Dishonest Arguments in Debate Games

Chiaki SAKAMA [1]
Wakayama University, Japan

Abstract. In this paper we consider a *debate game* between two players in which a player may provide false or inaccurate arguments as a tactic to win the game. We formulate a debate game using a formal argumentation framework and investigate situation where a player may provide dishonest arguments in the game. We also argue how a player can detect dishonest arguments of the opponent player.

Keywords. argumentation framework, debate game, dishonest argument

1. Introduction

In his essay *The Art Of Controversy* [1], the German philosopher Arthur Schopenhauer argues that it is our common practice to use dishonest arguments in a debate. One would use dishonest arguments to defend his/her position, and another would use dishonest arguments to beat the opponent. Schopenhauer says: "*Hence we make it a rule to attack a counter-argument, even though to all appearances it is true and forcible, in the belief that its truth is only superficial, and that in the course of the dispute another argument will occur to us by which we may upset it, or succeed in confirming the truth of our statement. In this way we are almost compelled to become dishonest; or, at any rate, the temptation to do so is very great*" [1]. He then introduces 38 strategies including dishonest tricks to win a debate.

Dishonest arguments can be considered in a formal argumentation framework. The following example is taken from [2]. Alex thinks that Hortis bank is on the brink of bankruptcy because it has massively invested in mortgage backed securities. Bob also thinks that Hortis is on the brink of bankruptcy, but has read in an interview in which the finance minister promises that the state will support Hortis if needed. However, Bob also knows that the liabilities of Hortis are so big that not even the state will be able to provide significant help to avert bankruptcy. The situation is represented by three arguments: A: "Hortis Bank is on the brink of bankruptcy because of the mortgage backed securities." B: "The state will save Hortis because the finance minister promised so." C: "Not even the state has the financial means to save Hortis." Here, the argument B attacks A, and the argument C attacks B. Then, A and C are accepted and B is rejected.

Suppose that Alex only has the argument A, while Bob has three arguments A, B and C together with attack relations among them. When Alex claims that A is accepted, the claim is also acceptable for Bob. On the other hand, suppose a situation that Alex and Bob are in an election debate and Bob wants to win a debate against Alex's argument

[1] Correspondence to: Department of Computer and Communication Sciences, Wakayama University, Sakaedani, Wakayama 640-8510, Japan; E-mail: sakama@sys.wakayama-u.ac.jp.

anyway. In this case, Bob could say that A is unacceptable because of the existence of the argument B. Since Alex cannot refute B, he is obliged to accept Bob's argument. In this debate, Bob *lies* to Alex on the argument B because he knows that B would be rejected by the argument C. The example shows several aspects of a debate. First, each player of a debate generally has different arguments at the beginning. Alex has A, while Bob has A, B and C. Second, each player would obtain new arguments during a debate. Alex learns the argument B after Bob's claim. Third, a player can select a honest/dishonest argument to win a debate. Bob lies on B to win the debate.

The purpose of this paper is to formulate such a debate using a formal argumentation framework. We introduce a *debate game* between two players such that (i) each player has its own argumentation framework as a subset of the *universal argumentation framework*, (ii) during a debate each player can *revise* his/her argumentation framework by new arguments provided by the opponent player, and (iii) each player may *lie* or *bullshit* on his/her arguments to win a game. We investigate situations in which a player has a chance to win a debate game using honest/dishonest arguments, and argue a best-practice strategy for a player. In the above example, Alex cannot detect Bob's dishonesty and loses the game. Then, our question is whether there is any possibility of detecting dishonest arguments in a debate game. We provide an answer to the question.

The rest of this paper is organized as follows. Section 2 reviews some basic notions in a formal argumentation framework. Section 3 introduces debate games between two players. Section 4 investigates situations in which a player provides dishonest arguments to win a game. Section 5 argues how a player can detect dishonest arguments of the opponent in a game. Section 6 discusses related issues and rounds off the paper. Due to the lack of space, we omit the proofs of propositions which can be found in the longer version of this paper that is available at the author's homepage.

2. Argumentation Framework

An *argumentation framework* (AF) [3] is a pair (Ar, att) where Ar is a set of arguments and $att \subseteq Ar \times Ar$. An argument A *attacks* an argument B iff $(A, B) \in att$. An argumentation framework (Ar, att) is associated with a directed graph (called an *argumentation graph*) in which vertices are arguments in Ar and directed arcs from A to B exist whenever $(A, B) \in att$.

Let (Ar, att) be an argumentation framework. A *labelling* [4] is a (total) function $\mathcal{L} : Ar \to \{\, \mathtt{in}, \mathtt{out}, \mathtt{undec} \,\}$. When $\mathcal{L}(A) = \mathtt{in}$ (resp. $\mathcal{L}(A) = \mathtt{out}$ or $\mathcal{L}(A) = \mathtt{undec}$) for $A \in Ar$, it is written as $\mathtt{in}(A)$ (resp. $\mathtt{out}(A)$ or $\mathtt{undec}(A)$). We call $\mathtt{in}(A)$, $\mathtt{out}(A)$ and $\mathtt{undec}(A)$ *labelled arguments*. A labelling $\mathcal{L}$ is called *complete labelling* [4] if for each argument $A \in Ar$, it holds that:

- $\mathcal{L}(A) = \mathtt{in}$ iff $\mathcal{L}(B) = \mathtt{out}$ for every $B \in Ar$ such that $(B, A) \in att$.
- $\mathcal{L}(A) = \mathtt{out}$ iff $\mathcal{L}(B) = \mathtt{in}$ for some $B \in Ar$ such that $(B, A) \in att$.
- $\mathcal{L}(A) = \mathtt{undec}$ iff $\mathcal{L}(A) \neq \mathtt{in}$ and $\mathcal{L}(A) \neq \mathtt{out}$.

The set $\{A \mid \mathcal{L}(A) = \mathtt{in}\}$ with a complete labelling $\mathcal{L}$ coincides with a *complete extension* of (Ar, att) [4]. Every argumentation framework has at least one complete labelling. In this paper, labelling means complete labelling unless stated otherwise.

3. Debate Game

Definition 1 (UAF, sub-AF) [2] The *universal argumentation framework* (UAF) is an AF which contains all arguments that can be constructed from all information that is available in the universe. Given $UAF = (Ar, att)$, $AF = (Ar', att')$ is called a *subargumentation framework* (sub-AF) of (Ar, att) if $Ar' \subseteq Ar$ and $att' \subseteq att$. Given $UAF = (Ar, att)$, a *player* P_i has his/her argumentation framework $AF_i = (Ar_i, att_i)$ as a sub-AF of the UAF such that $Ar_i \subseteq Ar$ and $att_i = att \cap (Ar_i \times Ar_i)$.

Note that labelling of the UAF (Ar, att) does not coincide with labelling of its sub-AF (Ar', att') for $A \in Ar'$ in general. A player has arguments and attack relations which are included in the UAF. If a player has two arguments $A, B \in Ar$, then he/she has any attack relation between them. If two players both have access to arguments A and B, then they agree on whether A and B attack each other. A player has no information on any argument $C \in Ar \setminus Ar_i$, hence does not have any attack relation in $att \setminus (Ar_i \times Ar_i)$.

Definition 2 (revision of AF) Let $UAF = (Ar, att)$, and $AF_i = (Ar_i, att_i)$ an argumentation framework of a player P_i. Given an argument $X \in Ar$, a *revision* of AF_i with X is defined as $AF_i \circ X = (Ar_i \circ X, att_i \circ X)$ where $Ar_i \circ X = Ar_i \cup \{X\}$ and $att_i \circ X = att_i \cup \{\, (X, Y),\ (Z, X) \mid Y,\ Z \in Ar_i \text{ and } (X, Y),\ (Z, X) \in att \setminus att_i \,\}$ if $X \in Ar \setminus Ar_i$; otherwise, $AF_i \circ X = AF_i$.

If a player revises his/her AF, then labelling changes accordingly.

Example 1 Let $UAF = (\{A, B, C\}, \{(C, B), (B, A)\})$ and $AF_i = (\{A, B\}, \{(B, A)\})$. Then, AF_i has the labelling $\{\, \mathtt{out}(A),\ \mathtt{in}(B) \,\}$. A revision of AF_i with C becomes $AF_i \circ C = (\{A, B, C\}, \{(C, B), (B, A)\})$ with the labelling $\{\, \mathtt{in}(A),\ \mathtt{out}(B),\ \mathtt{in}(C) \,\}$.

Definition 3 (claim) A *claim* is a pair of the form: $(\mathtt{in}(A), \_\_)$ or $(\mathtt{out}(A), \mathtt{in}(B))$ where A and B are different arguments. $(\mathtt{in}(A), \_\_)$ is read "A is labelled `in`", while $(\mathtt{out}(A), \mathtt{in}(B))$ is read "A is labelled `out` because B is labelled `in`". A claim $(\mathtt{in}(A), \_\_)$ (resp. $(\mathtt{out}(A), \mathtt{in}(B))$) by a player is *refuted* by the claim $(\mathtt{out}(A), \mathtt{in}(B))$ (resp. $(\mathtt{out}(B), \mathtt{in}(C))$) by another player.

Definition 4 (debate game) Let $UAF = (Ar, att)$, and $AF_1^0 = (Ar_1^0, att_1^0)$ and $AF_2^0 = (Ar_2^0, att_2^0)$ argumentation frameworks of two players. Then, an *admissible debate* is a sequence of claims $[(\mathtt{in}(X_0), \_\_), (\mathtt{out}(X_0), \mathtt{in}(Y_1)), (\mathtt{out}(Y_1), \mathtt{in}(X_1)), \ldots, (\mathtt{out}(X_i), \mathtt{in}(Y_{i+1})), (\mathtt{out}(Y_{i+1}), \mathtt{in}(X_{i+1})), \ldots]$ such that

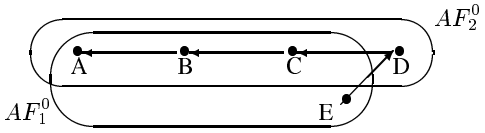
- $X_0 \in Ar_1^0$ and $X_k \in Ar_1^k$ where $AF_1^k = (Ar_1^k, att_1^k) = AF_1^{k-1} \circ Y_k$ $(k \geq 1)$.
- $Y_k \in Ar_2^k$ where $AF_2^k = (Ar_2^k, att_2^k) = AF_2^{k-1} \circ X_{k-1}$ $(k \geq 1)$.
- for each $\mathtt{out}(Z_j)$ in a claim by a player P_1 (resp. P_2), there is $\mathtt{in}(Z_i)$ $(i \leq j)$ in a claim by a player P_2 (resp. P_1).
- for each $(\mathtt{out}(U_i), \mathtt{in}(V_j))$, it holds that $(V_j, U_i) \in att$.

Let Δ_n $(n \geq 0)$ be any claim. A *debate game for an argument* X_0 is an admissible debate between two players where the initial claim is $\Delta_0 = (\mathtt{in}(X_0), \_\_)$. A debate game for an argument X_0 *terminates* with Δ_n if $[\Delta_0, \Delta_1, \ldots, \Delta_n]$ is an admissible debate and there is no claim Δ_{n+1} such that $[\Delta_0, \Delta_1, \ldots, \Delta_n, \Delta_{n+1}]$ is an admissible debate. In this case, we say that the player i who makes the last claim Δ_n *wins* the debate game for the argument X_0. (An old saying: "The one who has the last word laughs best" [3].)

[2] Martin Caminada developed a concept of UAF which is defined as the universe of all *valid* arguments [5]. By contrast, we take the UAF as consisting of all *possible* arguments. A similar notion is introduced in [6].

The player P_1 starts a debate with the claim $\Delta_0 = (\mathtt{in}(X_0), \_\!\_)$ based on its argumentation framework AF_1^0. The player P_2 then revises its argumentation framework AF_2^0 by X_0, and responds to the player P_1 with a counter-claim $\Delta_1 = (\mathtt{out}(X_0), \mathtt{in}(Y_1))$ based on the revised argumentation framework AF_2^1. A debate continues by iterating revisions and claims. By the third item of Definition 4, a player can refute not only the preceding claim of the opponent player, but any previous claim of the opponent. AF_i^k means an AF of a player P_i after k-th revision. We often omit k of AF_i^k and just call an argumentation framework AF_i of a player P_i when no confusion arises.

Example 2 Consider $UAF = (\{A, B, C, D, E\}, \{(E, D), (D, C), (C, B), (B, A)\})$, $AF_1^0 = (\{A, B, C, E\}, \{(C, B), (B, A)\})$ and $AF_2^0 = (\{A, B, C, D\}, \{(D, C), (C, B), (B, A)\})$ where the argumentation graph is on the right. AF_1^0 and AF_2^0 have the complete labellings: $\{\, \mathtt{in}(A),\ \mathtt{out}(B), \mathtt{in}(C),\ \mathtt{in}(E)\,\}$ and $\{\, \mathtt{out}(A),\ \mathtt{in}(B),\ \mathtt{out}(C),\ \mathtt{in}(D)\,\}$, respectively.



A debate game for the argument A between two players proceeds as follows:

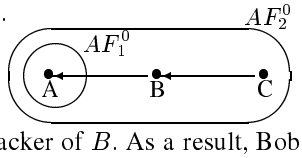
AF_1^0: $(\mathtt{in}(A), \_\!\_)$ "I claim that A is in."
AF_2^1: $(\mathtt{out}(A), \mathtt{in}(B))$ "A is out because B is in."
AF_1^1: $(\mathtt{out}(B), \mathtt{in}(C))$ "B is out because C is in."
AF_2^2: $(\mathtt{out}(C), \mathtt{in}(D))$ "C is out because D is in."
AF_1^2: $(\mathtt{out}(D), \mathtt{in}(E))$ "D is out because E is in."

Here, "AF_i^k: $(\mathtt{out}(X), \mathtt{in}(Y))$" means that a player P_i makes a claim $(\mathtt{out}(X), \mathtt{in}(Y))$ based on the argumentation framework AF_i^k. At first, the player P_1 has no information on the argument D, while the player P_2 has no information on the argument E. During the debate, the player P_1 learns the argument D by AF_2^2, then introduces it to AF_1^2 together with the attack relations (D, C) and (E, D). The player P_2 learns the argument E by AF_1^2 but cannot refute it. As a result, the player P_1 wins the game.

Note that each player revises his/her argumentation framework and labelling non-monotonically changes during a game (cf. Example 1). Then, a player may change his/her claim from $\mathtt{in}(X)$ to $\mathtt{out}(X)$ or from $\mathtt{out}(X)$ to $\mathtt{in}(X)$ for the same argument X in face of new evidences.

4. Dishonest Argumentation

Example 3 In the debate between Alex and Bob in Section 1, Alex has the argumentation framework $AF_1^0 = (\{A\}, \emptyset)$ and Bob has $AF_2^0 = (\{A, B, C\}, \{(C, B), (B, A)\})$. AF_1^0 has the labelling $\{\, \mathtt{in}(A)\,\}$, while AF_2^0 has the labelling $\{\, \mathtt{in}(A),\ \mathtt{out}(B),\ \mathtt{in}(C)\,\}$. When Alex claims $\mathtt{in}(A)$, it is acceptable for Bob. On the other hand, if Bob wants to win the debate game, he could provide *dishonest* arguments. Such a debate game would proceed in the following way:



AF_1^0: $(\mathtt{in}(A), \_\!\_)$
AF_2^1: $(\mathtt{out}(A), \mathtt{in}(B))$

Alex cannot refute Bob's claim because Alex knows no attacker of B. As a result, Bob wins the game. In this debate, however, Bob's claim at AF_2^1 is false because A is labelled in and B is labelled out in his complete labelling. In this case, we say that Bob is *lying*.

A player lies if he/she brings $\texttt{in}(A)$ while believing $\texttt{out}(A)$ or $\texttt{undec}(A)$ in his/her labelling. On the other hand, a player may bring an argument which is not in his/her AF. Such an argument is called *bullshit*. Lies and bullshit are dishonest arguments because information brought to the opposite player is false or inaccurate (in contrast to the reality as believed by the player). Honest or dishonest claims are defined as follows.

Definition 5 (honest/dishonest claim) Let $UAF = (Ar, att)$, and $AF = (Ar', att')$ an argumentation framework of a player. For any $A \in Ar'$,

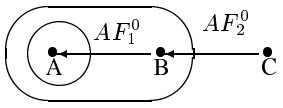
- a claim $(\texttt{in}(A), \_)$ is *honest* wrt AF if $\mathcal{L}(A) = \texttt{in}$ for some complete labelling $\mathcal{L}$ of AF; a claim $(\texttt{out}(A), \texttt{in}(B))$ is *honest* wrt AF if $B \in Ar'$ and $\mathcal{L}(B) = \texttt{in}$ for some complete labelling $\mathcal{L}$ of AF.
- a claim $(\texttt{in}(A), \_)$ is a *lie* wrt AF if $\mathcal{L}(A) \neq \texttt{in}$ for any complete labelling $\mathcal{L}$ of AF; a claim $(\texttt{out}(A), \texttt{in}(B))$ is a *lie* wrt AF if $B \in Ar'$ and $\mathcal{L}(B) \neq \texttt{in}$ for any complete labelling $\mathcal{L}$ of AF.
- a claim $(\texttt{in}(B), \_)$ or $(\texttt{out}(A), \texttt{in}(B))$ is *bullshit* wrt AF if $B \in Ar \setminus Ar'$.

A claim is called *dishonest* if it is either a lie or bullshit. A player in a debate game is called *honest* if every claim by the player is honest. Otherwise, a player is called *dishonest*. A labelled argument $\texttt{in}(X)$ (resp. $\texttt{out}(X)$) in a claim is also called *dishonest* if $\mathcal{L}(X) \neq \texttt{in}$ (resp. $\mathcal{L}(X) \neq \texttt{out}$) in any complete labelling $\mathcal{L}$ of AF.

We assume that a bullshitter understands what arguments are possible in the UAF but does not know whether it really holds or not. To allow the existence of dishonest players who bullshit, we need to modify Definition 4 of debate games such that each player may claim an argument which is not in his/her AF. Let $X_k \in {}^bAr_1^k$ and $Y_k \in {}^bAr_2^k$ where ${}^bAF_i^k = ({}^bAr_i^k, {}^batt_i^k) = AF_i^k \circ C$ for some $C \in Ar \setminus Ar_i^k$ $(i = 1, 2)$. In what follows, each player often plays a game based on ${}^bAF_i^k$, instead of AF_i^k.

Example 4 Suppose a different situation where Alex has the argumentation framework $AF_1^0 = (\{A\}, \emptyset)$ and Bob has $AF_2^0 = (\{A, B\}, \{(B, A)\})$ where $UAF = (\{A, B, C\}, \{(C, B), (B, A)\})$. Then AF_1^0 has the labelling $\{\, \texttt{in}(A) \,\}$, while AF_2^0 has the labelling $\{\, \texttt{out}(A),\ \texttt{in}(B) \,\}$. Suppose the following debate game:

AF_1^0: $(\texttt{in}(A), \_)$
AF_2^1: $(\texttt{out}(A), \texttt{in}(B))$
${}^bAF_1^1$: $(\texttt{out}(B), \texttt{in}(C))$



Bob cannot refute Alex's claim and Alex wins the game. In this debate, Alex's claim at ${}^bAF_1^1$ is bullshit because he has no belief on the truthfulness of C.

We next consider in which circumstances a player can win a game.

Definition 6 (upstream of an argument [5]) Given $AF = (Ar, att)$, the *upstream of an argument* $A \in Ar$ (written $UP_{AF}(A)$) is defined as the smallest set such that (i) $A \in UP_{AF}(A)$, and (ii) if $X \in UP_{AF}(A)$ and $(Y, X) \in att$, then $Y \in UP_{AF}(A)$.

Definition 7 (winning position) Let $UAF = (Ar, att)$, and $AF_i = (Ar_i, att_i)$ and $AF_j = (Ar_j, att_j)$ argumentation frameworks of two players. We say that AF_i is in a *winning position* of a debate game for an argument $X \in Ar_i$ if (i) $UP_{AF_j}(X) \subset UP_{AF_i}(X)$, and (ii) there is no $(Y, Z) \in att$ such that $Y, Z \in UP_{UAF}(X)$, $Y \in Ar \setminus Ar_i$ and $Z \in Ar_i$.

In Example 3, AF_2^0 is in a winning position for the argument A. In Example 4, on the other hand, no player is in a winning position for the argument A.

The winning position changes during a game. For instance, in Example 2, AF_1^0 is not in a winning position but the player P_1 gets a winning position at AF_1^2. If there is a debate game for an argument A in which a player can win the game using honest/dishonest claims, then we say that a player has a *chance* to win the game.

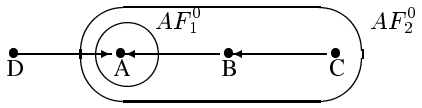
Proposition 1 *Suppose a debate game for the argument A. If AF_i is in a winning position for A, then a player P_i has a chance to win the game.*

The converse of Proposition 1 does not hold in general. For instance, Alex wins the game in Example 4, but he is not in a winning position for the argument A. From the moral viewpoint, a player wants to behave honestly as much as possible. From the practical viewpoint, dishonest arguments would cause personal discomfort and result in criticism if detected. Then, our question is in what circumstance a player has an incentive to use dishonest arguments. Let $AF = (Ar, att)$ and $A \in Ar$. An argument $X \in UP_{AF}(A)$ is *terminal* if there is no attack relation $(Y, X) \in att$ for any $Y \in Ar$. A *path* from X to Y is a directed path in an argumentation graph associated with AF. The *length* of a path from X to Y is the number of attack relations between the path from X to Y.

Proposition 2 *Suppose a debate game between two players P_1 and P_2 who have argumentation frameworks AF_1 and AF_2, respectively. Then, a honest player P_1 (resp. P_2) has a chance to win a debate game for the argument $A \in Ar_1$ if AF_1 (resp. AF_2) is in a winning position and the length of any path from any terminal $X \in UP_{AF_1}(A)$ to A is even (resp. some terminal $Y \in UP_{AF_2}(A)$ to A is odd).*

Proposition 2 provides a sufficient condition for an honest player to win a game. In other words, if the condition is not satisfied, a player has a reason to behave dishonestly. For instance, in Example 3, Bob (AF_2^0) is in a winning position but the length of a path from C to A is even. Then, Bob has an incentive of lying to win the game. On the other hand, the dishonest argument by Bob (AF_2^1) does not always succeed. If Alex has information on the argument C, then AF_2^1 is refuted by Alex and Bob loses the game. (Even if Alex has no information on C, he may bullshit $\mathtt{in}(C)$ and wins the game.) Thus, a player may lie or bullshit if he/she considers that there is no chance to win a game using honest claims only. Then, our next question is which dishonest claim a player should use if both a lie and bullshit are effective.

Example 5 Consider $UAF = (\{A, B, C, D\}, \{(C, B), (B, A), (D, A)\})$, $AF_1^0 = (\{A\}, \emptyset)$ and $AF_2^0 = (\{A, B, C\}, \{(C, B), (B, A)\})$. First, the player P_1 claims $(\mathtt{in}(A), \_)$ based on AF_1^0. To win the game, there are two possible claims for the player P_2. One is the lie $(\mathtt{out}(A), \mathtt{in}(B))$ and the other is the bullshit $(\mathtt{out}(A), \mathtt{in}(D))$.



In Example 5, the player P_2 can select either a lie or bullshit. Comparing these two options, we consider that a lie is less preferable than bullshit. This is because P_2 knows the falsehood of $(\mathtt{out}(A), \mathtt{in}(B))$, while he/she does not know the truthfulness of $(\mathtt{out}(A), \mathtt{in}(D))$.[3] The player P_2 does not know whether the player P_1 has information on C or D. If P_2 lies but P_1 has information on the argument C, then P_2 loses the game. Moreover, if P_1 knows that P_2 has information on C, then P_2 would be criticized for his/her lying. On the other hand, the player P_2 does not lose the game by bullshitting, and would not be criticized for his/her bullshitting even if the player P_1 has information

[3]The preference is stated as a postulate in [7] that *never lie if you can bullshit your way out of it.*

on D. (In this case, the initial claim by P_1 is a lie.) With these reasons, we introduce a *best-practice strategy* for a player to win a debate game.

Definition 8 (best-practice strategy) Let $UAF = (Ar, att)$, and $AF_1^0 = (Ar_1^0, att_1^0)$ and $AF_2^0 = (Ar_2^0, att_2^0)$ argumentation frameworks of two players P_1 and P_2, respectively. Let $(\mathtt{in}(A), \_\!\_)$ be the initial claim by AF_1^0.

1. If $\mathcal{L}(A) = \mathtt{in}$ in some complete labelling $\mathcal{L}$ of AF_1^k ($k \geq 1$), then the player P_1 makes an honest claim based on AF_1^k. Otherwise, (a) if there is an argument $C \in Ar \setminus Ar_1^k$ such that a claim $(\mathtt{out}(B), \mathtt{in}(C))$ with $B \in Ar_1^k$ refutes a claim by AF_2^k and $AF_1^k \circ C$ has a complete labelling $\mathcal{L}$ such that $\mathcal{L}(A) = \mathtt{in}$, then P_1 makes bullshit $(\mathtt{out}(B), \mathtt{in}(C))$ at ${}^bAF_1^k$; (b) otherwise, P_1 lies by $(\mathtt{out}(B), \mathtt{in}(D))$ for $B, D \in Ar_1^k$ at AF_1^k where $(\mathtt{out}(B), \mathtt{in}(D))$ refutes a claim by AF_2^k.

2. If $\mathcal{L}(A) = \mathtt{out}$ in some complete labelling $\mathcal{L}$ of AF_2^k ($k \geq 1$), then the player P_2 makes an honest claim based on AF_2^k. Otherwise, (a) if there is an argument $C \in Ar \setminus Ar_2^k$ such that a claim $(\mathtt{out}(B), \mathtt{in}(C))$ with $B \in Ar_2^k$ refutes a claim by AF_1^{k-1} and $AF_2^k \circ C$ has a complete labelling $\mathcal{L}$ such that $\mathcal{L}(A) = \mathtt{out}$, then P_2 makes bullshit $(\mathtt{out}(B), \mathtt{in}(C))$ at ${}^bAF_2^k$; (b) otherwise, P_2 lies by $(\mathtt{out}(B), \mathtt{in}(D))$ for $B, D \in Ar_2^k$ at AF_2^k where $(\mathtt{out}(B), \mathtt{in}(D))$ refutes a claim by AF_1^{k-1}.

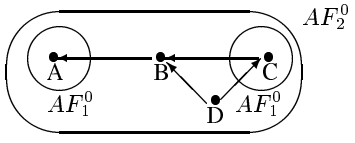
In Example 5, AF_2^0 has the labelling $\{\, \mathtt{in}(A),\ \mathtt{out}(B),\ \mathtt{in}(C)\,\}$. Then, the player P_2 claims the bullshit $(\mathtt{out}(A), \mathtt{in}(D))$ at ${}^bAF_2^1$ where $AF_2^1 \circ D$ has the labelling $\{\, \mathtt{out}(A),\ \mathtt{out}(B),\ \mathtt{in}(C),\ \mathtt{in}(D)\,\}$. Since the player P_1 cannot refute the claim at AF_1^1, the player P_2 wins the game.

5. Detecting Dishonest Arguments

> "*When your opponent makes a proposition, you must try to see whether it is not in some way – if needs be, only apparently – inconsistent with some other proposition which he has made or admitted, . . .*" [1, Stratagems XVI]

Example 6 Consider two argumentation frameworks $AF_1^0 = (\{A, C\}, \emptyset)$ and $AF_2^0 = (\{A, B, C, D\}, \{(D, B), (D, C), (C, B), (B, A)\})$, where AF_1^0 and AF_2^0 have the labellings: $\{\, \mathtt{in}(A),\ \mathtt{in}(C)\,\}$ and $\{\, \mathtt{in}(A),\ \mathtt{out}(B),\ \mathtt{out}(C), \mathtt{in}(D)\,\}$, respectively. Suppose the following debate game:

AF_1^0: $(\mathtt{in}(A), \_\!\_)$
AF_2^1: $(\mathtt{out}(A), \mathtt{in}(B))$
AF_1^1: $(\mathtt{out}(B), \mathtt{in}(C))$
AF_2^2: $(\mathtt{out}(C), \mathtt{in}(D))$



After AF_2^2, the player P_1 would reason that: (i) The player P_2 claimed $\mathtt{in}(B)$ at AF_2^1. Since I (i.e., the player P_1) provide no argument wrt B, P_2 should have $\mathtt{in}(B)$ in a labelling of AF_2^0. (ii) The player P_2 also claimed $\mathtt{in}(D)$ at AF_2^2. Since I provide no argument wrt D, P_2 should have $\mathtt{in}(D)$ in the same labelling of AF_2^0. (iii) The player P_2 has two arguments B and D at AF_2^0, then P_2 should have the attack relation (D, B) in AF_2^0. (iv) To have $\mathtt{in}(B)$ in a labelling, however, any attacker of B must be out. This contradicts the claim $\mathtt{in}(D)$. By this reasoning, the player P_1 concludes that the opponent P_2 has provided dishonest arguments in the game. In fact, AF_2^1 is a lie.

As remarked at the end of Section 3, a player may change labelling after revising his/her argumentation framework. Then, one cannot conclude that a player lies due to the fact that the player firstly claims $\mathtt{in}(X)$ and later claims $\mathtt{in}(Y)$ for some attacker

Y of X. In Example 6, however, the player P_1 conjectures that both $\mathtt{in}(B)$ and $\mathtt{in}(D)$ come from AF_2^0, and concludes that it is impossible to have them in the presence of the attack relation (D, B) in AF_2^0. Generally, a player can conclude that the opponent player provides dishonest arguments if there are two labelled arguments $\mathtt{in}(X)$ and $\mathtt{in}(Y)$ that stem from the opponent while they cannot coexist in any labelling of the argumentation framework that is known by two players during the game. The situation is formally presented as follows. We say that an argument Z is *originated* by a player P_1 (resp. P_2) if $\mathtt{in}(Z)$ is included in a claim made by AF_1^k (resp. AF_2^l) but not included in any claim made by AF_2^i (resp. AF_1^j) where $1 \leq i \leq k$ and $0 \leq j < l$.

Proposition 3 *Suppose the $UAF = (Ar, att)$ and a debate game for an argument X_0 between two players P_1 and P_2 who have AF_1^0 and AF_2^0, respectively. Then, P_1 at AF_1^k (resp. P_2 at AF_2^k) ($k \geq 2$) can conclude that P_2 (resp. P_1) has provided dishonest arguments in the game if there are two labelled arguments $\mathtt{in}(Y_i)$ and $\mathtt{in}(Y_j)$ $(0 < i < j \leq k)$ (resp. $\mathtt{in}(X_i)$ and $\mathtt{in}(X_j)$ $(0 \leq i < j < k)$) such that there is an odd-length path from Y_i to Y_j or from Y_j to Y_i (resp. from X_i to X_j or from X_j to X_i) in the argumentation graph associated with AF_1^k (resp. AF_2^k) in which every argument on the path is originated by P_2 (resp. P_1).*

6. Discussion

As argued by Schopenhauer [1], people frequently use dishonest arguments in daily life. In spite of this fact, formulation of dishonest arguments has received little attention in formal argumentation except the following studies. Caminada [2] argues how dishonesties are effectively used in formal argumentation, while the study focuses on conceptual issue on dishonesty and does not develop a formal theory of dishonest argumentation. Rahwan *et al.* [8] introduce a formal argumentation theory in which an agent may hide or lie about arguments. The purpose of their study is to develop a game-theoretic argumentation mechanism design and to investigate graph-theoretic conditions for strategy-proofness under the grounded semantics. A formal argumentation framework has been used for modelling dialogue games [9] or discussion games [10]. However, these studies do not consider dishonest arguments as a strategy for winning a game nor investigate possibilities of detecting dishonest arguments in a game.

References

[1] A. Schopenhauer. *The Art of Controversy*, Cosimo Classics, New York, 2007.

[2] M. Caminada. Truth, lies and bullshit, distinguishing classes of dishonesty, *SS@IJCAI*, 2009.

[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** (1995), 321–357.

[4] M. Caminada and D. Gabbay. A logical account of formal argumentation, *Studia Logica* **93** (2009), 109–145.

[5] M. Caminada. On the issue of argumentation and informedness, Research Report, 2011.

[6] N. D. Rotstein, M. O. Moguillansky, M. A. Falappa, A. J. García and G. R. Simari. Argument theory change: revision upon warrant, *Proc. COMMA 2008*, 336–347.

[7] C. Sakama, M. Caminada and A. Herzig. A logical account of lying, *Proc. 12th European Conference on Logics in Artificial Intelligence, Lecture Notes in AI*, vol. 6341, 286–299, Springer, 2010.

[8] I. Rahwan, K. Larson and F. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics, *Proc. IJCAI-09*, 251–256, 2009.

[9] H. Prakken. Coherence and flexibility in dialogue games for argumentation, *J. Logic and Computation* **15** (2005), 1009–1040.

[10] M. Caminada. Preferred semantics as Socratic discussion, *Argumentation in AI and Philosophy*, 2010.