September 2024

Argument and Belief

Chiaki SAKAMA [a,1]
[a] *Wakayama University, Japan*
ORCiD ID: Chiaki Sakama https://orcid.org/0000-0002-9966-3722

Abstract. Given an abstract argumentation framework $(\{p,q\},\{(p,q)\})$ in which an argument p attacks another argument q, argumentation semantics normally concludes that p is accepted and q is rejected. To reject p, on the other hand, a counter-argument attacking p is to be introduced. However, a player participating in an argumentation or a person in the audience of a public debate would have opinions such that "I do not believe p", "I still believe q", or "I do not believe that p attacks q" without any concrete grounds. In this study, we introduce the notions of *AF with beliefs* and *belief extensions* to represent interaction between arguments and beliefs. Those notions are used for modelling the audience of argumentation, dialogue between two agents, and inner conflict of an agent.

Keywords. argument, audience, belief, self-deception

1. Introduction

An *abstract argumentation framework* or simply *argumentation framework* (AF) [1] provides a simple framework for representing and reasoning about arguments. Given $AF = (\{p,q\},\{(p,q)\})$ in which an argument p attacks another argument q, argumentation semantics normally concludes that p is accepted and q is rejected. To reject p, on the other hand, a counter-argument attacking p is to be introduced. However, a player participating in an argumentation or a person in the audience of a public debate would have opinions such that "I do not believe p", "I still believe q", or "I do not believe that p attacks q" without any concrete grounds. In [2], the authors say:

> *Consider, for example, when a member of audience of a TV debate listens to the debate at home, she can produce the abstract argumentation graph based on the arguments and counterarguments exchanged. Then she can identify a probability function to represent the belief she has in each of the arguments. So she may disbelieve some of the arguments based on what she knows about the topic. Furthermore, she may disbelieve some of the arguments that are unattacked. As an extreme, she is at liberty of completely disbelieving all of the arguments (so assign probability 0 to all of them). If we want to model audiences, where the audience either does not want to or is unable to add counterarguments to an argument graph being constructed in some form of argumentation, we need to take the beliefs of the audience into account, and we need to consider which arguments they believe or disbelieve.*

[1] Corresponding Author: Chiaki Sakama, sakama@wakayama-u.ac.jp.

We share the motivation of this study with this statement, while we do not agree with the modelling in which a person in the audience identifies a probability function and computes probabilities of each argument. Rather, it would be natural to represent (dis)belief of arguments as formulas like Bp or $\neg Bp$ where B represents belief and p is an argument.

In this study, we introduce the framework of *AF with beliefs* (AFB) to represent interaction between arguments and beliefs. In AFB an agent's beliefs are added to the argumentation graph and interact with arguments. We introduce axioms for interlinking arguments and beliefs, and compute belief extensions that represent (dis)believed arguments as well as accepted arguments. We apply the framework to modelling the audience of argumentation, dialogue between two agents, and inner conflict of an agent. The rest of this paper is organized as follows. Section 2 introduces a framework of AF with beliefs. Section 3 applies the framework to dialogues between two agents. Section 4 uses the framework for representing inner conflict of an agent. Section 5 addresses related work and Section 6 summarizes the paper.

2. AF with Belief

We consider a language that contains a finite set of propositional variables (or atoms) $\mathscr{A} = \{p, q, r, \ldots\}$ and the logical connectives $\neg$, $\vee$, $\wedge$, $\supset$ and $\equiv$. An *argumentation framework* (AF) is a pair (A, R) where $A \subseteq \mathscr{A}$ is a finite set of *arguments* and $R \subseteq A \times A$ is an *attack relation*. For an AF (A, R), an argument p *attacks* an argument q if $(p, q) \in R$. An AF is represented as a directed graph in which nodes represent arguments and edges represent attacks. We write $p \to q$ iff $(p, q) \in R$.[2] $p \leftrightarrow q$ is an abbreviation of $p \to q$ and $q \to p$. A set S of arguments *attacks* an argument p iff there is an argument $q \in S$ that attacks p. A set S of arguments is *conflict-free* if there are no arguments $p, q \in S$ such that p attacks q. A set S of arguments *defends* an argument p if S attacks every argument that attacks p. We write $D(S) = \{\, p \mid S \text{ defends } p \,\}$.

The semantics of AF is defined as the set of designated *extensions*. The following four extensions are introduced in [1]. Given $AF = (A, R)$, a conflict-free set of arguments $S \subseteq A$ is: (i) a *complete extension* iff $S = D(S)$; (ii) a *stable extension* iff S attacks each argument in $A \setminus S$; (iii) a *preferred extension* iff S is a maximal complete extension of AF (wrt $\subseteq$); (iv) a *grounded extension* iff S is the minimal complete extension of AF (wrt $\subseteq$). We often abbreviate complete, stable, preferred, and grounded extensions as *co, st, pr*, and *gr*, respectively. When we refer to AF with σ extensions, it means AF with one of the above four extensions, i.e., $\sigma \in \{co, st, pr, gr\}$. An argumentation semantics σ is *universal* if any AF has at least one σ extension. Among the four semantics, *co*, *pr*, *gr* are universal but *st* is not.

In this paper we consider *agents* playing different roles: a person in the audience of a public debate, a person participating in a dialogue, or a person arguing with oneself at an intrapersonal level. An argumentation framework consists of arguments and attacks, then an agent possibly has two types of *beliefs*—belief on arguments and belief on attacks. If an agent *a believes* an argument p (resp. an attack $p \to q$) to be true, it is represented as $B_a p$ (resp. $B_a(p \to q)$). When the agent's identification is unimportant, a

[2] Throughout the paper, logical implication is represented by $\supset$ and is distinguished from the attack relation $\to$.

is omitted and it is simply written as Bp or $B(p \to q)$. An agent's disbelieving p (resp. $p \to q$) is represented by $\neg Bp$ (resp. $\neg B(p \to q)$). Beliefs are possibly nested, for instance, BBp represents that an agent believes that he/she believes p. In this paper, beliefs are nested at most two times. Technically, we handle $p \to q$, $p \leftrightarrow q$, $(\neg)Bp$, $(\neg)B(p \to q)$ or $(\neg)B(p \leftrightarrow q)$ as an atom, so B is not an operator in modal epistemic logic. For instance, $p \to q$ is considered an atom '$p\_attacks\_q$' and Bp (resp. $\neg B(p \to q)$) is considered an atom '$believe\_p$' (resp. '$disbelieve\_p\_attacks\_q$'). In this setting, the "atom" $\neg\neg Bp$ is identified with Bp. Note that atoms Bp and $\neg Bp$ are not related in the usual way of classical logic. To encode their semantic relation, we use attack relations $Bp \leftrightarrow \neg Bp$ (Definition 2.2). We do not consider the argument of the form $\neg p$, but it is also encoded as an argument q with the attack relations $q \leftrightarrow p$.

Given an argumentation framework $AF = (A,R)$, the set $\mathscr{B}_{AF}$ of *belief atoms over* AF is defined as $\mathscr{B}_{AF} = \{Bp, \neg Bp \mid p \in A\} \cup \{B(p \to q), \neg B(p \to q) \mid (p,q) \in R\}$.

Definition 2.1 (AF with belief) Let $AF = (A,R)$ be an argumentation framework. Then, *AF with belief (or AFB)* is defined as a triple $\Gamma = (A,R,S)$ where $S \subseteq \mathscr{B}_{AF}$. Γ is often written as (AF,S).

Definition 2.2 (attacks over beliefs) Let $AF = (A,R)$ be an argumentation framework. Then, define $R_B = R \cup \{(\neg Bp, p), (\neg Bp, Bp), (Bp, \neg Bp) \mid p \in A\}$.

R_B introduces attacks over belief atoms. Bp and $\neg Bp$ attack each other. In addition, if an agent does not believe p, she does not accept p. This is represented by the attack $\neg Bp \to p$.[3]

Definition 2.3 (attack axiom) Let p and q be arguments. Then

$$(\mathbf{AT}) \qquad Bp \wedge B(p \to q) \supset \neg Bq$$

is called the *attack axiom*.

The attack axiom (**AT**) says that if an agent believes an argument p and the attack relation $p \to q$, then the agent disbelieves the argument q. (**AT**) is rewritten as

$$Bq \wedge B(p \to q) \supset \neg Bp \qquad \text{or} \qquad Bp \wedge Bq \supset \neg B(p \to q).$$

Definition 2.4 ($cl_{AT}(S)$) Given $S \subseteq \mathscr{B}_{AF}$, define $cl_{AT}(S) \subseteq \mathscr{B}_{AF}$ as the smallest set of belief atoms satisfying the following conditions:

1. $S \subseteq cl_{AT}(S)$.
2. If $Bp \in cl_{AT}(S)$ and $B(p \to q) \in cl_{AT}(S)$, then $\neg Bq \in cl_{AT}(S)$.
3. If $Bq \in cl_{AT}(S)$ and $B(p \to q) \in cl_{AT}(S)$, then $\neg Bp \in cl_{AT}(S)$.
4. If $Bp \in cl_{AT}(S)$ and $Bq \in cl_{AT}(S)$, then $\neg B(p \to q) \in cl_{AT}(S)$.

The set $cl_{AT}(S)$ is *consistent* if it does not contain $\{Bp, \neg Bp \mid p \in \mathscr{A}\}$ nor $\{B(p \to q), \neg B(p \to q) \mid p,q \in \mathscr{A}\}$ as a subset. Given $AF = (A,R)$, define $cl_{AT}(S)_A = cl_{AT}(S) \cap \{Bp, \neg Bp \mid p \in A\}$ and $cl_{AT}(S)_R = cl_{AT}(S) \cap \{B(p \to q), \neg B(p \to q) \mid (p \to q) \in R\}$.

[3] We do not consider an argument of the form $\neg p$, so the attacks $Bp \to \neg p$ or $B\neg p \to p$ is not included.

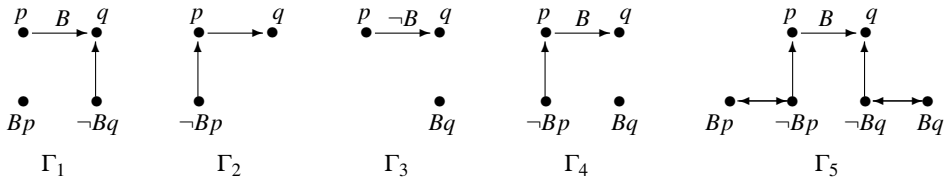

Figure 1. AFBs in Example 2.1

$cl_{AT}(S)$ represents a set of belief atoms closed under the application of the axiom (**AT**).

Definition 2.5 (belief extension) Let $\Gamma = (A, R, S)$ be an AFB. Then, a set E is a σ *belief extension* of Γ if E is a σ extension of $AF = (X, Y)$ where $X = A \cup cl_{AT}(S)_A$, $Y = ((X \times X) \cap R_B) \setminus \{(p \to q) \mid \neg B(p \to q) \in cl_{AT}(S)_R\}$, and $\sigma \in \{co, st, pr, gr\}$.

By definition, belief extensions are extensions of an argumentation graph that consists of arguments, belief over arguments, and attacks over them. Arguments in A and belief atoms in $cl_{AT}(S)_A$ possibly interact with each other in $AF = (X, Y)$. When an agent disbelieves an attack $p \to q$ in $cl_{AT}(S)_R$, the attack is cancelled and is removed from Y.

As $\neg Bp$ attacks p, no σ belief extension contains both p and $\neg Bp$ for any $p \in \mathscr{A}$. This means that AFB does not involve the Moore's paradox such that "an agent accepts an argument p but she disbelieves it". Also Bp and $\neg Bp$ mutually attack the other in R_B, so no σ belief extension contains both Bp and $\neg Bp$ for any $p \in \mathscr{A}$. On the other hand, there is a case that a σ belief extension contains neither Bp nor $\neg Bp$ for some $p \in \mathscr{A}$. This is because the formula $Bp \vee \neg Bp$ is *not* valid in our framework. That is, there may be an argument (or an attack) which an agent neither believes nor disbelieves. Technically, this is justified by handling Bp and $\neg Bp$ as atoms, i.e., the formula $Bp \vee \neg Bp$ is viewed as "$believe\_p \vee disbelieve\_p$" which is not a tautology.

Suppose an agent in the audience of a public debate. Then different belief states on an AF are represented by AFBs as follows.

Example 2.1 Let $AF = (\{p, q\}, \{(p, q)\})$ and $\sigma \in \{co, st, pr, gr\}$. Then,
(1) $\Gamma_1 = (AF, \{Bp, B(p \to q)\})$ has the σ belief extension $E_1 = \{p, Bp, \neg Bq\}$.
(2) $\Gamma_2 = (AF, \{\neg Bp\})$ has the σ belief extension $E_2 = \{\neg Bp, q\}$.
(3) $\Gamma_3 = (AF, \{Bq, \neg B(p \to q)\})$ has the σ belief extension $E_3 = \{p, q, Bq\}$.
(4) $\Gamma_4 = (AF, \{Bq, B(p \to q)\})$ has the σ belief extension $E_4 = \{\neg Bp, Bq, q\}$.
(5) $\Gamma_5 = (AF, \{Bp, Bq, B(p \to q)\})$ has the grounded belief extension $E_5 = \varnothing$; four stable (or preferred) belief extensions $E_6 = \{p, Bp, \neg Bq\}$, $E_7 = \{p, Bp, Bq\}$, $E_8 = \{q, \neg Bp, Bq\}$, and $E_9 = \{\neg Bp, \neg Bq\}$; and five complete belief extensions E_5–E_9.

Five different AFBs of Example 2.1 are illustrated in Figure 1. In Γ_1 an agent believes p and the attack $p \to q$, which implies $\neg Bq$ by (**AT**). The belief extension E_1 then represents that p is accepted and q is rejected, and the agent believes p but disbelieves q. In Γ_2, on the other hand, an agent disbelieves p. Then p is rejected and, as a result, q is accepted in E_2. In Γ_3 an agent believes q and disbelieves the attack $p \to q$. In this case, the attack is cancelled in Γ_3 and both p and q are accepted. By contrast, in Γ_4 an agent believes the attack $p \to q$. Then, Bq and $B(p \to q)$ deduce $\neg Bp$ by (**AT**), and $\neg Bp$ attacks p by $(\neg Bp, p)$ in R_B. As a result, E_4 represents that an agent believes q and disbelieves p,

so that p is rejected and q is accepted. In Γ_5, Bp and $B(p \to q)$ deduce $\neg Bq$, and Bq and $B(p \to q)$ deduce $\neg Bp$ by (**AT**). As a result, it becomes $cl_{AT}(S)_A = \{Bp, Bq, \neg Bp, \neg Bq\}$ which is inconsistent. In this case, the grounded belief extension becomes the empty set, while four different stable (preferred) belief extensions exist.

Since co, pr, gr are universal, $\Gamma = (AF, S)$ has a σ belief extension if $AF = (A, R)$ has a σ extension for $\sigma \in \{co, pr, gr\}$. For $\sigma = st$, on the other hand, when $AF = (A, R)$ has a stable extension, $\Gamma = (AF, S)$ may not have a stable extension; and when $AF = (A, R)$ has no stable extension, $\Gamma = (AF, S)$ may have a stable belief extension.

Example 2.2 (1) Consider $AF = (\{p, q\}, \{(p, q), (q, q)\})$ and $AFB = (AF, \{\neg Bp\})$. Then AF has the stable extension $\{p\}$, while AFB has no stable belief extension. (2) Consider $AF = (\{p\}, \{(p, p)\})$ and $AFB = (AF, \{\neg Bp\})$. Then AF has no stable extension, while AFB has the stable belief extension $\{\neg Bp\}$.

An AFB $\Gamma = (A, R, S)$ is *rational* if $cl_{AT}(S)$ is consistent. A rational AFB represents an agent who has a consistent belief over AF.

Proposition 2.1 *Let $\Gamma = (A, R, S)$ be a rational AFB. Then $cl_{AT}(S)_A \subseteq E$ holds for any σ belief extension E of Γ where $\sigma \in \{co, st, pr, gr\}$.*

Proof: Since $cl_{AT}(S)$ is consistent, each belief atom in $cl_{AT}(S)_A$ is not attacked by any argument in $A \cup cl_{AT}(S)_A$. Then those belief atoms are included in any σ belief extension of Γ. Hence, the result holds. □

Proposition 2.2 *Let $\Gamma = (A, R, S)$ be a rational AFB. If $B(p \leftrightarrow q)$ is in $cl_{AT}(S)$, there is no σ belief extension E such that $\{Bp, Bq\} \subseteq E$ where $\sigma \in \{co, st, pr, gr\}$.*

Proof: Suppose that there is a σ belief extension E such that $\{Bp, Bq\} \subseteq E$. Then $\{Bp, Bq\} \subseteq cl_{AT}(S)_A$ because any belief atom in E is an element of $cl_{AT}(S)_A$. Since $Bp \wedge B(p \to q)$ implies $\neg Bq$ by (**AT**), $\neg Bq \in cl_{AT}(S)_A$. This contradicts the assumption that $cl_{AT}(S)$ is consistent. □

Proposition 2.3 *Let $\Gamma = (A, R, S)$ be a rational AFB. If $B(p \to p)$ is in $cl_{AT}(S)$, there is no σ belief extension E such that $Bp \in E$ where $\sigma \in \{co, st, pr, gr\}$.*

Proof: The result is obtained by putting $p = q$ in Proposition 2.2. □

By Propositions 2.2 and 2.3, we can say that if a rational AFB has a σ belief extension such that $\{Bp, Bq\} \subseteq E$ (resp. $Bp \in E$) then an agent does not believe the attack $p \leftrightarrow q$ (resp. $p \to p$).

3. Dialogue

In this section, we consider dialogues between two agents a and b. Belief of each agent is represented by B_a and B_b, respectively. To distinguish arguments made by each agent, we often attach subscripts to arguments like p_a and q_b where p_a means that an argument p is made by an agent a.

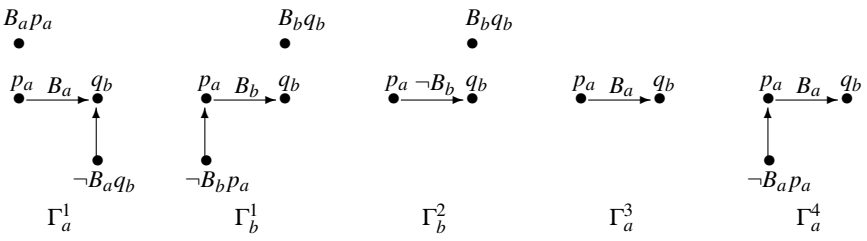

Figure 2. AFBs in Example 3.1

Definition 3.1 (dialogue) Let a and b be two agents. Then a *dialogue* between a and b is defined as a pair $\Delta = (\Gamma_a, \Gamma_b)$ where $\Gamma_a = (AF, S_a)$ and $\Gamma_b = (AF, S_b)$ are AFBs.

By definition, a dialogue consists of two AFBs such that each AFB represents belief of an agent wrt a common AF.

Definition 3.2 ((in)sincere agent) Let $\Gamma_a = (AF, S_a)$ be an AFB with $AF = (A, R)$. The agent a is *sincere* if $p_a \in A$ implies $B_a p_a \in S_a$. Otherwise, a is *insincere*.

Definition 3.2 presents that a sincere agent a makes an argument p_a only if she believes it. Attacks over beliefs (Definition 2.2) and the attack axiom (Definition 2.3) are respectively modified under the multiagent setting as follows. Given $AF = (A, R)$, $R_B = R \cup \{ (\neg B_i p_j, p_j), (\neg B_i p_j, B_i p_j), (B_i p_j, \neg B_i p_j) \mid p_j \in A \text{ and } i, j \in \{a, b\} \}$, and

$$(\mathbf{AT}) \quad B_i p_j \wedge B_i(p_j \to q_k) \supset \neg B_i q_k \text{ where } p_j, q_k \in A \text{ and } i, j, k \in \{a, b\}.$$

Definition 3.3 (static belief extension) Let $\Delta = (\Gamma_a, \Gamma_b)$ be a dialogue where $\Gamma_a = (AF, S_a)$ and $\Gamma_b = (AF, S_b)$. Then, a pair (E, F) is a *static σ belief extension* (or *σ-SBE* for short) of Δ if E (resp. F) is a σ extension of $AF = (X, Y)$ where $X = A \cup cl_{AT}(S_a)_A$ (resp. $X = A \cup cl_{AT}(S_b)_A$), $Y = ((X \times X) \cap R_B) \setminus \{(p \to q) \mid \neg B_a(p \to q) \in cl_{AT}(S_a)_R\}$ (resp. $Y = ((X \times X) \cap R_B) \setminus \{(p \to q) \mid \neg B_b(p \to q) \in cl_{AT}(S_b)_R\}$), and $\sigma \in \{co, st, pr, gr\}$.

A static belief extension represents belief states of each agent and accepted arguments.

Example 3.1 Let $AF = (\{p_a, q_b\}, \{(p_a, q_b)\})$ and $\sigma \in \{co, st, pr, gr\}$. Then,
(1) $\Delta_1 = (\Gamma_a^1, \Gamma_b^1)$ where $\Gamma_a^1 = (AF, \{B_a(p_a \to q_b), B_a p_a\})$ and $\Gamma_b^1 = (AF, \{B_b(p_a \to q_b), B_b q_b\})$ has the σ-SBE $(\{p_a, B_a p_a, \neg B_a q_b\}, \{q_b, B_b q_b, \neg B_b p_a\})$.
(2) $\Delta_2 = (\Gamma_a^1, \Gamma_b^2)$ where $\Gamma_b^2 = (AF, \{\neg B_b(p_a \to q_b), B_b q_b\})$ has the σ-SBE $(\{p_a, B_a p_a, \neg B_a q_b\}, \{p_a, q_b, B_b q_b\})$.
(3) $\Delta_3 = (\Gamma_a^3, \Gamma_b^1)$ where $\Gamma_a^3 = (AF, \{B_a(p_a \to q_b)\})$ has the σ-SBE $(\{p_a\}, \{q_b, B_b q_b, \neg B_b p_a\})$.
(4) $\Delta_4 = (\Gamma_a^4, \Gamma_b^1)$ where $\Gamma_a^4 = (AF, \{\neg B_a p_a, B_a(p_a \to q_b)\})$ has the σ-SBE $(\{\neg B_a p_a, q_b\}, \{q_b, B_b q_b, \neg B_b p_a\})$.

Five different AFBs used in dialogues of Example 3.1 are illustrated in Figure 2. In these dialogues, an agent b makes an argument q and an agent a makes a counter-argument p. In Δ_1, a believes her argument p_a and the attack $p_a \to q_b$, which results in disbelieving the argument q_b by (**AT**). Similarly, b believes his argument q_b and the attack $p_a \to q_b$, which results in disbelieving the argument p_a by (**AT**). The situation changes when the

agent b disbelieves the attack $p_a \to q_b$ in Δ_2. In this case, the attack is cancelled in Γ_b^2 and p_a is included in the σ belief extension of Γ_b^2. In Δ_1 and Δ_2, two agents are sincere. By contrast, Δ_3 and Δ_4 represent situations in which b is sincere but a is insincere. In Δ_3 an agent a makes an argument p_a but she has no belief on it (but she believes that p_a attacks q_b). In this case, the σ belief extension of Γ_a^3 contains no belief on arguments by a. In Δ_4, a makes an argument p_a but she disbelieves it. As $\neg B_a p_a$ attacks p_a, p_a is not included in the σ belief extension of Γ_a^4. Δ_3 and Δ_4 represent different types of dishonesty—Δ_3 represents *bluffing* or *bullshitting* and Δ_4 represents *lying* [3]. Such dishonest arguments appear in practice for the purpose of rejecting an unwanted argument by making up a fake argument.

Definition 3.3 characterizes a situation in which beliefs of agents do not change during a dialogue. On the other hand, belief of an agent may change during a dialogue when her argument is attacked by a counter-argument. We next characterize such a situation. In what follows, $B_a^t p$ (resp. $B_a^t(p \to q)$) means that a believes p (resp. $p \to q$) at time t where $t \geq 0$ is an integer representing discrete time steps. Given $AF = (A,R)$, define $\mathscr{B}_{AF}^T = \{ B_i^t p, \neg B_i^t p \mid p \in A \text{ and } t \in T \} \cup \{ B_i^t(p \to q), \neg B_i^t(p \to q) \mid (p,q) \in R \text{ and } t \in T \}$ where $i \in \{a,b\}$ and T is the set of integers representing time.

Definition 3.4 (belief change axiom) Let a be an agent and p, q arguments. Then,

$$(\mathbf{BC}) \qquad B_a^t p \wedge B_a^t(p \to q) \supset \neg B_a^{t+1} q \quad (t \in T)$$

is called the *belief change axiom*.

(**BC**) represents that when an agent a believes an argument p and the attack $p \to q$ at time t, a does not believe q at time $t+1$. (**BC**) represents a dynamic version of the attack axiom (**AT**). Like (**AT**), (**BC**) is rewritten as

$$B_a^{t+1} q \wedge B_a^t(p \to q) \ \supset \ \neg B_a^t p \qquad \text{or} \qquad B_a^t p \wedge B_a^{t+1} q \ \supset \ \neg B_a^t(p \to q).$$

Definition 3.5 (inertia rule) Let a be an agent and p, q arguments. The default rules:

$$(\mathbf{IR}) \qquad \frac{B_a^t \alpha : B_a^{t+1} \alpha}{B_a^{t+1} \alpha} \quad \text{and} \quad \frac{\neg B_a^t \alpha : \neg B_a^{t+1} \alpha}{\neg B_a^{t+1} \alpha} \quad (t \in T)$$

are called the *inertia rules*, where α is either an argument p or an attack $p \to q$.

(**IR**) are normal default rules in default logic [4] meaning that if $(\neg) B_a^t \alpha$ is the case and $(\neg) B_a^{t+1} \alpha$ is consistently assumed then conclude $(\neg) B_a^{t+1} \alpha$. Attacks over beliefs are modified for the dynamic setting as follows. Given $AF = (A,R)$, $R_D = R \cup \{ (\neg B_i^t p_j, p_j), (\neg B_i^t p_j, B_i^t p_j), (B_i^t p_j, \neg B_i^t p_j) \mid p_j \in A, i,j \in \{a,b\} \text{ and } t \in T \}$.

Definition 3.6 ($cl_D(S)$) Given $S \subseteq \mathscr{B}_{AF}^T$, define $cl_D(S) \subseteq \mathscr{B}_{AF}^T$ as the smallest set of belief atoms satisfying the following conditions:

1. $S \subseteq cl_D(S)$.
2. If $B_a^t p \in cl_D(S)$ and $B_a^t(p \to q) \in cl_D(S)$, then $\neg B_a^{t+1} q \in cl_D(S)$.
3. If $B_a^{t+1} q \in cl_D(S)$ and $B_a^t(p \to q) \in cl_D(S)$, then $\neg B_a^t p \in cl_D(S)$.
4. If $B_a^t p \in cl_D(S)$ and $B_a^{t+1} q \in cl_D(S)$, then $\neg B_a^t(p \to q) \in cl_D(S)$.

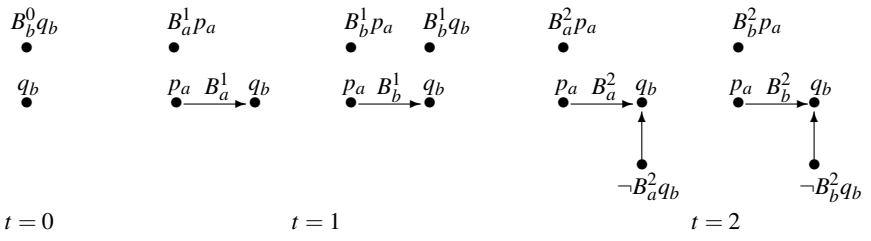

Figure 3. AFBs in Example 3.2

5. If $B_a^t \alpha \in cl_D(S)$ and $\{B_a^{t+1}\alpha\} \cup cl_D(S)$ is consistent, then $B_a^{t+1}\alpha \in cl_D(S)$.
6. If $\neg B_a^t \alpha \in cl_D(S)$ and $\{\neg B_a^{t+1}\alpha\} \cup cl_D(S)$ is consistent, then $\neg B_a^{t+1}\alpha \in cl_D(S)$.

Given $AF = (A,R)$, define $cl_D(S)_A = cl_D(S) \cap \{ B_i^t p, \neg B_i^t p \mid p \in A, i \in \{a,b\}, t \in T \}$ and $cl_D(S)_R = cl_D(S) \cap \{ B_i^t(p \to q), \neg B_i^t(p \to q) \mid (p \to q) \in R, i \in \{a,b\}, t \in T \}$.

$cl_D(S)$ represents a set of belief atoms closed under the application of the axiom (**BC**) and the inertia rule (**IR**).

Definition 3.7 (dynamic belief extension) Let $\Delta = (\Gamma_a, \Gamma_b)$ be a dialogue where $\Gamma_a = (AF, S_a)$ and $\Gamma_b = (AF, S_b)$. Then, a pair (E,F) is a *dynamic* σ *belief extension* (or σ-*DBE* for short) of Δ if E (resp. F) is a σ extension of $AF = (X,Y)$ where $X = A \cup cl_D(S_a)_A$ (resp. $X = A \cup cl_D(S_b)_A$), $Y = ((X \times X) \cap R_D) \setminus \{(p \to q) \mid \neg B_a^t(p \to q) \in cl_D(S_a)_R\}$ (resp. $Y = ((X \times X) \cap R_D) \setminus \{(p \to q) \mid \neg B_b^t(p \to q) \in cl_D(S_b)_R\}$), $t \in T$, and $\sigma \in \{co, st, pr, gr\}$.

Example 3.2 Consider a dialogue $\Delta = (\Gamma_a, \Gamma_b)$ with $\Gamma_a = (AF, \{B_a^1(p_a \to q_b), B_a^1 p_a\})$ and $\Gamma_b = (AF, \{B_b^1(p_a \to q_b), B_b^0 q_b, B_b^1 p_a\})$ where $AF = (\{p_a, q_b\}, \{(p_a, q_b)\})$. In this dialogue, b first makes an argument q_b at $t = 0$ and she believes it ($B_b^0 q_b$). Next, a makes a counter-argument p_a with the attack $p_a \to q_b$ at time $t = 1$, and he believes them ($B_a^1(p_a \to q_b)$ and $B_a^1 p_a$). b also believes the argument p_a and the attack $p_a \to q_b$ at $t = 1$ ($B_b^1(p_a \to q_b)$ and $B_b^1 p_a$). Then the belief state of each agent is computed as follows.
(1) $B_a^1 p_a$ and $B_a^1(p_a \to q_b)$ imply $\neg B_a^2 q_b$ by (**BC**).
(2) $B_b^1 p_a$ and $B_b^1(p_a \to q_b)$ imply $\neg B_b^2 q_b$ by (**BC**).
(3) $B_b^0 q_b$ implies $B_b^1 q_b$ by (**IR**).
(4) $B_a^1 p_a$ and $B_b^1 p_a$ respectively imply $B_a^2 p_a$ and $B_b^2 p_a$ by (**IR**).
(5) $B_a^1(p_a \to q_b)$ and $B_b^1(p_a \to q_b)$ respectively imply $B_a^2(p_a \to q_b)$ and $B_b^2(p_a \to q_b)$ by (**IR**).
(6) $B_b^1 q_b$ does *not* imply $B_b^2 q_b$ by (**IR**) and (2).

As a result, Δ has the σ-DBE (E,F) such that $E = \{p_a, B_a^1 p_a, B_a^2 p_a, \neg B_a^2 q_b\}$ and $F = \{p_a, B_b^0 q_b, B_b^1 q_b, B_b^1 p_a, B_b^2 p_a, \neg B_b^2 q_b\}$ where $\sigma \in \{co, st, pr, gr\}$.

Figure 3 illustrates the belief change of agents in Example 3.2. Note that $B_b^2 q_b$ is not implied in (6) of Example 3.2. This is because $\neg B_b^2 q_b$ is in $cl_D(S_b)$ by (2), so that $\{B_b^2 q_b\} \cup cl_D(S_b)$ is inconsistent.

When an agent gives credit to an argument made by another agent, lying or bluffing will succeed to *deceive* the other.

Example 3.3 Consider a dialogue $\Delta = (\Gamma_a, \Gamma_b)$ with $\Gamma_a = (AF, \{B_a^1(p_a \to q_b), \neg B_a^1 p_a\})$ and $\Gamma_b = (AF, \{B_b^1(p_a \to q_b), B_b^0 q_b, B_b^1 p_a\})$. In this dialogue, Γ_b is the same AFB as in

Example 3.2, while a disbelieves his argument p_a in Γ_a. Then the belief state of each agent is computed as follows.
(1) $B^1_b p_a$ and $B^1_b(p_a \to q_b)$ imply $\neg B^2_b q_b$ by (**BC**).
(2) $B^0_b q_b$ implies $B^1_b q_b$ by (**IR**).
(3) $\neg B^1_a p_a$ and $B^1_b p_a$ respectively imply $\neg B^2_a p_a$ and $B^2_b p_a$ by (**IR**).
(4) $B^1_a(p_a \to q_b)$ and $B^1_b(p_a \to q_b)$ respectively imply $B^2_a(p_a \to q_b)$ and $B^2_b(p_a \to q_b)$ by (**IR**).
(5) $B^1_b q_b$ does *not* imply $B^2_b q_b$ by (**IR**) and (1).

As a result, Δ has the σ-DBE (E,F) such that $E = \{\neg B^1_a p_a, \neg B^2_a p_a, q_b\}$ and $F = \{p_a, B^0_b q_b, B^1_b q_b, B^1_b p_a, B^2_b p_a, \neg B^2_b q_b\}$ where $\sigma \in \{co, st, pr, gr\}$.

Comparing the result of Example 3.3 with Example 3.1(4), b believes p_a and disbelieves q_b at time $t = 2$. As a result, b accepts the argument p_a and a successfully deceives b by lying. By contrast, a disbelieves p_a, so he does not accept p_a but accepts q_b.

4. Inner Conflict

In Section 2 we introduce the notion of a rational AFB that has a consistent set of beliefs $cl_{AT}(S)$. It may happen, however, that an agent has inconsistent beliefs on arguments and attacks. Such a situation is given as Γ_5 in Example 2.1(5) in which an agent has the belief $\{Bp, Bq, B(p \to q)\}$ that is inconsistent under the axiom (**AT**). Γ_5 has the four stable/preferred belief extensions that conflict with each other, while each extension is an alternative consistent set of beliefs of an agent.

In this section, we represent an inner conflict of beliefs of an agent such that "an agent believes that he believes an argument that in fact he does not believe" ($BBp \wedge \neg Bp$) or "an agent believes that he disbelieves an argument that in fact he believes" ($B\neg Bp \wedge Bp$). For example, suppose a patient who is advised by a doctor that he is alcohol dependent. Then the former represents a situation that the patient believes that he believes the fact that in fact he disbelieves, while the latter represents a situation that the patient believes that he disbelieves the fact that in fact he believes. Such belief states are known as *self-deception* [5,6]. Self-deception is captured as an AFB in which beliefs (and beliefs over beliefs) of an agent are conflicting.

To represent such belief states of an agent, we consider second-order nested beliefs. Given an argumentation framework $AF = (A,R)$, the set $\mathscr{NB}_{AF}$ of *nested belief atoms over AF* is defined as $\mathscr{NB}_{AF} = \{\, BBp, B\neg Bp \mid p \in A \,\}$. Define $\mathscr{B}^N_{AF} = \mathscr{B}_{AF} \cup \mathscr{NB}_{AF}$. For simplicity, we do not consider negation of nested beliefs like $\neg BBp$ or $\neg B\neg Bp$, nor nested beliefs over attacks like $(\neg)BB(p \to q)$ or $(\neg)B\neg B(p \to q)$. As before, nested beliefs are handled as atoms.

AFBs and attacks over beliefs are extended to handle nested beliefs.

Definition 4.1 (AF with nested belief) Let $AF = (A,R)$ be an argumentation framework. Then, *AF with nested belief (or AFNB)* is defined as a triple $\Lambda = (A,R,S)$ where $S \subseteq \mathscr{B}^N_{AF}$. Λ is often written as (AF,S).

Definition 4.2 (attacks over nested beliefs) Let $AF = (A,R)$ be an argumentation framework. Then, define $R_{NB} = R_B \cup \{\, (BBp, B\neg Bp), (B\neg Bp, BBp) \mid p \in A \,\}$ where R_B is the set defined in Definition 2.2.

In R_{NB}, $BBp \leftrightarrow B\neg Bp$ represents that it does not happen that an agent simultaneously believes both Bp and $\neg Bp$.

Next we introduce two axioms for nested beliefs.

Definition 4.3 (introspection axioms) Let p be an argument. Then, define

(PI) : $Bp \supset BBp$,
(NI) : $\neg Bp \supset B\neg Bp$.

(**PI**) and (**NI**) are called the *introspection axioms*.

Definition 4.4 ($cl_{APN}(S)$) Given $S \subseteq \mathscr{B}_{AF}$, define $cl_{APN}(S) \subseteq \mathscr{B}^{N}_{AF}$ as the smallest set of (nested) belief atoms satisfying the following conditions:

1. $S \subseteq cl_{AT}(S) \subseteq cl_{APN}(S)$.
2. If $Bp \in cl_{APN}(S)$, then $BBp \in cl_{APN}(S)$.
3. If $\neg Bp \in cl_{APN}(S)$, then $B\neg Bp \in cl_{APN}(S)$.

The set $cl_{APN}(S)$ is *consistent* if $\{B\alpha, \neg B\alpha\} \not\subseteq cl_{APN}(S)$ where α is either an argument or an attack. An AFNB $\Lambda = (A,R,S)$ is *rational* if $cl_{APN}(S)$ is consistent. Given $AF = (A,R)$, define $cl_{APN}(S)_A = cl_{APN}(S) \cap \{ Bp, \neg Bp, BBp, B\neg Bp \mid p \in A \}$ and $cl_{APN}(S)_R = cl_{APN}(S) \cap \{ B(p \to q), \neg B(p \to q) \mid (p \to q) \in R \}$.

$cl_{APN}(S)$ represents a set of (nested) belief atoms closed under the application of axioms (**AT**), (**PI**), and (**NI**).

Definition 4.5 (nested belief extension) Let $\Lambda = (A,R,S)$ be an AFNB. Then, a set E is a σ *nested belief extension* (or σ-NBE for short) of Λ if E is a σ extension of $AF = (X,Y)$ where $X = A \cup cl_{APN}(S)_A$, $Y = ((X \times X) \cap R_{NB}) \setminus \{(p \to q) \mid \neg B(p \to q) \in cl_{APN}(S)_R\}$, and $\sigma \in \{co, st, pr, gr\}$.

Example 4.1 Let $AF = (\{p\}, \varnothing)$ and $\sigma \in \{co, st, pr, gr\}$. Then,
(1) $\Lambda_1 = (AF, \{Bp\})$ has the σ-NBE $E_1 = \{p, Bp, BBp\}$.
(2) $\Lambda_2 = (AF, \{\neg Bp\})$ has the σ-NBE $E_2 = \{\neg Bp, B\neg Bp\}$.
(3) $\Lambda_3 = (AF, \{Bp, \neg Bp\})$ has the four stable (or preferred) NBEs: $E_3 = \{p, Bp, BBp\}$, $E_4 = \{p, Bp, B\neg Bp\}$, $E_5 = \{\neg Bp, BBp\}$, and $E_6 = \{\neg Bp, B\neg Bp\}$.

Three different AFNBs of Example 4.1 are illustrated in Figure 4. In all three AFNBs, Bp produces BBp by (**PI**) and $\neg Bp$ produces $B\neg Bp$ by (**NI**), where BBp and $B\neg Bp$ attack each other. Λ_1 and Λ_2 are the cases where no inner conflict arises. In Λ_3, on the other hand, an agent has conflicting beliefs Bp and $\neg Bp$, and there are consistent combinations of beliefs and nested beliefs as four stable (or preferred) NBEs. Of which, E_4 and E_5 represent inner conflicts of an agent and they correspond to the belief states of self-deception $Bp \wedge B\neg Bp$ and $\neg Bp \wedge BBp$ introduced at the beginning of this section. By definition, self-deception may arise when an agent has an inconsistent belief over AF.

Proposition 4.1 *If a rational AFNB $\Lambda = (A,R,S)$ has a σ-NBE E, then $\{BBp, \neg Bp\} \not\subseteq E$ and $\{B\neg Bp, Bp\} \not\subseteq E$ where $\sigma \in \{co, st, pr, gr\}$.*

Proof: When $\Lambda = (A,R,S)$ is rational, $\neg Bp \in cl_{APN}(S)$ implies $Bp \notin cl_{APN}(S)$ thereby $BBp \notin cl_{APN}(S)$. Hence, $\{BBp, \neg Bp\} \not\subseteq E$. $\{B\neg Bp, Bp\} \not\subseteq E$ is shown similarly. □

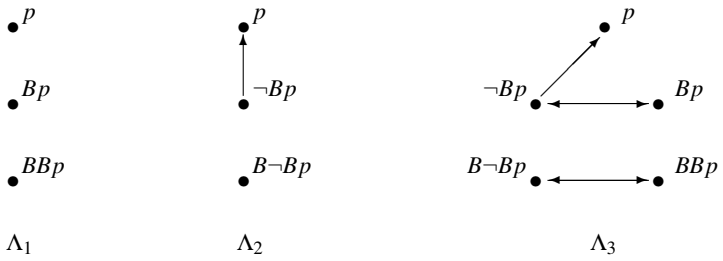

Figure 4. AFNBs in Example 4.1

5. Related Work

There are several studies targeting at arguments and beliefs. The studies [2,7] introduce a probabilistic semantics for abstract argumentation that assigns probabilities or degrees of belief to individual arguments. Given $AF = (A,R)$, introduce a probability function $P : 2^A \to [0,1]$ that assigns a probability to each extension E of AF. Then, for each $a \in A$, the probability $P(a)$ is defined as $P(a) = \sum_{a \in E \subseteq A} P(E)$ that represents the degree of belief that an argument a is in an extension of AF. In this setting, they introduce the notion of *epistemic extension* of P as the set S of arguments such that $P(a) > 0.5$ for any $a \in S$. They investigate conditions of probability functions that agree with intuition on the interrelationships of arguments and attacks. The approach based on probabilities provides a quantitative representation of beliefs over arguments, which is in contrast to our qualitative approach specifying beliefs by formulas.

Epistemic argumentation framework (EAF) [8] encodes the belief of an agent who reasons about arguments. An EAF is represented as a pair (AF, φ) where AF is an abstract argumentation framework and φ is an epistemic formula called an *epistemic constraint*. An EAF (AF, φ) represents the view of an agent who believes that φ is true in AF. The semantics of an EAF is given by an *ω-epistemic labelling set SL* where each $S \in SL$ is an ω-labelling of AF with $\omega \in \{co, st, gr, pr\}$ and SL is a $\subseteq$-maximal set of ω-labellings of AF that satisfy φ. By definition, an ω-epistemic labelling set is a collection of ω-labellings of an AF that reflect the beliefs of an agent. EAF and AFB share a common purpose to encode the belief of an agent on a given AF, while the semantic frameworks are different from each other. In EAF the epistemic formula φ works as an external constraint over the labellings of an AF. Then every element $S \in SL$ is an ω-labelling of the original AF, and it does not contain any epistemic formula. In AFB, on the other hand, beliefs of an agent are introduced to AF and interact with arguments in AF. As a result, σ belief extensions contain belief atoms in general.

Epistemic abstract argumentation framework (EAAF) [9] extends the abstract AF by introducing epistemic arguments and attacks. An epistemic attack from a to b is defined as: a defeats b if a occurs in at least one extension (*strong epistemic attack*) or in all extensions and at least one (*weak epistemic attack*). Arguments defeated through epistemic attacks are called epistemic arguments. The semantics of an EAAF (A, R, Ψ, Φ), where Ψ (resp. Φ) is a set of weak (resp. strong) epistemic attacks, is given by a set W of sets of arguments in A called *world view*. A world view represents epistemically acceptable arguments depending on the condition of their attacking arguments, while EAAF does not handle belief atoms in its framework.

The study [10] introduces a *debate game* between two players in which a player may provide false or inaccurate arguments as a tactic to win the game. A player lies if she makes an argument that she disbelieves, or a player bullshits if he makes an argument on which he has no belief. It investigates situations in which a player has a chance to win a game using (dis)honest arguments and argues the possibility of detecting dishonest arguments. The study formulates debate games in abstract argumentation frameworks, while it does not represent beliefs on arguments/attacks as done in this paper.

In this paper, we introduce dynamic belief extensions representing belief change of agents in a dialogue. Belief revision in argumentation has been studied in the literature [11] in which belief change is represented by the change of extensions over time. In our approach, belief change of an agent is represented in a single extension using belief atoms with time.

6. Conclusion

This paper introduced a framework that can represent interaction between arguments and beliefs. The AFB is used for representing belief states of players and the audience of argumentation. In two-persons dialogue, AFB can distinguish belief states of (in)sincere players. Belief change of a player is represented by dynamic belief extensions that can also model deceptive dialogues. Finally, inner conflicts of an agent are expressed using nested beliefs, and self-deception is realized by belief extensions of AFNB. Consideration on the epistemic aspect of argumentation has been done by several studies, but, to the best of our knowledge, this is the first attempt to formulate the interaction between argument and belief in the context of abstract argumentation frameworks. An interesting research issue is to represent and reason about argument and belief using structured argumentation.

References

[1] 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.

[2] Hunter, A., Thimm, M. Probabilistic reasoning with abstract argumentation frameworks. *Journal of Artificial Intelligence Research* 59:565–611, 2017.

[3] Sakama, C., Caminada, M., Herzig, A. A formal account of dishonesty. *Logic Journal of IGPL* 23:259–294, 2014.

[4] Reiter, R. A logic for default reasoning. *Artificial Intelligence* 13:81–132, 1980.

[5] Hintikka, J. *Knowledge and Belief – an Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca and London, 1962.

[6] Jones, A. J. I. On the logic of self-deception. *South American Journal of Logic* 1:387–400, 2015.

[7] Thimm, M. A probabilistic semantics for abstract argumentation. In: *Proc. 20th European Conference on Artificial Intelligence*, 750–755, 2012.

[8] Sakama, C., Son, T. C. Epistemic argumentation framework: theory and computation. *Journal of Artificial Intelligence Research* 69:1103–1126, 2020.

[9] Alfano, G., Greco, S., Parisi, F., Trubitsyna, I. Epistemic abstract argumentation framework: formal foundations, computation and complexity. In: *Proc. 22nd AAMAS*, 409–417, 2023.

[10] Sakama, C. Dishonest arguments in debate games. In: *Proc. 4th COMMA*, Frontiers in Artificial Intelligence and Applications 245, IOS Press, 177–184, 2012.

[11] Baroni, P., Fermé, E., Giacomin, M., Simari, G. R. Belief revision and computational argumentation: a critical comparison. *Journal of Logic, Language and Information* 31:555–589, 2022.