Abduction in Argumentation Frameworks and its Use in Debate Games

Chiaki Sakama

Department of Computer and Communication Sciences
Wakayama University, Sakaedani, Wakayama 640-8510, Japan
sakama@sys.wakayama-u.ac.jp

Abstract. This paper studies an *abduction* problem in formal argumentation frameworks. Given an argument, an agent verifies whether the argument is justified or not in its argumentation framework. If the argument is not justified, the agent seeks conditions to explain the argument in its argumentation framework. We formulate such abductive reasoning in argumentation semantics and provide its computation in logic programming. Next we apply abduction in argumentation frameworks to reasoning by players in *debate games*. In debate games, two players have their own argumentation frameworks and each player builds claims to refute the opponent. A player may provide false or inaccurate arguments as a tactic to win the game. We show that abduction is used not only for seeking counter-claims but also for building dishonest claims in debate games.

1 Introduction

Arguments and *explanations* play different roles in human reasoning and have been distinguished in philosophy of science. According to [17], "the purpose of an explanation is to show *why and how* some phenomenon occurred or some event happened; the purpose of an argument is to show *that* some view or statement is correct or true." In other words, "argument is the mechanism by which we produce knowledge" and "explanation is the mechanism by which we produce understanding" [22]. On the other hand, an argument is used for knowing whether an explanation is appropriate and an explanation is used for understanding how an evidence occurs in an argument. In this sense, arguments and explanations are mutually supportive, so "arguments and explanations have a complementary relationship and reasoning is normally perceived as incomplete when one occurs in the absence of the other" [22]. In the field of artificial intelligence, argumentation and *abduction* are implicitly related in [13] where Dung provides an argumentation-theoretic semantics of abductive logic programs. The framework has been later extended to *assumption-based argumentation* [9]. Dung also introduces *formal argumentation* [14] as an abstract framework for argumentative reasoning, and the framework has been extended in various ways to incorporate explanatory reasoning [5, 20, 28, 29].

This paper studies an abductive framework based on Dung's abstract argumentation. Different from previous studies, we combine an argumentation framework and *extended abduction* proposed by Inoue and Sakama [18]. In extended abduction, hypotheses can not only be added to background knowledge but also be removed from

it to explain (or unexplain) an observation. In the context of argumentation, extended abduction is used for verifying whether a particular argument is justified or not, and seeking conditions to explain a particular argument in an argumentation framework. We next apply the abductive framework to reasoning by players in *debate games* [26]. A debate game provides an abstract model of dialogue between two players based on a formal argumentation framework. A unique feature of debate games is that a player may claim false or inaccurate arguments as a tactic to win the game. The proposed framework combines abduction and argumentation in a way different from existing studies, and exploits a new application of abduction in a formal dialogue system based on argumentation frameworks.

The rest of this paper is organized as follows. Section 2 reviews abstract argumentation frameworks. Section 3 introduces abduction to argumentation frameworks, and Section 4 applies the framework to debate games. Section 5 discusses related issues and Section 6 concludes the paper.

2 Argumentation Framework

Definition 2.1 (argumentation framework). [10, 14] Let U be the universe of all possible arguments. An *argumentation framework* (AF) is a pair (Ar, att) where Ar is a finite subset of U and $att \subseteq Ar \times Ar$. An argument A *attacks* an argument B iff $(A, B) \in att$. A set $S \subseteq Ar$ is *conflict-free* if there is no $A, B \in S$ such that $(A, B) \in att$. A set $S \subseteq Ar$ is *admissible* iff it is conflict-free and for any $A \in S$ such that $(B, A) \in att$ for some $B \in Ar$, there is $C \in S$ such that $(C, B) \in att$.

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 $(A, B) \in att$. An argumentation framework is identified with its argumentation graph.

Definition 2.2 (labelling). [10] Let $AF = (Ar, att)$ be an argumentation framework. A *labelling* of AF 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 $\mathtt{undec}(A)$). In this case, the argument A is *accepted* (resp. *rejected* or *undecided*). We call $\mathtt{in}(A)$, $\mathtt{out}(A)$ and $\mathtt{undec}(A)$ *labelled arguments*.

Definition 2.3 (complete labelling). [10] Let $AF = (Ar, att)$ be an argumentation framework. A labelling $\mathcal{L}$ of AF is a *complete labelling* 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 $\mathcal{L}(A) \neq \mathtt{out}$.

Let $\mathtt{in}(\mathcal{L}) = \{A \mid \mathcal{L}(A) = \mathtt{in}\}$, $\mathtt{out}(\mathcal{L}) = \{A \mid \mathcal{L}(A) = \mathtt{out}\}$ and $\mathtt{undec}(\mathcal{L}) = \{A \mid \mathcal{L}(A) = \mathtt{undec}\}$.

Definition 2.4 (stable, semi-stable, grounded, preferred labelling). [10] Let AF be an argumentation framework and $\mathcal{L}$ a complete labelling of AF. Then, (1) $\mathcal{L}$ is a *stable labelling* iff $\texttt{undec}(\mathcal{L}) = \emptyset$. (2) $\mathcal{L}$ is a *semi-stable labelling* iff $\texttt{undec}(\mathcal{L})$ is minimal wrt set inclusion among all complete labellings of AF. (3) $\mathcal{L}$ is a *grounded labelling* iff $\texttt{in}(\mathcal{L})$ is minimal wrt set inclusion among all complete labellings of AF. (4) $\mathcal{L}$ is a *preferred labelling* iff $\texttt{in}(\mathcal{L})$ is maximal wrt set inclusion among all complete labellings of AF.

There is a one-to-one correspondence between the set $\texttt{in}(\mathcal{L})$ with a complete (resp. stable, semi-stable, grounded, preferred) labelling $\mathcal{L}$ of an argumentation framework AF and a *complete* (resp. *stable*, *semi-stable*, *grounded*, *preferred*) *extension* of AF [10, 14]. In this paper, the distinction between different labellings is often unimportant and $\mathcal{S}$*-labelling* means one of the five labellings introduced above.

Definition 2.5 (justify). [2] Let AF be an argumentation framework. Then, a labelled argument L is *skeptically* (resp. *credulously*) *justified* by AF under the $\mathcal{S}$-labelling if L is included in every (resp. some) $\mathcal{S}$-labelling $\mathcal{L}$ of AF.

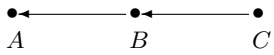
3 Abduction in Argumentation Framework

3.1 Explanations

Suppose the following dialogue between Alice and Bob:

Alice: "I think Mary can speak Japanese because she has stayed in Japan."
Bob: "I don't think so because her staying in Japan was too short to learn Japanese."

The situation is represented by the argumentation framework $AF = (\{A, B\}, \{(B, A)\})$ where A represents the argument "Mary speaks Japanese" by Alice and B represents the argument "Mary does not speak Japanese" by Bob. The AF has the complete labelling $\{\, \texttt{out}(A),\ \texttt{in}(B)\,\}$ which means that the argument A is rejected and the argument B is accepted. In another day, Bob observes that Mary speaks Japanese. To explain this, he assumes an argument C that Mary studied Japanese hard to be able to speak it well. The revised argumentation becomes $AF' = (\{A, B, C\}, \{(C, B),\ (B, A)\})$ and is represented by the argumentation graph below.



After introducing the new argument C, the situation changes: the revised AF' has the complete labelling $\{\, \texttt{in}(A),\ \texttt{out}(B),\ \texttt{in}(C)\,\}$, where A and C are now accepted and B is rejected. It illustrates the situation in which a new argument is introduced to explain a new observation. Suppose another dialogue such that

Alice: "I think the new iPhone will be selling well."
Bob: "I don't think so because few people will get interested in this new model."

The situation is represented by $AF = (\{A, B\}, \{(B, A)\})$ where A is rejected and B is accepted. Later it is observed that the new iPhone breaks the sales record. Bob then withdraws his argument B and the revised AF becomes $AF' = (\{A\}, \emptyset)$. Then, the argument A is now accepted in AF'. It illustrates the situation in which a previously believed argument is removed in face of a new observation.

To realize such explanatory reasoning in argumentation frameworks, it is necessary to introduce assumptions to an argumentation framework. In Definition 2.1, the set Ar of arguments is a subset of the universe U of all possible arguments. We then consider the notion of the universal argumentation framework which consists of the set of all possible arguments and attack relations over them.

Definition 3.1 (universal AF). The *universal argumentation framework* (UAF) is an argumentation framework (U, att_U) in which U is the set of all possible arguments and $att_U \subseteq U \times U$ is the set of fixed attack relations over U.

The UAF specifies a world which consists of arguments and attack relations over them. An *agent* has (partial) knowledge about the world as an argumentation framework $AF = (Ar, att)$ where $Ar \subseteq U$ is *finite* and $att = att_U \cap (Ar \times Ar)$. In this sense, AF is often called a *subargumentation framework* (*sub-AF* for short) of the UAF. The agent has a belief on the labelling of every argument in Ar based on the attack relations in att under the designated semantics $\mathcal{S}$. On the other hand, an agent can recognize the possibility of arguments in $U \setminus Ar$, but does not know whether those arguments are valid or not. The agent has no information on labelling of any argument in $U \setminus Ar$ and each argument in $U \setminus Ar$ is called a *hypothesis*. In what follows, an agent is identified with its AF.

Definition 3.2 (observation). Let $UAF = (U, att_U)$ and $AF = (Ar, att)$ a sub-AF. An *observation* O by AF is either $\mathtt{in}(A)$ or $\mathtt{out}(A)$ for some $A \in U$ such that $(A, A) \notin att_U$. When $O = \mathtt{in}(A)$ or $O = \mathtt{out}(A)$, define $arg(O) = A$.

When $O = \mathtt{in}(A)$ is observed, it means that there is an evidence for A. When $O = \mathtt{out}(A)$ is observed, on the other hand, it means that there is an evidence against A. In each case, an agent tries to skeptically or credulously justify O in his/her AF under a designated labelling. We consider that any meaningful observation contains no self-attacking argument, which is represented by the condition $(A, A) \notin att_U$.[1] If an agent fails to justify O in his/her AF, it implies that AF believed by the agent is inaccurate or incomplete. In this case, the agent performs *abduction* to explain O.

Definition 3.3 (explanation). Let $UAF = (U, att_U)$ and $AF = (Ar, att)$ a sub-AF. An observation O (by AF) is *skeptically* (resp. *credulously*) *explained* by $E = (I, J)$ under the $\mathcal{S}$-labelling of AF_E if O is included in every (resp. some) $\mathcal{S}$-labelling $\mathcal{L}_E$ of the argumentation framework $AF_E = (Ar_E, att_E)$ where $Ar_E = (Ar \setminus J) \cup I$, $I \subseteq U \setminus Ar$, $J \subseteq Ar$, and $att_E = att_U \cap (Ar_E \times Ar_E)$. In this case, E is called a

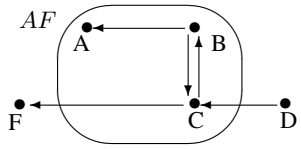
[1] A reviewer comments that "an argument A attacking itself is a very natural explanation for the observation that there is evidence against A, i.e. that A is out." However, A's attacking itself does not explain that "A is out" but explains that "A is *not* in." In fact, A is labelled $\mathtt{undec}$ in $AF = (\{A\}, \{(A, A)\})$ under the complete, semi-stable, grounded and preferred semantics. We exclude such "undecided" observations. (AF has no stable labelling.)

skeptical (resp. *credulous*) *explanation* of O (under the $\mathcal{S}$-labelling of AF_E), and we say that O has a skeptical (resp. credulous) explanation E in AF.

An explanation (I, J) of an observation O is *minimal* if $I' \subseteq I$ and $J' \subseteq J$ imply $I' = I$ and $J' = J$ for any explanation (I', J') of O. An explanation (I, J) is *empty* if $I = J = \emptyset$; otherwise, (I, J) is *non-empty*.

If E is a skeptical explanation of an observation O, then E is also a credulous explanation of O, but not vice versa. The notions of skeptical and credulous explanations coincide when AF_E has the unique $\mathcal{S}$-labelling. A skeptical/credulous explanation is simply called an *explanation* if the distinction between the two is unimportant in the context. In Definition 3.3, if $O = \mathtt{in}(A)$ (resp. $O = \mathtt{out}(A)$) for some argument A, the goal of abduction is to produce a labelling of AF in which A is labelled `in` (resp. `out`). To this end, arguments in J are removed from Ar and hypotheses in I are introduced to Ar to explain O. Removal of J means that an agent does not believe arguments in J anymore, or an agent has some reason to withdraw J. Introduction of I means that an agent learns new arguments in I. When O is observed by AF, it is either $arg(O) \in Ar$ or $arg(O) \in U \setminus Ar$. In case of $arg(O) \in Ar$, the argument $arg(O)$ is known by AF but its labelling in O may be different from the labelling of the argument in AF. In case of $arg(O) \in U \setminus Ar$, on the other hand, the argument $arg(O)$ is a hypothesis for AF and AF has no labelling of the argument.

Example 3.1. Let $UAF = (\{A, B, C, D, F\}, \{(B, A), (B, C), (C, B), (D, C), (C, F)\})$ and $AF = (\{A, B, C\}, \{(B, A), (B, C), (C, B)\})$ where AF has three complete labellings: $\mathcal{L}_1 = \{\, \mathtt{in}(A), \mathtt{out}(B), \mathtt{in}(C) \,\}$, $\mathcal{L}_2 = \{\, \mathtt{out}(A), \mathtt{in}(B), \mathtt{out}(C) \,\}$, and $\mathcal{L}_3 = \{\, \mathtt{undec}(A), \mathtt{undec}(B), \mathtt{undec}(C) \,\}$.



Then the following facts hold.

- Two observations $O_1 = \mathtt{in}(A)$ and $O_2 = \mathtt{out}(A)$ have the single minimal credulous explanation $E_0 = (\emptyset, \emptyset)$ under the complete labelling of $AF_{E_0} = AF$.
- $O_2 = \mathtt{out}(A)$ has two minimal skeptical explanations $E_1 = (\emptyset, \{C\})$ under the complete labelling of $AF_{E_1} = (\{A, B\}, \{(B, A)\})$, and $E_2 = (\{D\}, \emptyset)$ under the complete labelling of $AF_{E_2} = (\{A, B, C, D\}, \{(B, A), (B, C), (C, B), (D, C)\})$.
- $O_3 = \mathtt{in}(F)$ has two minimal skeptical explanations: $E_3 = (\{F\}, \{C\})$ under the complete labelling of $AF_{E_3} = (\{A, B, F\}, \{(B, A)\})$, and $E_4 = (\{D, F\}, \emptyset)$ under the complete labelling of $AF_{E_4} = UAF$.
- $O_4 = \mathtt{out}(D)$ has no credulous/skeptical explanation.

In Example 3.1, the observation $O_1 = \mathtt{in}(A)$ has the credulous empty explanation in AF. This means that the labelled argument $\mathtt{in}(A)$ is credulously justified in the argumentation framework AF under the complete labelling. On the other hand, $O_2 = \mathtt{out}(A)$ has two minimal skeptical explanations in AF and both of them are non-empty explanations. This means that the labelled argument $\mathtt{out}(A)$ is not skeptically justified in AF under the complete labelling. To skeptically justify $\mathtt{out}(A)$, it is necessary to remove the argument C from Ar or to introduce the hypothesis D to Ar in AF.

By Definitions 2.5 and 3.3, an observation O is skeptically (resp. credulously) justified by AF_E under the $\mathcal{S}$-labelling iff O has a skeptical (resp. credulous) explanation E

under the $\mathcal{S}$-labelling of AF_E. In particular, O is skeptically (resp. credulously) justified by AF under the $\mathcal{S}$-labelling iff O has the skeptical (resp. credulous) empty explanation under the $\mathcal{S}$-labelling of AF. An observation may have none, one, or multiple explanations in general. In particular, the next proposition holds.

Proposition 3.1 *Let $UAF = (U, att_U)$ and AF a sub-AF. For any $A \in U$,*

1. *an observation $O = \mathtt{in}(A)$ has a skeptical/credulous explanation in AF.*
2. *an observation $O = \mathtt{out}(A)$ has a credulous explanation in AF under the complete, (semi-)stable, preferred labelling iff there is an argument $B \in U$ such that $(B, A) \in att_U$ and $(B, B) \notin att_U$. Moreover, O has a skeptical explanation in AF under $\mathcal{S}$-labelling iff the additional condition $(A, B) \notin att_U$ is satisfied.*

Proof. (1) $\mathtt{in}(A)$ is included in every $\mathcal{S}$-labelling of the argumentation framework $AF_E = (\{A\}, \emptyset)$. Thus, O has the skeptical/credulous explanation $E = (\{A\}, Ar)$ in case of $A \notin Ar$; and $E' = (\emptyset, Ar \setminus \{A\})$ in case of $A \in Ar$. (2) If there is $B \in U$ such that $(B, A) \in att_U$ and $(B, B) \notin att_U$, then $\mathtt{out}(A)$ is included in some complete labelling of the argumentation framework $AF_E = (\{A, B\}, \{(B, A)\})$ in case of $(A, B) \notin att_U$; or $AF_E = (\{A, B\}, \{(B, A), (A, B)\})$ in case of $(A, B) \in att_U$. Thus, O has a credulous explanation $E = (\{A, B\}, Ar)$ under the complete, (semi-)stable, preferred labelling. In case of $(A, B) \notin att_U$, E is also a skeptical explanation under $\mathcal{S}$-labelling. The only-if part follows by definition. □

When an observation O does not have the empty skeptical/credulous explanation, O is not skeptically/credulously justified by AF under the $\mathcal{S}$-labelling. In this case, a non-empty explanation E is likely to change not only the labelling of the argument $arg(O)$ but the labellings of arguments other than $arg(O)$ in AF_E. In Example 3.1, for instance, the complete labelling $\mathcal{L}_1 = \{\, \mathtt{in}(A), \mathtt{out}(B), \mathtt{in}(C) \,\}$ of AF changes into $\{\, \mathtt{out}(A), \mathtt{in}(B) \,\}$ of AF_{E_1}. Thus, the explanation E_1 changes not only the labelling of A but also the labellings of B and C. The change of labellings between two argumentation frameworks is defined as follows.

Definition 3.4 (minimal change). Let $AF = (Ar, att)$ be an argumentation framework and $\mathcal{L}$ any $\mathcal{S}$-labelling of it. For any $\mathcal{S}$-labelling $\mathcal{L}_E$ of AF_E, define

$$\Delta(\mathcal{L}, \mathcal{L}_E) = \{\, A \mid \mathcal{L}(A) \neq \mathcal{L}_E(A) \text{ for } A \in Ar \,\} \cup \{\, A \mid A \in (Ar \backslash Ar_E) \cup (Ar_E \backslash Ar) \,\}.$$

A skeptical (resp. credulous) explanation E of an observation O *minimally changes* AF if for any skeptical (resp. credulous) explanation F of O in AF, the following condition is satisfied: for any $\mathcal{S}$-labelling $\mathcal{L}_F$ of AF_F which includes O, there is an $\mathcal{S}$-labelling $\mathcal{L}_E$ of AF_E which includes O such that $\Delta(\mathcal{L}, \mathcal{L}_F) \subseteq \Delta(\mathcal{L}, \mathcal{L}_E)$ implies $\Delta(\mathcal{L}, \mathcal{L}_E) \subseteq \Delta(\mathcal{L}, \mathcal{L}_F)$ for some $\mathcal{S}$-labelling $\mathcal{L}$ of AF.

If O has the empty explanation E in AF, then E minimally changes AF.

Example 3.2. In Example 3.1, the skeptical explanation $E_1 = (\emptyset, \{C\})$ of O_2 produces the complete labelling $\mathcal{L}_{E_1} = \{\, \mathtt{out}(A), \mathtt{in}(B) \,\}$, and the skeptical explanation $E_2 = (\{D\}, \emptyset)$ of O_2 produces the complete labelling $\mathcal{L}_{E_2} = \{\, \mathtt{out}(A), \mathtt{in}(B), \mathtt{out}(C), \mathtt{in}(D) \,\}$. Then, $\Delta(\mathcal{L}_1, \mathcal{L}_{E_1}) = \{A, B, C\}$ and $\Delta(\mathcal{L}_1, \mathcal{L}_{E_2}) = \{A, B, C, D\}$, so that E_1 minimally changes AF.

When an observation has more than one explanations, explanations that minimally change the labellings of arguments in AF are preferred.

Definition 3.5 (preferred explanation). Given an argumentation framework AF and an observation O, an explanation E is a *preferred explanation* of O if E minimally changes AF. A preferred explanation E is *most preferred* if it is also minimal (in the sense of Definition 3.3) among all of the preferred explanations of O.

Definition 3.5 says that there are two conditions for selecting the best explanations. The first condition requests that such explanations minimally change the labellings of the original AF. The second condition requests that the minimality of explanations. The first condition precedes the second one, that is, non-minimal preferred explanations are considered better than minimal non-preferred explanations. In particular, the empty explanation is always most preferred. By definition, we have the next result.

Proposition 3.2 *If an observation O has an explanation in an argumentation framework AF, then there is a most preferred explanation of O in AF.*

3.2 Computation

Next we provide a method of computing abduction in AF using *logic programming*. A *normal logic program* (or simply a *program*) is a set of rules of the form

$$A \leftarrow B_1, \ldots, B_m, \mathit{not}\, B_{m+1}, \ldots, \mathit{not}\, B_n$$

where A and B_i's are ground atoms ($n \geq m \geq 0$), and not represents the *negation as failure* operator. Let $\mathcal{B}_P$ be the Herbrand base of a program P. Then, a *3-valued interpretation* of a program P is defined as a pair $I = \langle T, F \rangle$ where T contains all ground atoms *true* in I, F contains all ground atoms *false* in I, and the remaining atoms in $W = \mathcal{B}_P \setminus (T \cup F)$ are *unknown*. Let $I(A) = 1$ (resp. $I(A) = \frac{1}{2}$, $I(A) = 0$) if $A \in T$ (resp. $A \in W$, $A \in F$), and $I(\mathit{not}\, A) = 1 - I(A)$. Then, a 3-valued interpretation I is a *model* of a program P if $I(A) \geq min\{I(L_i) \mid 1 \leq i \leq n\}$ holds for every rule $A \leftarrow L_1, \ldots, L_n$ in P where L_i is either B_i or $\mathit{not}\, B_i$. Among models of a program, the following models are important: *partial stable models*, *stable models*, *L-stable models*, *regular models*, and *well-founded models*. [2]

An argumentation framework $AF = (Ar, att)$ is transformed into the logic program P_{AF} by identifying each argument with a ground atom as follows [30]: $P_{AF} = \{\, A \leftarrow \mathit{not}\, B_1, \ldots, \mathit{not}\, B_n \mid A, B_1, \ldots, B_n \in Ar\, (n \geq 0)$ and $(B_i, A) \in att\ (1 \leq i \leq n) \,\}$. Then, there is a one-to-one correspondence between complete (resp. stable, semi-stable, grounded, preferred) labellings of AF and partial stable (resp. stable, L-stable, well-founded, regular) models of P_{AF} [11, 30]. We modify the transformation to characterize abduction in argumentation frameworks.

Definition 3.6 (transformation). Given $UAF = (U, att_U)$, the associated logic program P_{UAF} is defined as follows.

$$P_{UAF} = \{\, A \leftarrow \mathit{not}\, B_1, \ldots, \mathit{not}\, B_n, N_A \mid A, B_1, \ldots, B_n \in U\, (n \geq 0) \text{ and } (B_i, A) \in att_U\ (1 \leq i \leq n) \,\} \cup \{\, N_A \leftarrow \mathit{not}\, N'_A, \quad N'_A \leftarrow \mathit{not}\, N_A \mid A \in U \,\}$$

where N_A and N'_A are new ground atoms uniquely associated with each atom A.

[2] We refer the readers to the references in [11] for the precise definition of each semantics.

Each atom N_A or N'_A has one of the truth values *true*, *false* or *unknown*. If N_A is *true* (resp. *false*) in a partial stable model M of P_{UAF}, N'_A is *false* (resp. *true*) in M. Otherwise, both N_A and N'_A are *unknown* in M. If N_A is *true*, the rule $A \leftarrow not\, B_1, \ldots, not\, B_n, N_A$ is identified with $A \leftarrow not\, B_1, \ldots, not\, B_n$. In other words, by switching the truth values of N_A and N'_A, we can simulate introduction/removal of arguments $A, B_1, \ldots, B_n$ and attack relations (B_i, A) to/from a sub-AF of the UAF. For convenience, define $choice(U) = \{\, N_A \leftarrow not\, N'_A, \quad N'_A \leftarrow not\, N_A \mid A \in U\}$.

Example 3.3. Consider $UAF = (\{A, B, C\}, \{(C, B), (B, A)\})$ and $AF = (\{A, B\}, \{(B, A)\})$. Then, $P_{UAF} = \{A \leftarrow not\, B,\ N_A, \quad B \leftarrow not\, C,\ N_B, \quad C \leftarrow N_C\} \cup choice(\{A, B, C\})$ where the partial stable model $\langle\, \{B, N_A, N_B, N'_C\}, \{A, C, N'_A, N'_B, N_C\}\,\rangle$ corresponds to the complete labelling $\{\mathtt{out}(A),\ \mathtt{in}(B)\}$ of AF. On the other hand, the partial stable model $\langle\, \{A, C, N_A, N_B, N_C\}, \{B, N'_A, N'_B, N'_C\}\,\rangle$ corresponds to the complete labelling $\{\mathtt{in}(A),\ \mathtt{out}(B),\ \mathtt{in}(C)\}$ of $AF_E = UAF$ with $E = (\{C\}, \emptyset)$, and the partial stable model $\langle\, \{A, N_A, N'_B, N'_C\}, \{B, C, N'_A, N_B, N_C\}\,\rangle$ corresponds to the complete labelling $\{\mathtt{in}(A)\}$ of $AF_{E'} = (\{A\}, \emptyset)$ with $E' = (\emptyset, \{B\})$.

Lemma 3.3 [30] *Let $AF = (Ar, att)$ and P_{AF} its transformed logic program. If AF has a complete labelling $\mathcal{L}$, then $\langle\, T, F\,\rangle = \langle\, \mathtt{in}(\mathcal{L}), \mathtt{out}(\mathcal{L})\,\rangle$ where $\mathcal{B}_{P_{AF}} \setminus (T \cup F) = \mathtt{undec}(\mathcal{L})$ is a partial stable model of P_{AF}. Conversely, if $\langle\, T, F\,\rangle$ is a partial stable model of P_{AF}, then a labelling $\mathcal{L}$ such that $\mathtt{in}(\mathcal{L}) = T$, $\mathtt{out}(\mathcal{L}) = F$ and $\mathtt{undec}(\mathcal{L}) = \mathcal{B}_{P_{AF}} \setminus (T \cup F)$ is a complete labelling of AF.*

For a set S of atoms, let $\mathcal{N}_S = \{\, N_A \mid A \in S\,\}$; in particular, $\mathcal{N}_S = \emptyset$ if $S = \emptyset$.

Theorem 3.4. *Let $UAF = (U, att_U)$ and $AF = (Ar, att)$ a sub-AF. Also let $IN = \{N_A \mid A \in U \setminus Ar\}$ and $OUT = \{N_A \mid A \in Ar\}$. Then, an observation $O = \mathtt{in}(A)$ (resp. $O = \mathtt{out}(A)$) has a credulous explanation $E = (I, J)$ under a complete (or stable, semi-stable, grounded, preferred) labelling of AF_E iff P_{UAF} has a partial stable (or stable, L-stable, well-founded, regular) model $\langle\, T, F\,\rangle$ such that $A \in T$ (resp. $A \in F$), $\mathcal{N}_I = T \cap IN$ and $\mathcal{N}_J = F \cap OUT$. In particular, E is also a skeptical explanation of O iff $A \in T$ (resp. $A \in F$) for any $\langle\, T, F\,\rangle$ such that $\mathcal{N}_I = T \cap IN$ and $\mathcal{N}_J = F \cap OUT$.*

Proof. We show the result for complete labelling. If $O = \mathtt{in}(A)$ has a credulous explanation $E = (I, J)$ under a complete labelling of AF_E, then O is included in some complete labelling $\mathcal{L}_E$ of $AF_E = (Ar_E, att_E)$ where $Ar_E = (Ar \setminus J) \cup I$ with $I \subseteq U \setminus Ar$ and $J \subseteq Ar$. By Lemma 3.3, $\langle\, T, F\,\rangle$ with $T = \mathtt{in}(\mathcal{L}_E) \cup \{N_B \mid B \in Ar_E\} \cup \{N'_C \mid C \in U \setminus Ar_E\}$ and $F = \mathtt{out}(\mathcal{L}_E) \cup \{N'_B \mid B \in Ar_E\} \cup \{N_C \mid C \in U \setminus Ar_E\}$ becomes a partial stable model of P_{UAF}, and $A \in \mathtt{in}(\mathcal{L}_E)$. In this case, $\mathcal{N}_I = T \cap IN$ and $\mathcal{N}_J = F \cap OUT$ hold. In particular, if O is included in every complete labelling $\mathcal{L}_E$ of $AF_E = (Ar_E, att_E)$ with $E = (I, J)$, then $A \in T$ for any $\langle\, T, F\,\rangle$ such that $\mathcal{N}_I = T \cap IN$ and $\mathcal{N}_J = F \cap OUT$. The converse also holds by the fact that a partial stable model $\langle\, T, F\,\rangle$ of P_{UAF} is translated into a complete labelling $\mathtt{in}(\mathcal{L}_E) = \{A \mid A \in T \text{ and } N_A \in T\}$ and $\mathtt{out}(\mathcal{L}_E) = \{B \mid B \in F \text{ and } N_B \in T\}$ of AF_E. The results hold for (semi-)stable, grounded and preferred labelling using their equivalence to respective logic programming semantics [11]. The result of $O = \mathtt{out}(A)$ is shown in a similar way. □

Finally, we remark some complexity results on abduction in AF. By Proposition 3.1, an observation $O = \mathtt{in}(A)$ always has a skeptical/credulous explanation, and $O = \mathtt{out}(A)$ has a skeptical/credulous explanation if A is attacked by some argument B which satisfies simple conditions. Thus, deciding the existence of an explanation given an observation is trivial or done in polynomial time. On the other hand, given a pair of arguments $E = (I, J)$, the problem of deciding whether E is a credulous (or skeptical) explanation of an observation O under $\mathcal{S}$-labelling has different complexities under different semantics. In case of $O = \mathtt{in}(A)$, E is a credulous (resp. skeptical) explanation of O under $\mathcal{S}$-labelling of AF_E iff A is included in some (resp. every) $\mathcal{S}$-extension of AF_E. In case of $O = \mathtt{out}(A)$, put $UAF' = (U \cup \{X\}, att_U \cup \{(A, X)\})$ where X is a new argument such that $X \notin U$. For $AF = (Ar, att)$, put $AF' = (Ar \cup \{A, X\}, att \cup \{(A, X)\})$. Then, for any $A \in U$, E is a credulous (resp. skeptical) explanation of $O = \mathtt{out}(A)$ under $\mathcal{S}$-labelling of AF_E iff E is a credulous (resp. skeptical) explanation of $O' = \mathtt{in}(X)$ under $\mathcal{S}$-labelling of AF'_E. The next results hold by the complexity results in [16].

Theorem 3.5. *Let $UAF = (U, att_U)$ and $AF = (Ar, att)$ a sub-AF. Given $E = (I, J)$, deciding whether E is a credulous (resp. skeptical) explanation of an observation O under $\mathcal{S}$-labelling of AF_E is NP-complete (resp. polynomial) for complete labelling, NP-complete (resp. coNP-complete) for stable labelling, NP-complete (resp. Π_2^P-complete) for preferred labelling, and Σ_2^P-complete (resp. Π_2^P-complete) for semi-stable labelling. In case of grounded labelling, it is decided in polynomial time.*

4 Debate Games

Suppose a debate between a prosecutor (P) and a defense (D) in court.

P_1: The suspect is guilty because he had a grudge against the murder victim.
D_1: There is no evidence that the suspect killed the victim. No one is guilty until proven guilty.
P_2: There is an eyewitness who saw the suspect leaving the victim's apartment on the night of the crime.
D_2: The testimony is incredible because it was dark at night.

Given the argument P_1 by a prosecutor, the defense seeks an argument against P_1. Once the defense successfully refutes P_1 by the argument D_1, the prosecutor tries to refute D_1. A debate continues until one cannot refute the other. An appropriate modelling of debate should allow for the following three properties: (i) players have different beliefs and opinions in general; (ii) during a debate, each player may revise its own beliefs by new information provided by the opponent; (iii) a player may use inaccurate or even false arguments to win a debate [27].

Sakama [26] introduced a *debate game* based on an argumentation framework, which provides an abstract model of debates between two players and satisfies all three of the above requirements. We first review definitions of debate games. A *player* is an agent who has its own AF as a sub-AF of the given UAF.

Definition 4.1 (claim). [26] A *claim* is a pair of the form: $(\mathtt{in}(A), \_)$ or $(\mathtt{out}(B), \mathtt{in}(A))$ where A and B are different arguments. $(\mathtt{in}(A), \_)$ is read "A is labelled $\mathtt{in}$", while $(\mathtt{out}(B), \mathtt{in}(A))$ is read "B is labelled $\mathtt{out}$ because A is labelled $\mathtt{in}$". A claim $(\mathtt{in}(A), \_)$ or $(\mathtt{out}(B), \mathtt{in}(A))$ by a player is *refuted* by the claim $(\mathtt{out}(A), \mathtt{in}(C))$ with some argument C by another player.

Definition 4.2 (revision). [26] Let $UAF = (U, att_U)$ and $AF = (Ar, att)$ a sub-AF of the UAF. Then, a *revision* of AF with an argument $X \in U$ is defined as

$$AF \circ X = \begin{cases} (Ar \cup \{X\}, att \cup att_X) & \text{if } X \notin Ar \\ AF & \text{otherwise} \end{cases}$$

where $att_X = \{\, (X, Y),\ (Z, X) \mid Y,\ Z \in Ar \ \text{ and } \ (X, Y),\ (Z, X) \in att_U \setminus att \,\}$.

Definition 4.3 (debate game). [26] Let $UAF = (U, att_U)$, and $AF_1 = (Ar_1, att_1)$ and $AF_2 = (Ar_2, att_2)$ argumentation frameworks of two players P_1 and P_2, respectively. 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

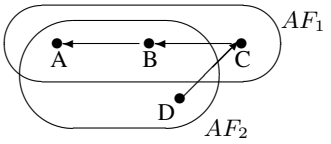
- $X_0 \in Ar_1$ and $X_k \in Ar_1^k$ where $AF_1^k = (Ar_1^k, att_1^k) = AF_1^{k-1} \circ Y_k$ $(k \geq 1)$ and $AF_1^0 = AF_1$.
- $Y_k \in Ar_2^k$ where $AF_2^k = (Ar_2^k, att_2^k) = AF_2^{k-1} \circ X_{k-1}$ $(k \geq 1)$ and $AF_2^0 = AF_2$.
- for each $\mathtt{out}(Z_j)$ in a claim by P_1 (resp.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

For a player P_1 (resp. P_2), the player P_2 (resp. P_1) is called the *opponent*.

Let Γ_n $(n \geq 0)$ be any claim. A *debate game* Δ (*for an argument* X_0) is an admissible debate between two players $[\Gamma_0, \Gamma_1, \ldots]$ where the initial claim is $\Gamma_0 = (\mathtt{in}(X_0), \_)$. A debate game Δ for an argument X_0 *terminates* with Γ_n if $\Delta = [\Gamma_0, \Gamma_1, \ldots, \Gamma_n]$ is an admissible debate and there is no claim Γ_{n+1} such that $[\Gamma_0, \Gamma_1, \ldots, \Gamma_n, \Gamma_{n+1}]$ is an admissible debate. In this case, the player P_i who makes the claim Γ_n *wins* the game.

The player P_1 starts a debate with the claim $\Gamma_0 = (\mathtt{in}(X_0), \_)$ based on its argumentation framework AF_1. The player P_2 then revises its argumentation framework AF_2 by X_0, and responds to the player P_1 with a counter-claim $\Gamma_1 = (\mathtt{out}(X_0), \mathtt{in}(Y_1))$ based on the revised argumentation framework AF_2^1. A debate continues by iterating revisions and claims. A debate game Δ terminates if each player does not repeat the same claim in the game ($\Gamma_i \neq \Gamma_{i+2k}$ $(k = 1, 2, \ldots)$ for any Γ_i $(i \geq 1)$ in Δ). 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 4.1. Let $UAF = (\{A, B, C, D\}, \{(D, C), (C, B), (B, A)\})$, $AF_1 = (\{A, B, C\}, \{(C, B), (B, A)\})$ and $AF_2 = (\{A, B, D\}, \{(B, A)\})$. AF_1 and AF_2 have the complete labellings: $\{\, \mathtt{in}(A), \mathtt{out}(B), \mathtt{in}(C) \,\}$ and $\{\, \mathtt{out}(A), \mathtt{in}(B), \mathtt{in}(D) \,\}$, respectively. The argumentation graph of two players is on the right.



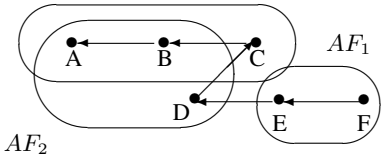
A debate game for the argument A between two players proceeds as follows:

AF_1: $(\mathtt{in}(A), \_)$ "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."

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 C. During the debate, the player P_2 learns the argument C by AF_1^1, then introduces it to AF_2^2 together with the attack relations (C, B) and (D, C). The player P_1 learns the argument D by AF_2^2 but cannot refute it. As a result, the player P_2 wins the game.

During a game, a player may make false or inaccurate claims to win the game.

Example 4.2. (1) Let $UAF = (\{A, B, C, D, E, F\}, \{(F, E), (E, D), (D, C), (C, B), (B, A)\})$, $AF_1 = (\{A, B, C, E, F\}, \{(F, E), (C, B), (B, A)\})$ and $AF_2 = (\{A, B, D\}, \{(B, A)\})$. AF_1 and AF_2 have the complete labellings: $\{\, \mathtt{in}(A), \mathtt{out}(B), \mathtt{in}(C), \mathtt{out}(E), \mathtt{in}(F) \,\}$ and $\{\, \mathtt{out}(A), \mathtt{in}(B), \mathtt{in}(D) \,\}$, respectively.

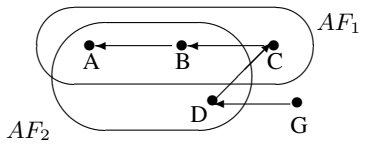


Consider a debate game for the argument A between two players as follows:

AF_1: $(\mathtt{in}(A), \_)$ "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."

The player P_2 cannot refute AF_1^2, then the player P_1 wins the game. In AF_1^2, however, P_1 provides a *false* claim on E because E is out in his/her labelling.

(2) Let $UAF = (\{A, B, C, D, G\}, \{(G, D), (D, C), (C, B), (B, A)\})$, $AF_1 = (\{A, B, C\}, \{(C, B), (B, A)\})$ and $AF_2 = (\{A, B, D\}, \{(B, A)\})$. AF_1 and AF_2 have the complete labellings: $\{\, \mathtt{in}(A), \mathtt{out}(B), \mathtt{in}(C) \,\}$ and $\{\, \mathtt{out}(A), \mathtt{in}(B), \mathtt{in}(D) \,\}$, respectively.



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

AF_1: $(\mathtt{in}(A), \_)$ "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: $(\mathtt{out}(D), \mathtt{in}(G))$ "D is out because G is in."

The player P_2 cannot refute AF_1^2, then the player P_1 wins the game. In AF_1^2, however, P_1 provides an *inaccurate* claim on G because G is not included in his/her labelling. In this sense, P_1 breaks the rule of admissibility of claims but P_2 cannot know it.

Definition 4.4 (honest/dishonest claim). [26] Let $UAF = (U, att_U)$ and $AF_i = (Ar, att)$ an argumentation framework of a player P_i in a debate game. Then,

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

A claim is called *dishonest* if it is either a lie or bullshit. A player is *honest* if every claim by the player is honest. Otherwise, a player is *dishonest*.[3]

A player P_i makes a claim under the complete labelling of his/her argumentation framework AF_i. A claim is honest if arguments included in the claim are credulously justified by AF_i. On the other hand, a player lies if he/she brings $\mathtt{in}(A)$ while believing $\mathtt{out}(A)$ or $\mathtt{undec}(A)$ in his/her labelling (AF_1^2 of Example 4.2(1)). A player bullshits if he/she brings $\mathtt{in}(A)$ while none of $\mathtt{in}(A)$, $\mathtt{out}(A)$ nor $\mathtt{undec}(A)$ is in his/her labelling (AF_1^2 of Example 4.2(2)). To allow the existence of dishonest players who may bullshit, Definition 4.3 of debate games is slightly modified in a way that each player may claim an argument which is not in his/her AF [26].

In a debate game, a player seeks a counter-claim which refutes a claim given by the opponent player. Viewing an argument given by the opponent player as an observation, computation of a counter-claim by a player is characterized by abduction as follows.

Theorem 4.1. *Let $UAF = (U, att_U)$ and $(\mathtt{out}(B), \mathtt{in}(A))$ (or $(\mathtt{in}(A), \_)$) be a claim made by a player P_1 under AF_1^k in a debate game.*

1. *If $O = \mathtt{out}(A)$ has the empty credulous explanation in AF_2^{k+1}, then a player P_2 can make an honest claim $(\mathtt{out}(A), \mathtt{in}(C))$ that refutes the claim by P_1.*
2. *Else if $O = \mathtt{out}(A)$ has no empty credulous explanation but has a non-empty credulous explanation E in AF_2^{k+1}, then a player P_2 cannot make an honest claim but can make a dishonest claim $(\mathtt{out}(A), \mathtt{in}(C))$ that refutes the claim by P_1.*
3. *Otherwise, if $O = \mathtt{out}(A)$ has no explanation, then P_2 cannot refute the claim by P_1 and loses the game.*

A similar result holds for a player P_1 against a claim made by a player P_2.

Proof. (1) If O has the empty credulous explanation in AF_2^{k+1}, then $\mathtt{out}(A)$ is credulously justified by AF_2^{k+1} under the complete labelling. In this case, P_2 can make an honest claim $(\mathtt{out}(A), \mathtt{in}(C))$ with an argument $C \in Ar_2^{k+1}$ such that $(C, A) \in att_2^{k+1}$. (2) Else if O has a non-empty credulous explanation $E = (I, J)$ in AF_2^{k+1}, then $\mathtt{out}(A)$ is credulously justified by $(AF_2^{k+1})_E = ((Ar_2^{k+1})_E, (att_2^{k+1})_E)$ under the complete labelling where $(Ar_2^{k+1})_E = (Ar_2^{k+1} \setminus J) \cup I$ and $(att_2^{k+1})_E = att_U \cap (Ar_2^{k+1})_E \times (Ar_2^{k+1})_E$. In this case, P_2 can make a dishonest claim $(\mathtt{out}(A), \mathtt{in}(C))$ with an argument $C \in (Ar_2^{k+1})_E$ such that $(C, A) \in (att_2^{k+1})_E$. (3) Otherwise, if O has no explanation in AF_2^{k+1}, P_2 cannot make a counter-claim $(\mathtt{out}(A), \mathtt{in}(C))$. □

[3] We use the notion of (dis)honest claims based on credulous justification under the complete labelling in [26], while alternative definitions are considered based on skeptical justification or different labellings.

In this characterization, an observation is always labelled $\mathtt{out}$. This is because the goal of a player is to justify $O = \mathtt{out}(A)$ or to explain it. When O has the empty credulous explanation, it is a most preferred explanation and a player makes an honest counter-claim. When O has multiple non-empty explanations, most preferred explanations are selected as best strategies. This is because a dishonest claim makes labellings of arguments deviate from those believed by the player. In Example 4.2(1), P_1 makes the dishonest claim $(\mathtt{out}(D), \mathtt{in}(E))$ but P_1 believes $\mathtt{in}(D)$ and $\mathtt{out}(E)$. A dishonest claim which increases such deviation is undesirable for a player because it would make difficult for the player to keep consistency during a debate and also increases the chance of dishonest claims being detected. However, selection of most preferred explanations as dishonest claims is not always successful. For instance, if the only explanation given for an observation needs to remove an argument that has already been used in the previous exchanges, then the player cannot hope to refute the opponent by hiding that argument. Comparing the lie $(\mathtt{out}(D), \mathtt{in}(E))$ by AF_1^2 in Example 4.2(1) with the bullshit $(\mathtt{out}(D), \mathtt{in}(G))$ by AF_1^2 in Example 4.2(2), lies are considered worse than bullshit. This is because the player P_1 knows the falsehood of $(\mathtt{out}(D), \mathtt{in}(E))$, while he/she does not know the truthfulness of $(\mathtt{out}(D), \mathtt{in}(G))$. There is no possibility of $\mathtt{in}(E)$ as far as F is $\mathtt{in}$, while there is a possibility of $\mathtt{in}(G)$ as far as there is no attacker of it. These behavioral rules are summarized as strategies of a player P_i as follows:

- If $O = \mathtt{out}(A)$ has the empty explanation in AF_i^k $(i = 1, 2;\ k \geq 1)$, then make an honest claim $(\mathtt{out}(A), \mathtt{in}(B))$ based on AF_i^k. Else if O has a preferred explanation E in AF_i^k then make a dishonest claim $(\mathtt{out}(A), \mathtt{in}(B))$ based on $(AF_i^k)_E$.
- If $O = \mathtt{out}(A)$ has non-empty multiple preferred explanations in AF_i^k, then select one $E = (I, J)$ such that for any $B \in J$, $\mathtt{in}(B)$ does not appear in any claim made by AF_i^j $(j < k)$.
- If $O = \mathtt{out}(A)$ has non-empty multiple preferred explanations in AF_i^k, then select one $E = (I, J)$ such that there is $B \in I \cap (U \setminus Ar_i^k)$ and $(B, A) \in att_U$ if any.

The first item says selecting honest claims at first. The other two items provide criteria for selecting dishonest claims. The second one is used for avoiding lie detection, while the third one presents preference of bullshit to lies.

5 Related Work

Abduction and argumentation have been combined in different ways in the literature. Dung [13] introduces the preferred extension semantics of abductive logic programs, which is defined as a maximally consistent set of hypotheses that contains its own defense against all attacks. The semantics is analyzed from the argumentation-theoretic viewpoint [19] and extended to *assumption-based argumentation* (ABA) [9]. In ABA an argument is a deduction of a conclusion (claim) c from a set of assumptions S represented as a tree, with c at the root and S at the leaves [15]. The goal of ABA is to construct an argument (tree) such that c is deduced from S using inference rules ($S \vdash c$). In ABA both a claim and assumptions are parts of an argument, which is different from our problem setting where arguments play the role of assumptions to explain another observed (labelled) argument.

Wakaki *et al.* [29] introduce hypothetical arguments to Dung's argumentation framework. They introduce *abductive argumentation framework* (AAF) which computes explanations to skeptically justify or not to credulously justify the argument supporting a claim. They consider concrete argumentation frameworks associated with *abductive logic programs* [19] under the answer set semantics. This is in contrast to our approach for abduction in abstract argumentation frameworks that have no restriction to any particular representation for arguments nor argumentation semantics. Moreover, in the AAF arguments are introduced to explain observations, while they cannot be removed from the knowledge base of an agent. In this sense, the AAF is based on the normal setting of abduction [19], while our current proposal is based on *extended abduction* of [18]. Extended abduction is particularly useful when a knowledge base is *nonmonotonic*. In nonmonotonic theories, deletion of formulas may introduce new formulas. Thus, addition and deletion of hypotheses play a complementary role in accounting for an observation in nonmonotonic theories. Since an argumentation framework is inherently nonmonotonic (i.e., introduction/removal of arguments changes labelling in general), the use of extended abduction is more natural and appropriate. Deletion of arguments would happen when one notices that his/her previous argument was incorrect (see the example at the beginning of Section 3). For another case, one would withdraw his/her argument and make a concession (to reach an agreement), even if he/she has a counter-argument against the opponent.

Kakas and Moraitis [20] use abduction to seek conditions to support arguments. An *argumentation theory* is defined as a pair (T, P) where T is a set of argument rules and P represents priorities over T. Then, a *supported argument* is defined as a tuple (Δ, S) where Δ is a set of argument rules from T and S is a set of hypothetical explanations. In their framework, an argument is a set of rules of the form $l_0 \leftarrow l_1, \ldots, l_n$ where l_i is a positive or negative literal. Each literal l_i $(1 \leq i \leq n)$ in the conditional part can be a hypothetical explanation but it is not an argument. This is different from our setting where explanations are also arguments. For another difference, abduction considered in their framework is normal setting of abduction, which is different from our setting of extended abduction. A supported argument is also used for building a proposal or responding to a proposal in argumentation-based negotiation [21]. Argumentation-based negotiation is studied by other researchers as well (for instance, [1]). A debate game is similar to argumentation-based negotiation in the sense that they use argumentation frameworks for formulating dialogues between competitive agents. However, the goal of negotiation is slightly different from debate—the goal of negotiation is to reach an agreement among players, while the goal of debate is to defeat the opponent player.

Šešelja and Straßer [28] integrate abduction and argumentation in their *explanatory argumentation framework* (EAF). An EAF is defined as a tuple $\langle \mathcal{A}, \chi, \rightarrow, \dashrightarrow, \sim \rangle$ where $\langle \mathcal{A}, \rightarrow \rangle$ is an AF, χ is a set of *explananda*, $\dashrightarrow$ is the *explanatory relation* over $\mathcal{A} \times (\mathcal{A} \cup \chi)$, and $\sim$ is the *incompatible relation* over $\mathcal{A} \times \mathcal{A}$. Thus, they distinguish attack relations and explanatory relations, and explananda and arguments. On the other hand, they do not distinguish arguments and hypotheses. Bex *et al.* [5] combine abduction and argumentation in the context of evidential reasoning. An argumentation framework is given as a pair (G, E) where G is a set of *evidential generalisations* and E is a set of *evidences*. The set O of observations is produced by applying evidential generalisa-

tions to evidences, and explanations (causal rules plus hypotheses) which account for the set of explananda $F \subseteq O$ are selected. In this study, argumentation and abduction are combined in a way different from ours: arguments are used for generating observations supported by evidences and justifying explanations against observations. Bex and Prakken [6] apply the framework to a formal dialogue game. In the game, players try to find a plausible and evidentially well-supported explanation for the explananda. None of the players wants to win, since they have the joint goal to find the best explanation of the explananda. This is in contrast with debate games where each player seeks explanations to justify its own individual argument to win a game.

Rotstein *et al.* [24] study argumentation theory change in abstract argumentation framework. A *dynamic argumentation framework* (DAF) has the universe U of arguments and the set $A \subseteq U$ of active arguments. Given an argument X to be warranted, a *dialectical tree* rooted in X is modified by activating nodes in $U \setminus A$ and by deactivating nodes in A to make X justified. They introduce argument change operators which expand the set A of arguments and contract some arguments from A. The goal of their study differs from ours in that their framework is dedicated to characterize dynamics of AF while abduction in AF is intended to reason explanations for/against a particular argument. Technically, their revision operators do not distinguish skeptical and credulous justifications. Baumann and Brewka [3] consider the problem of modifying an argumentation framework in a way that a desired set of arguments becomes an extension. To this end, they add new arguments and attack relations to an AF, while they do not delete arguments because one could delete everything and add the wanted arguments without any attacks. We consider deleting arguments (and corresponding attack relations) as well as introducing ones, while preferring explanations that minimally change the original AF. Baumann [4] enforces a desired set of arguments by adding/removing a minimal number of attack relations to an AF. He then introduces value functions to compute different types of modification. In this study, the distance between two argumentation frameworks is measured by counting added/removed attacks. On the other hand, we measure the distance by comparing labelling of arguments in two AFs. In this sense, minimal change considered in [4] is syntax-based, while minimal change considered in this paper is semantic-based. Boella *et al.* [7, 8] consider the effect of adding/removing arguments or attack relations under the grounded semantics. Cayol *et al.* [12] study the effect of an addition of an argument on the outcome of the argumentation semantics. The goal of these studies [7, 8, 12] is identifying possible changes of extensions after revising an argumentation framework, which is in contrast with our goal of identifying possible changes of an AF to have a particular outcome. Rahwan *et al.* [23] introduce a formal argumentation theory in which an agent may hide arguments or make up new arguments to accept a particular argument. The purpose of their study is to develop a game-theoretic argumentation mechanism design and to characterize strategy-proofness under graph-theoretic conditions. However, they do not provide any computational mechanism of dishonest arguments. We show the use of abduction in debate games based on formal argumentation frameworks, especially computing dishonest arguments. Extended abduction is also used for dishonest reasoning in logic programming [25]. In [25] an agent reasons dishonestly to have a particular goal at the individual level. The current study shows that extended abduction in AFs is used for computing dishonest arguments in debate games between two players.

6 Conclusion

We introduced extended abduction to abstract argumentation frameworks and provided its computational method in logic programming. Next we showed its application to computing (dis)honest claims in debate games. The result of this paper realizes extended abduction in argumentation frameworks, and provides a strong link between abduction, argumentative reasoning, and dishonest reasoning in a formal dialogue system based on AF. The abduction mechanism proposed in this paper will also be applied to revision of AF and will be realized in argumentation systems associated with logic programming. These issues are left for future work.

References

1. Amgoud, L., Dimopoulos, Y., Moraitis, P.: A unified and general framework for argumentation-based negotiation. In: Proc. AAMAS-07, pp. 1018–1025 (2007)
2. Baroni, P., Giacomin, M.: Semantics of abstract argument systems. In: Rahwan, I., Simari, G. R. (eds.), Argumentation in Artificial Intelligence, pp. 25–44. Springer (2009)
3. Baumann, R., Brewka, G.: Expanding argumentation frameworks: enforcing and monotonicity results. In: Proc. 3rd COMMA, Frontiers in AI, vol. 216, pp. 75–86. IOS Press (2010)
4. Baumann, R.: What does it take to enforce an argument? Minimal change in abstract argumentation. In: Proc. 20th European Conf. Artificial Intelligence, pp. 127–132. IOS Press (2012)
5. Bex, F. J., Prakken, H., Verheij, B.: Formalising argumentation story-based analysis of evidence. In: Proc. 11th Int'l Conf. Artificial Intelligence and Law, pp. 1–10 (2007)
6. Bex, F. J., Prakken, H.: Investigating stories in a formal dialogue game. In: Proc. 2nd Int'l Conf. Computational Models of Argument, pp. 73–84. IOS Press (2008)
7. Boella, G., Kaci, S., Van der Torre, L.: Dynamics in argumentation with single extensions: attack refinement and the grounded extension. In: Proc. AAMAS-09, pp. 1213–1214 (2009)
8. Boella, G., Kaci, S., Van der Torre, L.: Dynamics in argumentation with single extensions: abstract principles and the grounded extension. In: Proc. ECSQARU-09, LNCS (LNAI), vol. 5590, pp. 107–118. Springer, Heidelberg (2009)
9. Bondarenko, A., Dung, P. M., Kowalski, R. A., Toni, F.: An abstract, argumentation-theoretic approach to default reasoning. Artificial Intelligence 93, 63–101 (1997)
10. Caminada, M., Gabbay, D. M.: A logical account of formal argumentation. Studia Logica 93, 109–145 (2009)
11. Caminada, M., Sá, S., Alcântara, J.: On the equivalence between logic programming semantics and argumentation semantics. Technical Report ABDN-CS-13-01, University of Aberdeen, 2013. A shorter version in: Proc. ECSQARU-13, LNCS (LNAI), vol. 7958, pp. 97–108. Springer, Heidelberg (2013)
12. Cayrol, C., Dupin de Saint-Cyr, F., Lagasquie-Schiex, M.-C.: Change in abstract argumentation frameworks: adding an argument. J. Artificial Intelligence Research 38, 49–84 (2010)
13. Dung, P. M.: Negation as hypothesis: an abductive foundation for logic programming. In: Proc. ICLP, pp. 3–17. MIT Press (1991)
14. 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)
15. Dung, P. M., Kowalski, R. A., Toni, F.: Assumption-based argumentation. In: Rahwan, I., Simari, G. R. (eds.), Argumentation in Artificial Intelligence, pp. 199–218. Springer (2009)
16. Dvořák, W., Woltran, S.: On the intertranslatability of argumentation semantics. J. Artificial Intelligence Research 41, 445–475 (2011)

17. Hughes, W.: Critical Thinking: An introduction to the basic skills. Broadview Press (1992)
18. Inoue, K., Sakama, C.: Abductive framework for nonmonotonic theory change. In: Proc. IJCAI-95, pp. 204–210 (1995)
19. Kakas, A. C., Kowalski, R. A., Toni, F.: Abductive logic programming. J. Logic and Computation 2(6), 719–770 (1992)
20. Kakas, A. C., Moraitis, P.: Argumentative agent deliberation, roles and context. Electronic Notes in Theoretical Computer Science 70, 39–53 (2002)
21. Kakas, A. C., Moraitis, P.: Adaptive agent negotiation via argumentation. In: Proc. AAMAS-06, pp. 384–391 (2006)
22. Mayes, G. R.: Argument-explanation complementarity and the structure of informal reasoning. Informal Logic 30, 92–111 (2010)
23. Rahwan, I., Larson, K., Tohmé, F.: A characterisation of strategy-proofness for grounded argumentation semantics. In: Proc. IJCAI-09, pp. 251–256 (2009)
24. Rotstein, N. D., Moguillansky, M. O., Falappa, M. A., García, A. J., Simari, G. R.: Argument theory change: revision upon warrant. In: Proc. 2nd COMMA, pp. 336–347, IOS Press (2008)
25. Sakama, C.: Dishonest reasoning by abduction. In: Proc. IJCAI-11, pp. 1063–1068 (2011).
26. Sakama, C.: Dishonest arguments in debate games. In: Proc. 4th Int'l Conf. Computational Models of Argument, Frontiers in AI and Applications, vol. 245. IOS Press, pp. 177–184 (2012)
27. Schopenhauer, A.: The Art of Controversy. Originally published in 1896 and is translated by T. Bailey Saunders, Cosimo Classics, New York (2007)
28. Šešelja, D., Straßer, C.: Abstract argumentation and explanation applied to scientific debates. Synthese 190(12), 2195–2217 (2013)
29. Wakaki, T., Nitta, K., Sawamura, H.: Computing abductive argumentation in answer set programming. In: Proc. 6th ArgMas, LNCS (LNAI), vol. 6057, pp. 195–215. Springer, Heidelberg (2010)
30. Wu, Y., Caminada, M., Gabbay, D. M.: Complete extensions in argumentation coincides with 3-valued stable models in logic programming. Studia Logica 93(2-3), 383–403 (2009)