

Dishonest Arguments in Debate Games

Chiaki SAKAMA¹

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.

Key words. argumentation framework, debate game, dishonest argument

1. Introduction

“In itself Dialectic has nothing to do but to show how a man may defend himself against attacks of every kind, and especially against dishonest attacks; and, in the same fashion, how he may attack another man’s statement without contradicting himself, or generally without being defeated.”

– Arthur Schopenhauer, *The Art Of Controversy*

In his essay *The Art Of Controversy* [13], 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”* [13]. 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.

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

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 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.

2. Argumentation Framework

Definition 1 (argumentation framework [6]) An *argumentation framework* (AF) 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 $(A, B) \in att$.

Definition 2 (labelling [1]) Let (Ar, att) be an argumentation framework. A *labelling* is a (total) function $\mathcal{L} : Ar \rightarrow \{\text{in}, \text{out}, \text{undec}\}$.

When $\mathcal{L}(A) = \text{in}$ (resp. $\mathcal{L}(A) = \text{out}$ or $\mathcal{L}(A) = \text{undec}$) for $A \in Ar$, it is written as $\text{in}(A)$ (resp. $\text{out}(A)$ or $\text{undec}(A)$). We call $\text{in}(A)$, $\text{out}(A)$ and $\text{undec}(A)$ *labelled arguments*.

Definition 3 (complete labelling [1]) Let (Ar, att) be an argumentation framework. A labelling \mathcal{L} is called *complete labelling* if for each argument $A \in Ar$, it holds that:

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

The set $\{A \mid \mathcal{L}(A) = \text{in}\}$ with a complete labelling \mathcal{L} coincides with a *complete extension* of (Ar, att) [1]. Every argumentation framework has at least one complete labelling. In this paper, labelling means complete labelling unless stated otherwise.

3. Debate Game

A debate game has two players, a player P_1 and a player P_2 , who have their own argumentation frameworks AF_1 and AF_2 , respectively. The player P_1 starts with a particular argument X that is labelled *in* in some complete labelling of AF_1 . The player P_2 tries to refute *in*(X) based on an argument Y which attacks X and is labelled *in* in some complete labelling of AF_2 . The player P_1 then tries to refute *in*(Y) by finding an attacker of it. A debate game proceeds by repeating such arguments by two players. During a debate game, each player can learn a new argument that is posed by the opponent player. In this case, the player revises his/her argumentation framework by incorporating the new argument and corresponding attack relations. The player then continues the game based on the revised argumentation framework. We first introduce the universe of arguments, which contain every possible argument in the world.

Definition 4 (UAF, sub-AF)² The *universal argumentation framework* (UAF) is an argumentation framework which contains all arguments that can be constructed from all information that is available in the universe. Given the UAF (Ar, att) , an argumentation framework (Ar', att') is called a *subargumentation framework* (sub-AF, for short) of (Ar, att) if $Ar' \subseteq Ar$ and $att' \subseteq att$.

Given the 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. By the definition, 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 different players both have access to arguments A and B , then they agree on whether A and B attack each other. On the other hand, a player has no information on any argument $C \in Ar \setminus Ar_i$, hence does not have any attack relation in $att \setminus (Ar_i \times Ar_i)$. In what follows, we often identify a player with his/her argumentation framework.

Definition 5 (revision of AF) Let (Ar, att) be the UAF 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

²Martin Caminada developed a concept of UAF which is defined as the universe of all *valid* arguments [4]. By contrast, we take the UAF as consisting of all *possible* arguments. A similar notion is introduced in [?].

$att_i \circ X = att_i \cup \{ (X, Y), (Z, X) \mid Y, Z \in Ar_i \text{ and } (X, Y), (Z, X) \in att \setminus att_i \}$ if $X \in Ar \setminus Ar_i$; otherwise, $AF_i \circ X = AF_i$.

In the above definition, AF_i is revised by introducing a new argument and attack relations. Such a revision is made by each player whenever new arguments are revealed during a debate game. An argumentation framework of a player P_i after k -th revision is written by AF_i^k . 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 $\{\text{out}(A), \text{in}(B)\}$. A revision of AF_i with C becomes $AF_i \circ C = (\{A, B, C\}, \{(C, B), (B, A)\})$ which has the labelling $\{\text{in}(A), \text{out}(B), \text{in}(C)\}$.

Definition 6 (claim) A *claim* is a pair of the form: $(\text{in}(A), _)$ or $(\text{out}(A), \text{in}(B))$ where A and B are different arguments. $(\text{in}(A), _)$ is read “ A is labelled in”, while $(\text{out}(A), \text{in}(B))$ is read “ A is labelled out because B is labelled in”.

A claim $(\text{in}(A), _)$ by a player P_1 is *refuted* by the claim $(\text{out}(A), \text{in}(B))$ by a player P_2 . A claim $(\text{out}(A), \text{in}(B))$ by a player P_1 (resp. P_2) is *refuted* by the claim $(\text{out}(B), \text{in}(C))$ by a player P_2 (resp. P_1).

Definition 7 (debate game) Let (Ar, att) be the UAF, 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 $[(\text{in}(X_0), _), (\text{out}(X_0), \text{in}(Y_1)), (\text{out}(Y_1), \text{in}(X_1)), \dots, (\text{out}(X_i), \text{in}(Y_{i+1})), (\text{out}(Y_{i+1}), \text{in}(X_{i+1})), \dots]$ 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 $\text{out}(Z_j)$ in a claim by a player P_1 (resp. P_2), there is $\text{in}(Z_i)$ ($i \leq j$) in a claim by a player P_2 (resp. P_1).
- for each $(\text{out}(U_i), \text{in}(V_j))$, 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 $\Delta_0 = (\text{in}(X_0), _)$. A debate game for an argument X_0 *terminates* with Δ_n if $[\Delta_0, \Delta_1, \dots, \Delta_n]$ is an admissible debate and there is no claim Δ_{n+1} such that $[\Delta_0, \Delta_1, \dots, \Delta_n, \Delta_{n+1}]$ 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 .³

The player P_1 starts a debate with the claim $\Delta_0 = (\text{in}(X_0), _)$ 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 $\Delta_1 = (\text{out}(X_0), \text{in}(Y_1))$ based on the revised argumentation framework AF_2^1 . A debate continues by iterating revisions and claims (Figure 1). Note that by the third item of Definition 7, a player can refute not only the preceding claim of the opponent player, but any previous claim of the opponent. 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.

³An old saying: “The one who has the last word laughs best” [6].

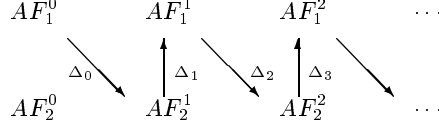
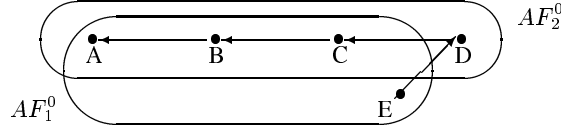


Figure 1. Debate game

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 $AF_2^0 = (\{A, B, C, D\}, \{(D, C), (C, B), (B, A)\})$. AF_1^0 and AF_2^0 have the complete labellings: $\{\text{in}(A), \text{out}(B), \text{in}(C), \text{in}(E)\}$ and $\{\text{out}(A), \text{in}(B), \text{out}(C), \text{in}(D)\}$, respectively. The argumentation graph of two players is described as follows.



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

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

Here, “ $AF_i^k: (\text{out}(X), \text{in}(Y))$ ” means that a player P_i makes a claim ($\text{out}(X), \text{in}(Y)$) 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.

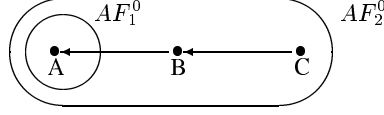
Note that each player revises his/her argumentation framework and labelling non-monotonically changes during a game (cf. Example 1). Then, a player may change his/her claim from $\text{in}(X)$ to $\text{out}(X)$ or from $\text{out}(X)$ to $\text{in}(X)$ for the same argument X in face of new evidences.

4. Dishonest Argumentation

“But even when a man has the right on his side, he needs Dialectic in order to defend and maintain it; he must know what the dishonest tricks are, in order to meet them; nay, he must often make use of them himself, so as to beat the enemy with his own weapons.” [13]

In a debate game, a player may provide dishonest arguments to win the game.

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 $AF_2^0 = (\{A, B, C\}, \{(C, B), (B, A)\})$. The argumentation graph of two players is described as follows.



Then AF_1^0 has the labelling $\{\text{in}(A)\}$, while AF_2^0 has the labelling $\{\text{in}(A), \text{out}(B), \text{in}(C)\}$. When Alex claims $\text{in}(A)$, 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 : $(\text{in}(A), _)$ “I claim that A is in.”
 AF_2^1 : $(\text{out}(A), \text{in}(B))$ “ A is out because B is in.”

Alex cannot refute Bob’s claim because Alex does not know any 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 $\text{in}(A)$ while believing $\text{out}(A)$ or $\text{undec}(A)$ in his/her labelling. On the other hand, a player may bring an argument which is not in his/her argumentation framework. 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 8 (honest/dishonest claim) Let (Ar, att) be the UAF and $AF = (Ar', att')$ an argumentation framework of a player. For any $A \in Ar'$,

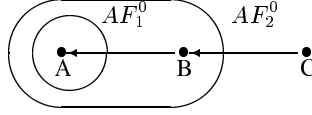
- a claim $(\text{in}(A), _)$ is *honest* wrt AF if $\mathcal{L}(A) = \text{in}$ for some complete labelling \mathcal{L} of AF ; a claim $(\text{out}(A), \text{in}(B))$ is *honest* wrt AF if $B \in Ar'$ and $\mathcal{L}(B) = \text{in}$ for some complete labelling \mathcal{L} of AF .
- a claim $(\text{in}(A), _)$ is a *lie* wrt AF if $\mathcal{L}(A) \neq \text{in}$ for any complete labelling \mathcal{L} of AF ; a claim $(\text{out}(A), \text{in}(B))$ is a *lie* wrt AF if $B \in Ar'$ and $\mathcal{L}(B) \neq \text{in}$ for any complete labelling \mathcal{L} of AF .
- a claim $(\text{in}(B), _)$ or $(\text{out}(A), \text{in}(B))$ 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 $\text{in}(X)$ (resp. $\text{out}(X)$) in a claim is also called *dishonest* if $\mathcal{L}(X) \neq \text{in}$ (resp. $\mathcal{L}(X) \neq \text{out}$) 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.⁴ To allow the existence of dishonest players who bullshit, we need to modify Definition 7 of debate games such that each player may claim an argument which is not in his/her argumentation framework. That is, any X_k or Y_k in a claim is not necessarily included in Ar_1^k or Ar_2^k but in Ar of the UAF (Ar, att) . 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, we consider a debate game in which each player often plays a game based on ${}^bAF_i^k$, instead of AF_i^k .

⁴A person who makes up some fact considers that it can happen in the world.

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 $UAF = (\{A, B, C\}, \{(C, B), (B, A)\})$. The argumentation graph is described as follows.



Then AF_1^0 has the labelling $\{\text{in}(A)\}$, while AF_2^0 has the labelling $\{\text{out}(A), \text{in}(B)\}$. Suppose the following debate game:

AF_1^0 : $(\text{in}(A), _)$ “I claim that A is in.”
 AF_2^0 : $(\text{out}(A), \text{in}(B))$ “ A is out because B is in.”
 ${}^bAF_1^0$: $(\text{out}(B), \text{in}(C))$ “ B is out because C is in.”

Bob cannot refute Alex’s claim and Alex wins the game. In this debate, Alex’s claim at ${}^bAF_1^0$ is bullshit because he has no belief on the truthfulness of C .

In Examples 3 and 4, a player cannot attack the dishonest argument provided by the opponent. We next consider in which circumstances a player can win a game.

Definition 9 (upstream of an argument [4]) Let $AF = (Ar, att)$ be an argumentation framework and $A \in Ar$. We define $upstream_{AF}(A)$ as the smallest set such that (i) $A \in upstream_{AF}(A)$, and (ii) if $X \in upstream_{AF}(A)$ and $(Y, X) \in att$, then $Y \in upstream_{AF}(A)$.

Definition 10 (winning position) Let (Ar, att) be the UAF, 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) $upstream_{AF_j}(X) \subset upstream_{AF_i}(X)$, and (ii) there is no $(Y, Z) \in att$ such that $Y, Z \in upstream_{UAF}(X)$, $Y \in Ar \setminus Ar_i$ and $Z \in Ar_i$.

Example 5 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.

Proof: If AF_i is in a winning position for A , there is an argument $X \in Ar_i$ such that $X \in upstream_{AF_i}(A) \setminus upstream_{AF_j}(A)$ by Definition 10(i). Then, AF_i can refute any claim made by the argumentation framework AF_j of the opponent. Moreover, by Definition 10(ii), any argument $Z \in upstream_{AF_i}(A)$ is attacked by no argument $Y \in Ar \setminus Ar_i$ in the UAF. Thus, AF_i can construct a claim that is not refuted by the opponent. Hence, P_i has a chance to win the game. \square

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 $X \in upstream_{AF}(A)$ is called *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.

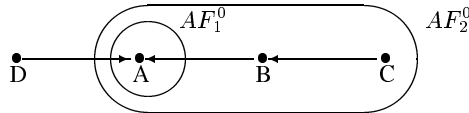
1. A honest player P_1 has a chance to win a debate game for the argument $A \in Ar_1$ if AF_1 is in a winning position and the length of any path from any terminal $X \in upstream_{AF_1}(A)$ to A is even.
2. A honest player P_2 has a chance to win a debate game for the argument $A \in Ar_1$ if AF_2 is in a winning position and the length of any path from some terminal $Y \in upstream_{AF_2}(A)$ to A is odd.

Proof: (1) By Proposition 1, P_1 has a chance to win a debate game for the argument $A \in Ar_1$ if he/she is in a winning position for A . Moreover, since the length of any path from any terminal $X \in upstream_{AF_1}(A)$ to A is even, AF_1 has a complete labelling such that $\mathcal{L}(A) = \text{in}$ and $\mathcal{L}(X) = \text{in}$.

(2) For a player P_2 , if the length of any path from some terminal $Y \in upstream_{AF_2}(A)$ to A is odd, AF_2 has a complete labelling such that $\mathcal{L}(A) = \text{out}$ and $\mathcal{L}(Y) = \text{in}$. \square

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 $\text{in}(C)$ 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 6 Consider $UAF = (\{A, B, C, D\}, \{(C, B), (B, A), (D, A)\})$, $AF_1^0 = (\{A\}, \emptyset)$ and $AF_2^0 = (\{A, B, C\}, \{(C, B), (B, A)\})$ with the argumentation graph:



First, the player P_1 claims $(\text{in}(A), _)$ based on AF_1^0 . To win the game, there are two possible claims for the player P_2 . One is the lie $(\text{out}(A), \text{in}(B))$ and the other is the bullshit $(\text{out}(A), \text{in}(D))$.

In Example 6, 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 the player P_2 knows the falsehood of $(\text{out}(A), \text{in}(B))$, while he/she does not know the truthfulness of $(\text{out}(A), \text{in}(D))$.⁵ The player P_2 does not know whether the player P_1 has information on C or D . If the player P_2 lies but the player P_1 has information on the argument C , then the player P_2 loses the game. Moreover, if the player P_1 knows that the player P_2 has information on C , then the player 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 on D . (In this case, the initial claim by the player P_1 is a lie.) With these reasons, we introduce a *best-practice strategy* for a player to win a debate game.

Definition 11 (best-practice strategy) Let (Ar, att) be the UAF, 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 $(\text{in}(A), _)$ be the initial claim on A by AF_1^0 .

1. If $\mathcal{L}(A) = \text{in}$ 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 $(\text{out}(B), \text{in}(C))$ 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 $\mathcal{L}(A) = \text{in}$, then P_1 makes bullshit $(\text{out}(B), \text{in}(C))$ at ${}^bAF_1^k$;
 - (b) otherwise, P_1 lies by $(\text{out}(B), \text{in}(D))$ for $B, D \in Ar_1^k$ at AF_1^k where $(\text{out}(B), \text{in}(D))$ refutes a claim by AF_2^k .
2. If $\mathcal{L}(A) = \text{out}$ 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 $(\text{out}(B), \text{in}(C))$ 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 $\mathcal{L}(A) = \text{out}$, then P_2 makes bullshit $(\text{out}(B), \text{in}(C))$ at ${}^bAF_2^k$;
 - (b) otherwise, P_2 lies by $(\text{out}(B), \text{in}(D))$ for $B, D \in Ar_2^k$ at AF_2^k where $(\text{out}(B), \text{in}(D))$ refutes a claim by AF_1^{k-1} .

In Example 6, AF_2^0 has the labelling $\{\text{in}(A), \text{out}(B), \text{in}(C)\}$. Then, the player P_2 claims the bullshit $(\text{out}(A), \text{in}(D))$ at ${}^bAF_2^1$ where $AF_2^1 \circ D$ has the labelling $\{\text{out}(A), \text{out}(B), \text{in}(C), \text{in}(D)\}$. Since the player P_1 cannot refute the claim at AF_1^1 , the player P_2 wins the game.

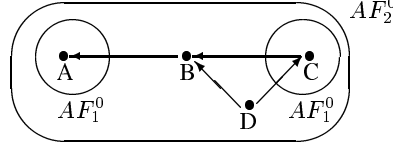
5. Detecting Dishonest Arguments

“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, . . .” [13, Stratagems XVI]

Generally, a player cannot detect dishonest arguments of the opponent player. In fact, Alex cannot detect Bob’s lying in Example 3, and Bob cannot detect Alex’s bullshit in Example 4. On the other hand, a player can detect dishonest arguments of the opponent if he/she finds contradiction in arguments provided by the opponent.

⁵The preference is stated as a postulate in [12] that *never lie if you can bullshit your way out of it*.

Example 7 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: $\{\text{in}(A), \text{in}(C)\}$ and $\{\text{in}(A), \text{out}(B), \text{out}(C), \text{in}(D)\}$, respectively. The argumentation graph is described as follows.



Suppose the following debate game:

- AF_1^0 : $(\text{in}(A), _)$ “I claim that A is in.”
 AF_2^0 : $(\text{out}(A), \text{in}(B))$ “ A is out because B is in.”
 AF_1^0 : $(\text{out}(B), \text{in}(C))$ “ B is out because C is in.”
 AF_2^0 : $(\text{out}(C), \text{in}(D))$ “ C is out because D is in.”

After AF_2^0 , the player P_1 would reason that

1. The player P_2 claimed $\text{in}(B)$ at AF_2^0 . Since I (i.e., the player P_1) provide no argument wrt B , P_2 should have $\text{in}(B)$ in a labelling of AF_2^0 .
2. The player P_2 also claimed $\text{in}(D)$ at AF_2^0 . Since I provide no argument wrt D , P_2 should have $\text{in}(D)$ in the same labelling of AF_2^0 .
3. 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 .
4. To have $\text{in}(B)$ in a labelling, however, any attacker of B must be out. This contradicts the claim $\text{in}(D)$.

By this reasoning, the player P_1 concludes that the opponent P_2 has provided dishonest arguments in the game. In fact, AF_2^0 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 $\text{in}(X)$ and later claims $\text{in}(Y)$ for some attacker Y of X . In Example 7, however, the player P_1 conjectures that both $\text{in}(B)$ and $\text{in}(D)$ 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 $\text{in}(X)$ and $\text{in}(Y)$ 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 $\text{in}(Z)$ 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.

1. P_1 at AF_1^k ($k \geq 2$) can conclude that P_2 has provided dishonest arguments in the game if there are two labelled arguments $\text{in}(Y_i)$ and $\text{in}(Y_j)$ ($0 < i < j \leq k$) such that there is an odd-length path from Y_i to Y_j or from Y_j to Y_i in the argumentation graph associated with AF_1^k in which every argument on the path is originated by P_2 .

2. P_2 at AF_2^k ($k \geq 2$) can conclude that P_1 has provided dishonest arguments in the game if there are two labelled arguments $\text{in}(X_i)$ and $\text{in}(X_j)$ ($0 \leq i < j < k$) such that there is an odd-length path from X_i to X_j or from X_j to X_i in the argumentation graph associated with AF_2^k in which every argument on the path is originated by P_1 .

Proof: (1) The player P_1 observes at AF_1^k that two labelled arguments $\text{in}(Y_i)$ and $\text{in}(Y_j)$ are claimed by the player P_2 at AF_2^i and AF_2^j , respectively. In addition, every argument appearing in the odd-length path from Y_i to Y_j or from Y_j to Y_i is not brought about by AF_1^l ($0 \leq l \leq k-1$). P_1 at AF_1^k then infers that P_2 at AF_2^0 must know that Y_i (resp. Y_j) directly/indirectly attacks Y_j (resp. Y_i). Thus, AF_2^l cannot have both $\text{in}(Y_i)$ and $\text{in}(Y_j)$ in any of its complete labelling. Hence, AF_1^k concludes that AF_2 has provided dishonest arguments at some AF_2^l ($1 \leq l \leq k$). (2) is shown in a similar way. \square

6. Discussion

As argued by Schopenhauer [13], 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] provides a comparative study between lies, bullshit, and deception, and argues how these different categories of dishonesties are effectively used in formal argumentation. In Example 3, Bob lies by claiming that A should be rejected. On the other hand, Caminada considers that Alex's claim that A should be accepted is more bullshit than Bob's claim that A should be accepted, since Alex's claim lacks proper knowledge about it. Caminada discusses conceptual issue on dishonesty, and does not develop a formal theory of dishonest argumentation as done in this paper. Rahwan *et al.* [10] introduce a formal argumentation theory in which an agent may hide or lie about arguments. Each agent has its own arguments (called a *type*), and is interested in the acceptance of a particular argument (called a *focal argument*). An agent maximizes utility if its focal argument is accepted. To maximize its utility, an agent may hide some of its arguments or may lie by making up new arguments that it does not have in its argument set. They characterize the strategy-proofness (i.e., truth-telling being a dominant strategy equilibrium) in formal argumentation systems when agents may hide and/or lie about arguments. Comparing their work with ours, there are some important differences. First, 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. By contrast, we study conditions under which dishonest arguments are effective in a debate game under the complete semantics. We develop a best-practice strategy for a player to win a game. Second, they consider multiple agents who simultaneously reveal their arguments. The situation is different from a debate game where two agents exchange arguments and revise their own argumentation frameworks during a game. Third, their notion of hiding arguments is a special case of our definition of lying. For instance, in Example 3, Bob lies by hiding the argument C . In Example 7, on the other hand, the player AF_2 is lying without hiding any argument. By contrast, their notion of lying corresponds to our definition of bullshit. According to Frankfurt, bullshit is a statement that "is grounded neither in a belief that it is true nor, as a lie must be, in a belief that it is not true" [7, p.33]. Then, the name

bullshit appears to be more appropriate for making up new arguments whose truthfulness are unknown to a player. Fourth, we investigate the possibility of detecting dishonest arguments in a debate game. The issue is not pursued in their paper. They argue the issue under the preferred semantics [8], while those differences presented above still exist.

A formal argumentation framework has been used for modelling dialogue games [9] or discussion games [3]. Prakken [9] develops an abstract framework of argumentation-based dialog systems in which players dialogue using a particular communication language under a specified protocol. During a game, a player may make a commitment that is not believed by the player. However, the study does not consider dishonest arguments as a strategy for winning a game nor investigate possibilities of detecting dishonest arguments in a game. Caminada [3] formulates a Socratic form of discussion, in which a proponent tries to avoid being lead to a contradiction by the opponent. The proponent of an argument wins the game iff he/she can survive the critical questions of the opponent without running into inconsistencies. In this study, it is assumed that two players share the same argumentation framework, so players do not revise the argumentation framework during a game. Moreover, a player does not use dishonest arguments as a tactic for winning a game. Belief revision of argumentation systems is studied in [5,11], but it is not used in dialogue games.

Acknowledgements

We thank Martin Caminada for his useful input and suggestions.

References

- [1] M. Caminada and D. Gabbay. A logical account of formal argumentation. *Studia Logica*, 93:109–145, 2009.
- [2] M. Caminada. Truth, lies and bullshit, distinguishing classes of dishonesty. In: *Proc. IJCAI Workshop on Social Simulation*, 2009.
- [3] M. Caminada. Preferred semantics as Socratic discussion. In: *Proc. 1st Workshop on Argumentation in Artificial Intelligence and Philosophy*, 2010.
- [4] M. Caminada. On the issue of argumentation and informedness. Research Report, University of Luxembourg, 2011.
- [5] C. Cayrol, F. D. de Saint Cyr, M. C. Lagasquie-Schiex. Revision of an argumentation system. In: *Proc. KR 2008*, pages 124–134, 2008.
- [6] P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n -person games. *Artificial Intelligence*, 77:321–357, 1995.
- [7] H. G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.
- [8] S. Pan, K. Larson and I. Rahwan. Argumentation mechanism design for preferred semantics. In: *Proc. 3rd International Conference on Computational Models of Argument*, pages 403–414, 2010.
- [9] H. Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of Logic and Computation*, 15(6):1009–1040, 2005.
- [10] I. Rahwan, K. Larson and F. Tohmé. A characterisation of strategy-proofness for grounded argumentation semantics. In: *Proc. IJCAI-09*, pages 251–256, 2009.
- [11] N. D. Rotstein, M. O. Moguillansky, M. A. Falappa, A. J. García and G. R. Simari. Argument theory change: revision upon warrant. In: *Proc. 2nd International Conference on Computational Models of Argument*, pages 336–347, 2008.
- [12] C. Sakama, M. Caminada and A. Herzig. A logical account of lying. In: *Proc. 12th European Conference on Logics in Artificial Intelligence, Lecture Notes in AI*, vol. 6341, pages 286–299, Springer, 2010.
- [13] A. Schopenhauer. *The Art of Controversy*. Originally published in 1896 and is translated by T. B. Saunders, Cosimo Classics, New York, 2007.