Beliefs supported by Arguments

Chenwei Shi[*] Sonja Smets[*][†]

Abstract

In this paper we explore the relation between an agent's doxastic attitude and her arguments in support of a given claim. Formally, we build further on Dung's argumentation framework in Dung (1995). We start by introducing a logic to reason about binary arguments which are either in favor or against a certain claim. Next we explore a number of notions from standard argumentation theory in our system, including the attack of an argument, the acceptability of an argument, the conflict-freeness of a set of arguments and its admissibility. Our setting will allow us to define new concepts, indicating when an argument perfectly defends a given claim P or when an argument only strategically defends a given claim P. The concept of strategic defensibility is then used to link an agent's arguments to her doxastic attitude. This setting offers a formal characterization of "argument"-based beliefs. As such we address an issue which was raised but not worked out in Dung (1995).

1 Introduction to the Framework

The literature on argumentation theory contains a number of different views on what an argument is. One of the common views takes "an argument as an attempt to present evidence for a conclusion" (Groarke, 2015). Taking this as the starting point of our approach, we view arguments as entities that support a given claim or proposition P. We view the given claim as the content that the argument supports. To express the content of an argument, we extend Dung's argumentation framework as follows:

Definition 1.1 (Argumentation-Support Frame (ASF))**.** An argumentation-support frame is a structure $\mathfrak{F} = \langle \mathcal{W}, \mathcal{AR}, \{\leftarrowtail^P\}^{P \subseteq \mathcal{W}}, f \rangle$ where

- $\mathcal{W} = \{u, v, w, \dots\}$ is a non-empty set of possible worlds and $\mathcal{AR} = \{s_1, s_2, s_3 \dots\}$ a non-empty set of arguments;
- $\leftarrowtail^P \subseteq \mathcal{AR} \times \mathcal{AR}$ is an attack relation labelled by a proposition $P \in \wp(\mathcal{W})$;
- $f : \mathcal{AR} \to \wp(\mathcal{W})$ assigns to each argument $s \in \mathcal{AR}$ a subset of $\mathcal{W}$ such that for any $s \in \mathcal{AR}$, $f(s) \neq \varnothing$;

This frame satisfies a number of conditions which we list here. For notational simplicity, we write $\mathcal{W} - P$ as $\overline{P}$.

[*] Institute for Logic, Language and Computation, University of Amsterdam

[†] S. Smets acknowledges the support from the European Research Council via ERC Grant no. 283963.

1. if $s_1 \leftarrowtail^P s_2$,
    (a) either $f(s_1) \subseteq P$ or $f(s_1) \subseteq \overline{P}$;
    (b) $f(s_1) \subseteq P$ then $f(s_2) \subseteq \overline{P}$;
2. $s_1 \leftarrowtail^P s_2$ if and only if $s_1 \leftarrowtail^{\overline{P}} s_2$;
3. if $s_1 \leftarrowtail^P s_2$ and $f(s_1) \subseteq Q \subseteq P$, then $s_1 \leftarrowtail^Q s_2$;

Explanation of Notation. Taking subsets of $\mathcal{W}$ as propositions, our notation $f(s) \subseteq P$ expresses that argument s supports proposition P. The condition that $f(s) \neq \varnothing$, which we impose on f, requires that there should not be any argument supporting a contradiction. We use the notation $s_1 \leftarrowtail^P s_2$ to express that argument s_1 is attacked by s_2 on whether P is the case. The idea of working with accessibility relations labelled by propositions is used in the literature on conditional logic and in particular in the context of logics for belief revision (Baltag and Smets, 2006a,b, 2008).

Explanation of Frame Conditions. Whether the attack relation holds between two arguments has something to do with the contents that are supported by the two arguments, as is indicated by the frame conditions in definition 1.1. The first condition says that the attack relation $\leftarrowtail_w^P$ only exists between two arguments which both take a different stands with regard to the claim P. The second condition captures the intuition that a debate about P is the same as a debate about $\overline{P}$. The third condition also specifies the relation between attack relations with respect to different topics. In particular, it says that if one argument is attacked on its claim that P is the case by certain argument, then it would be attacked by the same argument on its stronger claim.

Given the above frame, we build a model by adding a valuation-map V:

Definition 1.2 (Argumentation-Support model)**.** An argumentation-support model is a structure $\mathfrak{M} = \langle \mathcal{W}, \mathcal{AR}, \{\leftarrowtail^P\}^{P \subseteq \mathcal{W}}, f, V \rangle$ where

- $\langle \mathcal{W}, \mathcal{AR}, \{\leftarrowtail^P\}^{P \subseteq \mathcal{W}}, f \}$ is an argumentation-support frame.
- $V : \text{Prop} \to \wp(\mathcal{W})$ assigns a set of possible worlds to each atomic sentence from a given set of atomic propositions Prop.

Our models captures a dual perspective, we can either look at the set of arguments and the claims each of them supports or at the set of possible worlds and the arguments that support the propositions they satisfy. For the purpose of illustration, we formalize the following example in our setting:

Example 1.1. In front of a vague picture of an animal, some people are arguing:

- s_1: The animal in the picture has wings, so it is a bird;
- s_2: The animal looks like a bat, so it is not a bird, it is a mammal;
- s_3: The animal looks like a pterosaur, so it is neither a bird nor a mammal, it is a reptile;

Let $\text{Prop} = \{b, m, r\}$ where b says that the animal in the picture is a bird, m says that the animal in the picture is a mammal and r says that the animal in the picture is a reptile. Define a model $\mathfrak{M} = \langle \mathcal{W}, \mathcal{AR}, \{\leftarrowtail^P\}^{P \subseteq \mathcal{W}}, f, V \rangle$ where

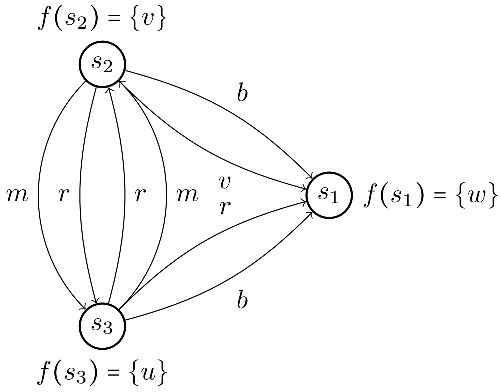

Figure 1.1: Graph for Example 1.1

- $\mathcal{W} = \{w, v, u\}$ and $\mathcal{AR} = \{s_1, s_2, s_3\}$;
- $V(b) = \{w\}, V(m) = \{v\}, V(r) = \{u\}$;
- $\leftarrowtail^{\mathcal{W}} = \leftarrowtail^{\varnothing} = \varnothing$, $\leftarrowtail^{\{w\}} = \leftarrowtail^{\{v,u\}} = \{(s_1, s_2), (s_1, s_3)\}$,
  $\leftarrowtail^{\{v\}} = \leftarrowtail^{\{w,u\}} = \{(s_2, s_3), (s_3, s_2), (s_1, s_2)\}$,
  $\leftarrowtail^{\{u\}} = \leftarrowtail^{\{w,v\}} = \{(s_2, s_3), (s_3, s_2), (s_1, s_3)\}$;
- $f(s_1) = \{w\}, f(s_2) = \{v\}, f(s_3) = \{u\}$.

Readers can check that this model is an argumentation-support model. (Also see the graph in Figure 1.1 for a more intuitive representation.) Note that the attack relation on m forms a cycle between s_2 and s_3, which is also the case for attack relation on r.

From Arguments to Beliefs. Given the above semantics, we can address the issue raised in Dung (1995) about the connection between an agent's doxastic state concerning P and her arguments in support of P. Dung (1995) does indicate the existence of a substantial relation between the acceptable arguments and the agent's belief formation:

> . . . a statement is believable if it can be argued successfully against attacking arguments. In other words, whether or not a rational agent believes in a statement depends on whether or not the argument supporting this statement can be successfully defended against the counterarguments. (Dung, 1995, p.323)

However, the formal framework in Dung (1995) does not make this relation between the doxastic state (the agent's beliefs) and the agent's arguments explicit. We have set it as our goal to address this question, namely how can the agent's beliefs be characterized on the basis of the agent's argumentation structure.

This question is in part also addressed in the work by Grossi and van der Hoek (2014), though our approaches are different. In Grossi and van der Hoek (2014), the authors start from a two-dimensional doxastic-argumentation structure and explore the connection between the two while our goal is to characterize the notion of belief directly in terms of the properties of the underlying arguments while assuming that the agent has an epistemic information state. In this way our setting resembles the semantic approach on evidence-based beliefs in van Benthem et al. (2012), we approach the topic of what we call 'argument-based beliefs' by taking our argument-support structures as the basic building block.

Without going into all the details of our formal setting we offer the main definitions and propositions of the argumentation theoretical notions that play a central role.

2 Formal Properties of Arguments

The following definitions are directly based on the related concepts in Dung (1995):

Definition 2.1 (Conflict-free and admissible set of arguments)**.** Given an ASF $\mathfrak{F}$ and a set of arguments $X \subseteq \mathcal{AR}$,

$$\begin{aligned} n^P(X) &= \{s \in \mathcal{AR} \mid \nexists\, s' \in X : s \leftarrowtail^P s'\} \\ d^P(X) &= \{s \in \mathcal{AR} \mid \forall (s, s') \in \leftarrowtail^P \exists s'' \in X : (s', s'') \in \leftarrowtail^P\}. \end{aligned}$$

X is conflict free for P if $X \subseteq n^P(X)$; X is admissible for P if $X \subseteq n^P(X)$ and $X \subseteq d^P(X)$.

Definition 2.2 (Defenders and attackers of an argument)**.** Given an ASF $\mathfrak{F}$ and an argument $s \in \mathcal{AR}$,

- $Def^P(s) = \{s_n \in \mathcal{AR} \mid \exists s_0, s_1, \ldots, s_n : \bigwedge_{i=0}^{n} s_i \leftarrowtail^P s_{i+1}$ where $s = s_0$ and $n \neq 0$ is an even number $\}$
- $Att^P(s) = \{s_n \in \mathcal{AR} \mid \exists s_0, s_1, \ldots, s_n : \bigwedge_{i=0}^{n} s_i \leftarrowtail^P s_{i+1}$ where $s = s_0$ and n is an odd number $\}$

Based on these notions, we define the main notion in this paper, which focuses on the status of one single argument rather than on a set of arguments.

Definition 2.3 (Defensibility)**.** Given an ASF $\mathfrak{F}$ and an argument $s \in \mathcal{AR}$,

1. s perfectly defends P if and only if $f(s) \subseteq P$ and $\{s\} \cup Def^P(s)$ is admissible for P;
2. s strategically defends P if and only if $f(s) \subseteq P$ and there is a subset S of $Def^P(s)$ such that $\{s\} \cup S$ is admissible for P.

Having assigned a content to each argument, we can ask how the status of one argument changes according to the change of debated claim.

Proposition 2.1. *Given an ASF $\mathfrak{F}$ and an argument $s \in \mathcal{AR}$,*

- *if s strategically defends P, then s strategically defends Q where $P \subseteq Q$;*
- *even if s perfectly defends P, s may not perfectly defend Q where $P \subseteq Q$;*
- *even if s perfectly/strategically defends P and also perfectly/strategically defends Q, s may not perfectly/strategically defend $P \cap Q$*

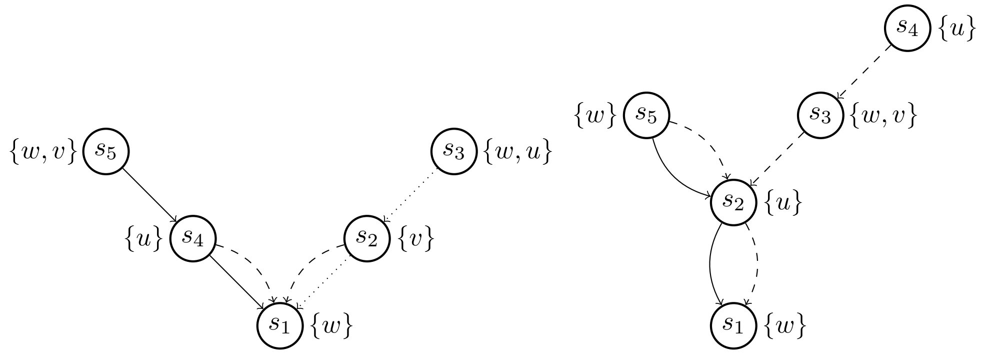

Figure 2.1: Counterexample for Proposition 2.1

Readers can check that in the graph on the right-hand side of Figure 2.1, s_1 perfectly defends $P = \{w\}$ but does not perfectly defend $Q = \{w,v\}$ (the solid arrow represents $\twoheadleftarrow^P$ and the dashed arrow represents $\twoheadleftarrow^Q$). However, s_1 can strategically defend Q. And in the graph of the left-hand side of Figure 2.1, s_1 perfectly defends $P = \{w,u\}$ (dotted arrow for $\twoheadleftarrow^P$) and $Q = \{w,v\}$ (solid arrow for $\twoheadleftarrow^Q$), but does not strategically defend $P \cap Q$ (dashed arrow for $P \cap Q$).

We say that P perfectly/strategically dominates a debate about P, if the claim P is perfectly/strategically defended by some argument s while $\overline{P}$ is not.

Definition 2.4 (Domination)**.** Given an ASF $\mathfrak{F}$, P perfectly/strategically dominates a debate about P if and only if there exists at least one argument which perfectly/strategically defends P but no argument which perfectly/strategically defends $\overline{P}$.

Although perfect defensibility implies strategic defensibility, strategic domination implies perfect domination:

Proposition 2.2. *Given an ASF $\mathfrak{F}$, if P strategically dominates, then P perfectly dominates, but not the other way around.*

Corresponding to Proposition 2.1, we have the following result for domination:

Proposition 2.3. *Given ASF $\mathfrak{F}$,*

- *if P strategically dominates, then Q strategically dominates where $P \subseteq Q$; however, this does not hold for perfect domination;*

- *even if both P and Q perfectly/strategically dominates, $P \cap Q$ may not perfectly/strategically dominate.*

Proposition 2.2 and Proposition 2.3 indicate that the strategic notions are better behaved. In order to characterize an agent's doxastic attitude on the basis of her arguments, we will work further with the notions of strategic defensibility and domination. Let us first offer the reader an important alternative characterization of the notion of 'strategic defensibility', based on the use of fixed point (in line with the use of greatest fixed point in van Benthem (2014) for defining solution concepts in strategic game-theoretic contexts):

Proposition 2.4. *Given an ASF $\mathfrak{F}$ and $s \in \mathcal{AR}$, there is a subset S of $Def^P(s)$ such that $\{s\} \cup S$ is admissible on P if and only if $s \in GFP.d^P$, where $GFP.d^P$ is the greatest fixed point of function d^P.*

So s strategically defends P if and only if $f(s) \subseteq P$ and $s \in GFP.d^P$. In the next section, the syntax and its truth conditions are proposed to reason about these strategic notions, which is inspired by the use of the two-dimensional logic and modal μ-calculus for studying argumentation theory in Grossi (2010, 2012); Grossi and van der Hoek (2014).

3 Logic for Argument-Based Beliefs

The syntax is given as follows

Definition 3.1. Let Prop = $\{p, q, r, \dots\}$ be a non-empty set of atomic propositions. $\mathcal{L}$ is the language generated by the following grammar:

$$\alpha ::= \top \mid p \mid \neg\alpha \mid \alpha \wedge \alpha \mid \boxminus\alpha \mid \boxbar\beta$$
$$\beta ::= \top \mid \Box\alpha \mid \neg\beta \mid \beta \wedge \beta \mid [\leftarrowtail^{\alpha}]\beta \mid \mathsf{Gfp}^{\alpha}$$

where $p \in$ Prop. The duals of the operators are defined as standard, such as $\Diamondbar$ for $\neg \boxbar \neg$ and $<\leftarrowtail^{\alpha}>$ for $\neg[\leftarrowtail^{\alpha}]\neg$.

The language is divided into two parts α and β. The α part is used to state facts about possible worlds, while β part is dedicated to the description of arguments. For example, the intuitive reading of $\Box\alpha$ says that the current argument supports α; $[\leftarrowtail^{\alpha}]\beta$ says that with respect to the debate about the proposition α: for all arguments directly attacking the current argument, β is the case; Gfp^{α} indicates that the current argument is acceptable in the sense of being in the the greatest fixed point of the function d with respect to α. $\boxbar$ is a universal operator ranging over the whole set of arguments, which is used to express whether arguments satisfying certain properties exist. And $\boxminus$ is the universal operator ranging over the whole set of possible worlds, which can be interpreted as a knowledge operator in this setting. We will call formulas that belong to the α part (β part) of this language α-formula (β-formula). When there is no need to make a distinction, φ is used to denote formulas in the whole language $\mathcal{L}$. The truth conditions follow below:

Let $\mathfrak{M}$ be an argumentation-support model, $[\![\alpha]\!]_{\mathfrak{M}} = \{w \in \mathcal{W} \mid \mathfrak{M}, (w, s) \vDash \alpha\}$. We omit the subscript $\mathfrak{M}$ whenever its use is clear from the context.. The truth of $\varphi \in \mathcal{L}$ is defined as follows:

Definition 3.2. Given an argumentation-claim model $\mathfrak{M}$,

- $\mathfrak{M}, (w, s) \vDash \top$
- $\mathfrak{M}, (w, s) \vDash p$ **iff** $w \in V(p)$
- $\mathfrak{M}, (w, s) \vDash \neg\alpha$ **iff** $\mathfrak{M}, (w, s) \nvDash \alpha$
- $\mathfrak{M}, (w, s) \vDash \alpha \wedge \alpha'$ **iff** $\mathfrak{M}, (w, s) \vDash \alpha$ and $\mathfrak{M}, (w, s) \vDash \alpha'$
- $\mathfrak{M}, (w, s) \vDash \boxminus\alpha$ **iff** for all $w' \in \mathcal{AR}, \mathcal{M}, (w', s) \vDash \alpha$

- $\mathfrak{M}, (w, s) \vDash \boxbar\beta$ **iff** for all $s' \in \mathcal{AR}, \mathcal{M}, (w, s') \vDash \beta$
- $\mathfrak{M}, (w, s) \vDash \Box\alpha$ **iff** $f(s) \subseteq [\![\alpha]\!]$
- $\mathfrak{M}, (w, s) \vDash \neg\beta$ **iff** $\mathfrak{M}, (w, s) \nvDash \beta$
- $\mathfrak{M}, (w, s) \vDash \beta \wedge \beta'$ **iff** $\mathfrak{M}, (w, s) \vDash \beta$ and $\mathfrak{M}, (w, s) \vDash \beta'$
- $\mathfrak{M}, (w, s) \vDash [\leftarrowtail^{\alpha}]\beta$ **iff** for any $s' \in \mathcal{AR}$ such that $s \leftarrowtail^{[\![\alpha]\!]} s', \mathfrak{M}, (w, s') \vDash \beta$.
- $\mathfrak{M}, (w, s) \vDash \mathsf{Gfp}^{\alpha}$ **iff** $s \in GFP.d^{[\![\alpha]\!]}$

Characterization of Strategic Defensibility and Domination. In this setting, strategic defensibility and domination can be defined:

$$\mathsf{Str}\,\alpha := \Box\,\alpha \wedge \mathsf{Gfp}^{\alpha}$$
$$\mathsf{Dom}\,\alpha := \diamondbar\,\mathsf{Str}\,\alpha \wedge \neg\,\diamondbar\,\mathsf{Str}\,\neg\alpha$$

Note that $\mathsf{Str}\,\alpha$ is a β-formula, while $\mathsf{Dom}\,\alpha$ is an α-formula.

Characterization of Belief. We propose to use $\mathsf{Dom}\,\alpha$ for the definition of the agent's belief:

$$B\alpha := \mathsf{Dom}\,\alpha.$$

According to Proposition 2.3, this belief operator is only closed upwards, i.e.

$$B\varphi \to B(\alpha \vee \alpha')$$

is valid in the class of argumentation-support frames, but it is not closed under conjunction. Moreover, in this setting,

$$B\alpha \to BB\alpha \text{ and}$$
$$\neg B\alpha \to B\neg B\alpha$$

are valid, since if $B\alpha$ ($\neg B\alpha$) is true somewhere in $\mathcal{W}$ then $[\![B\alpha]\!] = \mathcal{W}$ ($[\![\neg B\alpha]\!] = \mathcal{W}$). Recall that $\leftarrowtail^{\mathcal{W}}$ is empty, so the greatest fixed point of $d^{\mathcal{W}}$ is $\mathcal{W}$. Together with the fact that $[\![\Box\top]\!] = \mathcal{W}$, it implies that $\boxbar\,\mathsf{Str}\,B\alpha$ ($\boxbar\,\mathsf{Str}\,\neg B\alpha$).

We have shown that in the above syntax, the strategic notions can be expressed and and we have given an epistemic interpretation to these notions. The way we define belief based on arguments is closely related to the way van Benthem et al. (2012) define belief based on evidence. However, the logical properties of belief defined in this setting turn out to be different from those in van Benthem et al. (2012), where belief is not only closed upwards but also closed under conjunction. Moreover, $\neg B\bot$ is not valid in the evidence logic (van Benthem et al., 2012), while it is valid in our setting. Considering only these properties, our notion of 'argument based belief' has interesting features in common with certain concepts that are studied in the context of probabilistic belief (Halpern, 2003).

Future Outlook. In the full version of this paper, we provide an axiomatization for our logical system and show how this setting can be extended to a more general multi-agent system, in which we can reason about debates between agents and the role of agents' epistemic states in these debates.

References

Baltag, A. and Smets, S. (2006a). Conditional doxastic models: A qualitative approach to dynamic belief revision. In Mints, G. and de Queiroz, R., editors, *Proceedings of WOLLIC 2006, Electronic Notes in Theoretical Computer Science*, volume 165, pages 5–21.

Baltag, A. and Smets, S. (2006b). Dynamic belief revision over multi-agent plausibility models. In G. Bonanno, W. van der Hoek, M. W., editor, *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision (LOFT 2006)*, pages 11–16.

Baltag, A. and Smets, S. (2008). A qualitative theory of dynamic interactive belief revision. *Texts in logic and games*, 3:9–58.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77:321–357.

Groarke, L. (2015). Informal logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2015 edition.

Grossi, D. (2010). On the logic of argumentation theory. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*.

Grossi, D. (2012). Fixpoints and iterated updates in abstract argumentation. In *Proceedings of the 13th International Conferenceon Principles of Knowledge Representation and Reasoning*.

Grossi, D. and van der Hoek, W. (2014). Justified beliefs by justified arguments. In *Proceedings of the Fourteenth International Conference on Principles of Knowledge Representation and Reasoning*.

Halpern, J. Y. (2003). *Reasoning about Uncertainty*. MIT press.

van Benthem, J. (2014). *Logic in Games*. MIT press.

van Benthem, J., Fernández-Duque, D., and Pacuit, E. (2012). Evidence logic: A new look at neighborhood structures. In Bolander, T., Braüner, T., Ghilardi, S., and Moss, L., editors, *Proceedings of Advances in Modal Logic Volume 9*, volume 9, pages 97 – 118. King's College Press.