialogue d_k $(k \geq 0)$ between a persuader P and her opponent C on a subject argument A_0 *is a finite sequence of moves* $[m_0,\ldots,m_{k-1}]$ *where each* m_i $(0 \leq i \leq k-1)$ *is in the form of* (X_i,R_i,T_i) *and the following conditions are satisfied:*

$d_0 = [\,]$*;*
and if $k > 0$*,*
(i) $X_0 = P$*,* $R_0 = A_0$ *and* $T_0 =$ *assert.*
(ii) For each i $(0 \leq i \leq k-1)$*,* $X_i = P$ *if* i *is even,* $X_i = C$ *if* i *is odd.*
(iii) For each i $(0 \leq i \leq k-1)$*,* m_i *is one of the allowed moves.* An allowed move *is a move that obeys a dialogue protocol, as defined in Definition 3.4.*

For a dialogue $d_k = [m_0,\ldots,m_{k-1}]$, an argumentation framework of agent X for d_k is denoted by $\mathcal{AF}_X^{d_k}$; an agent X's prediction of Y's argumentation framework for d_k is denoted by $\mathcal{PAF}_Y^{d_k}$. They are defined constructively. $\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_Y^{d_0}$ are X's argumentation framework and her prediction of Y's argumentation framework given at an initial state where $A_0 \in \mathcal{AF}_{X_0}^{d_0}$.

A dialogue protocol is a set of rules for each act. An agent can give an argument contained in her argumentation framework at an instant. The preconditions of each act of agent X for d_k are formalized as follows. Hereafter, the symbol "_" in a move stands for anonymous.

Definition 3.4 (allowed move)**.** *Let* X,Y *be agents, and* $d_k = [m_0,\ldots,m_{k-1}]$ *be a dialogue. Let* $\mathcal{AF}_X^{d_k} = \langle AR_X^{d_k}, AT_X^{d_k}\rangle$ *and* $\mathcal{PAF}_Y^{d_k} = \langle PAR_Y^{d_k}, PAT_Y^{d_k}\rangle$ *be* X*'s argumentation framework and* X*'s prediction of* Y*'s argumentation framework for* d_k*, respectively. If a move* m_k *satisfies the precondition, then* m_k *is said to be* an allowed move for d_k*.*

When $k = 0$*,* $(X,A_0,assert)$ *is an allowed move where* A_0 *is a subject argument.*

When $k > 0$*, the precondition of each move is defined as follows.*

- $(X,A,assert)$*:*
  - $m_k \neq m_i$ *for* $\forall i$ $(0 \leq i < k)$
  - $m_{k-1} \neq (Y,\_,suspect)$
  - $\exists j$ $(0 \leq j < k)$*;* $m_j = (Y,A',\_)$ *and* $(A,A') \in AT_X^{d_k}$
- $(X,A,suspect)$*:*

- $m_{k-1} \neq (Y, \_, suspect)$
- $\exists j \ \ (0 \leq j < k)$; $m_j = (Y, A', \_)$ *and* $(A, A') \in PAT_Y^{d_k}$
- $\mathcal{L}^{\mathcal{PAF}_Y^{d_k}}(A) \neq out$

- $(X, A, excuse)$:
  - $m_{k-1} = (Y, A', suspect)$ *and* $(A, A') \in AT_X^{d_k}$ *and* $(\ \neg\exists(A_0, A_1, \ldots, A_n), \ (n > 1)$ *where* $A_0 = A_n = A$, $A_1 = A'$ *and* $(A_{i-1}, A_i) \in AT_X^{d_k}$ $(1 \leq \forall i \leq n)\ )$

- $(X, \_, pass)$

Agent can either give a counterargument A to an argument A' previously given by her opponent or just a pass. The same move of type *assert* is not allowed more than once throughout the dialogue.

A move of type *suspect* is to point out: "I suspect that you used argument A' while hiding another argument A." Y then has to demonstrate that they are not being deceptive by immediately giving a counterargument. This is a move of type *excuse*. As for *suspect*, a loop is avoided. An agent can give either a move of type *assert* and *suspect* on the same argument when both are allowed. An agent who is subjected to a move of type *excuse* is considered to have the burden of proof as Prakken et al. said (Prakken et al., 2005).

A move of type *pass* is passing on the turn, without giving any information. An agent can give it in two different ways: only when there is no other allowed moves (restricted use) or any time (free use). A *pass* move of a free use can be regarded as a kind of strategy of being silent and giving no information.

3.3 Update of Argumentation Frameworks

At each move, an argument in each agent's argumentation framework is disclosed. This may cause new arguments and new attacks to be put forward. A move of type *suspect* represents a suspicion on the previous argument, and generates no new arguments but for itself. This leads us to the following definition of an update of an argumentation framework with respect to a particular argument.

Definition 3.5 (update of argumentation framework)**.** *Let $\mathcal{UAF} = \langle UAR, UAT \rangle$ be a UAF. Let $\mathcal{AF} = \langle AR, AT \rangle$ be an argumentation framework, $A \in UAR$, and S be a set of arguments caused to be generated from A using a deductive inference, where the condition "if $A \in AR$ then $S \subseteq AR$" holds. Then, $\mathcal{AF}' = \langle AR \cup AR', AT \cup AT' \rangle$ is said to be* an argumentation framework of $\mathcal{AF}$ updated by A*, where* $AR' = \{A\} \cup S$ *and* $AT' = \{(B,C) | (B,C) \in UAT, (B \in AR', C \in AR) \vee (B \in AR, C \in AR') \vee (B \in AR', C \in AR')\}$ [2].

After the move $m_k = (X, R, T)$, the following updates are performed:

- d_{k+1} is obtained from d_k by adding m_k to its end
- $\mathcal{AF}_Y^{d_{k+1}}$, $\mathcal{PAF}_X^{d_{k+1}}$ and $\mathcal{PAF}_Y^{d_{k+1}}$ are argumentation frameworks of $\mathcal{AF}_Y^{d_k}$, $\mathcal{PAF}_X^{d_k}$ and $\mathcal{PAF}_Y^{d_k}$ updated by R, respectively
- $\mathcal{AF}_X^{d_k}$ remains unchanged

3.4 Dishonesty

Definition 3.6 (honest/dishonest move)**.** *For a dialogue $d_k = [m_0, \ldots, m_{k-1}]$ where $m_k = (X, R, T)$, if $\mathcal{L}^{\mathcal{AF}_X^{d_k}}(R) = in$, then m_k is said to be X*'s honest move *and R is said to be* an honest argument*; otherwise, m_k is said to be X*'s dishonest move *and R is said to be* a dishonest argument.

Definition 3.7 (suspicious move)**.** *For a dialogue $d_k = [m_0, \ldots, m_{k-1}]$ where $m_{k-1} = (X, R, assert)$ or $m_{k-1} = (X, R, excuse)$, if $\mathcal{L}^{\mathcal{PAF}_X^{d_k}}(R) \neq in$, then m_{k-1} is said to be* a suspicious move for *Y, and R is said to be* a suspicious argument.

Intuitively, honest move means that an agent gives an argument that she believes, and suspicious move means that she cannot believe her opponent argument. Note that "honest" is a concept for the persuader, whereas "suspicious" is that for her opponent. Hence, a dishonest argument is not always a suspicious argument and a suspicious argument is not always a dishonest argument.

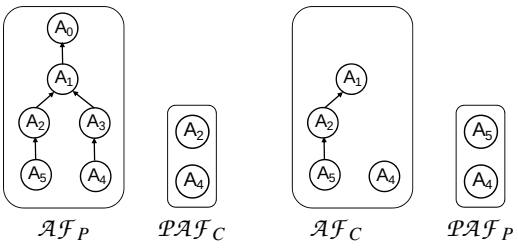
3.5 An Example

Consider the Labo-Selection Example shown in Section 1. Alice is a persuader P and Bob is a persuadee C. Let $A_0, A_1, \ldots, A_5$ be the following arguments.

A_0:	apply to Charlie's laboratory
A_1:	do not apply to Charlie's laboratory
A_2:	apply to Charlie's laboratory because he is generous
A_3:	apply to Charlie's laboratory because he is strict
A_4:	do not apply to Charlie's laboratory because he is strict
A_5:	Charlie is not generous

[2] $\mathcal{AF}'$ can be calculated without assuming $\mathcal{UAF}$ and S, if we handle an argumentation framework instantiated with logical formulas. In this case, we construct an argumentation framework by logical deduction from a given set of formulas (Amgoud et al., 2000; Yokohama and Takahashi, 2016).

We show one example of initial argumentation frameworks in Figure 2. Assume that $\mathcal{AF}_P$ is the same with a given UAF.

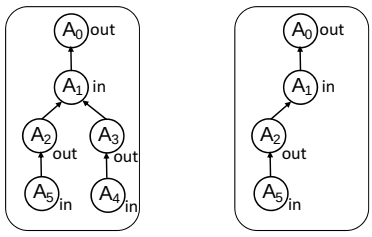


$\mathcal{AF}_P$ $\mathcal{PAF}_C$ $\mathcal{AF}_C$ $\mathcal{PAF}_P$

Figure 2: Initial state of argumentation frameworks.

A dialogue proceeds as follows:
$m_0 = (P, A_0, assert)$
$m_1 = (C, A_1, assert)$
$m_2 = (P, A_2, assert)$
$m_3 = (C, A_5, suspect)$

In this case, since $\mathcal{L}^{\mathcal{AF}_P^{d_3}}(A_2) = out$ (Figure 3(a)), m_2 is P's dishonest move, and since $\mathcal{L}^{\mathcal{PAF}_P^{d_3}}(A_2) = out$ (Figure 3(b)), m_2 is a suspicious move for C which causes to give m_3, a move of type *suspect*.



(a) $\mathcal{AF}_P^{d_3}$ (b)$\mathcal{PAF}_P^{d_3}$

Figure 3: Argumentation frameworks for d_3.

3.6 Termination

There are two ways to terminate a persuasive dialogue with a dishonest argument. In the first case, the agent cannot make an excuse when their opponent pointed out their deception. In this case, the agent is regarded as dishonest because she cannot answer her opponent's challenge, regardless of whether she actually made a dishonest move. In the second case, there exists d_k such that neither agent can make a move of type *assert* or *suspect*. In this case, we say that *the persuasion of P on subject argument* A_0 *succeeds* if $\mathcal{L}^{\mathcal{AF}_C^{d_k}}(A_0) = in$ holds; and *fails*, otherwise. After one agent has made a *pass* move, the other agent may present additional arguments, until neither agent has any further arguments.

Definition 3.8 (win/lose of persuasion)**.** *If a persuasion succeeds or the persuadee is regarded as dishonest, then it is said that* the persuader wins*. If persuasion fails or the persuader is regarded as dishonest, then it is said that* the persuader loses*.*

If we consider a dialogue as a debate game, win/lose of the game is judged by the arguments disclosed so far. We construct *a committed argumentation framework (CAF)* in addition to agents' inner argumentation frameworks.

Definition 3.9 (committed argumentation framework)**.** *For a dialogue* $d_k = [m_0, \ldots, m_{k-1}]$ *where* $m_k = (X, A, T)$*,* committed argumentation framework $\mathcal{CAF}^{d_k} = \langle AR^{d_k}, AT^{d_k} \rangle$ *is defined as follows.*

$$\mathcal{CAF}^{d_0} = \langle \emptyset, \emptyset \rangle$$

$$\mathcal{CAF}^{d_k} = \begin{cases} \langle AR^{d_{k-1}} \cup \{A\}, AT^{d_{k-1}} \cup (A, A') \rangle & (T \neq pass) \\ \langle AR^{d_{k-1}}, AT^{d_{k-1}} \rangle & (T = pass) \end{cases}$$

where $k > 0$ *and A attacks the previous argument* A'*.*

Definition 3.10 (win/lose of debate game)**.** *Let* $\mathcal{CAF}^{d_k}$ *be the CAF at the termination of the dialogue. It is said that the agent who proposed the subject argument* wins *if* $\mathcal{L}^{\mathcal{CAF}^{d_k}}(A_0) = in$ *holds; and* loses*, otherwise.*

4 EXPERIMENTAL RESULTS

4.1 Condition

A dialogue proceeds between a persuader P and a persuadee C according to the dialogue model described in Section 3.

P aims to persuade C to accept a subject argument, whereas C makes allowed moves but does not have a goal.

We take an arbitrary UAF with a tree structure that satisfies the following conditions[3]:

- The root node is a subject argument.
- The number of nodes is less than 20.
- The number of child nodes for each node is at most three.
- The number of leaf nodes is five to seven.

For a given UAF $\mathcal{UAF}$, we make four initial argumentation frameworks: $\mathcal{AF}_P$, $\mathcal{AF}_C$, $\mathcal{PAF}_C$ and $\mathcal{PAF}_P$. These frameworks all satisfy the inclusion relationships.

[3]These conditions are based on real argumentation experiments regarding a social issue conducted in a certain laboratory.

To simplify the problem, we assume that one of $\mathcal{AF}_P$ and $\mathcal{AF}_C$ is the biggest, that is, identical to $\mathcal{UAF}$; and the other is about half its size, that is, it consists of about half the number of arguments and half the number of attacks. In addition, to invoke a dialogue, we assume that the smaller one inherits half of the paths from the $\mathcal{UAF}$, and is denoted *half*. As for $\mathcal{PAF}_P$ and $\mathcal{PAF}_C$, we assume that one of them is the biggest, that is, *half*; and the other is the smallest, that is, the empty set. Under these conditions, we set the initial set of argumentation frameworks for a given UAF as one of the following four types in Table 1.

Table 1: Argumentation frameworks for agents.

Type	$\mathcal{AF}_P$	$\mathcal{AF}_C$	$\mathcal{PAF}_C$	$\mathcal{PAF}_P$
I	$\mathcal{UAF}$	*half*	*half*	$\emptyset$
II	*half*	$\mathcal{UAF}$	$\emptyset$	*half*
III	$\mathcal{UAF}$	*half*	$\emptyset$	*half*
IV	*half*	$\mathcal{UAF}$	*half*	$\emptyset$

We defined ten sets of argumentation frameworks for each type and then simulated all possible dialogues for five different UAFs. As a result, 50 cases are investigated for each type.

We implemented seven dialogue models (a)∼(g) that use different protocols: honest and dishonest strategies, and a range of strategies for making *pass* moves. We executed argumentations for each model and compared their results.

4.2 Results

We count the number of *P*'s win and that of *P*'s lose in all dialogues in each case. If the number of wins is greater than or equal to the number of losses, we say that *P is dominant* in the case; otherwise, we say that *C is dominant*. For example, if *P* wins 125 dialogues, and loses 67, then *P* is considered to be dominant in this case. Table 2 and Table 3 show the ratio of cases in which *P* was dominant out of the 50 cases tested for each type. We investigated persuasion dialogue and debate game, respectively.

In the following tables, the term 'honest' means that an agent only makes honest moves, while 'dishonest' means that an agent makes both honest and dishonest moves. The notation for honesty is as follows: 'dd' means that *P* and *C* are both dishonest, 'hd' means *P* is honest and *C* is dishonest, and 'dh' means *P* is dishonest and *C* is honest. The notation for *pass* moves is as follows: 'free' means that *pass* moves are allowed at any time, whereas 'rst' means that they are restricted when an agent cannot make any other move. The term 'termination' indicates *C*'s termination strategy: 'both' means that if *P* makes a *pass* move, then *C* also makes a *pass* move immediately, and 'one' means that *C* continues to make an allowed move as possible as *C* can.

Table 2 shows the results of persuasion, and compares the effects of dishonesty and *pass* moves. Comparing the result of models (a), (b) and (c), *P* is not dominant more often in (b) than that in (a) and (c). It shows that the number of dialogues which *P* wins increases by giving dishonest arguments. Comparing Type I and III, *P* is dominant in fewer cases when *P* has more predictions, which is against our expectation. It is because the forms of $\mathcal{AF}_C$ differ between Type I and III. It follows that the result of the dialogue depends on the form of the initial argumentation frameworks. We did not find any significant differences between honesty and dishonesty, or the two ways of making a *pass* move.

Table 2: The ratio of cases in which *P* is dominant (%): persuasion.

Type	model	(a)	(b)	(c)	(d)
	honesty	dd	hd	dh	dd
	pass	free	free	free	rst
I		36	32	36	36
II		40	40	40	40
III		62	60	62	50
IV		40	40	40	64

Table 3 shows the results of the debate game. We investigated the effect of the different *pass* move strategies. Both agents are dishonest. *P*'s *pass* move strategy is fixed as follows: when *C* gives a *pass* move, then *P* also makes a *pass* move immediately even if she still has allowed moves; in the other situations, she can give a *pass* move only when she has no other allowed moves. We investigated the effect of the different *pass* move strategies by varying *C*'s *pass* move strategy.

Table 3 shows the effect of *C*'s strategy. In model (g), if *C* makes a *pass* move, then *P* makes a *pass* move. This terminates the argumentation immediately. At that time, the label of the subject argument in CAF is *in*, which means that *P* has won in the debate game. *C* can make a *pass* move at any time, which causes *P* to be dominant more often than in cases (e) and (f).

Next, we investigated the effect of deception by examining each dialogue in specific cases. We compared the result of model (b) in which *P* is honest, to the others in which *P* is dishonest, for each initial set of argumentation frameworks. There exists no set of argumentation frameworks for which *P* is dominant in (b) and *C* is dominant in the other models. In addition, we have found two cases in Type I in wich *P* has

no possibility to win in model (b) while C is regarded as dishonest (P wins) in some dialogues in the other models. It is because the increase of P's arguments in number by giving dishonest arguments causes to increase a chance to reveal her opponent dishonesty. In these cases, P can win by making appropriate moves if she is dishonest, but not if she is honest.

Table 3: The ratio of cases in which P is dominant (%): debate game.

	model	(e)	(f)	(g)
Type	pass	rst	rst	free
	termination	one	both	both
I		34	34	68
II		22	22	24
III		58	58	62
IV		16	44	32

Table 4 shows the number of dialogues in which P reveals C's dishonesty and its ratio against all the dialogues for two specific cases. For example, in case 1, P loses all dialogues in model (b), while C is regarded as dishonest (P wins) in 22 dialogues, which is 2.0 percent of all dialogues and P loses the remaining dialogues in model (a).

Table 4: The number of dialogues in which P reveals C's dishonesty and its ratio.

	case 1				
model	(b)	(a)	(d)	(e)	(f)
number	0	22	12	2	12
ratio (%)	0	2.0	2.5	11.1	25

	case 2				
model	(b)	(a)	(d)	(e)	(f)
number	0	2	0	1	2
ratio (%)	0	7.4	0	25	33.3

4.3 Discussions

Even if an agent does not have any predictions, she may reveal the dishonesty of her opponent. This is counter-intuitive because the agent does not seem to make a move of type *suspect*. This is because the same argument can be given with a different act such as $(X, assert, A)$ and $(X, suspect, A)$. In this case, her own argumentation framework originally contains argument A. If she makes the move $(X, assert, A)$, then argument A is added to her prediction. Later, together with arguments presented by her opponent, her predictions have accumulated such that the label of A is now *out*. As a result, she may make the move $(X, suspect, A)$.

An agent loses if she cannot return a move of type *excuse* immediately upon receiving a *suspect* move. It follows that, when possible, it is more advantageous to make a move in the form $(X, A, suspect)$ than $(X, A, assert)$.

The ratio of P's dominance appears to depend on the form of the initial argumentation frameworks. However, our simulations did not show that making a dishonest argument or a *pass* move has any significant effect, regardless of the initial argumentation frameworks. Therefore, whether there is a relationship between the initial argumentation frameworks and the outcome of the argument remains an open question. The next point to consider will be how to determine moves strategically. We need to conduct more experiments and further analysis to address these points.

5 RELATED WORKS

Parsons et al. investigated the relationships between agents' initial knowledge and the outcome of the dialogue (Parsons et al., 2003). They clarified their characteristics and examined the effect of agents' tactics theoretically.

Thimm provided an excellent survey of strategic argumentation (Thimm, 2014). He classified the treatment of strategic argumentation from a variety of viewpoints, including game theory, opponent model, an so on.

In this paper, we assumed that an opponent model is given in advance, which is a similar assumption in most other works. On the other hand, some studies have investigated the process of updating the opponent model during the dialogue. Hunter studied persuasive dialogues, and how to evaluate them using an opponent model (Hunter, 2015). He proposed an asymmetric model between a system and a user, where a system makes moves such as $inform$ and *challenge* and receives simple yes/no answer from the user. Considering a user's reply, a user model that a system has is updated using probability. Since it is asymmetric, a user's argument is so restricted, and he does not focus on the strategy itself.

Rienstra el al.'s work is the most relevant to ours. They proposed several kinds of opponent models and presented experimental results from arguments based on these models. The prediction in our paper corresponds to the 'simple model' in their paper (Rienstra et al., 2013). They evaluated each dialogue upon termination and updated the opponent model probabilistically, while we don't use probability. In addition, they do not handle a dishonest argument and semantics of an argument.

Several platforms have been developed to compare agents' strategic arguments, and there are strategic argument competitions (Yuan et al., 2008). However, these do not include any perspectives on dishonest arguments.

Sakama formalized dishonesty using an argumentation framework as a debate game (Sakama, 2012; Sakama et al., 2015). Different from persuasion, they judged the outcome of the dialogue by committed argumentation framework, thus each agent needs not estimate an opponent model. He also investigated some properties of his model theoretically but did not formalize the detection of, or excuses for, deception.

6 CONCLUSIONS

We have presented the results of simulations of dishonest argumentation based on an opponent model. This is the first attempt to present an evaluation of dishonest argumentation. The results show that the use of dishonest arguments affects the chances of successfully persuading an opponent, or winning a debate games. But we could not identify a relationship between the result of a dialogue and the argumentation frameworks of agents.

As this is a preliminary report, only simple cases are handled. In future, we should perform more experiments on various types of argumentation frameworks that include cyclic structures, and facilitate more precise analysis. We will also investigate the results under different semantics. since concepts regarding dishonesty depend on the semantics.

REFERENCES

Amgoud, L. and de Saint-Cyr, F. (2013). An axiomatic approach for persuasion dialogs. In *ICTAI 2013*, pages 618–625.

Amgoud, L., Maudet, N., and Parsons, S. (2000). Modeling dialogues using argumentation. In *ICMAS2000*, pages 31–38.

Baroni, P., Caminada, M., and Giacomin, G. (2011). An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410.

Bench-Capon, T. (2003). Persuasion in practice argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.

Black, E. and Hunter, A. (2015). Reasons and options for updating an opponent model in persuasion dialogues. In *TAFA2015*.

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

Hadjinikolis, C., Siantos, Y., Modgil, S., Black, E., and McBurney, P. (2013). Opponent modelling in persuasion dialogues. In *IJCAI2013*, pages 164–170.

Hunter, A. (2015). Modelling the persuadee in asymmetric argumentation dialogues for persuasion. In *IJCAI2015*, pages 3055–3061.

Parsons, S., Wooldridge, M., and Amgoud, L. (2003). On the outcomes of formal inter-agent dialogues. In *AAMAS2003*, pages 616–623.

Prakken, H. (2006). Formal systems for persuasion dialogue. *The Knowledge Engineering Review*, 21(2):163–188.

Prakken, H., Reed, C., and Walton, D. (2005). Dialogues about the burden of proof. In *ICAIL2005*, pages 115–124.

Rahwan, I., Lason, K., and Tohm'e, F. (2009). A characterization of strategy-proofness for grounded argumentation semantics. In *IJCAI2009*, pages 251–256.

Rahwan, I. and Simari, G. (2009). *Argumentation in Artificial Intelligence*. Springer.

Rienstra, T., Thimm, M., and Oren, N. (2013). Opponent models with uncertainty for strategic argumentation. In *IJCAI2013*, pages 332–338.

Sakama, C. (2012). Dishonest arguments in debate games. In *COMMA2012*, pages 177–184.

Sakama, C., Caminada, M., and Herzig, A. (2015). A formal account of dishonesty. *The Logic Journal of the IGPL*, 23(2):259–294.

Takahashi, K. and Yokohama, S. (2017). On a formal treatment of deception in argumentative dialogues. In *EUMAS-AT2016, Selected papers*, pages 390–404.

Thimm, M. (2014). Strategic argumentation in multi-agent systems. *Kunstliche Intelligenz*, 28(3):159–168.

Yokohama, S. and Takahashi, K. (2016). What should an agent know not to fail in persuasion? In *EUMAS-AT2015, Selected papers*, pages 219–233.

Yuan, T., Schulze, J., Devereux, J., and Reed, C. (2008). Towards an arguing agents competition: Building on argumento. In *CMNA*.