On a Formal Treatment of Deception in Argumentative Dialogues

Kazuko Takahashi and Shizuka Yokohama*

School of Science&Technology, Kwansei Gakuin University,
2-1, Gakuen, Sanda, 669-1337, JAPAN
ktaka@kwansei.ac.jp, rec2016yoko@gmail.com

Abstract. This paper formalizes a dialogue that includes dishonest arguments in persuasion. We propose a dialogue model that uses a predicted opponent model and define a protocol using this prediction with an abstract argumentation framework. We focus on deception as dishonesty; that is, the case in which an agent hides her knowledge. We define the concepts of dishonest argument and suspicious argument by means of the acceptance of arguments in this model. We show how a dialogue including dishonest arguments proceeds according to the protocol and discuss a condition for a dishonest argument to be accepted without being revealed.

Keywords: argumentation, dialogue, persuasion, dishonesty, opponent model

1 Introduction

Persuasion is a popular form of dialogue that can help in reaching an agreement between agents. It is considered to be a process of solving inconsistency between agents' beliefs. Dialogue systems based on argumentation frameworks have been studied because argumentation is an efficient technique for handling inconsistency [14]. Several strategies are used to succeed in persuasion, and agents may sometimes lie or hide information that is disadvantageous to them to succeed in persuasion. However, few studies have examined dialogue that includes such dishonest arguments.

Dishonesty in argumentation frameworks was studied by Caminada. He classified dishonesty in dialogues into three types [7]: giving the negation of her belief (lie), generating an argument of which she does not know the truth (bullshit), and hiding an argument that she knows (deception). Sakama formalized the former two types using argumentation frameworks [17]. This formalization was made from the viewpoint of the agent who offers a dishonest argument, and not from that of the agent receiving it. That is, the dialogue proceeds without the receiver knowing what is going on, and she does not suspect her opponent's argument or reveal its dishonesty. Basically, to suspect the opponent's argument

* Currently, NEC Co., Ltd.

or reveal its dishonesty, especially to point out a deception, an agent should know, or at least predict, the opponent's belief.

Consider the following situation in which students are selecting a research laboratory.

> Alice tries to persuade Bob to apply to the same laboratory. Alice knows that Professor Charlie is strict, as well as 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, and Alice knows his intention.

Alice probably says, "Let's apply to Charlie's laboratory, because he is generous," hiding the fact that Charlie is strict, to persuade Bob. If Bob does not know of Charlie's reputation, he does not suspect Alice and accepts her argument.

However, assume that Bob knows both that (i) Charlie is strict and about Alice's knowledge (ii) Alice not only knows that Charlie is strict but also that Bob does not like strict professors. If Alice says, "Let's apply to Charlie's laboratory because he is generous," then he suspects its truth, and may say, "You know that Charlie is strict, and you also know that I do not like strict professors. Don't try to persuade me by hiding that fact." Alice deceives Bob, and it is based on the fact that Bob knows about Alice's knowledge whether he suspects her argument and points out her deception.

In the previous work [20], we formalized a persuasion dialogue using a predicted opponent model. We proposed a strategy and discussed what should be in a predicted opponent model so that persuasion does not fail. However, dishonest arguments were not discussed there.

In this paper, we modified our protocol to admit dishonest arguments of deception and formalize the mechanism used for giving a dishonest argument, suspecting an argument, pointing out a deception, and making an excuse.

In our dialogue model, each agent has two argumentation frameworks: her own and the prediction of her opponent's. A dialogue protocol is defined based on these frameworks. A dishonest argument and a suspicious argument are defined using the labelling semantics. The argumentation frameworks are updated as a dialogue proceeds. Accepted arguments in the current argumentation framework are considered to be her current beliefs. When her opponent gives an argument that is not accepted in her prediction of the opponent's argumentation framework, then she can point out the fact that the argument is suspicious. When an agent points out a suspicious argument, the opponent will make an excuse, if possible. An excuse may be accepted or suspected again. Also, if an excuse cannot be given, the suspect of the argument is not cleared. An agent sometimes succeeds in persuasion by accumulating dishonest arguments, and sometimes fails with the revelation of those dishonest arguments.

We illustrate how the defined protocol works and show that an excuse can be finally accepted after repetitive excuses if the agent always gives honest arguments. Furthermore, we discuss conditions on the agents' argumentation frameworks so that an agent succeeds in persuasion using dishonest arguments.

The rest of the paper is organized as follows. Section 2 describes the argumentation framework on which our model is based. Section 3 formalizes our dialogue protocol and concepts regarding dishonesty. Section 4 shows how this protocol works. Section 5 discusses the properties of the model. Section 6 compares our approach with other approaches. Finally, Section 7 presents 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 [8].

Definition 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'*.*

We define inclusions between argumentation frameworks.

Definition 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* $\mathcal{AF}_1 \subseteq \mathcal{AF}_2$*.*

For a given argumentation framework, we give its semantics based on labelling [5].

Definition 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 one abstains from an opinion as to whether the argument is accepted or rejected.

Definition 4 (complete labelling). *Let* $\mathcal{AF} = \langle AR, AT\rangle$ *be an argumentation framework and* $\mathcal{L}^{\mathcal{AF}}$ *its labelling. If the following condition holds for each* $A \in AR$*, then* $\mathcal{L}^{\mathcal{AF}}$ *is said to be* a complete labelling on $\mathcal{AF}$*.*

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 $\mathcal{L}^{\mathcal{AF}}(A) = in$.

Definition 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.*

A unique grounded labelling exists for any argumentation framework. For argumentation framework $\mathcal{AF}$ and its complete/grounded labelling $\mathcal{L}^{\mathcal{AF}}$, the set of arguments labelled *in* coincides with a complete/grounded extension of $\mathcal{AF}$ in extension-based semantics [5]. There are various semantics based on labelling, but here, we use the term "labelling" to mean grounded labelling.

Additionally, we define several other concepts used in Section 5 where we discuss the properties of this model.

Definition 6 (argumentation framework on an argument). *Let $\mathcal{AF} = \langle AR, AT \rangle$ be an argumentation framework, and $A \in AR$ be an argument. A sub-AF $\mathcal{AF}' = \langle AR', AT' \rangle$ that satisfies the following conditions is called an* argumentation framework of $\mathcal{AF}$ on A*:*

- $A \in AR'$
- *If $B \in AR'$ and $(C, B) \in AT$, then $C \in AR'$ and $(C, B) \in AT'$*

If an argumentation framework is a tree, it is said to be an *argumentation tree*. In an argumentation tree, the depth of the root node is 0, and a node at which the depth is even/odd is called an *even/odd node*.

Definition 7 (strong argumentation framework). *Let $\mathcal{TAF}_1$ and $\mathcal{TAF}_2$ be argumentation trees of which the root nodes correspond to the same argument, and $\mathcal{TAF}_1 \subseteq \mathcal{TAF}_2$. For any argument A of a leaf that is an odd node in $\mathcal{TAF}_1$, there exists an argument A' that attacks A in $\mathcal{TAF}_2$. Then it is said that $\mathcal{TAF}_2$ is stronger than $\mathcal{TAF}_1$.*

We can divide argumentation tree into a finite number of strategic argumentation trees.

Definition 8 (strategic argumentation tree). *For an argumentation tree, its* strategic argumentation tree *is its sub-AF containing all the child nodes of each even node and exactly one child node of each odd node.*

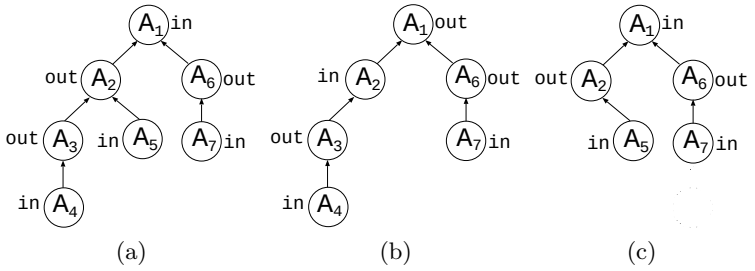


Fig. 1. Argumentation tree and its strategic argumentation trees with their labels

For example, Figure 1(a) shows an argumentation tree and Figure 1(b)(c) show its strategic argumentation trees.

3 Argumentative Dialogue Model

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* $\mathcal{UAF}$ which contains all arguments that can be constructed from all information that is available in the universes [17]. We naturally assume that $\mathcal{UAF}$ does not include an attack from an argument to itself. Let $\mathcal{AF}_X$ and $\mathcal{AF}_Y$ be argumentation frameworks of X and Y, respectively, where $\mathcal{AF}_X, \mathcal{AF}_Y \subseteq \mathcal{UAF}$; $\mathcal{PAF}_Y$ 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. Additionally, 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.

We introduce acts in a persuasion dialogue. The act *assert* is asserting an argument, *suspect* is pointing out a suspicious argument, and *excuse* is giving an excuse for it.

Definition 9 (act). *An act is assert, suspect, or excuse.*

Definition 10 (move). A move *is a triple* (X, R, T)*, where* X *is an agent,* R *is an argument, and* T *is an act.*

Definition 11 (dialogue). A dialogue d_k $(k > 0)$ 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:*

(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 below.*

For a dialogue $d_k = [m_0, \ldots, m_{k-1}]$ $(k > 0)$, 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 $\mathcal{PAF}_Y^{d_k}$. $\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_Y^{d_0}$ are X's argumentation framework and her prediction of Y's argumentation framework given at an initial state.

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 12 (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* $\forall i$ $(0 \leq i < k)$
    (It is not allowed more than once throughout the dialogue.)
  - $m_{k-1} \neq (Y, \_, suspect)$
    (The act immediately before the move is not suspect.)
  - $\exists j$ $(0 \leq j < k)$*;* $m_j = (Y, A', \_)$ *and* $(A, A') \in AT_X^{d_k}$
    (It is a counterargument to an argument previously given.)
- $(X, A, suspect)$*:*
  - $m_{k-1} \neq (Y, \_, suspect)$
    (The act immediately before the move is not suspect.)
  - $\exists j$ $(0 \leq j < k)$*;* $m_j = (Y, A', \_)$ *and* $(A, A') \in PAT_Y^{d_k}$
    (It is a counterargument to an argument previously given in her prediction.)
  - $\mathcal{L}^{\mathcal{PAF}_Y^{d_k}}(A) \neq out$
    (The label is not out in her prediction.)
- $(X, A, excuse)$*:*
  - $m_{k-1} = (Y, A', suspect)$ *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* $(A_{i-1}, A_i) \in AT_X^{d_k}$ $(1 \leq \forall i \leq n)$)
    (The act immediately before the move is suspect, a counterargument to the argument given immediately before, and there is no cycle of attacks including (A, A').)

Basically, an agent can give either a move of $(X, \_, assert)$ or $(X, \_, suspect)$ when both are allowed. However, we give priority to the move of *suspect* because here we are interested in dishonest arguments and it is not suitable to leave a suspicious argument.

A move of *suspect* is to point out, "I suspect that you used argument A' while hiding another argument A." Then Y has to give a counterargument immediately after this, demonstrating that it is not a deception. This is an excuse. Intuitively, when X thinks that Y tells what Y does not believe, X suspects; then Y immediately excuses to appeal that she believes it.

At each move, an argument in each agent's argumentation framework is disclosed. It may cause the generation of new arguments and new attacks. An act

suspect represents a suspicion on the argument previously given and generates no other arguments but for itself. As a result, an argumentation framework is updated with respect to the argument.

Definition 13 (update of argumentation framework). *Let* $\mathcal{UAF} = \langle UAR, UAT \rangle$ *be a universal argumentation framework. 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 by* A*, where 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* $AT' = \{(B, C) | (B, C) \in UAT, (B \in AR', C \in AR) \vee (B \in AR, C \in AR') \vee (B \in AR', C \in AR')\}$ [1].

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.

Deception is giving an argument while hiding an argument that attacks it, and "dishonesty" in this paper means deception.

Definition 14 (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 15 (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.

Definition 16 (cleared suspicious argument). *If* $m_{k-1} = (X, R, T)$ *is a suspicious move for* Y*, and there exists* h*;* $k < h$ *and* $\mathcal{L}^{\mathcal{PAF}_X^{d_h}}(R) = in$*, then* R *is said to be a* cleared suspicious argument *for* Y *at* d_h*, and it is said that* a suspicious argument R for Y is cleared at d_h*.*

Note that "honest" is a concept for the persuader, whereas "suspicious" is that for her opponent. It means that a dishonest argument is not always a suspicious argument and that a suspicious argument is not always a dishonest argument.

If neither agent has an allowed move, then the dialogue terminates. There are two types of termination. The first case is the one in which an agent cannot make an excuse when her opponent points out her deception. In this case, she is

[1] $\mathcal{AF}'$ can actually be calculated without assuming $\mathcal{UAF}$ and S, if we handle an argumentation framework instantiated with logical formulas. In that case, we construct an argumentation framework from a given set of formulas as a knowledge base by logical deduction [2, 20]. Strictly speaking, not an argument itself but formulas included in the argument may cause to generate new arguments.

regarded as dishonest because she cannot make an excuse, regardless of whether she actually made a dishonest move. The second case is the one in which there exists d_k such that neither agent can make an *assert* or *suspect* move. In this case, it is said that *persuasion of X by a subject argument A_0 succeeds* if $\mathcal{L}^{\mathcal{AF}_Y^{d_k}}(A_0) = in$ holds; *persuasion by a subject argument fails*, otherwise.

4 Examples of Dishonest Dialogues

We consider three scenarios in which suspicious moves occur. In these scenarios, persuader X gives dishonest arguments so that she tries to make the opponent Y believe a subject argument. The opponent Y may suspect X's argument and point out the deception, X tries to give an excuse against Y's pointing out. These scenarios show how the opponent Y reveals X's dishonest arguments using her prediction.

Let $\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_X^{d_0}$ be X's argumentation framework and Y's prediction of X's argumentation framework given at an initial state. For simplicity, we assume that no new arguments are caused to be generated but only a given argument in each move may be added, in these scenarios.

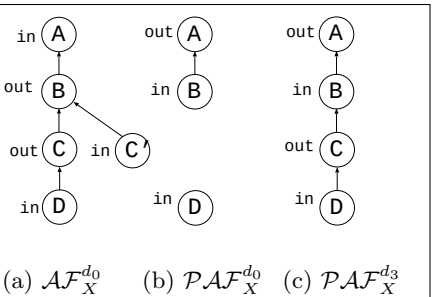
Scenario 1:

(X,A,assert): "Let's apply to Charlie's laboratory, because he is generous."
(Y,B,suspect): "You know that Charlie is strict, and you also know that I do not like strict professors. Don't try to persuade me by hiding that fact."
(X,C,excuse): "No, he is not strict, because I got an excellent grade last year, although my report was not very good."
(Y,D,suspect): "I don't think so, because I heard that not a few students failed. Don't try to persuade me by hiding that fact."

$\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_X^{d_0}$ are shown in Figure 2(a) and Figure 2(b), respectively. In this case, $\mathcal{AF}_X^{d_0}$ is unchanged and $\mathcal{PAF}_X^{d_0}$ is changed after the move m_2 (Figure 2(c)).

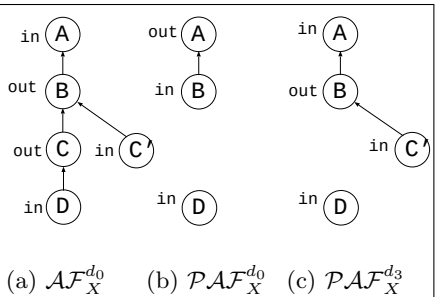
A dialogue proceeds as follows.

1. $m_0 = (X, A, assert)$: The first move. It is a suspicious move for Y because $\mathcal{L}^{\mathcal{PAF}_X^{d_1}}(A) \neq in$.
2. $m_1 = (Y, B, suspect)$: An allowed move because $\mathcal{L}^{\mathcal{PAF}_X^{d_1}}(B) \neq out$ and B attacks A in $\mathcal{PAF}_X^{d_1}$.
3. $m_2 = (X, C, excuse)$: An allowed move because C attacks B in $\mathcal{AF}_X^{d_2}$. An attack (D, C) is included as an attack of $\mathcal{UAF}$, because $\mathcal{AF}_X^{d_0} \subseteq \mathcal{UAF}$. Therefore, $\mathcal{PAF}_X^{d_2}$ is updated by C to get $\mathcal{PAF}_X^{d_3}$. It is also a suspicious move for Y because $\mathcal{L}^{\mathcal{PAF}_X^{d_3}}(C) \neq in$ (Figure 2(c)).
4. $m_3 = (Y, D, suspect)$: An allowed move because $\mathcal{L}^{\mathcal{PAF}_X^{d_3}}(D) \neq out$ and D attacks C in $\mathcal{PAF}_X^{d_3}$.
5. X cannot give an *excuse* and the dialogue terminates at d_4.



(a) $\mathcal{AF}_X^{d_0}$ (b) $\mathcal{PAF}_X^{d_0}$ (c) $\mathcal{PAF}_X^{d_3}$

Fig. 2. Argumentation frameworks in Scenario 1



(a) $\mathcal{AF}_X^{d_0}$ (b) $\mathcal{PAF}_X^{d_0}$ (c) $\mathcal{PAF}_X^{d_3}$

Fig. 3. Argumentation frameworks in Scenario 2

When C is given by X, it causes Y to create a new chance of an attack against X. Note that X has C' as a counterargument to B. However, the move containing C' is not allowed as m_4, because X should make an excuse for D in m_3 immediately.

In X's viewpoint, she gave two arguments, A and C. A is an honest argument, because $\mathcal{L}^{\mathcal{AF}_X^{d_0}}(A) = in$ and C is a dishonest argument, because $\mathcal{L}^{\mathcal{AF}_X^{d_2}}(C) \neq in$. In Y's viewpoint, both arguments are suspicious arguments for Y, and neither is cleared at d_4, because $\mathcal{L}^{\mathcal{PAF}_Y^{d_4}}(A) \neq in$ and $\mathcal{L}^{\mathcal{PAF}_Y^{d_4}}(C) \neq in$.

This scenario shows that X really makes a deception, and that it is revealed.

Scenario 2:

The third argument in Scenario 1 is replaced by the following argument.

(X,C',excuse): "You should apply to Charlie's lab., despite the fact that he is strict, because he has a strong connection to your promotion."

$\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_X^{d_0}$ are shown in Figure 3(a) and Figure 3(b), respectively. This is the same situation as that of Scenario 1. However, suspicious argument A is cleared by giving C' as a first excuse. $\mathcal{AF}_X^{d_0}$ is unchanged and $\mathcal{PAF}_X^{d_0}$ is changed after the move m_2 (Figure 3(c)).

A dialogue proceeds as follows. Moves m_0 and m_1 are the same as those in Scenario 1.

3. $m_2 = (X, C', excuse)$: An allowed move because C' attacks B in $\mathcal{AF}_X^{d_2}$, $\mathcal{PAF}_X^{d_2}$ is updated by C' to get $\mathcal{PAF}_X^{d_3}$, and as a result, m_2 is not a suspicious move for Y because $\mathcal{L}^{\mathcal{PAF}_X^{d_3}}(C') = in$ (Figure 3(c)).
4. Y cannot give *suspect* any more.

From X's viewpoint, she gave two arguments A and C', both of which were honest because $\mathcal{L}^{\mathcal{AF}_X^{d_0}}(A) = in$ and $\mathcal{L}^{\mathcal{AF}_X^{d_2}}(C') = in$. From Y's viewpoint, A is a suspicious argument for Y but finally cleared at d_3. C' is not a suspicious argument intrinsically. Agent Y may have more arguments, because $\mathcal{PAF}_X^{d_3} \subseteq$

$\mathcal{AF}_Y^{d_3}$. Therefore, if Y has an allowed move for d_3, then the dialogue continues by giving a move of *assert*; otherwise, it terminates, and X succeeds in persuading Y, because $\mathcal{L}^{\mathcal{AF}_Y^{d_3}}(A) = in$.

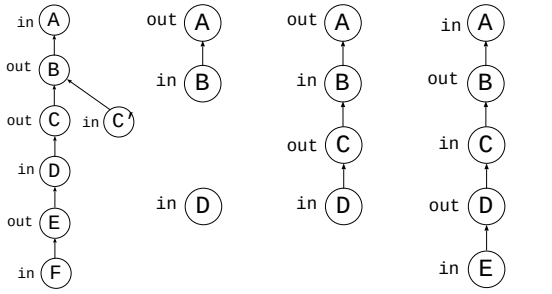
This scenario shows that X is always honest and even if her moves are suspicious, they are finally cleared.

Scenario 3:

The following argument is added to the end of a dialogue in Scenario 1.

(X,E,excuse): "It's just a rumor. I found that all the students passed the exam on the publication board.

$\mathcal{AF}_X^{d_0}$ and $\mathcal{PAF}_X^{d_0}$ are shown in Figure 4(a) and Figure 4(b), respectively. It is a modified version of Scenario 1. The difference is that $\mathcal{AF}_X^{d_0}$ has more arguments E and F. $\mathcal{AF}_X^{d_0}$ is unchanged and $\mathcal{PAF}_X^{d_0}$ is changed after the moves m_2 and m_4, respectively (Figure 3(c)(d)).



(a) $\mathcal{AF}_X^{d_0}$ (b) $\mathcal{PAF}_X^{d_0}$ (c) $\mathcal{PAF}_X^{d_3}$ (d) $\mathcal{PAF}_X^{d_5}$

Fig. 4. Argumentation frameworks in Scenario 3

A dialogue proceeds as follows. Moves from m_0 to m_3 are the same as those in Scenario 1.

5. $m_4 = (X, E, excuse)$: An allowed move because E attacks D in $\mathcal{AF}_X^{d_4}$, $\mathcal{PAF}_X^{d_4}$ is updated by E, and as a result, m_4 is not a suspicious move for Y because $\mathcal{L}^{\mathcal{PAF}_X^{d_5}}(E) = in$ (Figure 4(d)).
6. Y cannot give *suspect* any more.

From X's viewpoint, she gave three arguments A, C and E. A is an honest argument, because $\mathcal{L}^{\mathcal{AF}_X^{d_0}}(A) = in$ whereas C and E are dishonest arguments, because $\mathcal{L}^{\mathcal{AF}_X^{d_2}}(C) \neq in$ and $\mathcal{L}^{\mathcal{AF}_X^{d_4}}(E) \neq in$. From Y's viewpoint, A and C

are suspicious arguments for Y and cleared at d_5 because $\mathcal{L}^{\mathcal{PAF}_X^{d_5}}(A) = in$ and $\mathcal{L}^{\mathcal{PAF}_X^{d_5}}(C) = in$. E is not a suspicious argument intrinsically. Similar to the case in Scenario 2, X succeeds in persuasion depending on $\mathcal{AF}_Y^{d_5}$.

This scenario shows that X deceives repetitively, and that it is not revealed.

5 Properties of the Model

We discuss two properties that hold in our dialogue model. The first one shows that an excuse can be finally accepted after repetitive excuses if the agent always gives honest arguments. The second one shows a condition in which a suspicious argument is finally cleared.

Lemma 1. *For a dialogue d_k ($k \geq 0$), $\mathcal{PAF}_X^{d_k} \subseteq \mathcal{AF}_X^{d_k}$ holds.*

Proof. We prove this by induction. Let $\mathcal{AF}_X^{d_k} = \langle AR_X^{d_k}, AT_X^{d_k} \rangle$ and $\mathcal{PAF}_X^{d_k} = \langle PAR_X^{d_k}, PAT_X^{d_k} \rangle$. $\mathcal{PAF}_X^{d_0} \subseteq \mathcal{AF}_X^{d_0}$ holds. For $k > 0$, let $m_{k-1} = (X_{k-1}, R, T)$ and S be a set of arguments caused to be generated by R. $PAR_X^{d_k} = PAR_X^{d_{k-1}} \cup \{R\} \cup S$. If $X_{k-1} = X$, $AR_X^{d_{k-1}} = AR_X^{d_k}$. Here, $R \in AR_X^{d_{k-1}}$ and $S \subseteq AR_X^{d_{k-1}}$. From an induction hypothesis, $PAR_X^{d_{k-1}} \subseteq AR_X^{d_{k-1}}$. Therefore, $PAR_X^{d_k} \subseteq AR_X^{d_k}$. If $X_{k-1} = Y$, $AR_X^{d_k} = AR_X^{d_{k-1}} \cup \{R\} \cup S$. Therefore, $PAR_X^{d_k} \subseteq AR_X^{d_k}$. Thus, $PAR_X^{d_k} \subseteq AR_X^{d_k}$ for every $k \geq 0$. From the definition of attacks, it is trivial that $PAT_X^{d_k} \subseteq AT_X^{d_k}$. Thus, $\mathcal{PAF}_X^{d_k} \subseteq \mathcal{AF}_X^{d_k}$ holds. □

Lemma 2. *For a dialogue $d_{h+1} = [m_0, m_1, \ldots, m_k, \ldots, m_h]$, if $m_{k-1} = (X, R, T)$ is a suspicious move for Y, R is a cleared suspicious argument at d_h but not cleared at d_i ($k \leq i < h$), then $\mathcal{AF}_X^{d_i}$ is unchanged for all i; $k \leq i < h$.*

Proof. The act of the move m_i is either *excuse* or *suspect*. We prove the lemma depending on these acts. First, consider the case of $m_i = (X, B, excuse)$. $\mathcal{AF}_X^{d_{i+1}} = \mathcal{AF}_X^{d_i}$ from the definition of update. Second, consider the case of $m_i = (Y, B, suspect)$. Let $\mathcal{AF}_X^{d_i} = \langle AR_X^{d_i}, AT_X^{d_i} \rangle$ and $\mathcal{PAF}_X^{d_i} = \langle PAR_X^{d_i}, PAT_X^{d_i} \rangle$. Here, $B \in PAR_X^{d_i}$, and $PAR_X^{d_i} \subseteq AR_X^{d_i}$ from Lemma 1. $AR_X^{d_{i+1}} = AR_X^{d_i} \cup \{B\}$ holds, since B does not cause to generate new arguments. Thus, $AR_X^{d_{i+1}} \subseteq AR_X^{d_i}$ holds. Similarly, $AT_X^{d_{i+1}} \subseteq AT_X^{d_i}$ holds. Thus, $\mathcal{AF}_X^{d_{i+1}} \subseteq \mathcal{AF}_X^{d_i}$ holds. On the other hand, $\mathcal{AF}_X^{d_i} \subseteq \mathcal{AF}_X^{d_{i+1}}$ holds from the definition of update. Hence, $\mathcal{AF}_X^{d_{i+1}} = \mathcal{AF}_X^{d_i}$ holds. □

Proposition 1. *For a dialogue $d_{k+1} = [m_0, \ldots, m_{k-1}, m_k]$, let $m_{k-1} = (X, A, T)$ and $m_k = (Y, B, suspect)$. If $m_{k-1} = (X, A, T)$ is an honest move, then X can give an honest move $m_{k+1} = (X, C, excuse)$.*

Proof. Let $\mathcal{AF}_X^{d_k} = \langle AR_X^{d_k}, AT_X^{d_k} \rangle$ and $\mathcal{PAF}_X^{d_k} = \langle PAR_X^{d_k}, PAT_X^{d_k} \rangle$. Since m_{k-1} is an honest move and X's argumentation framework does not change after

giving m_{k-1}, $\mathcal{L}^{\mathcal{AF}_X^{d_k}}(A) = in$. On the other hand, since $m_k = (Y, B, suspect)$, $\mathcal{L}^{\mathcal{PAF}_X^{d_k}}(B) \neq out$ and $(B, A) \in PAT_X^{d_k}$. Here, $(B, A) \in AT_X^{d_k}$, because $\mathcal{PAF}_X^{d_k} \subseteq \mathcal{AF}_X^{d_k}$ from Lemma 1. Therefore, $\mathcal{L}^{\mathcal{AF}_X^{d_k}}(B) = out$, which means that there exists an argument C such that $(C, B) \in AT_X^{d_k}$ and $\mathcal{L}^{\mathcal{AF}_X^{d_k}}(C) = in$. $\mathcal{L}^{\mathcal{AF}_X^{d_{k+1}}}(C) = in$, because $\mathcal{AF}_X^{d_k} = \mathcal{AF}_X^{d_{k+1}}$ from Lemma 2. Thus, $(C, B) \in AT_X^{d_{k+1}}$ and $\mathcal{L}^{\mathcal{AF}_X^{d_{k+1}}}(C) = in$. Thus, $m_{k+1} = (X, C, excuse)$ is X's allowed move and an excuse for m_k. □

Next, we consider the condition on which a suspicious argument is finally cleared.

We can decide it not by surveying all possible dialogues, but only from the argumentation frameworks at the state in which the suspicious argument occurs. We use strategic argumentation trees on a subject argument. Intuitively, for an argumentation framework on a subject argument, each strategic argumentation tree shows a set of possible dialogues on a persuader's specific moves.

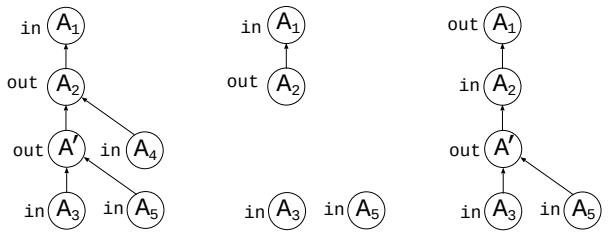
The condition should be that the opponent has no attack to persuader's argument in the final move of *excuse* in her prediction. Since the opponent's prediction is a subset of the persuader's argumentation framework, it means that all leaf nodes in the persuader's argumentation framework are labelled *in*. This condition is rather strict and can be loosened so that: first, the labels of the leaf nodes are not necessarily *in*, second, it is enough to consider only one strategic argumentation tree.

Before describing the condition, we introduce the concept of a complemented argumentation framework (compl-AF). When an agent is given a new argument from her opponent, a new attack may be generated from the existing arguments to the new argument by the update process. Complemented argumentation framework shows possible results of succeeding updates afterwards.

Definition 17 (compl-AF). *Let $\mathcal{AF} = \langle AR, AT \rangle$ be an argumentation framework and $\mathcal{TAF}$ be an argumentation tree, such that $\mathcal{AF} \subseteq \mathcal{TAF}$. If there exists a branch $(A_0, \ldots, A_{n-1}, A', A_n)$ in $\mathcal{TAF}$ such that $A_0, \ldots, A_{n-1}, A_n \in AR$ and $A' \notin AR$, then $AR' = AR \cup \{A'\}$, $AT' = AT \cup \{(A_n, A'), (A', A_{n-1})\}$. An argumentation framework $\langle AR', AT' \rangle$ obtained by doing this update for all such arguments A' is said to be* a complemented argumentation framework (compl-AF, in short) of $\mathcal{AF}$ wrt $\mathcal{TAF}$.

Compl-AF is an argumentation tree. Figure 5 shows its example.

Proposition 2. *For a dialogue d_k $(k > 0)$, assume that $m_{k-1} = (X, A, \_)$ and $m_k = (Y, B, suspect)$ are given. Let $\mathcal{TAF}_X^{d_k}$ be an argumentation framework of $\mathcal{AF}_X^{d_k}$ on a subject argument. (i) If $\mathcal{TAF}_X^{d_k}$ is an argumentation tree, and (ii) if there exist a strategic argumentation tree $\mathcal{SS}_X$ of $\mathcal{TAF}_X^{d_k}$ and a strategic argumentation tree $\mathcal{SS}_Y$ of the compl-AF of $\mathcal{PAF}_X^{d_k}$ wrt $\mathcal{TAF}_X^{d_k}$, such that $\mathcal{SS}_X$ is stronger than $\mathcal{SS}_Y$, then there exists h such that $k < h$, $\mathcal{L}^{\mathcal{PAF}_X{}^{d_h}}(A) = in$.*



(a) $\mathcal{TAF}$ (b) $\mathcal{AF}$ (c) Compl-AF of $\mathcal{AF}$ wrt $\mathcal{TAF}$

Fig. 5. Complemented argumentation framework

Sketch of Proof.

$\mathcal{SS}_X$ is a sub-AF of $\mathcal{AF}_X^{d_k}$ and it is an argumentation tree. The idea of the proof is to show that X can proceed a dialogue along a branch of $\mathcal{SS}_X$ and give *excuse* whenever Y gives *suspect*. For each k', such that $k < k'$, X can give a move $m_{k'+1} = (X, D, excuse)$ for a move $m_{k'} = (Y, C, suspect)$ as follows.

- If there is an attack (D, C) in $\mathcal{PAF}_X^{d'_k}$, then it is also an attack in $\mathcal{AF}_X^{d'_k}$, and so $m_{k'+1}$ is an allowed move.
- Else if there is an attack (D, C) in $\mathcal{SS}_Y$, it is also an attack in $\mathcal{SS}_X$, and so $m_{k'+1}$ is an allowed move.
- Otherwise, there exists an attack (D, C) in $\mathcal{SS}_X$, where C is an argument of an odd node of $\mathcal{SS}_Y$, because $\mathcal{SS}_X$ is stronger than $\mathcal{SS}_Y$, and so $m_{k'+1}$ is an allowed move.

There exists d_h for which there is no allowed move. In the third case, argument D that is not included in $\mathcal{SS}_Y$ appears in the move. Assume that $\mathcal{SS}'_Y$ is obtained by adding all such arguments appeared in $m_{k+1}, \ldots, m_h$, to the odd nodes of $\mathcal{SS}_Y$. Then, all of the leaves of $\mathcal{SS}'_Y$ are even nodes, and $\mathcal{L}^{\mathcal{SS}'_Y}(A) = in$ and $\mathcal{L}^{\mathcal{SS}'_Y}(B) = out$. Thus, there exists an argument E that attacks B such that $\mathcal{L}^{\mathcal{SS}'_Y}(E) = in$. Considering $\mathcal{PAF}_X^{d'_k}$ is updated in the above second and third cases, $\mathcal{SS}'_Y \subseteq \mathcal{PAF}_X^{d_h}$ holds. Since $\mathcal{SS}'_Y$ is an update of a strategic argumentation tree $\mathcal{SS}_Y$, the argument E attacks B also in $\mathcal{PAF}_X^{d_h}$, and $\mathcal{L}^{\mathcal{PAF}_X^{d_h}}(E) = in$. Thus, $\mathcal{L}^{\mathcal{PAF}_X^{d_h}}(B) = out$, and finally we get $\mathcal{L}^{\mathcal{PAF}_X^{d_h}}(A) = in$. □

This proposition shows a condition on which a suspicious argument is cleared if the agent selects a proper move under the specified condition. For example, suspicious arguments are cleared in Scenario 2, but not cleared in Scenario 1.

From this property, when a persuader has enough arguments in her argumentation framework that can attack whatever argument her opponent gives, she may succeed in persuasion without her dishonesty being revealed.

6 Related Works

There have been a few works on dishonest argumentation. Caminada proposed a classification of dishonesty occurring in multi-agent systems as well as human society, and described the relationship with argumentation [7]. Sakama formalized an untrusted argumentation including a lie and bullshit [17]. His formalization is from the viewpoint of the agent who gives a dishonest argument, and not from the agent that receives it. He did not define a protocol for pointing out a lie or one for making an excuse. On the other hand, we consider the situation from the viewpoints of both agents, and define protocols for pointing out a deception and making excuses. Additionally, his model is simpler in which only one argument is added at each move, while we consider the case where more arguments are caused to be generated. Rahwan et al. discussed hiding and lying in argumentation using game-theory techniques [15]. The most significant difference between our work and these other works is the usage of a predicted argumentation framework.

It is essential to consider an opponent's beliefs, especially when handling a strategic dialogue. Several works have examined this issue. Thimm et al. studied a strategy that reflected an opponent's belief [19], but they did not relate the belief to an acceptance of an argumentation framework. Rienstra et al. presented a strategy for selecting the best move from multiple opponent models with probability [16], and Hadjinikolis el al. showed an approach for augmenting opponent models from accumulated dialogues with an agent's likelihood [9]. They evaluated their approaches experimentally, whereas we focused on protocols more theoretically. Black et al. investigated the usage and maintenance of opponent models formally, illustrating a simple persuasion dialogue with different types of persuader [6]. These works also did not discuss dishonesty.

Prakken et al. studied the "burden of proof" in legal persuasion dialogues [12]. They focused on the issue which agent has to prove a subject or an argument depending on the protocols. It is considered that an agent that is given a move of *suspect* has a burden of proof and she makes an excuse in our persuasion dialogue model. Different from our model, they discussed on protocol level without considering argumentation frameworks of agents.

7 Conclusions

We have formalized a dialogue that includes dishonest arguments in persuasion. Deception is a technique often used in a real society, which is not regarded as dishonest at a glance. We formalized an argumentative dialogue that includes it. To this end, we proposed a dialogue model that uses a prediction of the opponent's argumentation framework. This is the first attempt at formalizing deception in this manner in the treatment of argumentative dialogues. Extension of this model should be considered so that it can handle other types of dishonest arguments, such as lies and bullshit.

We have also discussed the conditions for an agent to succeed in persuasion without her dishonesty being revealed. Generalization of these conditions is one of our future works.

Furthermore, we assume that a predicted argumentation framework is included within an actual one in this paper. The properties of models without this assumption should also be investigated.

References

1. Amgoud,L. and Cayrol,C.: On the acceptability of arguments in preference-based argumentation. UAI 1998, pp.1-7 (1998)
2. Amgoud,L, Maudet,N. and Parsons,S.: Modeling dialogues using argumentation. ICMAS2000, pp.31-38 (2000)
3. Amgoud,L. and Maudet,N.: Strategical considerations for argumentative agents (Preliminary Report). NMR2002, pp.399-407 (2002)
4. Bench-Capon,T.: Persuasion in practice argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3), 429-448 (2003)
5. Baroni,P., Caminada,M. and Giacomin,G.: An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4), pp.365-410 (2011)
6. Black,E. and Hunter,A.: Reasons and options for updating an opponent model in persuasion dialogues. TAFA2015 (2015)
7. Caminada,M.: Truth, Lies and Bullshit; distinguishing classes of dishonesty. IJCAI Workshop on Social Simulation, pp.39-50.
8. Dung,P.M.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77, 321-357 (1995)
9. Hadjinikolis,C., Siantos,Y. Modgil,S., Black.E. and McBurney,P.: Opponent modelling in persuasion dialogues. IJCAI2013, pp.164-170 (2013)
10. Modgil,S.: Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 195(0), 361-397 (2013)
11. Parsons,S., Wooldridge,M. and Amgoud,L.: On the outcomes of formal inter-agent dialogues. AAMAS2003, pp.616-623 (2003)
12. Prakken,H., Reed,C. and Walton,D.: Dialogues about the burden of proof. ICAIL2005, pp. 115-124 (2005).
13. Prakken,H.: An abstract framework for argumentation with structured arguments. *Arguments and Computation*, 1(2), 93-124 (2010)
14. Rahwan,I. and Simari,G.(eds.): *Argumentation in Artificial Intelligence*, Springer (2009)
15. Rahwan,I, Lason,K. and Tohmé,F.: A characterization of strategy-proofness for grounded argumentation semantics. IJCAI2009, pp.251-256 (2009)
16. Rienstra,T. Thimm,M. and Oren,N.: Opponent models with uncertainty for strategic argumentation. IJCAI2013, pp.332-338 (2013)
17. Sakama,C.: Dishonest arguments in debate games. COMMA2012, pp.177-184 (2012)
18. Sakama,C., Caminada,M. and Herzig,A.: A Formal Account of Dishonesty. *The Logic Journal of the IGPL*, 23(2) 259-294 (2015)
19. Thimm,M. and García,A.: On strategic argument selection in structured argumentation systems. ArgMAS2010, pp.286-305 (2010)
20. Yokohama,S. and Takahashi,K.: "What should an agent know not to fail in persuasion?" EUMAS-AT2015, Selected papers, pp. 219-233, LNCS 9571 (2016)