Formal Argumentation and Modal Logic

Carlos Iván Chesñevar
Universidad Nacional del Sur & Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
cic@cs.uns.edu.ar

Jürgen Dix
Technische Universität Clausthal, Germany
dix@tu-clausthal.de

Beishui Liao
Zhejiang University (ZJU), China
baiseliao@zju.edu.cn

Jieting Luo
Zhejiang University (ZJU), China
luojieting@zju.edu.cn

Carlo Proietti
Consiglio Nazionale delle Ricerche (CNR), Italy
carlo.proietti@ilc.cnr.it

Antonio Yuste-Ginel
Universidad Complutense de Madrid (UCM), Spain
antoyust@ucm.es

Abstract

The interrelationship between defeasible argumentation and modal logic is rooted in their shared goal of capturing and modelling reasoning under uncertainty and changing conditions. In the last years, researchers have explored different ways to combine these two formalizations to create more robust systems for handling complex reasoning tasks, in which modal operators can be incorporated into argumentation systems.

Vol. 12 No. 3 2025
Journal of Applied Logics — IfCoLog Journal of Logics and their Applications

In this article we analyse three different lines of work to combine modal logic and argumentation: a) a logic-based framework that combines dynamic logic and argumentation for value-based planning; b) alternating-time temporal logic extended with coalitional argumentation; c) different combined approaches for integrating epistemic logics and argumentation. These three alternatives will help the reader to understand different interplays that can take place when combining argumentation and modal logic. On the one hand, we show that argumentation systems can be combined with very different readings of modal operators (i.e., dynamic, temporal and epistemic). On the other hand, modal logic and argumentation can be used in different relative positions. When representing and reasoning about plans, modal logic is applied for the reasoning on the object level and a structured argumentation framework is built on the meta-level over modal logic. When epistemically reasoning about opponents' argumentative information, modal logic can be built over argumentation. For checking the strategic properties of coalitions of agents, argumentation is put inside modal logic so that the coalition can enlarge according to the theory of coalitional argumentation.

1 Introduction

While formal argumentation captures diverse kinds of reasoning and dialogue activities with uncertainty and conflicting information, modal logic plays a major role in philosophy and related fields as a tool for understanding and reasoning about concepts such as knowledge, obligation, time, and actions. The combination of argumentation and modal logic has been of interest for some time [8]. As we can find a large and heterogeneous body of literature on this subject, the first thing is to propose a criterion according to which one can divide and categorise the different works.

One possible such criterion is classifying the different approaches according to the *modal operators* that they use (e.g., temporal, epistemic, dynamic, etc). The most direct way in which one can relate argumentation and modal logic is by noting that an abstract argumentation framework [41] is nothing but a Kripke frame (that is, a basic semantic structure in modal logic). A natural research enterprise is then to use modal techniques to study abstract argumentation. This was done in a series of papers by Davide Grossi and more people, e.g. [33; 53; 54; 55; 47; 52; 56] (and cf. [33, Sect. 4.4] for an alternative approach following the same basic idea). In these approaches, the attack relation becomes an accessibility relation that one can use to interpret different modalities, and different types of extensions can then be defined in modal-logic-based languages. This is what might be called an *argumentative approach* to the combination of formal argumentation and modal

logic, and it is well studied in [8, Sect. 3.2].

In a different vein, one can use a non-argumentative interpretation of modal operators for increasing the original expressive power of argumentation systems, so as to jointly reason about argumentation and some other relevant cognitive dimension. Thus, for instance, the use of modalities can be *temporal*. In [28], Alternating-time Temporal Logic is used to reason about what properties a coalition can enlarge to enforce, extended with argumentation to provide how this very coalition is formed. Yet another usual interpretation of modal operators, *dynamic logic*, consists in understanding that they quantify over possible executions of programs or, more in general, over actions. In [72], the arrows of the underlying modal structure are associated with actions that an artificial agent may execute. Moreover, each arrow is also possibly associated with a set of values that they promote/demote. Argumentation frameworks are then used in their value-based version to let the agent decide what is the best available plan to reach a given goal according to her value scale. Finally, there is a branch of the literature that works on the combination of formal argumentation and *epistemic logic*, where we can distinguish two main lines of work. The first line concerns using arguments to determine beliefs. This is the main intuition underlying a series of papers [87; 90; 92; 91; 89], which focus on an argumentative extension of topological epistemic models. In a different technical setting, [30; 31] syntactically capture the relation between argumentation and belief using awareness epistemic logic and ASPIC$^+$ arguments. The main idea of the second line is to have beliefs about my opponent's argumentative information, which plays a crucial role in the choice of my moves during a dialogue. It was first treated from an epistemic logic perspective in the work of Schwarzentruber et al. [86], and later on in [81; 82]. Using these epistemic-argumentative models, we are able to reason about higher-order and unquantified uncertainty about argumentation frameworks, which is in turn a key feature in strategical settings for argumentation.

Another criterion that we can employ for categorising the combination of argumentation and modal logic is their relative positions. Modal logic can be used for reasoning on the object level, while structured argumentation can be built on the meta-level over modal logic. As an example, in [72], arguments for plan selection are constructed using a planner agent's beliefs that are expressed in modal logic. In contrast, modal logic can be built over argumentation. For example, [86] and [81; 82] see argumentative information as the object that agents reason about using modal logic. Different from the above two ways, [28] puts argumentation inside modal logic: given a coalition of agents, the framework can be used to check its strategic properties, allowing the coalition to be enlarged according to the theory of coalitional argumentation.

The rest of this article is organised as follows. We provide the minimal necessary

background on modal logic and abstract argumentation frameworks in Section 2. Then, in Sections 3, 4 and 5 we cover the three different combinations of modal operators and argumentation systems that we have just introduced (dynamic operators, temporal operators and epistemic operators, respectively). Finally, Sections 6 and 7 close the paper by giving some pointers to further literature on the topic and sketching out current trends and challenges at the intersection between modal logic and formal argumentation.

2 Formal preliminaries

In this section, we will present definitions for different core concepts in modal logic and argumentation, which are shared to some extent by the formalisms introduced in later sections. Other concepts (such as more specific semantics and modalities) defined later can be related to these core concepts.

2.1 Modal logic

Here we provide the general definitions of the syntactic and semantic notions for modal logic that we will use in the rest of the article. We are going to work with different interpretations of modal logic and, therefore, with different interpretations of modalities. Hence, to keep things abstract enough, we assume as given a finite set of *generic labels* $\mathbb{L} = \{l_1, ..., l_n\}$. Depending on the context of application, elements of $\mathbb{L}$ may denote actions, action profiles, agents or sets of agents. Moreover, we assume as fixed from now on a denumerable set of *atomic propositions* $\Phi = \{p, q, ...\}$.

Definition 2.1 (Labelled multi-modal language)**.** *The language* $\mathcal{L}(\Phi, \mathbb{L})$ *is given by the following BNF*

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box_l\varphi \qquad p \in \Phi, l \in \mathbb{L}.$$

That is, the language for propositional logic enriched with a set of $\Box_l$-modalities. We will often employ $\Diamond_l$ as the dual of $\Box_l$, defined as $\neg\Box_l\neg$ (and we sometimes take $\Diamond_l$ as the primitive operator and define $\Box_l$ as $\neg\Diamond_l\neg$ instead). The rest of Boolean operators are defined as usual using $\wedge$ and $\neg$. $\mathcal{L}(\Phi)$ denotes the propositional fragment of $\mathcal{L}(\Phi, \mathbb{L})$ (i.e., the result of dropping the clause $\Box_l$ from the previous grammar).

A multi-modal language of this kind is typically interpreted on a *labelled transition system* (a.k.a. a *multi-relational Kripke frame*), defined as follows.

Definition 2.2 (Labelled Transition Systems and Models)**.** *A labelled transition system over $\mathbb{L}$ is a tuple $T = \langle S, \mathcal{R} \rangle$, where*

- *S is a finite, non-empty set of states; and*
- *$\mathcal{R} \subseteq S \times \mathbb{L} \times S$ is a transition relation between states labelled with elements of $\mathbb{L}$. We use $\mathcal{R}_l$ to denote the relation $\{\langle x, y \rangle \in S^2 \mid \langle x, l, y \rangle \in \mathcal{R}\}$.*

Further, a model (over Φ and $\mathbb{L}$) –sometimes called an interpreted labelled transition system or, more extensively, a multi-relational Kripke model– is defined as a pair $\mathcal{M} = \langle T, \mathcal{V} \rangle$ where:

- *T is a labelled transition system; and*
- *$\mathcal{V}$ is a propositional valuation $\mathcal{V} : S \to 2^{\Phi}$ that assigns each state s with the subset of atomic propositions which are true at state s; thus for each $s \in S$ we have $\mathcal{V}(s) \subseteq \Phi$.*

For notational convenience, we sometimes unravel the content of T and write $\mathcal{M} = \langle S, \mathcal{R}, \mathcal{V} \rangle$. Here again, the labels may stand for different kinds of transition relations, e.g. $\langle s, t \rangle \in \mathcal{R}_l$ may denote the execution of an action changing the system from state s to state t. Or else, when l stands for an agent, it may indicate that agent l considers t as an alternative to s.

Once we define a valuation for propositional atoms, the next fundamental step is to define the full notion of truth with respect to satisfaction relation $\models$, more precisely to specify under which conditions a given formula φ is true at a given state s in a model $\mathcal{M}$ (denoted $\mathcal{M}, s \models \varphi$). The following definition does it in a recursive way.

Definition 2.3 (Truth)**.** *Formulas of the labelled multi-modal language are interpreted in pointed models recursively as follows:*

$$
\begin{array}{lll}
\mathcal{M}, s \models p & \textit{iff} & p \in \mathcal{V}(s) \\
\mathcal{M}, s \models \neg\varphi & \textit{iff} & \mathcal{M}, s \not\models \varphi \\
\mathcal{M}, s \models (\varphi \wedge \psi) & \textit{iff} & \mathcal{M}, s \models \varphi \textit{ and } \mathcal{M}, s \models \psi \\
\mathcal{M}, s \models \Box_l\varphi & \textit{iff} & \textit{for all } t \in S, \textit{ } s\mathcal{R}_l t \textit{ implies } \mathcal{M}, t \models \varphi
\end{array}
$$

We say that a formula φ is valid in a model $\mathcal{M} = \langle S, \mathcal{R}, \mathcal{V} \rangle$ iff $\mathcal{M}, s \models \varphi$ for every $s \in S$, and that a formula φ is valid in a transition system T iff it is valid in every model $\mathcal{M}$ based on T. Further, φ is valid in a class of transition systems iff it is valid in every element in the class.

Some formulas – more precisely some schemes, i.e. general forms of formula – are valid only in classes of systems where the transition relations satisfy some specific property. In such case we say that a formula φ *defines* the class of frames satisfying this property or, more briefly, that it defines this property. Many such formulas work as *axiom schemes* for different axiomatic calculi of modal logic. For example the general scheme (K) $= \Box_l(\varphi \rightarrow \psi) \rightarrow (\Box_l\varphi \rightarrow \Box_l\psi)$, which is in fact valid in all systems, serves to axiomatise the most basic calculus of modal logic. Some such schemes, particularly relevant in what follows, are written in the table below, together with the property of $\mathcal{R}_l$ they define.

	Axiom scheme	Property of $\mathcal{R}_l$
(K)	$\Box_l(\varphi \rightarrow \psi) \rightarrow (\Box_l\varphi \rightarrow \Box_l\psi)$	
(PF)	$\Diamond_l\varphi \rightarrow \Box_l\varphi$	Partial Functionality
(D)	$\Box_l\varphi \rightarrow \Diamond_l\varphi$	Seriality
(T)	$\Box_l\varphi \rightarrow \varphi$	Reflexivity
(4)	$\Box_l\varphi \rightarrow \Box_l\Box_l\varphi$	Transitivity
(5)	$\neg\Box_l\varphi \rightarrow \Box_l\neg\Box_l\varphi$	Euclideanity

Before ending this subsection we define some normal modal logics that will be used in other parts of the paper. The minimal modal system K is the smallest set of formulas containing all instances of the axiom scheme (K), all the valid formulas of propositional calculus, and closed under both Modus Ponens — if $\varphi, \varphi \rightarrow \psi \in K$, then $\psi \in K$ — and the Necessitation Rule — if $\varphi \in K$, then $\Box\varphi \in K$. Extensions of K are defined by adding more formulas to the basic generating set of K and closing again the resulting set under Modus Ponens and the Necesitation Rule. This is expressed as $K + (S_1) + ... + (S_n)$ where S_1,...,S_n are the new schemata. The following table defines some well known extensions of K:

Modal system	Definition
T	K + (T)
S4	K + (T) + (4)
S5	K + (T) + (4) + (5)
KD45	K + (D) + (4) + (5)

2.2 Abstract argumentation

Argumentation frameworks, the general structures for abstract argumentation, are defined as follows:

Definition 2.4 (Abstract argumentation framework [41])**.** *An Abstract Argumentation Framework (AF) is a directed graph $AF = \langle Ar, att\rangle$ where Ar is a set of elements called arguments, and $att \subseteq Ar \times Ar$ is binary relation over arguments.*

Although Dung called *att* an attack relation, it is sometimes clearer to interpret it as a defeat relation. Roughly speaking, an argument a attacks another argument b if they are incompatible (they cannot be jointly accepted); while a defeats b if a attacks b and a is at least as strong as b. This distinction (attack vs. defeats) emerges from the literature on structured argumentation [22; 35] and it will be exemplified in several parts of this article, where the expression "be as at least strong as" will be attributed precise formal meanings.

Argumentation frameworks are, in their bare bones, nothing more than directed graphs. What is fundamental is the specification of their *semantics* – sometimes also called *solution concepts* – which encode different criteria of justification for (sets of) arguments. The following definition provides the original ones by Dung [41], which are the ones used in this article.

Definition 2.5 (Argumentation semantics)**.** *Given* $AF = \langle Ar, att\rangle$ *and* $\mathcal{E} \subseteq Ar$*,*

- $\mathcal{E}$ *is conflict-free iff there does not exist* $a, b \in \mathcal{E}$ *such that* $\langle a, b\rangle \in att$*.*
- *An argument* $a \in Ar$ *is acceptable w.r.t. a set* $\mathcal{E}$ *(a is defended by* $\mathcal{E}$*), iff* $\forall\langle b, a\rangle \in att$*,* $\exists c \in \mathcal{E}$ *such that* $\langle c, b\rangle \in att$*.*
- *A conflict-free set of arguments* $\mathcal{E}$ *is admissible iff each argument in* $\mathcal{E}$ *is acceptable w.r.t.* $\mathcal{E}$*.*
- $\mathcal{E}$ *is a complete extension of AF iff* $\mathcal{E}$ *is admissible and each argument in Ar that is acceptable w.r.t.* $\mathcal{E}$ *is in* $\mathcal{E}$*.*
- $\mathcal{E}$ *is the grounded extension of AF iff* $\mathcal{E}$ *is the minimal (w.r.t. set inclusion) complete extension.*
- $\mathcal{E}$ *is the preferred extension of AF iff* $\mathcal{E}$ *is the maximal (w.r.t. set inclusion) complete extension.*
- $\mathcal{E}$ *is a stable extension of AF iff* $\mathcal{E}$ *is conflict-free and* $\forall b \in Ar\backslash\mathcal{E}$*,* $\exists a \in \mathcal{E}$ *such that* $\langle a, b\rangle \in att$*.*

Let $AF = \langle Ar, att\rangle$ be an AF, let $\mathcal{S} \in \{\mathcal{CO}, \mathcal{GR}, \mathcal{PR}, \mathcal{ST}\}$ (where $\mathcal{CO}$ stands for complete, $\mathcal{GR}$ for grounded, $\mathcal{PR}$ for preferred, and $\mathcal{ST}$ for stable), we denote by $\mathcal{E}_{\mathcal{S}}(AF)$ the set of $\mathcal{S}$-extensions of AF.

For a detailed study of these and further semantics, the reader is referred to [17].

3 Argumentation and dynamic logic for value-based planning

Autonomous agents are supposed to be able to perform value-based ethical reasoning based on their value systems in order to distinguish moral from immoral behavior. Existing work on value-based practical reasoning such as [11; 20; 69] demonstrates how an agent can reason about what he should do among alternative action options that are associated with value promotion or demotion. More than that, agents are supposed to be able to finish tasks or achieve goals that are assigned by their users through performing a sequence of actions. Classical planning concerns finding a successful sequence of actions achieving a goal. Since there might exist multiple plans that an agent can follow and each plan might promote or demote different values along each action, the agent should be able to resolve the conflicts between them and evaluate which plan he should follow. If the decision-making problem concerns choosing a plan instead of an action, then we first need to know how an agent can see whether he can follow a particular plan to achieve his goal. Modal logic allows us to represent and verify whether a goal can be achieved by executing a plan under specific conditions such as norm compliance assumptions [1; 65; 2], namely telling agents whether a plan works or not, but cannot tell agents whether it is the best option. Certainly, agents can collect the representation results regarding whether a plan promotes or demotes a specific set of values and then compare different plans using lifting approaches as what has been done in [74]. However, the order lifting problem is a major challenge in many areas of AI and no approach is ultimately "correct". Moreover, the agent in our setting needs to lift the preference over values to the preference over plans with respect to value promotion and demotion, which even complicates the problem. Therefore, we need a more natural and intuitive approach to deal with representation results.

It has been shown that argumentation provides a useful mechanism to model and resolve conflicts [41], and particularly can be used for the decision making of artificial intelligence in a dialectical way, and provides explanation for that [80; 10; 83; 71]. In this section, we develop a logic-based framework that combines modal logic and argumentation for value-based planning. In the first part, modal logic is used as a technique to represent and verify agents' belief in terms of whether a plan with its local properties of value promotion or demotion can be followed to achieve an agent's goal. Using the representation results to construct arguments, we then propose an argumentation framework that allows an agent to reason about his plans in the form of support and objection. We prove that our framework satisfies a set of properties consistent with our understanding of rational decision making. Our

preliminary idea has been presented in [73], where arguments are constructed with a value for promotion or demotion. However, we notice that arguing about plans using this way of argument construction is in fact equivalent to arguing about plans using arguments that are constructed with a set of values for promotion or demotion in the democratic lifting way. We thus in this version construct arguments with a set of values for promotion or demotion and allow more lifting ways for comparing sets of values.

3.1 Modal logic for representation

The basic semantic structure of our approach is a transition system (Def. 2.2) where the set of generic labels $\mathbb{L}$ is understood as a set of actions $Act = \{\alpha_1, \alpha_2, ..., \alpha_n\}$ that are executable by an agent. This way of looking at transitions systems represents the computational behavior of a system caused by an agent's actions in the agent's subjective view. Hence, vertices S corresponds to possible states of the system, and the relation $\mathcal{R} \subseteq S \times Act \times S$ represents the possible transitions of the system. When a certain action $\alpha \in Act$ is performed, the system might progress from a state s to a different state s' in which different propositions hold. Moreover, some restrictions are imposed on relation $\mathcal{R}$ in order to capture some intuitions. Recall that $\Phi = \{p, q, ...\}$ is a set of atomic propositions.

Definition 3.1 (Action Transition Models)**.** *An action transition model is a interpreted labelled transition system (i.e., a model) $T = \langle S, \mathcal{R}, \mathcal{V}\rangle$ (Def. 2.2) where the set of labels $\mathbb{L}$ represents a set of actions $Act = \{\alpha_1, ..., \alpha_n\}$. Moreover, it is assumed that*

- *for all $s \in S$ there exists an action $a \in Act$ and a state $s' \in S$ such that $\langle s, \alpha, s'\rangle \in \mathcal{R}$;*
- *we restrict actions to be deterministic, that is, if $\langle s, \alpha, s'\rangle \in \mathcal{R}$ and $\langle s, \alpha, s''\rangle \in \mathcal{R}$, then $s' = s''$.*

Since the relation $\mathcal{R}$ is partially functional, we write $s[\alpha]$ to denote the state s' for which it holds that $\langle s, \alpha, s'\rangle \in \mathcal{R}$. We also use $s[\alpha_1, \ldots, \alpha_n]$ to denote the resulting state for which a sequence of actions $\alpha_1, \ldots, \alpha_n$ successively execute from state s. A pointed action transition model is a pair $\langle T, s\rangle$ such that T is an action transition model, and $s \in S$ is a state from T. Adopted from [66; 67], the language $\mathcal{L}(\Phi, Act)$ is just our generic labelled language $\mathcal{L}(\Phi, \mathbb{L})$ (Def. 2.1) with $\mathbb{L} = Act$. For convenience, we take $\Diamond_\alpha$ instead of $\Box_\alpha$ to be the primitive modal operator. The notion of truth in a pointed action transition model is then also the generic one (Sect. 2.1). We just make explicit the cause for $\Diamond_\alpha$:

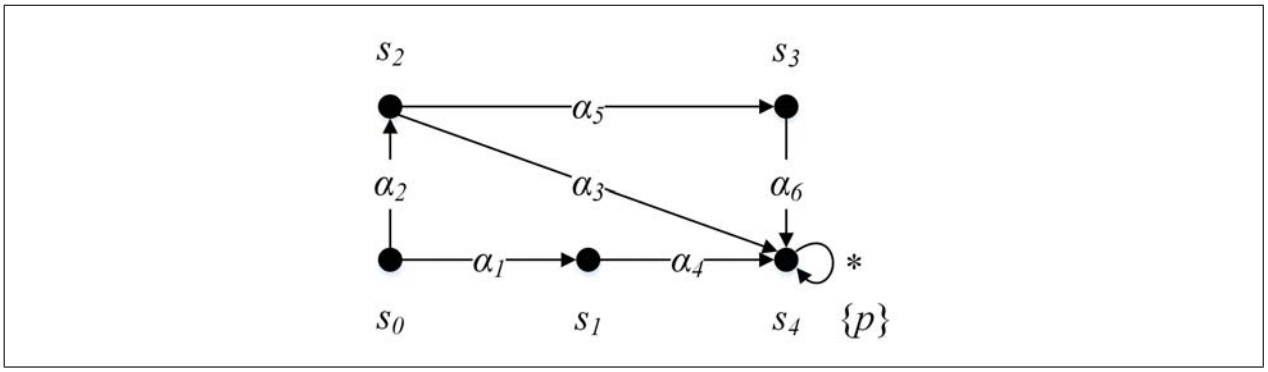


Figure 1: Transition system T. The star loop around state s_4 means that the agent stays in state s_4 whatever he does.

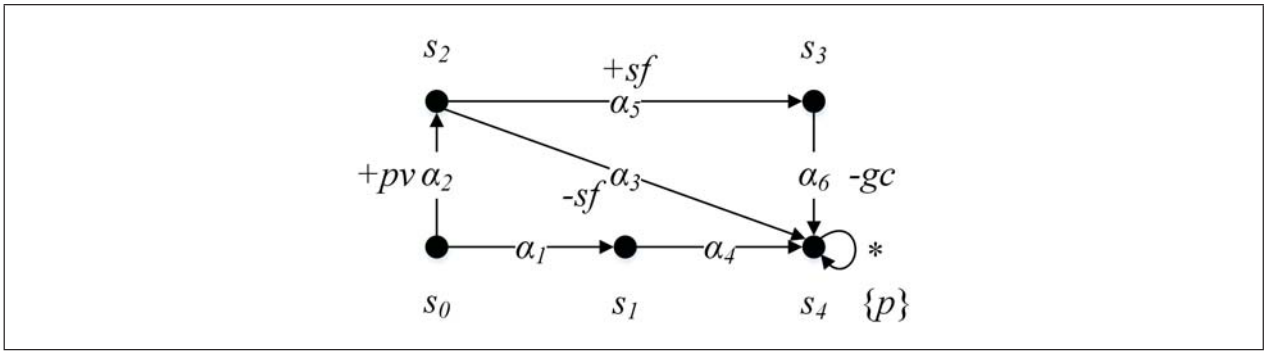


Figure 2: A value-based action transition model VT. The star loop around state s_4 means that the agent stays in state s_4 whatever he does.

$$T, s \models \Diamond_\alpha \varphi \text{ iff } s[\alpha] \text{ exists and } T, s[\alpha] \models \varphi.$$

Given a pointed action transition model $\langle T, s\rangle$, we say that a sequence of actions $\lambda = \alpha_1 \ldots \alpha_n$ brings about a φ-state if and only if $T, s \models \Diamond_{\alpha_1} \ldots \Diamond_{\alpha_n} \varphi$. In the rest of the section, we will sometimes write $\Diamond_\lambda \varphi$ instead of $\Diamond_{\alpha_1} \ldots \Diamond_{\alpha_n} \varphi$ for short.

A action transition model represents how a system progresses by an agent's actions. Besides, an agent in the system is assumed to have his own goal, which is a formula expressed in propositional logic $\mathcal{L}(\Phi)$. It is indeed possible for an agent to have multiple goals and his preference over different goals. For example, a goal hierarchy is defined in [1] to represent increasingly desired properties that the agent wishes to hold. However, we find that the setting about whether the agent has one goal or multiple goals is in fact not essential for our analysis, so we simply assume that the agent only has a goal for simplifying our presentation.

Example 3.2. *Consider the action transition model T in Figure 1, which represents how an agent can get to a pharmacy to buy medicine for his user. State s_0 is the*

initial state, representing staying at home, and proposition p, representing arriving at a pharmacy, holds in state s_4. The agent can perform actions α_1 to α_6 in order to get to state s_4. From this action transition model, the following formulas hold:

$$T, s_0 \models \Diamond_{\alpha_1}\Diamond_{\alpha_4}p,$$
$$T, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_3}p,$$
$$T, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_5}\Diamond_{\alpha_6}p,$$

which means that the agent can first perform action α_1 and then action α_4, or action α_2 followed by action α_3, or action α_2 followed by actions α_5 and α_6, to get to the pharmacy.

It is important for an agent not only to achieve his goal, but also to think about how to achieve his goal. As we can see from the running example, there are multiple ways for the agent to get to the pharmacy, and the agent needs to evaluate which one is the best to choose. In this section, agents are able to perform value-based practical reasoning in terms of planning their actions to achieve their goals. We first assume that an agent has a set of values. A value can be seen as an abstract standard according to which agents have their preferences over options. For instance, if we have a value denoting *equality*, we prefer the options where equal sharing or equal rewarding hold. Unlike [74] where a value is interpreted as a state formula, we simply assume a value as a primitive structure without considering how it is defined. Moreover, agents can always compare any two values, so we define an agent's value system as a total pre-order (instead of a strict total order) over a set of values, representing the degree of importance of something.

Definition 3.3 (Value Systems)**.** *A value system $V = \langle \text{Val}, \succsim \rangle$ is a tuple consisting of a finite set of values $\text{Val} = \{v_1, ..., v_k\}$ together with a total pre-ordering $\succsim$ over* Val*. When $v_i \succsim v_j$, we say that value v_j is at least as important as value v_i. As is standard, we define $v_i \sim v_j$ to mean $v_i \succsim v_j$ and $v_j \succsim v_i$, and $v_i \prec v_j$ to mean $v_i \succsim v_j$ and $v_i \not\sim v_j$.*

We label some of the transitions with the values promoted and demoted by moving from a starting state to a ending state. Notice that not every transition can be labeled, as some transitions may not be relevant to any value in an agent's value system. Formally, a function $\delta : \{+, -\} \times \text{Val} \to 2^{\mathcal{R}}$ is a valuation function over T which defines the status (promoted (+) or demoted (-)) of a value $v \in \text{Val}$ ascribed to a set of transitions. We then define a value-based action transition model VT as a action transition model together with a value system V and a function δ.

Definition 3.4 (Value-based Transition Model)**.** *A value-based action transition model is defined by a triple $VT = \langle T, V, \delta \rangle$, where T is an action transition model, V is a value system and δ is a valuation function that assigns value promotion or demotion to a set of transitions.*

Given a sequence of actions with respect to a value-based action transition model, we then express whether the performance of the sequence in a state promotes or demotes a specific value, which can be done by extending our language $\mathcal{L}(\Phi, Act)$ with new modalities of the form promoted$(v, \alpha_1 \dots \alpha_n)$ and demoted$(v, \alpha_1 \dots \alpha_n)$. The formula promoted$(v, \alpha_1 \dots \alpha_n)$ (resp. demoted$(v, \alpha_1 \dots \alpha_n)$) should be intuitively read as there exists an action that promotes (resp. demotes) value v in the sequence of actions $\alpha_1, \dots, \alpha_n$. Given a pointed value-based action transition model (VT, s) and a value $v \in$ Val, the satisfaction relation $VT, s \models \psi$ is extended with the following new semantics:

- $VT, s \models$ promoted$(v, \alpha_1 \dots \alpha_n)$ iff there exists $1 \leq m \leq n$ such that

$$(s[\alpha_1, \dots, \alpha_{m-1}], \alpha_m, s[\alpha_1, \dots, \alpha_m]) \in \delta(+, v);$$

- $VT, s \models$ demoted$(v, \alpha_1 \dots \alpha_n)$ iff there exists $1 \leq m \leq n$ such that

$$(s[\alpha_1, \dots, \alpha_{m-1}], \alpha_m, s[\alpha_1, \dots, \alpha_m]) \in \delta(-, v).$$

Notice that the formula only expresses the local property of a sequence of actions in terms of value promotion or demotion by an action within the sequence. Thus, it is possible that an action within the sequence promotes value v but it gets demoted by another action within the sequence, meaning that both $VT, s \models$ promoted$(v, \alpha_1 \dots \alpha_n)$ and $VT, s \models$ demoted$(v, \alpha_1 \dots \alpha_n)$ hold at the same time. Since a sequence of actions is denoted as λ, we will sometimes write promoted(v, λ) instead of promoted$(v, \alpha_1 \dots \alpha_n)$ and demoted(v, λ) instead of demoted$(v, \alpha_1 \dots \alpha_n)$ for short. Having the above formulas, the agent is then aware of which value gets promoted or demoted along a sequence of actions. We continue our running example to illustrate how to use our logical language to express and verify properties of sequences of actions.

Example 3.5. *Suppose the ethical agent has privacy (pv), safety (sf) and good conditions (gc) as his values and a value system as $pv \prec gc \prec sf$. As in Figure 2, some of the transitions have been labeled with value promotion or demotion with respect to the agent's values. Taking action α_2 in state s_0 is interpreted as asking for the permission of taking a private path, which promotes the value of privacy. Taking*

action α_3 means crossing the road without using the crosswalk, which demotes the value of safety of the agent, and conversely taking action α_4 in state s_2 promotes the value of safety of the agent. Finally, performing action α_5 in state s_3 means stepping into water. As the agent is a robot, which should avoid getting wet, this choice will demote the value of maintaining good conditions of the agent. The agent can verify whether he can achieve his goal while promoting or demoting a specific value by performing a sequence of actions. The verification results are listed below:

$$
\begin{aligned}
&VT, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_3} p \wedge \text{promoted}(pv, \alpha_2\alpha_3)\\
&VT, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_3} p \wedge \text{demoted}(sf, \alpha_2\alpha_3)\\
&VT, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_5}\Diamond_{\alpha_6} p \wedge \text{promoted}(pv, \alpha_2\alpha_5\alpha_6)\\
&VT, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_5}\Diamond_{\alpha_6} p \wedge \text{promoted}(sf, \alpha_2\alpha_5\alpha_6)\\
&VT, s_0 \models \Diamond_{\alpha_2}\Diamond_{\alpha_5}\Diamond_{\alpha_6} p \wedge \text{demoted}(gc, \alpha_2\alpha_5\alpha_6)
\end{aligned}
$$

3.2 Value-based planning: an argumentative approach

Given a action transition model and an agent's goal, modal logic allows an agent to represent and verify whether he can achieve his goal while promoting or demoting a specific value by performing a sequence of actions. Since following different plans might promote or demote different sets of values, next question is how the agent *internally* decides what to do given the representation results. In this section, we propose to use argumentation as a technique for an agent's planning. Formal argumentation is a nonmonotonic formalism for representing and reasoning about conflicts based on the construction and the evaluation of interacting arguments [41]. In particular, it has been used in practical reasoning, concerned with reasoning about what agents should do, given different alternatives and outcomes they bring about [20; 7]. Since argumentation resolves conflicts in a dialectical way, it also provides justification and explanation to the final solution. In general, epistemic planning considers the following problem [25; 26]: Given my current state of belief, and a desirable state of belief, how do I get from one to the other? In particular, each plan is labeled with a set of values that are promoted or demoted along the plan. The agent needs to look for a plan that is not only feasible but also optimal with respect to value promotion and demotion. We first define the notion of plans. A plan is defined as a finite sequence of actions that will bring about the agent's goal in the underlying action transition model.

Definition 3.6 (Plans)**.** *Given a value-based action transition model VT, a state s and a formula $g \in \mathcal{L}(\Phi)$ as an agent's goal, a sequence of actions λ over Act is said*

to be a plan w.r.t s and g, denoted as $\lambda_{s,g}$, iff $VT, s \models \Diamond_\lambda g$. Sometimes, we write λ for $\lambda_{s,g}$ if it is clear from the context.

A sequence of actions is denoted as λ if it is a plan. Given a set of available plans, the agent can construct arguments to support or oppose the execution of a plan. The reason to supporting a plan is that the plan promotes a set of values, and the reason to opposing a plan is that the plan demotes a set of values, which can be expressed as formulas in our language $\mathcal{L}(\Phi)$. We define two types for arguments for planning: an ordinary argument supports the performance of a plan, while a blocking argument opposes the performance of a plan.

Definition 3.7 (Ordinary Arguments for Planing)**.** *Given a value-based action transition model VT, a state s, a goal g and a plan λ w.r.t. s and g,*

- *let $A \subseteq$ Val be a set of values, a non-empty ordinary argument is a pair $\langle +A, \lambda \rangle$, read as "plan λ should be selected because it promotes values A", iff*

$$VT, s \models \bigwedge_{v \in A} \text{promoted}(v, \lambda),$$

- *an empty ordinary argument is a pair $\langle -\emptyset, \lambda \rangle$, read as "plan λ should be selected because it does not demote any values", such that*

$$VT, s \models \bigwedge_{v \in \text{Val}} \neg\, \text{demoted}(v, \lambda).$$

Definition 3.8 (Blocking Arguments for Planning)**.** *Given a value-based action transition model VT, a state s and a plan λ,*

- *let $A \subseteq$ Val be a set of values, a non-empty blocking argument is a pair $\langle -A, \neg\lambda \rangle$, read as "plan λ should not be selected because it demotes values A", iff*

$$VT, s \models \bigwedge_{v \in A} \text{demoted}(v, \lambda);$$

- *an empty blocking argument is a pair $\langle +\emptyset, \neg\lambda \rangle$, read as "plan λ should not be selected because it does not promote any values", such that*

$$VT, s \models \bigwedge_{v \in \text{Val}} \neg\, \text{promoted}(v, \lambda).$$

We use $\mathcal{A}_o^p$ to denote the set of ordinary arguments and use $\mathcal{A}_b^p$ to denote the set of blocking arguments for planning, and $\mathcal{A}^p = \mathcal{A}_o^p \cup \mathcal{A}_b^p$ to denote the set of two types of arguments. In the following text, unless it is addressed clearly, an ordinary argument can refer to a non-empty ordinary argument or an empty ordinary argument, and a blocking argument can refer to a non-empty blocking argument or an empty blocking argument. Both an ordinary argument and a blocking argument correspond to representation results. Conventionally, we represent an argument using an alphabet $a, b, \ldots$ and thus the plan that it supports or opposes is denoted $\lambda_a, \lambda_b, \ldots$ and the set of promoted or demoted values is denoted as uppercase letters V_a, V_b, etc.

Example 3.9 (Ordinary arguments and blocking arguments)**.** *The value-based action transition model in Fig. 2 shows that the agent is aware of three plans $\alpha_1\alpha_4$, $\alpha_2\alpha_3$ and $\alpha_2\alpha_5\alpha_6$. Plan $\alpha_2\alpha_3$ promotes value pv but demote value sf, plan $\alpha_1\alpha_4$ does not promote or demote any value, and plan $\alpha_2\alpha_5\alpha_6$ promote values pv and sf, but demote value gc. Based on the representation results, the agent can construct the following ordinary arguments and blocking arguments: $\langle +\emptyset, \neg\alpha_1\alpha_4\rangle$, $\langle +\{pv\}, \alpha_2\alpha_3\rangle$, $\langle -\{sf\}, \neg\alpha_2\alpha_3\rangle$, $\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6\rangle$ and $\langle -\{gc\}, \neg\alpha_2\alpha_5\alpha_6\rangle$.*

When we get to choose a plan to follow, there are conflicts between the alternatives as they cannot be followed all at the same time. The conflicts are interpreted as attacks between two ordinary arguments supporting different plans and one ordinary argument and one blocking argument supporting and objecting to the same plan respectively.

Definition 3.10 (Attacks for Planning)**.** *Given a set of ordinary arguments $\mathcal{A}_o^p$ and a set of blocking arguments $\mathcal{A}_b^p$,*

- *for any two ordinary arguments $a, b \in \mathcal{A}_o^p$, a attacks b iff $\lambda_a \neq \lambda_b$;*
- *for any ordinary argument $a \in \mathcal{A}_o^p$ and any blocking argument $b \in \mathcal{A}_b^p$,*
  - *a attacks b iff $\lambda_a = \lambda_b$;*
  - *b attacks a iff $\lambda_a = \lambda_b$.*

The set of attacks (an attack relation) over $\mathcal{A}^p$ are denoted as att^p.

It is obvious that our attack relation is mutual. It should be noticed that there is no attack between two blocking arguments, as a blocking argument only functions as blocking the conclusion of an ordinary argument but does not make a conclusion by itself.

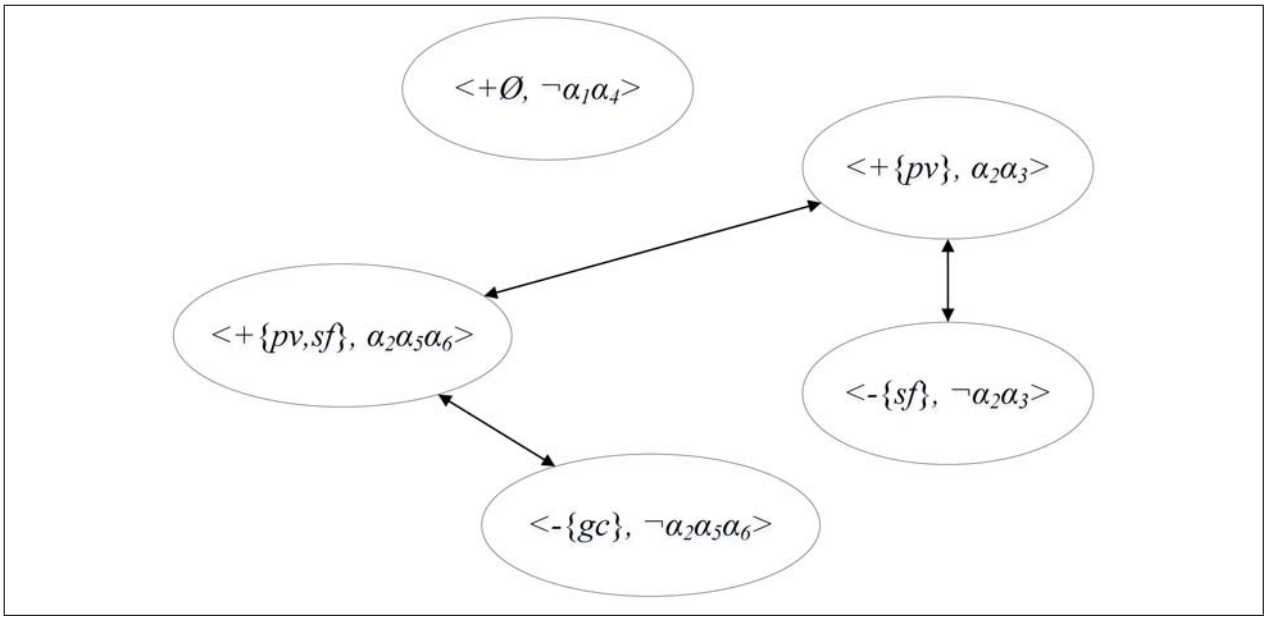


Figure 3: Attack relation between ordinary arguments and blocking arguments.

Example 3.11. *In the running example, there are three ordinary arguments and three blocking arguments. The attack relation is depicted in Figure 3, where any two ordinary arguments with different plans mutually attack (for instance,* $\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6\rangle$ *and* $\langle +\{pv\}, \alpha_2\alpha_3\rangle$*), and any ordinary argument and blocking argument with the same plan are mutually attacked (for instance,* $\langle +\{pv\}, \alpha_2\alpha_3\rangle$ *and* $\langle -\{sf\}, \neg\alpha_2\alpha_3\rangle$*).*

The attack relation represents conflicts between plans. However, the notion of attack may not be sufficient for modeling conflicts between arguments, as an agent has his preference over the values that are promoted or demoted by different plans. In structured argumentation frameworks such as ASPIC$^+$ [76], an argument a can be used as a counter-argument to another argument b, if a successfully attacks, i.e. defeats, b. Whether an attack from a to b (on its sub-argument b') succeeds as a defeat, may depend on the relative strengths of a and b, which is a preference over arguments a and b based on the preferences over their constituent ordinary premises and defeasible rules. Here we use the same approach to decide an attack succeeds as a defeat. Recall that an agent has a value system, which was defined as a total pre-order over a set of values. So there needs to be a lifting way that allows the planning agent to lift the preference over values to preferences over arguments. Two lifting ways are commonly used in structured argumentation: the so called *Elitist* and *Democratic* ways. Eli (denoted as $\trianglelefteq_E$) compares sets on their minimal and Dem (denoted as $\trianglelefteq_D$) on their maximal elements.

Definition 3.12 (Lifting)**.** *Given two set of values A and B, $\trianglelefteq_E$ is defined as follows:*

$$A \trianglelefteq_E B \text{ iff there exists } v_a \in A \text{ s.t. for all } v_b \in B : v_a \precsim v_b.$$

$\trianglelefteq_D$ is defined as follows:

$$A \trianglelefteq_D B \text{ iff for all } v_a \in A \text{ there exists } v_b \in B : v_a \precsim v_b.$$

We use $\trianglelefteq \in \{\trianglelefteq_E, \trianglelefteq_D\}$ to denote an arbitrary lifting approach of the above. We define $A \simeq B$ to mean $A \trianglelefteq B$ and $B \trianglelefteq A$, and $A \triangleleft B$ to mean $A \trianglelefteq B$ and it is not the case that $A \simeq B$.

It is easy to prove that $\trianglelefteq$ is reflexive and transitive. We can then determine the defeat relation over two arguments based on the value system. The notion of defeat combines the notions of attack and preference.

Definition 3.13 (Defeats for Planning)**.** *Given a set of arguments $\mathcal{A}^p$, a set of attacks $\mathcal{R}^p$ over $\mathcal{A}^p$ and a value system V, for any two arguments $a, b \in \mathcal{A}^p$, a defeats b iff a attacks b and it is not the case that $V_a \trianglelefteq V_b$ or b is an empty argument. The set of defeats (a defeat relation) over $\mathcal{A}^p$ based on an attack relation att^p, a value system V and a lifting $\trianglelefteq$ is denoted as $\mathcal{D}^p(att^p, V, \trianglelefteq)$. We write $\mathcal{D}^p$ for short if it is clear from the context.*

In words, given mutual attacks between two arguments, the attack from the argument with less preferred value set to the argument with a more preferred value set does not succeed as a defeat, and the empty argument is always defeated. One might ask whether it is more convenient to combine the notions of attack relation and defeat relation. We argue that two notions represent the relation between two arguments from different perspectives, one for the conflicts between plans and the other for the preferences over values. Because of that, defining these two notions separately can make our framework more clear, even though technically it is possible to combine them. Here are several properties that characterize our defeat relation.

Proposition 3.14. *Given two ordinary arguments $a, b \in \mathcal{A}^p_o$, a and b defeat each other iff $\lambda_a \neq \lambda_b$ and $A \simeq B$. Given an ordinary argument $a \in \mathcal{A}^p_o$ and a blocking argument $b \in \mathcal{A}^p_b$, a and b defeat each other iff $\lambda_a = \lambda_b$ and ($A \simeq B$ or both a and b are empty arguments).*

Proof. Proof follows from Definition 3.13 directly. □

Proposition 3.15. *Given a set of arguments $\mathcal{A}^p$, a defeat relation $\mathcal{D}^p$ on $\mathcal{A}^p$ never forms any pure odd cycles.*

Proof. According to Definition 3.13, in order for an argument a to defeat another argument b, value set A must be not less preferred than B or B is an empty argument. Given three non-empty arguments a, b, c, since the preference order over sets of values is transitive, if a defeats non-empty argument b and b defeats c, then a also defeats c. For the case where the set of values are equally preferred, because of Proposition 3.14, any odd cycles that are formed by $\mathcal{D}^p$ are always together with two-length cycles, which are not pure odd cycles. For the case where there exists empty arguments, if a is an non-empty argument and b is an empty argument, then c is also empty. As a is an non-empty argument and c is empty, c cannot defeat a. □

Proposition 3.16. *Given a set of arguments $\mathcal{A}^p$, a defeat relation $\mathcal{D}^p$ on $\mathcal{A}^p$ is irreflexive.*

Proof. It is a special case of Proposition 3.15 for the number of arguments in the odd cycle being one. □

We are now ready to construct a Dung-style abstract argumentation framework with ordinary arguments, blocking arguments and the defeat relation on them.

Definition 3.17 (Argumentation Frameworks for Planning)**.** *Given a pointed value-based action transition model (VT, s) and a formula $g \in \mathcal{L}(\Phi)$ as an agent's goal, an argumentation framework for planning over (VT, s) and g is a pair $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$, where $\mathcal{A}^p$ is a set of arguments and $\mathcal{D}^p$ is a defeat relation on $\mathcal{A}^p$.*

Example 3.18. *In our running example, the agent has a value system as $pv \prec gc \prec sf$, which means that safety is more important than keeping good condition, and keeping good condition is more important than privacy. With lifting $\trianglelefteq_D$, we then can see some of the attacks in Figure 3 do not succeed as defeats. For example, argument $\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6 \rangle$ and argument $\langle -\{gc\}, \neg\alpha_2\alpha_5\alpha_6 \rangle$ are mutually attacked, but since $\{gc\} \trianglelefteq_D \{pv, sf\}$, only the attack from argument $\langle +\{pv, sf\}, \neg\alpha_2\alpha_5\alpha_6 \rangle$ to argument $\langle -\{gc\}, \alpha_2\alpha_5\alpha_6 \rangle$ becomes a defeat. Notice that argument $\langle +\emptyset, \neg\alpha_1\alpha_4 \rangle$ do not receive any defeats or defeat any arguments because there is no ordinary argument with plan $\alpha_1\alpha_4$.*

With lifting $\trianglelefteq_E$, we should notice the defeat between argument $\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6 \rangle$ and argument $\langle -\{gc\}, \neg\alpha_2\alpha_5\alpha_6 \rangle$. Since $\{pv, sf\} \trianglelefteq_E \{gc\}$, argument $\langle -\{gc\}, \neg\alpha_2\alpha_5\alpha_6 \rangle$ to argument $\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6 \rangle$. All other defeats remain the same as with lifting $\trianglelefteq_D$. See the defeat relation in Figure 4 and Figure 5 with different lifting ways.

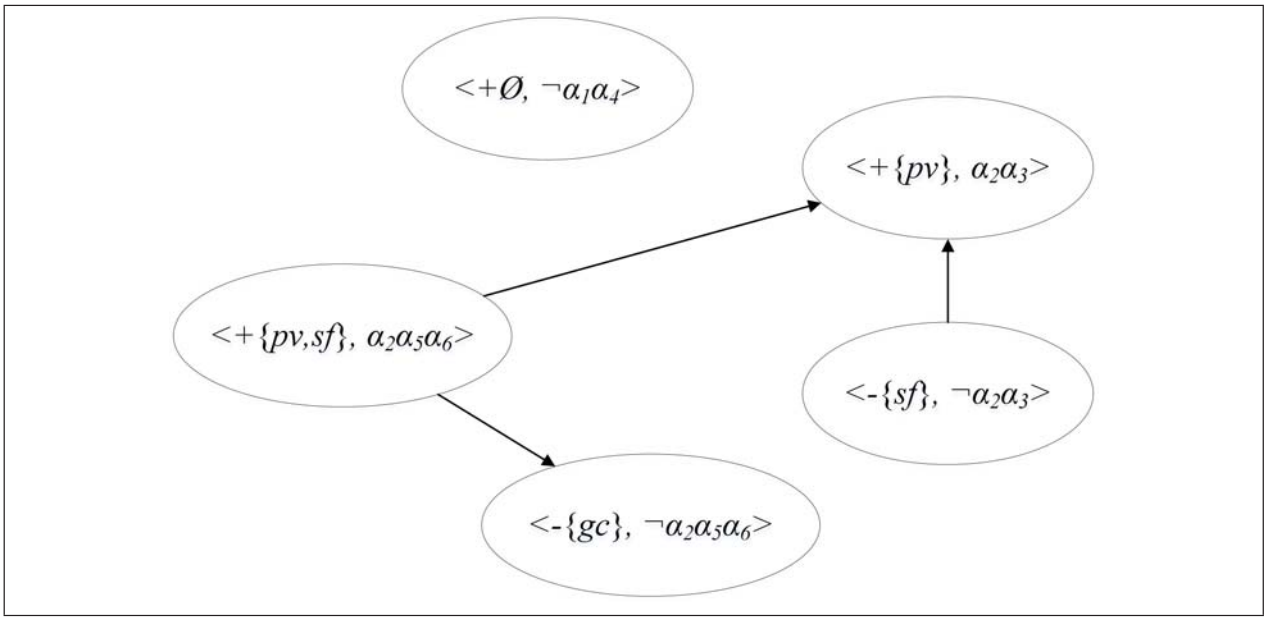


Figure 4: An argumentation framework for planning with lifting $\trianglelefteq_D$.

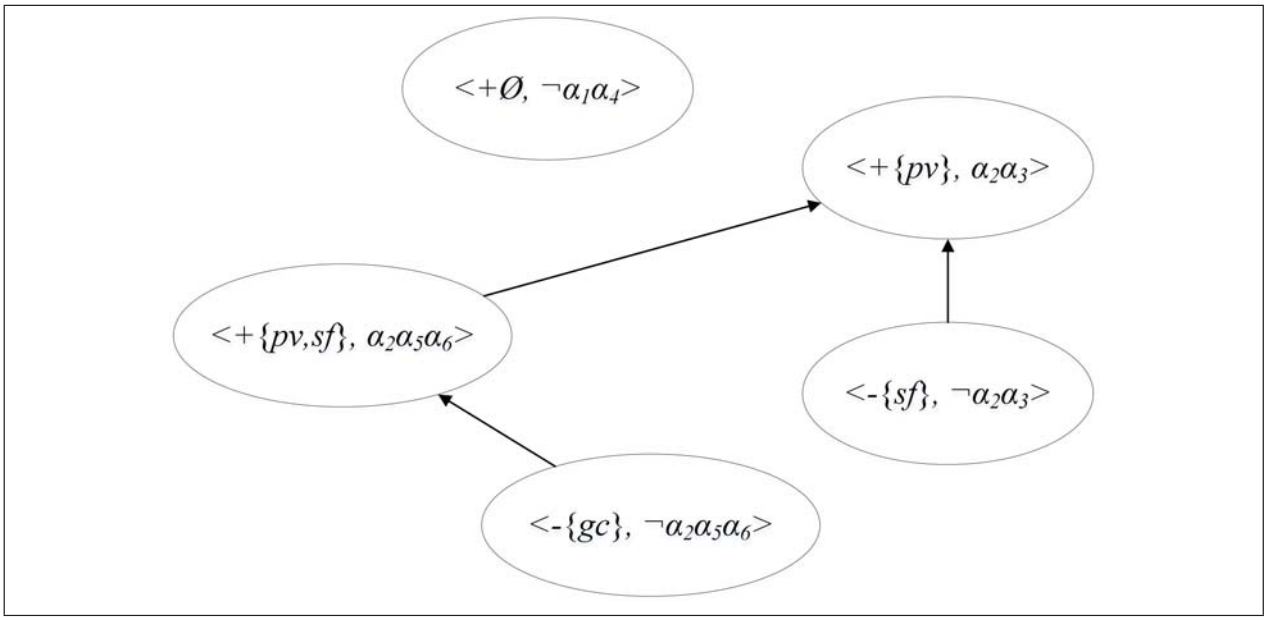


Figure 5: An argumentation framework for planning with lifting $\trianglelefteq_E$.

Given an argumentation framework for planning PAF, the status of arguments is evaluated, producing sets of arguments that are acceptable together, which are based on the notions of conflict-freeness, acceptability and admissibility. The well-known argumentation semantics are listed in Definition 2.5, each of which provides a pre-defined criterion for determining the acceptability of arguments in a PAF [41]. We use $\mathcal{S} \in \{\mathcal{CO}, \mathcal{PR}, \mathcal{GR}, \mathcal{ST}\}$ to denote the complete, preferred, grounded and stable semantics, respectively, and $\mathcal{E}_{\mathcal{S}}(PAF)$ to denote the set of extensions of PAF under a semantics in $\mathcal{S}$. The following propositions characterize our argumentation framework in terms of Dung's semantics.

Proposition 3.19. *Given $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$, $\mathcal{E}_{\mathcal{PR}}(PAF) = \mathcal{E}_{\mathcal{ST}}(PAF)$.*

Proof. Since our defeat relation $\mathcal{D}^p$ never forms any pure odd cycles by Proposition 3.15, which means that PAF is limited controversial, each preferred extension of PAF is stable. Detailed proof can be found in [41]. □

Proposition 3.20. *Given $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$ and the grounded extension E of PAF, if E contains an ordinary argument, then $\mathcal{E}_{\mathcal{PR}}(PAF) = \mathcal{E}_{\mathcal{GR}}(PAF)$.*

Proof. Suppose $\mathcal{E}_{\mathcal{PR}}(PAF) \neq \mathcal{E}_{\mathcal{GR}}(PAF)$, which means that there is more than one preferred extension. Since an ordinary argument is contained in the grounded extension E, it should also be contained in each preferred extension. However, each preferred extension indicates a distinct plan, which will be later proved by Proposition 3.22 and its implication. Contradiction! □

The notion of optimal plans is then defined under a specific semantics in Definition 2.5. Similarly to [70], given an argument a, we write concl(a) for the conclusion of argument a, and Oplans($PAF, \mathcal{S}$) for the set of conclusions of ordinary arguments from the extensions under a specific semantics.

Definition 3.21 (Optimal Plans)**.** *Given $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$ and a semantics $\mathcal{S}$, a set of optimal plans, written as* Oplans($PAF, \mathcal{S}$)*, are the conclusions of the ordinary arguments within extensions under semantics $\mathcal{S}$.*

$$\text{Oplans}(PAF, \mathcal{S}) = \{\text{concl}(a) \mid a \in \mathcal{A}^p_o, a \in E \text{ and } E \in \mathcal{E}_{\mathcal{S}}(PAF)\}$$

We show that the results of our approach are consistent with the rationality of decision-making through the following propositions. Firstly, all the accepted arguments within an extension indicate the same plan.

Proposition 3.22. *Given an argumentation framework for planning $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$ and an extension E of PAF under a specific semantics as defined in Definition 2.5,*

1. *for any two ordinary arguments $a, b \in E$, it is the case that $\lambda_a = \lambda_b$;*

2. *for any ordinary argument $a \in E$ and any blocking argument $b \in E$, $\lambda_a \neq \lambda_b$.*

Proof. For any extension E under a specific semantics, it is required that all the arguments in E should be conflict-free. 1. By Definition 3.13, we can derive two cases: either there is no attack between these two arguments, or one argument attacks the other but does not succeed as a defeat. For the former case, two arguments contain the same plan. For the latter case, since any attack between two arguments is mutual, if an attack from argument a to argument b fails to be a defeat because $A \lhd B$ or argument a is an empty argument while argument b is a non-empty argument, the attack from argument b to argument a will succeed to be a defeat. That means that the second case is impossible and only the first case holds. Hence, the two arguments have the same plan. 2. We can prove in a similar way that for any ordinary argument $a \in E$ and any blocking argument $b \in E$, $\lambda_a \neq \lambda_b$, □

From that we can see, if there are multiple preferred extensions, then each of them indicates a distinct plan. Secondly, when using lifting $\trianglelefteq_D$, our argumentation-based approach always accepts the argument with the most preferred value. Because of that, the plan that promotes the most preferred value will be accepted and the plan that demotes the most preferred value will be rejected.

Proposition 3.23. *Given an argumentation framework for planning $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$ with lifting $\trianglelefteq_D$, let $v \in$ Val be a value such that for all arguments $a \in \mathcal{A}^p$ and all values $v' \in V_a$ it is the case that $v' \precsim v$, then an argument with value v is in a preferred extension. If it is not in a cycle, then it is also in the grounded extension.*

Proof. Because $v' \precsim v$, according to Definition 3.13 and lifting $\trianglelefteq_D$, an argument with value v only gets defeated by an argument with value v' that satisfies $v \sim v'$ or $v = v'$. In such a case, the defeats are mutual so argument a is self-defended. Thus, it is contained in a preferred extension. If it is not in a cycle, which means that it is not self-defended but only defeats other arguments, then it is in the grounded extension. □

When using lifting $\trianglelefteq_E$, our argumentation-based approach always rejects the argument with the least preferred value. Because of that, the plan that promotes the least preferred value will be rejected.

Proposition 3.24. *Given an argumentation framework for planning $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$ with lifting $\trianglelefteq_E$, let $v \in$ Val be a value such that for all arguments $a \in \mathcal{A}^p$ and all values $v' \in V_a$ it is the case that $v \precsim v'$, then an argument with value v is rejected under any semantics.*

Proof. In order for an argument with v to be accepted, there must be another accepted argument that defends the an argument with v. However, since for all arguments $a \in \mathcal{A}^p$ and all values $v' \in V_a$ it is the case that $v \precsim v'$, this argument will also defeat the argument with v, making it rejected. □

Because of the above three propositions, the agent can conclude to follow an optimal plan to achieve his goal. However, the notion of optimal plans is defined as the set of conclusions of ordinary arguments from the extensions, so the set of optimal plans becomes empty if an extension does not contain any ordinary arguments. The following proposition indicates the conditions for which the set of optimal plans is not empty.

Proposition 3.25. *Given an argumentation framework for planning* $PAF = \langle \mathcal{A}^p, \mathcal{D}^p \rangle$, $\text{Oplans}(PAF, \mathcal{S}) \neq \emptyset$ *iff there exists an ordinary argument* a *such that it is not defeated by a blocking argument* b *with* $V_a \prec V_b$.

Proof. Having $\text{Oplans}(PAF, \mathcal{S}) \neq \emptyset$ means that there is at least one extension which contains at least one ordinary argument. $\Rightarrow$: Suppose there does not exists an ordinary argument a such that it is not defeated by a blocking argument b with $V_a \prec V_b$, which means that all the ordinary arguments (if exist) are defeated by a blocking argument and not self-defended against a blocking argument. In such a case, there exists a blocking argument that does not receive any defeats, which makes all the ordinary arguments rejected. Contradiction! $\Leftarrow$: If there exists an ordinary argument such that it is not defeated by a blocking argument with $V_a \prec V_b$, then (1) the ordinary argument does not receive any defeats and thus it should be contained in the grounded extension, or (2) the ordinary argument is in a two-length cycle with a blocking argument and thus it should be contained in a preferred extension, or (3) the ordinary argument receives defeats from other ordinary arguments and thus there is always an ordinary argument accepted. Hence, $\text{Oplans}(PAF, \mathcal{S})$ is not an empty set. □

Example 3.26. *The argumentation framework for planning* PAF *with lifting* $\trianglelefteq_D$ *can be represented as Fig. 4. Because*

$$\mathcal{E}_{\mathcal{PR}}(PAF) = \mathcal{E}_{\mathcal{GR}}(PAF) = \mathcal{E}_{\mathcal{ST}}(PAF) = \{\{\langle +\{pv, sf\}, \alpha_2\alpha_5\alpha_6\rangle, \langle +\emptyset, \neg\alpha_1\alpha_4\rangle, \langle -\{sf\}, \neg\alpha_2\alpha_3\rangle\}\}$$

and thus $\text{Oplans}(PAF, \mathcal{S}) = \{\alpha_2\alpha_5\alpha_6\}$*, the agent can follow plan* $\alpha_2\alpha_5\alpha_6$ *to get to a pharmacy. The argumentation framework for planning* PAF *with lifting* $\trianglelefteq_E$ *can be represented as Fig. 5. Because*

$$\mathcal{E}_{\mathcal{PR}}(PAF) = \mathcal{E}_{\mathcal{GR}}(PAF) = \mathcal{E}_{\mathcal{ST}}(PAF) = \{\{\langle -\{gc\}, \alpha_2\alpha_5\alpha_6\rangle, \langle +\emptyset, \neg\alpha_1\alpha_4\rangle, \langle -\{sf\}, \neg\alpha_2\alpha_3\rangle\}\}$$

and thus $\text{Oplans}(PAF, \mathcal{S}) = \emptyset$.

When making plans, an agent must evaluate the available options based on his value system. Representation results express all the available plans with value promotion and demotion, and the agent has a preference order over values as part of the agent's value system. Intuitively, the agent can establish preferences over plans from representation results and preferences over values. However, since each plan has a set of promoted values and a set of demoted values, the agent must specify their preferences over plans from both aspects, which traditional lifting approaches cannot accommodate. In structured argumentation, like ASPIC+, people use lifting approaches to determine the defeat between two arguments based on preferences over rules and premises in each argument. Drawing inspiration from this, we suggest constructing both ordinary and blocking arguments for the execution of a plan in order to account for the promoted and demoted values associated with it. The success of one argument in defeating another depends on the preference order between the two sets of values pertaining to the arguments. In essence, rather than directly translating preferences over values into preferences over plans, we translate preferences over values into preferences over sets of values when determining the defeat relation between arguments, ultimately leading to accepted plans. This demonstrates that our argumentation-based approach serves as a dialogical justification for the use of lifting approaches and as a mediating mechanism between preferences over values and preferences over plans.

4 Argumentation and temporal logic for coalition formation

Argumentation has proven useful to provide a sound model to conceptualize reasoning processes related to *coalition formation* in multiagent systems [5; 6]. The underlying approach is based on using conflict and preference relationships among coalitions to determine which coalitions should be adopted by the agents according to a particular argumentation semantics, which can then be computed using a suitable proof theory.

A variant of modal logic suitable for temporal reasoning called Alternating-time Temporal Logic (ATL) [4] can provide a further development on the above concept,

making it possible to reason about the behavior and abilities of agents under various rationality assumptions [63; 64; 29]. In ATL the key construct has the form $\langle\langle A\rangle\rangle\phi$, which expresses that a coalition A of agents can *enforce* the formula ϕ. Under a model theoretic viewpoint, $\langle\langle A\rangle\rangle\phi$ holds whenever the agents in A have a winning strategy for ensuring the satisfiability of ϕ (independently of the behavior of A's opponents). However, this operator accounts only for the *theoretical existence* of such a strategy, not taking into account whether the coalition A can be actually formed. Indeed, in order to join a coalition, agents usually require some kind of *incentive* (e.g. sharing common goals, getting rewards, etc.), since usually forming a coalition does not come for free (fees have to be paid, communication costs may occur, etc.). Consequently, several possible coalition structures among agents may arise, from which the best ones should be adopted according to some rationally justifiable procedure.

In this section we present an argumentative approach to extend ATL for modelling coalitions. We provide a formal extension of ATL, CoalATL, by including a new construct $\langle\!|A|\!\rangle\phi$ which denotes that *the group A of agents is able to build a coalition B, $A \cap B \neq \emptyset$, such that B can enforce ϕ.* That is, it is assumed that agents in A work together and try to form a coalition B. The actual computation of the coalition is modelled in terms of a given argumentation semantics [41] in the context of coalition formation [5]. In a second step, we enrich CoalATL with goals. We address the question *why* agents should cooperate. Goals refer to agents' subjective incentive to join coalitions. We show that the proof theory for modelling coalitions in our framework can be embedded as a natural extension of the model checking procedure used in ATL.

4.1 Alternating-time Temporal Logic in a nutshell

Alternating-time Temporal Logic (ATL) [4] enables reasoning about temporal properties and strategic abilities of agents. The language of ATL is defined as follows.

Definition 4.1 ($\mathcal{L}_{ATL}$ [4]). *Let $\mathbb{A}\mathrm{gt} = \{a_1, \ldots, a_k\}$ be a nonempty finite set of all agents, and Φ be a set of propositions (with typical element p). We denote by "a" a typical agent, and by "A" a typical group of agents from $\mathbb{A}\mathrm{gt}$. $\mathcal{L}_{ATL}(\mathbb{A}\mathrm{gt}, \Phi)$ is defined by the following grammar: $\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\langle A\rangle\rangle \bigcirc \varphi \mid \langle\langle A\rangle\rangle \Box\, \varphi \mid \langle\langle A\rangle\rangle \varphi\, \mathcal{U}\, \varphi$.*

Informally, $\langle\langle A\rangle\rangle\varphi$ expresses that agents A have a *collective strategy to enforce* φ. ATL formulae include the usual temporal operators: $\bigcirc$ ("in the next state"), $\Box$ ("always from now on") and $\mathcal{U}$ (strict "until"). Additionally, $\Diamond$ ("now or sometime in the future") can be defined as $\Diamond\, \varphi \equiv \top\, \mathcal{U}\, \varphi$.

The semantics of ATL is defined by *concurrent game structures*. We recall that $\Phi = \{p, q, r, ...\}$ denotes a set of atomic propositions, $\mathbb{A}\text{gt} = \{a_1, \ldots, a_k\}$ is a set of *agents*, and $Act = \{\alpha_1, ..., \alpha_n\}$ is a set of actions.

Definition 4.2 (CGS [4])**.** *A* concurrent game structure (CGS) *is a tuple* $\mathcal{M} = \langle S, \mathcal{V}, d, o\rangle$*, where each of the components is defined as follows:*

- S *is a set of* states*.*
- $\mathcal{V} : S \to 2^{\Phi}$ *is a* valuation function*.*
- $d : \mathbb{A}\text{gt} \times S \to 2^{Act}$ *is a function that indicates the actions available to agent* $a \in \mathbb{A}\text{gt}$ *in state* $q \in S$*. We often write* $d_a(q)$ *instead of* $d(a, q)$*, and use* $d(q)$ *to denote the set* $d_{a_1}(q) \times \cdots \times d_{a_k}(q)$ *of* action profiles *in state* q*.*
- *Finally,* o *is a* transition function *which maps each state* $q \in S$ *and action profile* $\vec{\alpha} = \langle \alpha_1, \ldots, \alpha_k\rangle \in d(q)$ *to another state* $q' = o(q, \vec{\alpha})$*.*

Note that these structures can be seen as a special case of our generic labelled transition systems (Definition 2.2) where the set of labels is instantiated a the set of all action profiles. Moreover, the underlying relation $\mathcal{R}$ (here represented with o) is partially functional (just as in Definition 3.1). Nota also that "q" is not to be confused with a propostional variable, it is simply a state in which certain propositional variables are true (determined by $\mathcal{V}$).

A *path* $\lambda = q_0 q_1 \cdots \in S^{\omega}$ is an infinite sequence of states such that there is a transition between each q_i, q_{i+1}.We define $\lambda[i] = q_i$ to denote the i-th state of λ. The set of all paths starting in q is defined by $\Lambda_{\mathcal{M}}(q)$.

A (memoryless) *strategy* of agent a is a function $s_a : S \to Act$ such that $s_a(q) \in d_a(q)$. We denote the set of such functions by Σ_a. A *collective strategy* s_A for team $A \subseteq \mathbb{A}\text{gt}$ specifies an individual strategy for each agent $a \in A$; the set of A's collective strategies is given by $\Sigma_A = \prod_{a \in A} \Sigma_a$ and $\Sigma := \Sigma_{\mathbb{A}\text{gt}}$.

The *outcome* of strategy s_A in state q is defined as the set of all paths that may result from executing s_A: $out(q, s_A) = \{\lambda \in \Lambda_{\mathcal{M}}(q) \mid \forall i \in \mathbb{N}_0\ \exists \vec{\alpha} = \langle \alpha_1, \ldots, \alpha_k\rangle \in d(\lambda[i])\ \forall a \in A\ (\alpha_a = s_A^a(\lambda[i]) \wedge o(\lambda[i], \vec{\alpha}) = \lambda[i+1])\}$, where s_A^a denotes agent a's part of the collective strategy s_A.

Definition 4.3 (ATL Semantics)**.** *Let a* CGS $\mathcal{M} = \langle S, \mathcal{V}, d, o\rangle$ *and* $q \in S$ *be given. The semantics is given by a satisfaction relation* $\models$ *as follows:*

$\mathcal{M}, q \models p$ *iff* $p \in \mathcal{V}(q)$

$\mathcal{M}, q \models \neg\varphi$ *iff* $\mathcal{M}, q \not\models \varphi$

$\mathcal{M}, q \models \varphi \wedge \psi$ *iff* $\mathcal{M}, q \models \varphi$ *and* $\mathcal{M}, q \models \psi$

$\mathcal{M}, q \models \langle\langle A \rangle\rangle \bigcirc \varphi$ *iff there is* $s_A \in \Sigma_A$ *st.* $\mathcal{M}, \lambda[1] \models \varphi$ *for all* $\lambda \in out(q, s_A)$

$\mathcal{M}, q \models \langle\langle A \rangle\rangle \Box\, \varphi$ *iff there is* s_A *st.* $\mathcal{M}, \lambda[i] \models \varphi$ *for all* $\lambda \in out(q, s_A)$ *and* $i \in \mathbb{N}_0$

$\mathcal{M}, q \models \langle\langle A \rangle\rangle \varphi \mathcal{U} \psi$ *iff there is* $s_A \in \Sigma_A$ *st., for all* $\lambda \in out(q, s_A)$*, there is* $i \in \mathbb{N}_0$ *with* $\mathcal{M}, \lambda[i] \models \psi$*, and* $\mathcal{M}, \lambda[j] \models \varphi$ *for all* $0 \leq j < i$*.*

We note that the given semantics aligns well with Definition 4.1 and all the formulae introduced there.

4.2 Coalitions and argumentation

In this subsection, we provide an argument-based characterization of coalition formation that will be used later to extend ATL. We follow an approach similar to [5], where an argumentation framework for generating coalition structures is defined, generalizing the framework of Dung for argumentation [41], [1] extended with a *preference relation*. The basic notion is that of a *coalitional framework*, which contains a set of elements $\mathfrak{C}$ (usually seen as agents or coalitions), an attack relation (for modelling conflicts among elements of $\mathfrak{C}$), and a preference relation between elements of $\mathfrak{C}$ (to describe favorite agents/coalitions).

Definition 4.4 (Coalitional framework [5])**.** *A* coalitional framework *is a triple* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *where* $\mathfrak{C}$ *is a non-empty set of elements,* $att \subseteq \mathfrak{C} \times \mathfrak{C}$ *is an attack relation, and* $\prec$ *is a preorder on* $\mathfrak{C}$ *representing preferences on elements in* $\mathfrak{C}$*.*

Let S *be a non-empty set of elements.* $\mathbb{CF}(S)$ *denotes the set of all coalitional frameworks where elements are taken from the set* S*, i.e. for each* $\langle \mathfrak{C}, att, \prec \rangle \in \mathbb{CF}(S)$ *we have that* $\mathfrak{C} \subseteq S$*.*

The set $\mathfrak{C}$ in Definition 4.4 is intentionally generic, accounting for various possibilities. One is to consider $\mathfrak{C}$ as a set of agents $\mathbb{A}\mathrm{gt} = \{a_1, \ldots, a_k\}$: $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle \in \mathbb{CF}(\mathbb{A}\mathrm{gt})$. Then, a *coalition* is given by $C = \{a_{i_1}, \ldots, a_{i_l}\} \subseteq \mathfrak{C}$ and "agent" can be used as an intuitive reference to elements of $\mathfrak{C}$. Another possibility is to use a coalitional framework $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ based on $\mathbb{CF}(2^{\mathbb{A}\mathrm{gt}})$. Now elements of $\mathfrak{C} \subseteq 2^{\mathbb{A}\mathrm{gt}}$ are *groups* or *coalitions* (where we consider singletons as groups too) of agents. Under this interpretation a coalition $C \subseteq \mathfrak{C}$ is a *set of sets* of agents. Although "coalition" is already used for $C \subseteq \mathfrak{C}$, we also use the intuitive reading "coalition" or "group"

[1] The reader is referred to Section 2.2 for further details about Dung's approach to abstract argumentation.

Figure 6: Figure (a) (resp. (b)) corresponds to the coalitional frameworks defined in Example 4.5 (resp. 4.14 (b)). Nodes represent agents and arrows between nodes stand for the attack relation.

to address elements in $\mathfrak{C}$.[2] Yet another way is not to use the specific structure for elements in $\mathfrak{C}$, assuming it just consists of abstract elements, e.g. c_1, c_2, etc. One may think of these elements as individual agents or coalitions. This approach is followed in [5]. From now on we will mainly follow the first alternative when informally speaking about coalitional frameworks (i.e., we consider $\mathfrak{C}$ as a set of agents).

Example 4.5. *Consider the following two coalitional frameworks: (i)* $\mathcal{CF}_1 = \langle \mathfrak{C}, att, \prec \rangle$ *where* $\mathfrak{C} = \{a_1, a_2, a_3\}$, $att = \{\langle a_3, a_2\rangle, \langle a_2, a_1\rangle, \langle a_1, a_3\rangle\}$ *and agent* a_3 *is preferred over* a_1, *i.e.* $a_1 \prec a_3$; *and (ii)* $\mathcal{CF}_2 = \langle \mathfrak{C}', att', \prec' \rangle$ *where* $\mathfrak{C}' = \{\{a_1\}, \{a_2\}, \{a_3\}\}$, $att' = \{\langle \{a_3\}, \{a_2\}\rangle, \langle \{a_2\}, \{a_1\}\rangle, \langle \{a_1\}, \{a_3\}\rangle\}$ *and group* $\{a_3\}$ *is preferred over* $\{a_1\}$, *i.e.* $\{a_1\} \prec' \{a_3\}$. *They capture the same scenario and are isomorphic but* $\mathcal{CF}_1 \in \mathbb{CF}(\{a_1, a_2, a_3\})$ *and* $\mathcal{CF}_2 \in \mathbb{CF}(\mathcal{P}(\{a_1, a_2, a_3\}))$; *that is, the first framework is defined regarding single agents and the latter over (trivial) coalitions. Figure 6 (a) shows a graphical representation of the first coalitional framework.*

Let $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ be a coalitional framework. For $C, C' \in \mathfrak{C}$, we say that C *attacks* C' iff $C\, att\, C'$. The attack relation represents conflicts between elements of $\mathfrak{C}$; for instance, two agents may rely on the same (unique) resource or they may have disagreeing goals, which prevent them from cooperation. However, the notion of attack may not be sufficient for modelling conflicts, as some elements (resp. coalitions) in $\mathfrak{C}$ may be preferred over others. This leads to the notion of *defeater* which combines the notions of attack and preference. Following Dung's approach to abstract argumentation (see Section 2.2), members in a coalition may prevent attacks to members in the same coalition, *defending* each other. This prompts the following definitions:

Definition 4.6 (Defeater)**.** *Let* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *be a coalitional framework and let*

[2] The first interpretation is a special case of the second (coalitional frameworks are members $\mathbb{CF}(2^{\mathbb{A}\mathrm{gt}})$).

$C, C' \in \mathfrak{C}$. *We say that* C defeats C' *if, and only if,* C *attacks* C' *and* C' *is not preferred over* C *(i.e., not* $C \prec C'$*). We also say that* C *is a* defeater *for* C'.

Definition 4.7 (Defence)**.** *Let* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *be a coalitional framework and* $C, C' \in \mathfrak{C}$. *We say that* C' defends itself against C *if, and only if,* C' *is preferred over* C*, i.e.,* $C \prec C'$*, and* C' defends itself *if it defends itself against any of its attackers. Furthermore,* C *is* defended by *a set* $\mathfrak{S} \subseteq \mathfrak{C}$ *of elements of* $\mathfrak{C}$ *if, and only if, for all* C' *defeating* C *there is a coalition* $C'' \in \mathfrak{S}$ *defeating* C'.

In other words, if an element C' defends itself against C then C may attack C' but C is not allowed to defeat C'. A minimal requirement one should impose on a coalition is that its members do not defeat each other; otherwise, the coalition may be unstable and break up sooner or later because of conflicts among its members. This is formalised in the next definition.

Definition 4.8 (Conflict-free)**.** *Let* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *be a coalitional framework and* $\mathfrak{S} \subseteq \mathfrak{C}$ *a set of elements in* $\mathfrak{C}$*. Then,* $\mathfrak{S}$ *is called* conflict-free *if, and only if, there is* no $C \in \mathfrak{S}$ *defeating some member of* $\mathfrak{S}$.

It should be remarked that our notions of "defence" and "conflict-free" are defined in terms of "defeat" rather than "attack".[3] Given a coalitional framework $\mathcal{CF}$ we will use argumentation to compute coalitions with desirable properties. In argumentation theory, many different semantics have been proposed to define ultimately accepted arguments [41].

Definition 4.9 (Coalitional framework semantics)**.** *A* semantics *for a coalitional framework* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *is a (isomorphism invariant) mapping* $\mathcal{E}$ *which assigns to a given coalitional framework* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *a set of subsets of* $\mathfrak{C}$*, i.e.,* $\mathcal{E}(\mathcal{CF}) \subseteq \mathcal{P}(\mathfrak{C})$.

Let $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ be a coalitional framework. To formally characterize different semantics we will define a function $\mathcal{F}_{\mathcal{CF}} : \mathcal{P}(\mathfrak{C}) \to \mathcal{P}(\mathfrak{C})$ which assigns to a set of coalitions $\mathfrak{S} \in \mathcal{P}(\mathfrak{C})$ the coalitions defended by $\mathfrak{S}$.

Definition 4.10 (Characteristic function $\mathcal{F}$)**.** *Let* $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ *be a coalitional framework and* $\mathfrak{S} \subseteq \mathfrak{C}$*. The function* $\mathcal{F}$ *defined by*

$$\mathcal{F}_{\mathcal{CF}} : \mathcal{P}(\mathfrak{C}) \to \mathcal{P}(\mathfrak{C})$$
$$\mathcal{F}_{\mathcal{CF}}(\mathfrak{S}) = \{C \in \mathfrak{C} \mid C \text{ is defended by } \mathfrak{S}\}$$

is called characteristic function.[4]

[3] In [5; 6] these notions are defined the other way around, resulting in a different characterization of stable semantics.

[4] We omit the subscript $\mathcal{CF}$ if it is clear from context.

$\mathcal{F}$ can be applied recursively to coalitions resulting in new coalitions. For example, $\mathcal{F}(\emptyset)$ provides all undefeated coalitions and $\mathcal{F}^2(\emptyset)$ constitutes the set of all elements of $\mathfrak{C}$ which members are undefeated *or* are defended by undefeated coalitions.

Example 4.11. *Consider again the coalitional framework $\mathcal{CF}_1$ given in Example 4.5. The characteristic function applied on the empty set results in $\{a_3\}$ since the agent is undefeated, $\mathcal{F}(\emptyset) = \{a_3\}$. Applying $\mathcal{F}$ on $\mathcal{F}(\emptyset)$ determines the set $\{a_1, a_3\}$ because a_1 is defended by a_3. It is easy to see that $\{a_1, a_3\}$ is a fixed point of $\mathcal{F}$.*

We now introduce the first concrete semantics called coalition structure semantics, which was originally defined in [5].

Definition 4.12 (Coalition structure $\mathcal{E}_{\mathcal{CS}}$ [5])**.** *Let $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ be a coalitional framework. Then*

$$\mathcal{E}_{\mathcal{CS}}(\mathcal{CF}) := \left\{ \bigcup_{i=1}^{\infty} \mathcal{F}^i_{\mathcal{CF}}(\emptyset) \right\}$$

is called coalition structure semantics *or just* coalition structure *for $\mathcal{CF}$.*

For a coalitional framework $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ with a finite set $\mathfrak{C}$[5] the characteristic function $\mathcal{F}$ is continuous [41, Lemma 28]. Since $\mathcal{F}$ is also monotonic it has a least fixed point given by $\mathcal{F}(\emptyset) \uparrow^{\omega}$ (according to Knaster-Tarski). We have the following straightforward properties of coalition structure semantics.

Proposition 4.13 (Coalition structure)**.** *Let $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ be a coalitional framework with a finite set $\mathfrak{C}$. There is always a unique coalition structure for $\mathcal{CF}$. Furthermore, if no element of $C \in \mathfrak{C}$ defends itself then the coalitional structure is empty, i.e. $\mathcal{E}_{\mathcal{CS}}(\mathcal{CF}) = \{\emptyset\}$.*

Example 4.14. *The following situations illustrate the notion of coalitional structure:*

(a) Consider Example 4.11. Since $\{a_1, a_3\}$ is a fixed point of $\mathcal{F}_{\mathcal{CF}_1}$ the coalitional framework $\mathcal{CF}_1$ has $\{a_1, a_3\}$ as coalitional structure.

(b) $\mathcal{CF}_3 := \langle \mathfrak{C}, att, \prec \rangle \in \mathbb{CF}(\{a_1, a_2, a_3\})$ (shown in Figure 6(b)), is a coalitional framework with $\mathfrak{C} = \{a_1, a_2, a_3\}$, $att = \{ \langle a_1, a_2 \rangle, \langle a_1, a_3 \rangle , \langle a_2, a_1 \rangle, \langle a_2, a_3 \rangle, \langle a_3, a_1 \rangle\}$ and a_3 is preferred over a_2, $a_2 \prec a_3$, has the empty coalition as associated coalition str., i.e. $\mathcal{E}_{\mathcal{CS}}(\mathcal{CF}) = \{\emptyset\}$.

[5] Actually, it is enough to assume that $\mathcal{CF}$ is finitary (cf. [41, Def. 27]).

Since the coalition structure is often very restrictive, it seems reasonable to introduce other less restrictive semantics, following Dung's approach to abstract argumentation (see Section 2.2). We redefine these semantics in terms of the characteristic function $\mathcal{F}$:

Definition 4.15 (Argumentation Semantics)**.** *Let $\langle \mathfrak{C}, att, \prec \rangle$ be a coalitional framework, $\mathfrak{S} \subseteq \mathfrak{C}$ a set of elements of $\mathfrak{C}$. $\mathfrak{S}$ is called (a)* admissible extension *iff $\mathfrak{S}$ is conflict-free and $\mathfrak{S}$ defends all its elements, i.e. $\mathfrak{S} \subseteq \mathcal{F}(\mathfrak{S})$; (b)* complete extension *iff $\mathfrak{S}$ is conflict-free and $\mathfrak{S} = \mathcal{F}(\mathfrak{S})$; (c)* grounded extension *iff $\mathfrak{S}$ is the smallest (wrt. to set inclusion) complete extension; (d)* preferred extension *iff $\mathfrak{S}$ is a maximal (wrt. to set inclusion) admissible extension; (e)* stable extension *iff $\mathfrak{S}$ is conflict-free and it defeats all arguments not in $\mathfrak{S}$. Let $\mathcal{E}_{\mathcal{CS}}(\mathcal{CF})$ (resp. $\mathcal{E}_{\mathcal{CO}}(\mathcal{CF})$, $\mathcal{E}_{\mathcal{GR}}(\mathcal{CF})$, $\mathcal{E}_{\mathcal{PR}}(\mathcal{CF})$ and $\mathcal{E}_{\mathcal{ST}}(\mathcal{CF})$) denote the semantics which assigns to a coalitional structure $\mathcal{CF}$ all its admissible (resp. complete, grounded, preferred, and stable) extensions.*

There is only one unique coalition structure (possibly the empty one) for a given coalitional framework, but there can be several stable and preferred extensions. The existence of at least one preferred extension is assured which is not the case for the stable semantics. Thus, the possible coalitions very much depend on the used semantics.

Example 4.16. *For $\mathcal{CF}_3$ from Example 4.14 the following holds:*

$$\mathcal{E}_{\mathcal{CS}}(\mathcal{CF}) = \{\emptyset\}$$
$$\mathcal{E}_{\mathcal{AD}}(\mathcal{CF}) = \{\{a_1\}, \{a_2\}, \{a_3\}, \{a_2, a_3\}\}$$
$$\mathcal{E}_{\mathcal{CO}}(\mathcal{CF}) = \mathcal{E}_{\mathcal{GR}}(\mathcal{CF}) = \{\{a_1\}, \{a_2, a_3\}\} =$$
$$\mathcal{E}_{\mathcal{PR}}(\mathcal{CF}) = \mathcal{E}_{\mathcal{ST}}(\mathcal{CF}) = \{\{a_1\}, \{a_2, a_3\}\}$$

Analogously, for the coalitional framework $\mathcal{CF}_1$ from Example 4.5 there exists one complete extension $\{a_1, a_3\}$ which is also a grounded, preferred, and stable extension.

4.3 Coalitional ATL

In this section we combine *argumentation for coalition formation* and ATL and introduce *Coalitional* ATL (CoalATL). This logic extends ATL by new operators $\langle\!|A|\!\rangle$ for each subset $A \subseteq \mathbb{A}\mathrm{gt}$ of agents. These new modalities combine, or rather integrate, coalition formation into the original ATL cooperation modalities $\langle\langle A \rangle\rangle$. The intended reading of $\langle\!|A|\!\rangle\varphi$ is that the group A of agents *is able to form a coalition $B \subseteq \mathbb{A}\mathrm{gt}$ such that some agents of A are also members of B*, $A \cap B \neq \emptyset$, and *B can enforce φ*. Coalition formation is modelled by the formal argumentative approach in the

context of coalition formation, as described in Section 4.2, based on the framework developed in [5].

Our main motivation for this logic is to make it possible to reason about the ability of building coalition structures, and not only about an *a priori* specified group of agents (as it is the case for $\langle\langle A\rangle\rangle\varphi$). The new modality $\langle\!|A|\!\rangle$ provides a rather subjective view of the agents in A and their power to create a group B, $A \cap B \neq \emptyset$, which in turn is used to reason about the ability to enforce a given property.

The language of CoalATL is as follows.

Definition 4.17 ($\mathcal{L}_{ATL^c}$)**.** *Let* $\mathbb{A}\text{gt} = \{a_1, \ldots, a_k\}$ *be a finite, nonempty set of agents, and* Φ *be a set of propositions (with typical element* p*). We use the symbol "*a*" to denote a typical agent, and "*A*" to denote a typical group of agents from* $\mathbb{A}\text{gt}$*. The logic* $\mathcal{L}_{ATL^c}(\mathbb{A}\text{gt}, \Phi)$ *is defined by the following grammar:*

$$\begin{aligned}\varphi ::= & p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \langle\langle A\rangle\rangle \bigcirc \varphi \mid \langle\langle A\rangle\rangle \Box\, \varphi \mid \langle\langle A\rangle\rangle \varphi\, \mathcal{U}\, \varphi \mid \\ & \langle\!|A|\!\rangle \bigcirc \varphi \mid \langle\!|A|\!\rangle \Box\, \varphi \mid \langle\!|A|\!\rangle \varphi\, \mathcal{U}\, \varphi\end{aligned}$$

We extend concurrent game structures by means of *coalitional frameworks* and an *argumentative semantics*. A coalitional framework is assigned to each state of the model capturing the current "conflicts" among agents. In doing so, we allow that conflicts can change over time, being thus *state dependent*. Moreover, we assume that coalitional frameworks are agent-dependent. Thus, two intial groups of agents may have different skills to form coalitions. Consider for instance the following example.

Example 4.18. *Imagine two agents* a_1 *and* a_2 *which are not able (because they do not have the money) to convince* a_3 *and* a_4 *to join. But* a_1*,* a_2 *and* a_3 *together have the money and all four can enforce a property* φ*. So* $\{a_1, a_2\}$ *are not able to build a greater coalition to enforce* φ*; but* $\{a_1, a_2, a_3\}$ *are. So we are not looking at coalitions per se, but how they evolve from others.*

We assume that the argumentative semantics is the same for all states.

Definition 4.19 (CGM)**.** *A* coalitional game model *(CGM) is given by a tuple*

$$\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S}\rangle$$

where $\langle S, \mathcal{V}, d, o\rangle$ *is a* CGS*,* $\zeta : 2^{\mathbb{A}\text{gt}} \to (S \to \mathbb{CF}(\mathbb{A}\text{gt}))$ *is a function which assigns a coalitional framework over* $\mathbb{A}\text{gt}$ *to each state of the model subjective to a given group of agents, and* $\mathcal{S}$ *is an (argumentative) semantics as defined in Definition 4.9. The set of all such models is given by* $\mathbb{M}(S, \mathbb{A}\text{gt}, \Phi, \zeta, \mathcal{S})$*.*

A model provides an argumentation semantics $\mathcal{S}$ which assigns all formable coalitions to a given coalitional framework. As argued before we require from a valid coalition that it is not only justified by the argumentation semantics but that it is also not disjunct with the predetermined starting coalition. This leads to the notion *valid coalition*.

Definition 4.20 (Valid coalition)**.** *Let A, $B \subseteq \mathbb{A}\text{gt}$ be groups of agents, $\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S}\rangle$ be a* CGM *and $q \in S$. We say that B is a* valid coalition *with respect to A, q, and $\mathcal{M}$ whenever $B \in \mathcal{E}_{\mathcal{S}}(\zeta(A)(q)))$ and $A \cap B \neq \emptyset$. Furthermore, we use $\text{VC}_{\mathcal{M}}(A, q)$ to denote the set of all valid coalitions regarding A, q, and $\mathcal{M}$ (subscript $\mathcal{M}$ is omitted if clear from the context).*

Remark 4.21. *In [27] we assume that the members of the initial group A work together, whatever the reasons might be. So group A was added to the semantics. This ensured that agents in A can enforce ψ on their own, if they are able to do so. Even if A is* not *accepted originally by the argumentation semantics, i.e. $A \notin \mathcal{E}_{\mathcal{S}}(\zeta(A)(q))$. Here, we drop this requirement. As pointed out in [27] the "old" semantics is just a special case of this new one: The operator from [28] can be defined as $\langle\!|A|\!\rangle\gamma \vee \langle\langle A\rangle\rangle\gamma$.*

Moreover, we changed the condition that the predefined group given in the coalitional operator must be a subset of the formed coalition, $A \subseteq B$, to the weaker requirement that only some member of the inital coalition should be in the new one, $A \cap B \neq \emptyset$.

The semantics of the new modality is given by

Definition 4.22 (CoalATL Semantics)**.** *Let a* CGM *$\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S}\rangle$ a group of agents $A \subseteq \mathbb{A}\text{gt}$, and $q \in S$ be given. The semantics of Coalitional* ATL *extends that of* ATL*, given in Definition 4.3, by the following rule ($\langle\!|A|\!\rangle\psi \in \mathcal{L}_{ATL^c}(\mathbb{A}\text{gt}, \Phi)$):*

$\mathcal{M}, q \models \langle\!|A|\!\rangle\psi$ *iff there is a coalition $B \in \text{VC}_{\mathcal{M}}(A, q)$ such that $\mathcal{M}, q \models \langle\langle B\rangle\rangle\psi$.*

Remark 4.23 (Different Semantics, $\models_{\mathcal{S}}$)**.** *We have just defined a whole class of semantic rules for modality $\langle\!| \cdot |\!\rangle$. The actual instantiation of the semantics $\mathcal{S}$, for example $\mathcal{ST}$, $\mathcal{PR}$, and $\mathcal{CS}$ defined in Section 4.2, affects the semantics of the cooperation modality.*

For the sake of readability, we sometimes annotate the satisfaction relation $\models$ with the presently used argumentation semantics. That is, given a CGM *$\mathcal{M}$ with an argumentation semantics $\mathcal{S}$ we write $\models_{\mathcal{S}}$ instead of $\models$.*

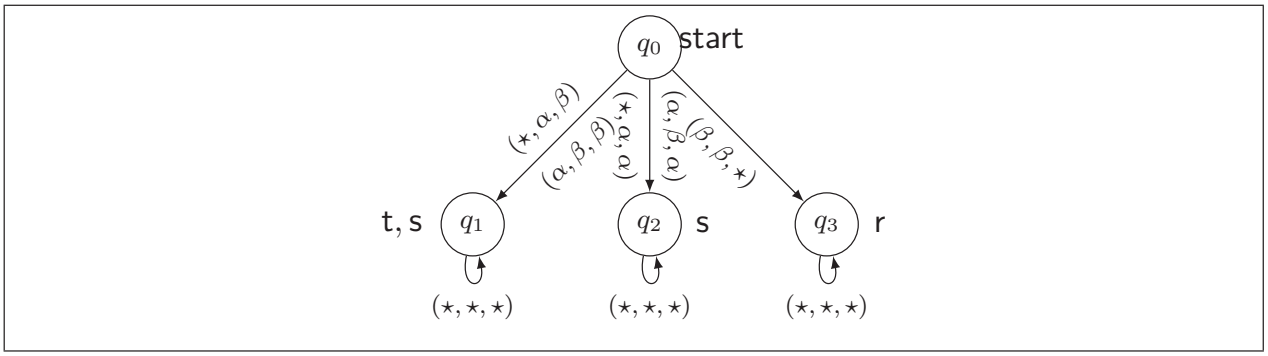


Figure 7: A simple CGS defined in Example 4.25.

The underlying idea of the semantic definition of $\langle\!|A|\!\rangle\psi$ is as follows. A given (initial) group of agents $A \subseteq \mathbb{A}\mathrm{gt}$ is able to form a *valid coalition* B (where A and B must not be disjoint), with respect to a given coalitional framework $\mathcal{CF}$ and a particular semantics $\mathcal{S}$, such that B can enforce ψ.

Similarly to the different possibilities in our definition of valid coalitions there are other sensible semantics for CoalATL. The semantics we presented here is not particularly dependent on time; i.e., except from the selection of a valid coalition B at the initial state there is no further interaction between time and coalition formation. We have chosen this simplistic definition to present our main idea—the connection of ATL and coalition formation by means of argumentation–as clear as possible. A precise approach dealing with time, however, is beyond the scope of this article.

Proposition 4.24 ([28])**.** *Let $A \subseteq \mathbb{A}\mathrm{gt}$ and $\langle\!|A|\!\rangle\psi \in \mathcal{L}_{ATL^c}(\mathbb{A}\mathrm{gt}, \Phi)$. Then it holds that $\langle\!|A|\!\rangle\psi \rightarrow \bigvee_{B \in 2^{\mathbb{A}\mathrm{gt}}, A \subseteq B} \langle\langle B \rangle\rangle \psi$ is a validity with respect to CGM's.*

Compared to ATL, a formula like $\langle\!|A|\!\rangle\varphi$ does *not* refer to the ability of A to enforce φ, but rather to the ability of A to *constitute* a coalition B, such that $A \cap B \neq \emptyset$, and then, in a second step, to the ability of B to enforce φ. Thus, two different notions of ability are captured in these new modalities. For instance, $\langle\langle A \rangle\rangle \psi \wedge \neg \langle\!|\mathbb{A}\mathrm{gt}|\!\rangle \psi$ expresses that group A of agents can enforce ψ, but there is no *reasonable* coalition at all which can enforce ψ (particularly not A, although they possess the theoretical power to do so).

The next example motivates the usefulness of the new modality. Classic ATL can only consider sets of agents that can enforce something, but it can not take into account whether such sets can indeed by formed (are allowed in the coalitional framework). The new modality, however, allows to model such situations.

Example 4.25. *There are three agents a_1, a_2, and a_3 which prefer different outcomes. Agent a_1 (resp. a_2, a_3) desires to get outcome* r *(resp.* s, t*). One may*

assume that all outcomes are distinct; for instance, a_1 is not satisfied with an outcome x whenever $x \neq \mathsf{r}$. Each agent can choose to perform action α or β. Action profiles and their outcomes are shown in Figure 7. The $\star$ is used as a placeholder for any of the two actions, i.e. $\star \in \{\alpha, \beta\}$. For instance, the profile $(\beta, \beta, \star)$ leads to state q_3 whenever agent a_1 and a_2 perform action β and a_3 either does α or β.

According to the scenario depicted in the figure, a_1 and a_2 cannot commonly achieve their goals. The same holds for a_1 and a_3. On the other hand, there exists a situation, q_1, in which both agents a_2 and a_3 are satisfied. One can formalise the situation as the coalitional game $\mathcal{CF} = \langle \mathfrak{C}, att, \prec \rangle$ given in Example 4.14(b), that is, $\mathfrak{C} = \mathbb{A}\mathrm{gt}$, $att = \{(a_1, a_2), (a_1, a_3), (a_2, a_1), (a_2, a_3), (a_3, a_1)\}$ and $a_2 \prec a_3$.

We formalise the example as the CGM $\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S} \rangle$ where $\mathbb{A}\mathrm{gt} = \{a_1, a_2, a_3\}$,$S = \{q_0, q_1, q_2, q_3\}$, $\Phi = \{\mathsf{r}, \mathsf{s}, \mathsf{t}\}$, and $\zeta(A)(q) = \mathcal{CF}$ for all states $q \in S$ and groups $A \subseteq \mathbb{A}\mathrm{gt}$. Transitions and the state labeling can be seen in Figure 7. Furthermore, we do not specify a concrete semantics $\mathcal{S}$ yet, and rather adjust it in the remainder of the example.

We can use pure ATL formulas, i.e. formulas not containing the new modalities $\langle\!| \cdot |\!\rangle$, to express what groups of agents can achieve. We have, for instance, that agents a_1 and a_2 can enforce a situation which is undesirable for a_3: $\mathcal{M}, q_0 \models \langle\!\langle a_1, a_2 \rangle\!\rangle \bigcirc \mathsf{r}$. Indeed, $\{a_1, a_2\}$ and the grand coalition $\mathbb{A}\mathrm{gt}$ (since it contains $\{a_1, a_2\}$) are the only coalitions which are able to enforce $\bigcirc \mathsf{r}$; we have

$$\mathcal{M}, q_0 \models \neg \langle\!\langle X \rangle\!\rangle \bigcirc \mathsf{r} \tag{1}$$

for all $X \subset \mathbb{A}\mathrm{gt}$ and $X \neq \{a_1, a_2\}$. Outcomes s or t can be enforced by a_2: $\mathcal{M}, q_0 \models \langle\!\langle a_2 \rangle\!\rangle \bigcirc (\mathsf{s} \vee \mathsf{t})$. Agents a_2 and a_3 also have the ability to enforce a situation which agrees with both of their desired outcomes: $\mathcal{M}, q_0 \models \langle\!\langle a_2, a_3 \rangle\!\rangle \bigcirc (\mathsf{s} \wedge \mathsf{t})$

These properties do not take into account the coalitional framework, that is, whether specific coalitions can be formed or not. By using the coalitional framework, we get

$$\mathcal{M}, q_0 \models_{\mathcal{S}} \langle\!\langle a_1, a_2 \rangle\!\rangle \bigcirc \mathsf{r} \wedge \neg \langle\!| a_1 |\!\rangle \bigcirc \mathsf{r} \wedge \neg \langle\!| a_2 |\!\rangle \bigcirc \mathsf{r}$$

for any semantics $\mathcal{S}$ introduced in Definition 4.9 and calculated in Example 4.16. The possible coalition (resp. coalitions) containing a_1 (resp. a_2) is $\{a_1\}$ (resp. are $\{a_2\}$ and $\{a_2, a_3\}$). But neither of these can enforce $\bigcirc \mathsf{r}$ (in q_0) because of (1). Thus, although it is the case that the coalition $\{a_1, a_2\}$ has the theoretical ability to enforce r in the next moment (which is a "losing" situation for a_3), a_3 should not consider it as sensible since agents a_1 and a_2 would not agree to constitute a coalition (according to the coalitional framework $\mathcal{CF}$).

The decision for a specific semantics is a crucial point and depends on the actual application. The next example shows that with respect to a particular argumenta-

tion semantics, agents are able to form a coalition which can successfully achieve a given property, whereas another argumentative semantics does not allow that.

Example 4.26. CoalATL *can be used to determine whether a coalition for enforcing a specific property exists. Assume that* $\mathcal{S}$ *represents the grounded semantics. For instance, the statement*

$$\mathcal{M}, q_0 \models_{\mathcal{E}_{\mathcal{GR}}} \langle\!|\emptyset|\!\rangle \bigcirc \mathtt{t}$$

expresses that there is a grounded coalition (i.e. a coalition wrt to the grounded semantics) which can enforce $\bigcirc \mathtt{t}$*, namely the coalition* $\{a_2, a_3\}$*. This result does not hold for all semantics; for instance, we have*

$$\mathcal{M}, q_0 \not\models_{\mathcal{E}_{\mathcal{CS}}} \langle\!|\emptyset|\!\rangle \bigcirc \mathtt{t}$$

with respect to the coalition structure semantics, since the coalition structure is the empty coalition and $\mathcal{M}, q_0 \not\models \langle\langle\emptyset\rangle\rangle \bigcirc \mathtt{t}$*.*

Note that it is easily possible to extend the language by an *update mechanism*, in order to compare different argumentative semantics using formulae inside the object language.

4.4 Cooperation and Goals

Why should agents join coalitions? Up to now we did not address this question and focussed on *why not* to cooperate. Often cooperation does not come for free and it requires some kind of incentive (i.e. sharing common goals or getting rewards) to offer one's ability in order to support other agents. Coalitional frameworks, however, were mainly used to model conflicts between agents, and therewith, avoid cooperation. In [28] the authors propose *goals* as agents' incentives to join coalitions; the following is based on that work.

We are now incorporating a *goal framework* into CoalATL models. First of all, each agent is equipped with a *set of goals* $\mathcal{G}_a$ where $a \in \mathbb{A}\mathrm{gt}$ and $\mathcal{G} := \bigcup_{a \in \mathbb{A}\mathrm{gt}} \mathcal{G}_a$. Goals are formulated as ATL-path formulae or conjunctions of them. An agent, say Bill, might have the goal—or rather a dream—that it will sometimes be able to buy a new car *without* asking other people (e.g. its wife Ann). Such a goal can be formulated as $\Diamond \langle\langle \mathrm{Bill} \rangle\rangle \bigcirc \mathsf{buyNewCar}$. Sometimes Bill would like to *enforce* to buy a new car in the next moment. To assign goals to agents a CGM is extended by a *goal mapping*.

Definition 4.27 (Goal mapping $\mathfrak{g}$)**.** *A* goal mapping *is a function* $\mathfrak{g} : \mathbb{A}\mathrm{gt} \to (S^+ \to \mathcal{P}(\mathcal{G}))$ *assigning a set of goals to a given sequence of states and an agent.*

So, a goal mapping assigns a set of goals to a *history*. This is needed to introduce goals into CGM's. The history dependency can be used, for instance, to model when a goal should be removed from the list: An agent having a goal $\Diamond$ s may drop it after reaching a state in which s holds.

So far, we did not say how goals can be actually used to form coalitions. We assume, given some task, that agents having goals satisfied or partly satisfied by the outcome of the task are willing to cooperate to bring about the task. Consider, for instance, the ATL formula $\langle\langle A\rangle\rangle\gamma$. It says that A can enforce γ—the *objective*. In the context of Coalitional ATL it is even more intuitive: $\langle\!|A|\!\rangle\gamma$ means that A is able to from a coalition B which can enforce the objective γ. Of course, rational agents should have reasons to bring about γ in order to work towards γ. In the following we will use the notion *objective* (or objective formula) to refer to both the task itself and the outcome of it. A typical objective is written as o. Agents which have goals fulfilled or at least partly fullfilled by objective o are possible candidates to participate in a coalition aiming at o. We consider *CoalATL objectives* which are CoalATL path formulae.

We say that an objective o *satisfies* goal g, $o \hookrightarrow g$, if the goal g is fulfilled after o has been accomplished. Intuitively, an objective $\Box$ t satisfies goal $\Box$ (t $\vee$ s) and supports goal $\Diamond$ t.

4.5 Coalitional ATL with Goals

In this section, we merge together Coalitional ATL with the goal framework described above. The syntax of the logic is given as in Definition 4.17. The necessary change takes place in the semantics. We redefine what it means for a coalition to be *valid*.

Up to now, valid coalitions were solely determined by coalitional frameworks. Conflicts represented by such frameworks are a coarse, but necessary, criterion for a successful coalition formation process. However, nothing is said about incentives *to join* coalitions, only why coalitions should *not* be joined.

Goals allow us to capture the first issue. For a given objective formula o and a finite sequence of states, called *history*, we only consider agents which have some goal supported by the current objective. *CGM's with goals* are given as a straightforward extension of CGM's (cf. Definition 4.19).

Definition 4.28 (CGM with goals)**.** *A* CGM with goals *(CGMG) $\mathcal{M}$ is given by a model of $\mathbb{M}(S, \mathbb{A}\mathrm{gt}, \Phi, \mathcal{S}, \zeta)$ extended by a set of goals $\mathcal{G}$ and a goal mapping $\mathfrak{g}$ over $\mathcal{G}$. The set of all such models is denoted by $\mathbb{M}^g(S, \mathbb{A}\mathrm{gt}, \Phi, \mathcal{S}, \zeta, \mathcal{G}, \mathfrak{g})$ or just $\mathbb{M}^g$ if we assume standard naming.*

To define the semantics we need some additional notation. Given a path $\lambda \in S^\omega$

we use $\lambda[i,j]$ to denote the sequence $\lambda[i]\lambda[i+1]\ldots\lambda[j]$ for $i,j \in \mathbb{N}_0 \cup \{\infty\}$ and $i \leq j$. A *history* is a finite sequence $h = q_1 \ldots q_n \in S^+$, $h[i]$ denotes state q_i if $n \geq i$, q_n for $i \geq n$, and ε for $i < 0$ where $i \in \mathbb{Z} \cup \{\infty\}$. Furthermore, given a history h and a path or history λ the combined path/history starting with h extended by λ is denoted by $h \circ \lambda$.

Finally, we present the semantics of CoalATL *with* goals. It is similar to Definition 4.22. Here, however, it is necessary to keep track of the steps (visited states) made to determine the goals of the agents. The finite list of steps already taken is denoted by τ.

Definition 4.29 (Goal-based semantics of $\mathcal{L}_{ATL^c}$)**.** *Let $\mathcal{M}$ be a* CGMG*, q a state, and $i,j \in \mathbb{N}_0$. Let $\tau \in S+$, any finite sequence of states already visited. The goal-based semantics of $\mathcal{L}_{ATL^c}$ formulae is given as follows:*

$\mathcal{M}, q, \tau \models p$ *iff* $p \in \mathcal{V}(q)$

$\mathcal{M}, q, \tau \models \varphi \wedge \psi$ *iff* $\mathcal{M}, q, \tau \models \varphi$ *and* $\mathcal{M}, q, \tau \models \psi$

$\mathcal{M}, q, \tau \models \neg\varphi$ *iff not* $\mathcal{M}, q, \tau \models \varphi$

$\mathcal{M}, q, \tau \models \langle\langle A \rangle\rangle \bigcirc \varphi$ *iff there is* $s_A \in \Sigma_A$ *such that* $\mathcal{M}, \lambda[1], \tau \circ \lambda[1] \models \varphi$ *for all* $\lambda \in out(q, s_A)$

$\mathcal{M}, q, \tau \models \langle\langle A \rangle\rangle \Box\, \varphi$ *iff there is* s_A *such that* $\mathcal{M}, \lambda[i], \tau \circ \lambda[1,i] \models \varphi$ *for all* $\lambda \in out(q, s_A)$ *and* $i \in \mathbb{N}_0$

$\mathcal{M}, q, \tau \models \langle\langle A \rangle\rangle \varphi \mathcal{U}\, \psi$ *iff there is* $s_A \in \Sigma_A$ *such that, for all* $\lambda \in out(q, s_A)$*, there is* $i \in \mathbb{N}_0$ *with* $\mathcal{M}, \lambda[i], \tau \circ \lambda[1,i] \models \psi$*, and* $\mathcal{M}, \lambda[j], \tau \circ \lambda[1,j] \models \varphi$ *for all* $0 \leq j < i$.

$\mathcal{M}, q, \tau \models \langle\!| A |\!\rangle \bigcirc \varphi$ *iff there is* $s_A \in \Sigma_A$ *such that* $\mathcal{M}, \lambda[1], \tau \circ \lambda[1] \models \varphi$ *for all* $\lambda \in out(q, s_A)$

$\mathcal{M}, q, \tau \models \langle\!| A |\!\rangle \Box\, \varphi$ *iff there is* s_A *such that* $\mathcal{M}, \lambda[i], \tau \circ \lambda[1,i] \models \varphi$ *for all* $\lambda \in out(q, s_A)$ *and* $i \in \mathbb{N}_0$

$\mathcal{M}, q, \tau \models \langle\!| A |\!\rangle \varphi \mathcal{U}\, \psi$ *iff there is* $s_A \in \Sigma_A$ *such that, for all* $\lambda \in out(q, s_A)$*, there is* $i \in \mathbb{N}_0$ *with* $\mathcal{M}, \lambda[i], \tau \circ \lambda[1,i]$ *models* ψ*, and* $\mathcal{M}, \lambda[j], \tau \circ \lambda[1,j]$ *models* φ *for all* $0 \leq j < i$.

Ultimately, we are interested in $\mathcal{M}, q \models \varphi$ *defined as* $\mathcal{M}, q, q \models \varphi$.

All the new functionality provided by goals is captured by the new valid coalition function VC^g.

Definition 4.30 (Valid coalitions,$\text{VC}^g(q, A, o, \tau)$)**.** *Let $\mathcal{M} \in \mathbb{M}^g$, $\tau \in S^+$, $A, B \subseteq \mathbb{A}\text{gt}$, o an* CoalATL *objective.*

We say that B is a valid coalition *after τ with respect to A, o, and $\mathcal{M}$ if, and only if,*

1. *$B \in \mathcal{E}(\zeta(\tau[\infty])(A))$, $A \cap B \neq \emptyset$, and*
2. *there are goals $g_{b_i} \in \mathfrak{g}_{b_i}(\tau)$, one per agent $b_i \in B$, such that $o \hookrightarrow_{\mathcal{M},\tau,B} g_{b_1} \wedge \cdots \wedge g_{b_{|B|}}$*

The set $\text{VC}^g(q, A, o, \tau)$ consists of all such valid coalitions wrt to $\mathcal{M}$.

Thus, for the definition of valid coalitions among other things, a goal mapping, a function ζ and a sequence of states τ are required. The intuition of τ is that it represents the history (the sequence of states visited so far including the current state). So, τ is used to determine which goals of the agents are still active.

Finally, we have to define when a goal is satisfied.

Definition 4.31 (Satisfaction of goals)**.** *Let g be an* ATL*-goal, o an $\mathcal{L}_{ATL^c}$-objective, and $\tau \in S^+$. We say that objective o* satisfies *g, for short $o \hookrightarrow_{\mathcal{M},\tau,B} g$, with respect to $\mathcal{M}, \tau$, and B if, and only if, there is a strategy $s_B \in \Sigma_B$ such that*

1. *for all $\lambda \in out(\tau[\infty], s_B)$: $\mathcal{M}, \lambda, \tau \models o$ implies $\mathcal{M}, \lambda \models g$, and*
2. *there is some path $\lambda \in out(\tau[\infty], s_B)$ with $\mathcal{M}, \lambda, \tau \models o$.*

A goal is satisfied by an objective if each path (enforceable by B) that satisfies the objective does also satisfy the goal. That is, satisfaction of the objective will guarantee that the goal becomes true. The second condition ensures that the coalition actually has a way to bring about the goal. However, in [28] it is shown that the second condition is superfluous.

4.6 Model Checking ATL^c

In this section, we present an algorithm for model checking CoalATL formulae. The model checking problem is given by the question whether a given CoalATL formula follows from a given CGM $\mathcal{M}$ and state q, i.e. whether $\mathcal{M}, q \models \varphi$ [36]. In [4] it is shown that model checking ATL is **P**-complete, with respect to the number of transitions of $\mathcal{M}$, m, and the length of the formula, l, and can be done in time $\mathcal{O}(m \cdot l)$.

For CoalATL we also have to treat the new coalitional modalities in addition to the normal ATL constructs. Let us consider the formula $\langle\!\langle A \rangle\!\rangle \psi$. According to the semantics of $\langle\!\langle A \rangle\!\rangle$, given in Definition 4.22, we must check whether there is a coalition

B such that (i) $A \cap B \neq \emptyset$, (ii) B is acceptable by the argumentation semantics, and (iii) $\langle\langle B \rangle\rangle \psi$. The number of possible candidate coalitions B which satisfy (i) and (ii) is bounded by $|2^{\mathbb{A}\mathrm{gt}}|$. Thus, in the worst case there might be *exponentially* many calls to a procedure checking whether $\langle\langle B \rangle\rangle \psi$. Another source of complexity is the time needed to compute the argumentation semantics. In [43], for instance, it is stated that credulous acceptance[6] using preferred semantics is **NP**-complete.

Both considerations together suggest that the model checking complexity has two computationally hard parts: exponentially many calls to $\langle\langle A \rangle\rangle \psi$ and the computation of the argumentation semantics. Indeed, Theorem 4.32 will support this intuition. However, we show that it is possible to "combine" both computationally hard parts to obtain an algorithm which is in $\mathbf{\Delta_2^P} = \mathbf{P^{NP}}$, if the computational complexity to determine whether a given coalition is acceptable are not harder than **NP**.

For the rest of this section, we will denote by $\mathcal{L}_{\mathcal{S},\mathcal{CF}}$ the set of all coalitions A such that A is acceptable according to the coalitional framework $\mathcal{CF}$ and the argumentation semantics $\mathcal{S}$, i.e. $\mathcal{L}_{\mathcal{S},\mathcal{CF}} := \{A \mid A \in \mathcal{E}(\mathcal{CF})\}$.

Given some complexity class $\mathcal{C}$, we use the notation $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathcal{C}$ to state that the word problem of $\mathcal{L}_{\mathcal{S},\mathcal{CF}}$, i.e., whether A is a member of $\mathcal{L}_{\mathcal{S},\mathcal{CF}}$, is in $\mathcal{C}$. Actually in [28] it was stated that $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathbf{P}$ for all semantics $\mathcal{S}$ defined in Definition 4.15. In Figure 8 we propose a model checking algorithm for CoalATL. The complexity result given in the next theorem is modulo the complexity needed to compute membership in $\mathcal{L}_{\mathcal{S},\mathcal{CF}}$.

Theorem 4.32 (Model checking CoalATL [28])**.** *Let a CGM $\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S} \rangle$ be given, $q \in S$, $\varphi \in \mathcal{L}_{ATL^c}(\mathbb{A}\mathrm{gt}, \Phi)$, and $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathcal{C}$. Model checking CoalATL with respect to the argumentation semantics $\mathcal{S}$[7] is in $\mathbf{P^{NP^C}}$.*

The last theorem gives an upper bound for model checking CoalATL with respect to an arbitrary but fixed semantics $\mathcal{S}$. A finer grained classification of the computational complexity of $\mathcal{L}_{\mathcal{S},\mathcal{CF}}$ allows to improve the upper bound given in Theorem 4.32. Assume that $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathbf{P}$ and consider the last case of function *mcheck* in Figure 8 labelled by $(\star)$, $\varphi \equiv \langle\!|A|\!\rangle \psi$. First, a coalition $B \in 2^{\mathbb{A}\mathrm{gt}}$ is non-deterministically chosen and then, it is checked whether B satisfies the three conditions (1-3) in $(\star)$. Each of the three tests can be done in deterministic polynomial time. Hence, the verification of $\mathcal{M}, q \models \langle\!|A|\!\rangle \psi$, in the last case, meets the "guess and verify" principle which is characteristic for problems in **NP**. This brings the overall complexity of the algorithm to $\mathbf{\Delta_2^P}$. More surprisingly, the same result holds even for the case where $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathbf{NP}$.

[6] That is, whether an argument is in *some* preferred extension.

[7] That is, whether $\mathcal{M}, q \models_{\mathcal{S}} \varphi$.

function $mcheck(\mathcal{M}, q, \varphi)$**;**
Given a CGM $\mathcal{M} = \langle S, \mathcal{V}, d, o, \zeta, \mathcal{S}\rangle$, a state $q \in S$, and $\varphi \in \mathcal{L}_{ATL^c}(\mathbb{A}\mathrm{gt}, \Phi)$ the algorithm returns $\top$ if, and only if, $\mathcal{M}, q \models_{\mathcal{S}} \varphi$.

case φ **contains no** $\langle\!|B|\!\rangle$**: if** $q \in mcheck_{\mathrm{ATL}}(\mathcal{M}, \varphi)$ **return** $\top$ **else** $\bot$

case φ **contains some** $\langle\!|B|\!\rangle$**:**

case $\varphi \equiv \neg\psi$**: return** $\neg(\mathcal{M}, q, \psi)$

case $\varphi \equiv \psi \vee \psi'$**: return** $mcheck(\mathcal{M}, q, \psi) \vee mcheck(\mathcal{M}, q, \psi')$

case $\varphi \equiv \langle\langle A\rangle\rangle T\psi$**:** Label all states q' where $mcheck(\mathcal{M}, q', \psi) == \top$ with a new proposition yes and return $mcheck(\mathcal{M}, q, \langle\langle A\rangle\rangle T\mathsf{yes})$; T stands for $\Box$ or $\bigcirc$.

case $\varphi \equiv \langle\langle A\rangle\rangle \psi\, \mathcal{U}\, \psi'$**:** Label all states q' where $mcheck(\mathcal{M}, q', \psi) == \top$ with a new proposition yes_1, all states q' where $mcheck(\mathcal{M}, q', \psi') == \top$ with a new proposition yes_2 and return $mcheck(\mathcal{M}, q, \langle\!|A|\!\rangle \mathsf{yes}_1\, \mathcal{U}\, \mathsf{yes}_2)$

case $\varphi \equiv \langle\!|A|\!\rangle T\psi$**,** ψ **contains some** $\langle\!|C|\!\rangle$**:** Label all states q' where $mcheck(\mathcal{M}, q', \psi) == \top$ with a new proposition yes and return $mcheck(\mathcal{M}, q, \langle\!|A|\!\rangle T\mathsf{yes})$; T stands for $\Box$ or $\bigcirc$.

case $\varphi \equiv \langle\!|A|\!\rangle \psi\, \mathcal{U}\, \psi'$**,** ψ **or** ψ' **contain some** $\langle\!|C|\!\rangle$**:** Label all states q' where $mcheck(\mathcal{M}, q', \psi) == \top$ with a new proposition yes_1, all states q' where $mcheck(\mathcal{M}, q', \psi') == \top$ with a new proposition yes_2 and return $mcheck(\mathcal{M}, q, \langle\!|A|\!\rangle \mathsf{yes}_1\, \mathcal{U}\, \mathsf{yes}_2)$

case $\varphi \equiv \langle\!|A|\!\rangle \psi$ **and** ψ **contains no** $\langle\!|C|\!\rangle$**:** Non-deterministically choose $B \in 2^{\mathbb{A}\mathrm{gt}}$

if

(1) $B \in (\mathcal{E}(\zeta(A)(q)))$,
(2) $A \cap B \neq \emptyset$, and $\qquad (\star)$
(3) $q \in mcheck_{\mathrm{ATL}}(\mathcal{M}, \langle\langle B\rangle\rangle \psi)$

then return $\top$ **else** $\bot$

function $mcheck_{\mathbf{ATL}}(\mathcal{M}, \varphi)$**;**
Given a CGS $\mathcal{M} = \langle S, \mathcal{V}, d, o\rangle$ and $\varphi \in \mathcal{L}_{ATL}(\mathbb{A}\mathrm{gt}, \Phi)$, the standard ATL model checking algorithm (cf. [4]) returns all states q with $\mathcal{M}, q \models_{\mathrm{ATL}} \varphi$.

return $\{q \in S \mid \mathcal{M}, q \models_{\mathrm{ATL}} \varphi\}$

Figure 8: A model checking algorithm for CoalATL

Corollary 4.33 ([28]). *If $\mathcal{L}_{\mathcal{S},\mathcal{CF}} \in \mathbf{NP}$ (resp. $\mathbf{NP}$-complete) then model checking* CoalATL *is in $\mathbf{\Delta_2^P}$ (resp. $\mathbf{\Delta_2^P}$-complete) with respect to $\mathcal{E}$.*

In [43] the complexity of credulous reasoning with respect to the preferred and stable extensions is analyzed and determined to be **NP**-complete. This is in the line with our result: there can be a polynomial number of calls to $mcheck(\mathcal{M}, q, \langle\!|A|\!\rangle\psi)$ (where ψ does not contain any cooperation modality $\langle\!| \cdot |\!\rangle$). Now, the problem of checking whether $mcheck(\mathcal{M}, q, \langle\!|A|\!\rangle\psi)$ holds is very similar to checking whether some argument is credulously accepted. In both cases we have to ask for the existence of a set X with specific properties (in our framework we refer to X as a coalition and in [43] as an argument) which can be validated in polynomial deterministic time.

Corollary 4.34 ([28]). *Model checking* CoalATL *is in $\mathbf{\Delta_2^P}$ for all semantics defined in Definition 4.15.*

5 Argumentation and epistemic logic

Doxastic and epistemic logics are the branches of modal logics that investigate the properties of belief (*dóxa* in ancient Greek) and knowledge (*epistēmē*), both in single and multi-agent contexts. There are several connections between argumentation and the analysis of knowledge and belief, that one can abridge as an influence in both directions. On the one hand, arguments inform our beliefs about the world and provide the grounds for many things we claim to know. Conversely, our beliefs and the things we know shape the way we produce and put forward arguments. The potential of combining analytic tools from doxastic/epistemic logic and formal argumentation can be easily argued for in many areas of application. Yet, such a combination is a relatively recent endeavour, most of the work dating back only to the last decade or so.

In what follows, we present the aim and rationale of the most salient efforts in this sense, and situate them along to the two just mentioned directions of influence. The rest of this section proceeds as follows. We first provide some essential background on epistemic and doxastic logics (Section 5.1) and hint at some advances in their field that are relevant for combining them with argumentation. In Section 5.2 we provide a more articulated description of the rationale for combining tools from these disciplines in both directions of influence, i.e. from arguments to knowledge and belief in Section 5.2.1 and from knowledge and belief to arguments in Section 5.2.2. Finally, in Section 5.3 we overview recent works exploring the first direction of influence, and we do the same in Section 5.4 for works along the second direction.[8]

[8]Part of the content of the whole section builds upon previous work of Antonio Yuste-Ginel

5.1 Epistemic logic and reasoning about knowledge

Fundamental philosophical questions such as 'what is to know something?' and 'how knowledge differs from mere belief?' can be traced back, in the western philosophical tradition, at least to Plato's *Theaetetus*. Epistemic and doxastic logics as an axiomatic deductive rendering of the notions of, respectively, knowledge and belief, have a much more recent history. These fields date back to the seminal work of von Wright[101] and the subsequent more systematic treatment by Hintikka [62], which introduced relational (Kripke) models as their standard semantics.

In this framework, knowledge and belief are interpreted as universal modalities (expressed by a $\Box$-operator) where modal notions such as 'it is known that ϕ' (resp. 'it is believed that ϕ') are interpreted as 'ϕ is the case in all states that are accessible from the actual one'. In most of what follows, to keep things simple, we treat knowledge and belief as separate and independent modalities, specifying the interpretation of $\Box$ in each context. It should be noticed though that this is not the only possible option. Indeed, a long tradition in epistemology, dating back to Plato, identifies knowledge as a derivative notion, i.e. as some form of true belief. As a consequence, modal approaches inspired by this tradition formalise belief and knowledge as interdependent modalities, most of the time with belief as a primitive modality and knowledge as defined by it.[9] Yet another option, that we will touch upon in what follows, is to treat both belief and knowledge as derivative modalities, e.g., by grounding them on the arguably more primitive concept of *evidence*.

At an intuitive level, knowledge and belief have different properties which translate into specific axioms. Knowledge is usually required to satisfy *factivity*: to know that ϕ implies that ϕ is true, which arguably does not hold in the case of simple belief. Factivity is expressed by the axiom schema (T) $\Box\phi \rightarrow \phi$ (Section 2.1), which defines reflexivity at the level of structures. Belief is instead often required to satisfy the condition that it is not possible to believe a contradiction, expressed by the schema $\neg\Box \bot$. The latter is equivalent to schema (D) $\Box\phi \rightarrow \Diamond\phi$. In fact, both formulas define *seriality* of the accessibility relation, i.e. for any state s there is always some state t accessible from s. Since reflexivity entails seriality, the doxastic interpretation of $\Box$ puts a weaker constraint on the accessibility relation than the epistemic interpretation.

Additional properties for both knowledge and belief are so-called *positive* and

and Carlo Proietti [30; 31; 81; 82; 107]. The exposition of the material is inspired by the PhD dissertation of the first author [106, Chp. 5], although novel approaches are discussed here, and the presentation has been systematically improved and expanded.

[9] The converse option to treat belief as derived from knowledge as primitive has also been put forward in recent epistemological discussion [103; 104] or in well-known approaches to the dynamics of epistemic attitudes [16].

negative introspection. Positive introspection postulates that anything that is known (resp. believed) is also known to be known (resp. believed to be believed), and is captured by the axiom schema (4) $\Box\phi \rightarrow \Box\Box\phi$, which defines transitivity. Negative introspection instead means that anything that is not known (resp. not believed) is also known to be not known (resp. believed to be not believed) and is expressed by (5) $\neg\Box\phi \rightarrow \Box\neg\Box\phi$, which defines *euclidianity*: any two states that are accessible from a third one have access to each other. The more or less 'standard' calculus for doxastic logic is KD45, i.e. the system K of normal modal logic augmented with axioms (D), (4) and (5). The status of both axioms (4) and (5) is instead debated with regard to the epistemic reading of $\Box$. Many philosophers tend to reject both of them, assuming (T) as the only valid axiom schema for knowledge. On the other hand, computer scientists usually accept both, taking the system S5 (i.e. K + (T) + (4) + (5)) as a viable axiomatization of epistemic logic, i.e. one where the accessibility relation is an equivalence relation.

In general, both knowledge and belief may have different meanings depending on the context of application. Consequently, their modal rendering as a $\Box$-operator may obey different properties, which entails the validity or invalidity of different axiom schemas. In this sense, even the axioms and rules of the basic system of normal modal logic T may be disputed. For example, accepting the necessitation rule N – inferring $\vdash \Box\phi$ from $\vdash \phi$ – entails that all logical validities are known. The latter seems fine when modelling what an agent *implicitly* knows, or can infer in principle, but is inadequate when modelling the *explicit* knowledge of agents with limited computational resources. This is known as the problem of *logical omniscience*. Otherwise, in a doxastic context, we may read the belief operator $\Box\phi$ as the 'agent assigns high probability to ϕ being true'. Here, the formula $(\Box\phi \wedge \Box\psi) \rightarrow \Box(\phi \wedge \psi)$ fails to hold in general: the fact that two separate events are highly probable does not entail that their conjunction is. However, this formula is a logical consequence of the axioms of K.

In cases like these, there are two main strategies of approach. On the one hand, one can add new operators to the basic language in order to express more nuanced epistemic or doxastic concepts. This is, for example, the strategy of *awareness logics* [45; 46], where an awareness operator $A\phi$ – meaning 'the agent is aware of ϕ' – is added to the language and used to define explicit knowledge in conjunction with $\Box$. On the other hand, one can weaken the basic logic and, consequently, change its semantics. This is the case of *neighbourhood semantics* [34] – that we will encounter in Section 5.3 – which do not validate all axioms and inference rules of K, as the ones mentioned in the previous paragraph. The same outcome may be obtained by defining the $\Box$-operator of knowledge or belief on top of a different operator with a neighbourhood semantics, in our case an *evidence* modality [78]. All these strategies

have been applied at the interplay between argumentation and epistemic logic as we will see in what follows.

5.2 A twofold influence

5.2.1 Influence 1: From arguments to knowledge and beliefs

Our beliefs about the world are shaped by the evidence we encounter, which can be either direct (e.g., *seeing*) or indirect (e.g., by *testimony* or by *inference*). Such evidence is often of an argumentative nature. I may believe that Jones owns a Ford because I have seen him riding one (direct evidence). Yet, this belief may be defeated by Smith telling me that Jones is around with his company car, which makes an argument to the conclusion that I have not seen him riding his own car. In recent years, doxastic and epistemic logics have been combined with abstract argumentation with the aim to explore the many senses in which belief can be supported or defeated by arguments. In Section 5.3 we present some of these approaches [30; 31; 57; 87; 90]. This line of investigation has a strong link with central issues in epistemology. One of them is the debate around the so-called JTB thesis, according to which knowledge is to be defined as *justified true belief*. This thesis has been harshly debated since Edmund Gettier raised a number of famous objections against it in a famous paper [51]. The core of the issue lies in the fact that the central notion of justification needs specification. In fact, abstract argumentation provides a full theory of justification (as defence against counterarguments). In this respect, it naturally works as a tool to assess the JTB theory. A first approach along these lines is to be found in [89; 91] and will be presented in Section 5.3.2.

5.2.2 Influence 2: From knowledge and beliefs to arguments

Regarding the second direction of influence, everyone agrees that our beliefs have a strong impact on the type of arguments we are prone to endorse. Trivially, if I compare two arguments a and b and I believe that the premisses of a are true, while I am unsure whether one of the premisses of b holds, then I should conceive, *ceteris paribus*, argument a as strictly stronger than b. Some of the works we mentioned in the previous paragraph (e.g., [30]) do take care of these kinds of principles operating in epistemic argument evaluation. Furthermore, the arguments we produce in a social context are influenced by the beliefs and knowledge we attribute to our audience. For instance, I may easily fool a child with some argument that I wouldn't use in other contexts. In general, arguing requires a theory of other minds and has many strategic aspects that link argumentation

to the study of persuasion techniques. Perhaps, this is what determined the development of argumentation and rhetoric as separate from logic *stricto sensu*.[10] Yet, the formal approach to the aspects of strategic argumentation becomes nowadays more and more relevant for the purposes of human-machine interaction and the goal of building intelligent debaters. This is indeed what motivates *opponent modelling* in formal argumentation, today a fairly active area of research, as witnessed by an increasing number of works over the last years [84; 96; 58; 75; 3]. Here again, combining formal argumentation with tools from (dynamic) epistemic logics provides a general tool to categorize different approaches to opponent modelling and to inspire further developments. In Section 5.4, we illustrate work in this direction and their link to applications.

5.3 From arguments to knowledge and belief

The works presented in this section are those exploring the first direction of influence, from arguments to knowledge and belief (Section 5.2.1). We proceed from the most natural and simpler approach by Grossi and van der Hoek [57], which simply fuses standard modal logic and abstract argumentation. We then go towards the one initiated by [87], displaying more complex (topological) models for modal logic in order to account for a notion of argument-based evidence enabling to formalize the JTB theory. We finally present the most articulated approach, enriching both the formalism for modal logic, by means of awareness logics, and the one for argumentation, exploiting the richer ASPIC$^+$ formalism for structured argumentation. This allows for a finer granularity when representing concepts in argumentation, e.g. different types of attacks among arguments (such as *rebuttal*, *undermining* or *undercut* from [79]), and therefore the possibility of encoding more articulated types of argument-based beliefs.

5.3.1 Product models for argumentation and belief

As seen in Section 2.1, Kripke semantics for modal logic provides a natural tool to talk about graphs and, therefore, to reason about abstract argumentation and its solution concepts [53; 54; 55; 33]. As illustrated in Section 5.1, they are also the primary tool for doxastic and epistemic logics. Therefore, combining the respective Kripke semantics is perhaps the most natural approach for fusing these two different

[10] It should be noticed that in Aristotle's *Organon*, argumentation, or *dialectic*, was intended to be a branch of logic – which constitutes the object of the *Topics* and *Sophistical Refutations* – the main difference being that the object of dialectic is syllogisms with uncertain or generally assumed premises (*endoxa*) rather than true ones.

frameworks. The work by Grossi and van der Hoek [57] proceeds along these lines and is one of the first combining epistemic logic and abstract argumentation to analyse the interactions between beliefs and argumentation. The keystone of the work is indeed the use of *product models* [50; 68]. Here, possible worlds are pairs $\langle s, a\rangle$ with s a doxastic state and a a given argument. Intuitively, s is the 'actual' state of affairs and a is the 'currently entertained' argument. This allows, among other things, the definition and formal analysis of several forms of justified belief. This and related work by Shi et al. [92] are covered in detail in [8, Sect. 3.2.2.], so we skip a full presentation to avoid overlapping.

5.3.2 Topo-argumentative models for argument-based epistemic attitudes

The notion of *evidence* is a central one for epistemology and lies behind that of *justified belief*, i.e. a belief supported by strong or undefeated evidence. There are many ways to frame the interplay between evidence, belief and knowledge in a modal logical setting. As mentioned, one of them consists in modelling evidence as a primitive notion by means of neighbourhood semantics, with knowledge and belief as derived concepts [98; 97; 12].[11] Arguments are typical sources of evidence and solution concepts from abstract argumentation can shed some light on the way they may serve to justify a belief. The line of work of [88; 87; 90; 89; 91] brings together all these insights by combining abstract argumentation with so-called *topological models* of evidence for doxastic logics.[12]

The main point of departure of this approach is to understand pieces of evidence as members of a topological structure. A *topology* τ over a non-empty set S is a set of sets $\tau \subseteq 2^S$, such that: (i) $\emptyset, S \in \tau$ (the unit and the empty set are its elements); (ii) if $A, B \in \tau$, then $A \cap B \in \tau$ (closure under finite intersections); and (iii) for any –possibly infinite– family $\{A_x\}_{x\in X} \subseteq \tau$, we have that $\bigcup_{x\in X} A_x \in \tau$ (closure under arbitrary unions). The topology generated by a family of sets $B \subseteq 2^S$ is the smallest topology τ_B such that $B \subseteq \tau_B$. Given a topology τ, its elements are usually called *opens*. These opens represent pieces of evidence in topological models for epistemic logic [12], and they will be the arguments of the topological argumentation model we are about to present:

[11] Another approach runs by adding specific *justification* terms to the language of doxastic-epistemic logics [14; 15; 44], a general strategy borrowed from *justification logics* [9].

[12] Note that [8] mentions some of these works, but focus on the presentation of a different approach by the same authors [92] that works without topology and it is somehow closer to [57], so we devote some space for the introduction of topological tools into the modelling of argument-based beliefs.

Definition 5.1 (TA-models)**.** *A* Topological-Argumentation model (TA-model) *for a countable set of atomic variables* Φ *is a tuple* $M = \langle S, E_0, \tau_{E_0}, \rightsquigarrow, \mathcal{V} \rangle$*, where*

- $S \neq \emptyset$ *is a set of* possible worlds*.*
- $E_0 \subseteq 2^S \setminus \{\emptyset\}$ *is a collection of* basic pieces of evidence*.*
- τ_{E_0} *is the topology generated by* E_0*.*
- $\rightsquigarrow \subseteq \tau_{E_0} \times \tau_{E_0}$ *is a* defeat relation *satisfying:*
  - *for every* $A \in \tau_{E_0} \setminus \{\emptyset\}$*, we have* $\langle A, \emptyset \rangle \in \rightsquigarrow$ *and* $\langle \emptyset, A \rangle \notin \rightsquigarrow$*.*
  - *for every* $A, B \in \tau_{E_0}$*, we have* $A \cap B = \emptyset$ *iff either* $\langle A, B \rangle \in \rightsquigarrow$ *or* $\langle B, A \rangle \in \rightsquigarrow$*.*
- $\mathcal{V} : S \to 2^{\Phi}$*.*

The idea of modelling basic pieces of evidence as sets of possible worlds (elements of the collection $E_0 \subseteq 2^S \setminus \{\emptyset\}$) can be traced back to [98; 97]. The main assumption behind it is that *evidence* is understood as *information-as-range* [91], so that if S represents all the epistemic alternatives of the agent, a piece of evidence $A \subseteq S$ tells the agent that the actual world is in A (and hence $S \setminus A$ should be disregarded according to A). Note, however, how $A \in E_0$ does not informally mean that the agent accepts A, but she rather takes it as a starting point for reasoning.

The topological structure τ_{E_0} represents the possible ways in which the agent can logically combine her basic pieces of evidence. Importantly, here the elements of τ_{E_0} play the role of *arguments* (see [77] for a discussion) and $\rightsquigarrow$ represents a *defeat* relation among them. The idea behind this specific definition is that there is a defeat from A to B only when A and B are incompatible pieces of evidence and A is 'as least as strong as' B. In sum, $\rightsquigarrow$ functions as a way of modelling how incompatible pieces of evidence are weighted. This process of evaluation is modelled through the conflict calculus introduced by [41] (Section 2.2).[13]

In particular, two forms of argument-based belief are defined over TA-models in [91]. Both of them are based in the grounded semantics for abstract argumentation

[13]The relation $\rightsquigarrow$ is deemed an "attack" relation in [87] and the subsequent works. However, we believe that the notion of defeat makes better sense in the current context (see Section 2.2 for the distinction between attacks and defeats). In this sense, 'A attacks B' is best understood as the (symmetric) incompatibility relation $A \cap B = \emptyset$, i.e. the second precondition of $A \rightsquigarrow B$ in definition 5.1, while further properties of $\rightsquigarrow$ (e.g. the first precondition in definition 5.1) act as symmetry-breaking constraints to assess the relative weight of A and B. We thank one of the anonymous reviewers of this article for asking us to clarify this point.

frameworks (see Definition 2.5). Recall that $\mathcal{E}_{\mathcal{GR}}(\langle Ar, \rightsquigarrow\rangle)$ is used to denote the set of all grounded extensions of $\langle Ar, \rightsquigarrow\rangle$ (Section 2.2), and hence $\bigcup \mathcal{E}_{\mathcal{GR}}(\langle Ar, \rightsquigarrow\rangle)$ denotes the grounded extension, since it is unique (as shown by [41]). Let $M = \langle S, E_0, \tau_{E_0}, \rightsquigarrow, \mathcal{V}\rangle$ be a TA-model, and let $P \subseteq S$ be a proposition, then:

- the agent has a *grounded belief* on P iff there is an $A \in \bigcup \mathcal{E}_{\mathcal{GR}}(\langle \tau_{E_0}, \rightsquigarrow\rangle)$ such that $A \subseteq P$.
- the agent has a *fully grounded belief* on P iff for every $A \in \bigcup \mathcal{E}_{\mathcal{GR}}(\langle \tau_{E_0}, \rightsquigarrow\rangle)$, there is an $A' \in \bigcup \mathcal{E}_{\mathcal{GR}}(\langle \tau_{E_0}, \rightsquigarrow\rangle)$, such that $A' \subseteq A$ and $A' \subseteq P$.

The authors' choice of employing only the grounded semantics to define belief may have several reasons. First, the sceptic flavour of grounded semantics is particularly significant in the context of epistemic as opposed to practical reasoning, i.e. reasoning about what to believe as opposed to reasoning about what to do [80]. Second, as mentioned, the grounded extension is always unique. This dodges the discussion that would arise if a semantics that returns multiple extensions were used, namely, which of the (mutually incompatible) extension should be the one that actually serves the agent to ground her/his beliefs. Finally, as pointed out by [91], the grounded extension is never empty in the current setting, as it can be shown that W is always undefeated. This guarantees, among other things, that valid propositions are always groundly believed. However, it seems worthy to investigate whether and how other semantics sharing the properties of uniqueness and non-emptiness could work in this framework.

Both notions, grounded belief and fully grounded belief, are possible formalizations of the first type of influence, i.e. of how arguments determine specific types of belief. Curiously, while fully grounded beliefs satisfy KD45 axioms, grounded beliefs fail to be closed under conjunction (and hence do not satisfy the (K) axiom).[14] Moreover, pairwise consistency among groundly believed propositions is guaranteed, but this is not the case when we consider sets of groundly believed propositions with more than two elements. In terms of the literature about rationality postulates for argumentation systems [32], grounded belief is directly consistent but not indirectly consistent. In contrast, fully grounded belief is indirectly (i.e. totally) consistent. Finally, fully grounded belief is strictly stronger than grounded belief, so that the former implies the latter but not vice versa. This brief comparison among both notions makes explicit the existing tension between *believing more* (or more informatively) and believing *more consistently* (see [88] and [91] for a detailed discussion on such a tension).

[14] Or, in other words, the grounded extension of $\langle \tau_{E_0}, \rightsquigarrow\rangle$ is not always closed under intersections. See [91] for conditions under which this is actually the case.

Along the same lines, [89] provides an argumentative account of the JTB characterization of knowledge. Here too, the author defines different notions of knowledge, namely K_1,K_2 and K_3, ranging from weaker to stronger and investigates their logical properties and the conditions for their equivalence. The weakest K_1 is defined simply as:

- the agent *knows* P (K_1P) iff it has a grounded belief on P and P is true at the world of evaluation.

In this version of JTB, justified belief is therefore grounded belief. It is also shown that (grounded) belief implies believing to know ($\models B\phi \rightarrow BK_1\phi$), and that many other properties postulated by standard logics for knowledge and belief hold [95].

Before closing this subsection, we mention a related, interesting work by Wang and Li [102]. This paper introduces a generalization of neighbourhood models where arguments are understood as sets of propositions, and propositions are represented semantically as sets of possible worlds. The main reason for us to leave a detailed review of this work out of this article is that the approach lacks an explicit modelling of conflicts among arguments, which is an essential feature of all the logical frameworks introduced here, and of formal models of argumentation in general.

5.3.3 A more syntactic approach

In [30; 31], the authors adopt a more syntactic approach to modelling the interaction between arguing and believing. In short, it is based on the combination of *awareness epistemic models* [45] with ASPIC$^+$ arguments [76]. The main reason behind this move is to bridge epistemic logic with the field of *structured argumentation* [22], where arguments are essentially understood as composite entities, with premises, conclusions, inferential links between the two, and which may encompass subarguments as their parts.

Syntax. Unlike previously reviewed works, here arguments are, together with other formulas, first-class syntactic citizens, as specified by the definition of the language.

Definition 5.2 (Language for Awareness of Structured Arguments)**.** *The language* $\mathcal{L}_{ASA}$[15] *is the pair* $\langle \mathsf{F}, Ar \rangle$ *of formulas and arguments, which are defined by mutual recursion as follows:*

[15]The language is simply denoted $\mathcal{L}$ in [31]. We use *ASA* as an abbreviation of 'Awareness of Structured Arguments' in order to distinguish it from the rest of the languages that appear in this article.

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \Box\varphi \mid \mathsf{aware}(\alpha) \mid \mathsf{conc}(\alpha) = \varphi \mid$$
$$\mid \mathsf{strict}(\alpha) \mid \mathsf{undercuts}(\alpha, \alpha) \mid \mathsf{wellshap}(\alpha) \qquad p \in \Phi, \alpha \in Ar.$$
$$\alpha ::= \langle\varphi\rangle \mid \langle\alpha_1, ..., \alpha_n \twoheadrightarrow \varphi\rangle \mid \langle\alpha_1, ..., \alpha_n \Rightarrow \varphi\rangle \qquad \varphi \in \mathsf{F}, n \geq 1.$$

As in ASPIC$^+$, the grammar defines three types of arguments. $\langle\varphi\rangle$ is an *atomic argument* whose sole premise and conclusion is φ. $\langle\alpha_1, ..., \alpha_n \twoheadrightarrow \varphi\rangle$ is the argument that *deductively* concludes φ from the conclusions of subarguments $\alpha_1, ..., \alpha_n$. $\langle\alpha_1, ..., \alpha_n \Rightarrow \varphi\rangle$ is the argument that *presumtively/defeasibly* concludes φ from the conclusions of subarguments $\alpha_1, ..., \alpha_n$. The modal operator $\Box$ denotes (implicit) belief. Concerning other formulas, $\mathsf{aware}(\alpha)$ reads "the agent is aware of argument α"; $\mathsf{conc}(\alpha) = \varphi$ reads "the conclusion of α is φ"; $\mathsf{strict}(\alpha)$ reads "argument α is strict (i.e. it contains no defeasible rule)". $\mathsf{undercuts}(\alpha, \beta)$ means that argument α undercuts argument β (i.e. α attacks the last rule employed in the construction of β). Finally, $\mathsf{wellshap}(\alpha)$ means that α is well-shaped or well-constructed, in the sense that it only uses either valid deductive rules or accepted defeasible rules in its construction.

Semantics. The semantics of $\mathcal{L}_{ASA}$ is strongly meta-syntactic, meaning that its model theory relies on functions ranging over language constructions. Let us introduce some of them. $\mathsf{SEQ}(\mathsf{F})$ is used to denote the *set of all finite sequences over* F and $\langle\langle\varphi_1, ..., \varphi_n\rangle, \varphi\rangle$ denotes an arbitrary element of $\mathsf{SEQ}(\mathsf{F})$ with $n+1$ elements. These sequences are used to represent defeasible inference steps in the meta-language. Moreover, the following ASPIC$^+$'s functions are used to analyse **arguments' structures**:

- $\mathsf{Prem}(\alpha)$ returns the **premisses** of α and it is defined as follows: $\mathsf{Prem}(\langle\varphi\rangle) := \{\varphi\}$, $\mathsf{Prem}(\langle\alpha_1, ..., \alpha_n \hookrightarrow \varphi\rangle) := \mathsf{Prem}(\alpha_1) \cup ... \cup \mathsf{Prem}(\alpha_n)$ where $\hookrightarrow \in \{\twoheadrightarrow, \Rightarrow\}$.

- $\mathsf{Conc}(\alpha)$ returns the **conclusion** of α and it is defined as follows $\mathsf{Conc}(\langle\varphi\rangle) := \{\varphi\}$ and $\mathsf{Conc}(\langle\alpha_1, ..., \alpha_n \hookrightarrow \varphi\rangle) := \{\varphi\}$ where $\hookrightarrow \in \{\twoheadrightarrow, \Rightarrow\}$.

- $\mathsf{sub_A}(\alpha)$ returns the **subarguments** of α and it is defined as follows: $\mathsf{sub_A}(\langle\varphi\rangle) := \{\langle\varphi\rangle\}$ and $\mathsf{sub_A}(\langle\alpha_1, ..., \alpha_n \hookrightarrow \varphi\rangle) := \{\langle\alpha_1, ..., \alpha_n \hookrightarrow \varphi\rangle\} \cup \mathsf{sub_A}(\alpha_1) \cup ... \cup \mathsf{sub_A}(\alpha_n)$ where $\hookrightarrow \in \{\twoheadrightarrow, \Rightarrow\}$.

- $\mathsf{TopRule}(\alpha)$ returns the **top rule** of α, i.e. the last rule applied in the formation of α. It is defined as follows: $\mathsf{TopRule}(\langle\varphi\rangle)$ is left undefined, $\mathsf{TopRule}(\langle\alpha_1, ..., \alpha_n \twoheadrightarrow \varphi\rangle) = \mathsf{TopRule}(\langle\alpha_1, ..., \alpha_n \Rightarrow \varphi\rangle) := \langle(\mathsf{Conc}(\alpha_1), ..., \mathsf{Conc}(\alpha_n)), \varphi\rangle$.

- $\mathsf{DefRule}(\alpha)$ returns the set of **defeasible rules** of α and it is defined as $\mathsf{DefRule}(\langle\varphi\rangle) := \emptyset$, $\mathsf{DefRule}(\langle\alpha_1, ..., \alpha_n \twoheadrightarrow \varphi\rangle) := \mathsf{DefRule}(\alpha_1) \cup ... \cup \mathsf{DefRule}(\alpha_n)$ and $\mathsf{DefRule}(\langle\alpha_1, ..., \alpha_n \Rightarrow \varphi\rangle) := \{\langle(\mathsf{Conc}(\alpha_1), ..., \mathsf{Conc}(\alpha_n)), \varphi\rangle\} \cup \mathsf{DefRule}(\alpha_1) \cup ... \cup \mathsf{DefRule}(\alpha_n)$.

Let us also define **semantic propositional negations**, for any $\varphi, \psi \in \mathsf{F}$: $\varphi =\sim \psi$ abbreviates $\mathsf{wellshap}(\langle\langle\varphi\rangle \twoheadrightarrow \neg\psi\rangle) \wedge \mathsf{wellshap}(\langle\langle\psi\rangle \twoheadrightarrow \neg\varphi\rangle)$.

Definition 5.3 ($\mathcal{L}_{ASA}$-models)**.** *A* ***model*** *for $\mathcal{L}_{ASA}$ is a tuple $M = \langle S, \mathcal{R}_{\mathcal{B}}, \mathcal{O}, \mathcal{D}, \mathfrak{n}, \mathcal{V}\rangle$ where:*

- *$\langle S, \mathcal{R}_{\mathcal{B}}\rangle$ is a doxastic structure (i.e., $\mathcal{R}_{\mathcal{B}}$ is serial, transitive and euclidean).*[16]
- *$\mathcal{O} \subseteq Ar$ the* awareness set of the agent*.*
- *$\mathcal{D} \subseteq \mathsf{SEQ}(\mathsf{F})$ is a set of* accepted defeasible rules*. These rules are assumed to be consistent and invalid according to classic propositional logic.*
- *$\mathfrak{n} : \mathsf{SEQ}(\mathsf{F}) \to \Phi$ is a (possibly partial)* naming function *for rules, where $\mathfrak{n}(R)$ informally means "the rule R is applicable".*
- *$\mathcal{V}$ is and an* atomic valuation*, i.e. a function $\mathcal{V} : S \to 2^{\Phi}$.*

Let $M = \langle S, \mathcal{R}_{\mathcal{B}}, \mathcal{O}, \mathcal{D}, \mathfrak{n}, \mathcal{V}\rangle$ be a model for $\mathcal{L}_{ASA} = \langle\mathsf{F}, Ar\rangle$. We use $\vdash_0$ to denote logical consequence in classical propositional logic. The **set of well-shaped arguments** $WS^M \subseteq Ar$ (depending on $\mathcal{D}$ in M) is the smallest set fulfilling the following conditions:

1. $\langle\varphi\rangle \in WS^M$ for any $\varphi \in \mathsf{F}$.
2. $\langle\alpha_1, ..., \alpha_n \twoheadrightarrow \varphi\rangle \in WS^M$ iff both $\alpha_i \in WS^M$ for every $1 \leq i \leq n$ and $\{\mathsf{Conc}(\alpha_1), ..., \mathsf{Conc}(\alpha_n)\} \vdash_0 \varphi$.
3. $\langle\alpha_1, ..., \alpha_n \Rightarrow \varphi\rangle \in WS^M$ iff both $\alpha_i \in WS^M$ for every $1 \leq i \leq n$ and $\langle\langle\mathsf{Conc}(\alpha_1), ..., \mathsf{Conc}(\alpha_n)\rangle, \varphi\rangle \in \mathcal{D}$.

Definition 5.4 (Truth in $\mathcal{L}_{ASA}$-models)**.** *Let $\langle M, w\rangle$ be a pointed model for $\mathcal{L}_{ASA}$, that is, $M = \langle S, \mathcal{R}_{\mathcal{B}}, \mathcal{O}, \mathcal{D}, \mathfrak{n}, \mathcal{V}\rangle$ is a model and $w \in S$. The* ***truth*** *relation, relating pointed models and formulas, is given by:*

[16] Recall that this is an instance of Definition 2.2 with a single label representing the modelled agent.

$$
\begin{array}{rll}
M, w \models \Box\varphi & \textit{iff} & \textit{for all } w' \in S\textit{: } (w, w') \in \mathcal{R}_{\mathcal{B}} \\
 & & \textit{implies } M, w' \models \varphi. \\
M, w \models \mathsf{aware}(\alpha) & \textit{iff} & \alpha \in \mathcal{O}. \\
M, w \models \mathsf{conc}(\alpha) = \varphi & \textit{iff} & \mathsf{Conc}(\alpha) = \varphi. \\
M, w \models \mathsf{strict}(\alpha) & \textit{iff} & \mathsf{DefRule}(\alpha) = \emptyset. \\
M, w \models \mathsf{undercuts}(\alpha, \beta) & \textit{iff} & \beta = \langle \beta_1, ..., \beta_n \Rightarrow \psi \rangle \textit{ and} \\
 & & \mathsf{Conc}(\alpha) = \neg\mathfrak{n}(\mathsf{TopRule}(\beta)). \\
M, w \models \mathsf{wellshap}(\alpha) & \textit{iff} & \alpha \in WS^M.
\end{array}
$$

Types of beliefs. $\mathcal{L}_{ASA}$ is rich enough to distinguish several kinds of belief. A general distinction is made between basic and argument-based beliefs (mimicking the one among intuitive and inferential beliefs in cognitive sciences [94]). Basic beliefs are based on non-inferential information, such as observation or trusted testimonies. There are, in turn, two subtypes of basic beliefs, inherited from the epistemic modal logic tradition: implicit and explicit ones. Basic-implicit beliefs are the ideal beliefs of a perfect reasoner and are captured by the primitive operator $\Box$. Its explicit counterparts are defined *à la* [45], using atomic arguments for simulating awareness of sentences: $\Box^e\varphi := \Box\varphi \wedge \mathsf{aware}(\langle\varphi\rangle)$.

There are also two types of argument-based beliefs: *deductive* beliefs and *grounded* beliefs. Deductive beliefs are defined as $\mathsf{B}^{\mathsf{D}}(\alpha, \varphi) := \mathsf{accept}(\alpha) \wedge \mathsf{aware}(\alpha) \wedge \mathsf{conc}(\alpha) = \varphi \wedge \mathsf{strict}(\alpha) \wedge \mathsf{wellshap}(\alpha)$ (to be read as "the agent has a deductive belief that φ based on argument α"), where $\mathsf{accept}(\alpha) := \bigwedge_{\varphi \in \mathsf{Prem}(\alpha)} \Box\varphi$ stands for argument doxastic acceptance. Note that the definition of argument doxastic acceptance (accept) includes a very simple instance of Influence 2: the only arguments taken into consideration by the agent are those whose premisses are believed. Moreover, the definition of deductive beliefs ($\mathsf{B}^{\mathsf{D}}(\alpha, \varphi)$) is a clear instance of Influence 1: the arguments that the agent is aware of influence her (deductive) beliefs.

The definition of grounded beliefs needs some preliminary argumentative notions. The first one is a notion of binary preference among arguments. Many preference relations are definable in $\mathcal{L}_{ASA}$, capturing different versions of Influence 2. As an illustration, a very simple notion of preference, that assumes that the agent only takes into account doxastically accepted arguments, consists in preferring strict arguments over non-strict ones: $\alpha \geq \beta := \mathsf{strict}(\alpha) \vee \neg\mathsf{strict}(\beta)$.

Argument attacks and defeats (= successful attacks), notions imported from ASPIC$^+$, are definable in $\mathcal{L}_{ASA}$:

- **Undercutting a subargument** $\mathsf{undercuts}^*(\alpha, \beta) := \bigvee_{\beta' \in \mathsf{sub_A}(\beta)} \mathsf{undercuts}(\alpha, \beta')$.
- **Unrestricted successful rebuttal**

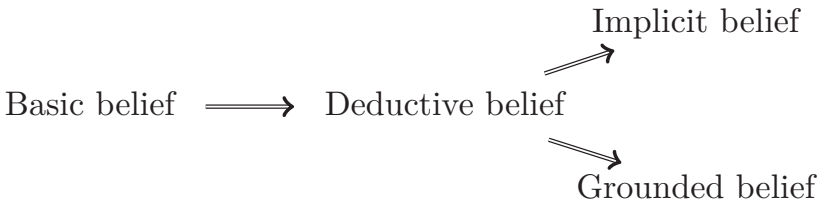
$\mathsf{Urebuts}(\alpha, \beta) := \neg\mathsf{strict}(\beta) \wedge \bigvee_{\beta' \in \mathsf{sub_A}(\beta)}(\mathsf{conc}(\alpha) = \varphi \wedge \mathsf{conc}(\beta') = \psi \wedge \varphi =\sim \psi \wedge \alpha \geq \beta')$.

- **Defeat** $\mathsf{defeat}(\alpha, \beta) := \mathsf{undercuts}^*(\alpha, \beta) \vee \mathsf{Urebuts}(\alpha, \beta)$.

Let $\langle M, w\rangle$ be a pointed model for $\mathcal{L}_{ASA} = \langle \mathsf{F}, Ar\rangle$, we define its **associated argumentation framework** as $AF^M := \langle Ar^M, \rightsquigarrow\rangle$, where $Ar^M := \{\alpha \in Ar \mid M, w \models \mathsf{aware}(\alpha) \wedge \mathsf{wellshap}(\alpha) \wedge \mathsf{accept}(\alpha)\}$ and $\rightsquigarrow \subseteq Ar^M \times Ar^M$ is given by $\alpha \rightsquigarrow \beta$ iff $M, w \models \mathsf{defeat}(\alpha, \beta)$. Finally, we expand our language with the grounded belief operator $\mathsf{B}(\alpha, \varphi)$, interpreted in pointed models as follows:

$$M, w \models \mathsf{B}(\alpha, \varphi) \quad \text{iff} \quad \alpha \in \bigcup \mathcal{E}_{\mathcal{GR}}(AF^M) \quad \text{and} \quad \mathsf{Conc}(\alpha) = \varphi.$$

The relative strength of the four kinds of belief is given by the following diagram:



5.4 From knowledge and belief to arguments

This section introduces formal work exploring the second direction of influence: how knowledge and beliefs, especially those about other minds, influence the use of argument. Historically, the first work combining epistemic logic and abstract argumentation in this sense is [86], followed by [81; 82], where the original framework was systematically expanded and dynamified. Unlike the previous section, here, we will not treat different works separately, but rather introduce some of their high-level ideas as well as their applications.

One key feature that differentiates the logical frameworks presented here from those in the previous section is the multi-agent perspective, due to the essential role played in this context by reasoning about others' beliefs and knowledge. Therefore, a preliminary step, before introducing modalities, is to combine abstract argumentation frameworks and multi-agency. Many options are explored in the literature (see also [82] for a review). Here, we introduce one of them, probably the most popular (see [105] for a principle-based analysis of its semantics).

Definition 5.5 (Multi-agent AF)**.** *Let* $\mathbb{A}\mathrm{gt}$ *and* $\mathbf{A}$ *be two finite, non-empty and disjoint sets (agents and arguments, respectively), a* multi-agent AF *(MAF) is a tuple* $\langle Ar, \{Ar_i \mid i \in \mathbb{A}\mathrm{gt}\}, att\rangle$ *where* $Ar \subseteq \mathbf{A}$*;* $Ar_i \subseteq Ar$*; and* $att \subseteq Ar \times Ar$*. We*

say that the MAF $\langle Ar, \{Ar_i \mid i \in \mathbb{A}\text{gt}\}, att\rangle$ *is based on* Ar*. We use* $\mathsf{MAF}(Ar)$ *to denote the set of all MAFs based on* Ar*.*

Intuitively, in such a MAF, $\langle Ar, att\rangle$ represents all potentially relevant arguments and their interactions, while Ar_i is the set of arguments that agent i is aware of. We further define $att_i = att \cap (Ar_i \times Ar_i)$ as the set of attacks that agent i is aware of. Note that the definition of att_i implies that an agent is aware of an attack whenever it is aware of the arguments involved. Therefore, attacks in a MAF have an 'objective' meaning, since no agent can be 'mistaken' about them. Nonetheless, uncertainty and incomplete knowledge about attacks can still be represented at the modal level, as we shall see in what follows.

We now plug MAFs into worlds of a multi-agent doxastic model. The ideas underlying the following definition can be traced back to [86]. We get rid of some of the assumptions presented there and some others introduced in [82].[17]

Definition 5.6 (EA-models)**.** *An Epistemic-Argumentative model (EA-model) for* Ar *is a tuple* $M = \langle S, \mathcal{R}, \mathcal{D}\rangle$ *s.t.*

- $\langle S, \mathcal{R}\rangle$ *is a Kripke frame over* $\mathbb{A}\text{gt}$*, where* S *is the set of possible worlds and* $\mathcal{R}$ *specifies the epistemic accessibility relations of different agents (Def. 2.2), and*
- $\mathcal{D} : S \to \mathsf{MAF}(Ar)$ *is a function specifying a multi-agent AF (Def. 5.5) for each possible world.*

Let $w \in S$*, we denote* $\mathcal{D}(w)$ *as* $\langle Ar, \{Ar_i(w) \mid i \in \mathbb{A}\text{gt}\}, att(w)\rangle$*.*

Some interesting properties relating $\mathcal{R}$ and $\mathcal{D}$ are summarised in Table 1.

Let us illustrate the previous definition through a simple example.

Example 5.7 (An EA-model)**.** *Let us consider the EA-model* $M = \langle S, \mathcal{R}, \mathcal{D}\rangle$ *for* $\mathbb{A}\text{gt} = \{1, 2\}$ *and* $Ar = \{a, b, c, d\}$ *where:* $S = \{w_0, w_1\}$*;* $\mathcal{R}(1) = \{\langle w_0, w_1\rangle, \langle w_1, w_1\rangle\}$*,* $\mathcal{R}(2) = \{\langle w_0, w_0\rangle, \langle w_1, w_1\rangle\}$*; and* $\mathcal{D}$ *is defined for each world:* $Ar_1(w_0) = \{a, b, c, d\}$*,* $Ar_2(w_0) = \{a, b, c\}$*,* $att(w_0) = \{\langle a, b\rangle, \langle b, a\rangle, \langle c, b\rangle, \langle c, d\rangle, \langle d, b\rangle, \langle d, c\rangle\}$*;* $Ar_1(w_1) = \{a, b, c, d\}$*,* $Ar_2(w_1) = \{a, b\}$*,* $att(w_1) = \{\langle a, b\rangle, \langle b, a\rangle, \langle c, b\rangle, \langle c, d\rangle, \langle d, c\rangle\}$*. This model is represented graphically in Figure 9. Intuitively, at the actual world* (w_0) *agent 2 is aware of arguments* a*,* b *and* c *(red-circled area in* w_0*). However, agent 1 does not know that agent 2 is aware of* c*, for she has access only to* w_1 *where the awareness*

[17] The reasons for doing so is that we seek maximum generality here and that some of these assumptions were introduced for mere technical reasons that are not relevant at the current level of abstraction.

M satisfies...	iff for every $i \in \mathbb{A}\text{gt}$, $w, u \in S$...
Positive knowledge of attacks	$w\mathcal{R}_i u \Rightarrow att(w) \subseteq att(u)$
Negative knowledge of attacks	$w\mathcal{R}_i u \Rightarrow att(u) \subseteq att(w)$
Positive introspection of arguments	$w\mathcal{R}_i u \Rightarrow Ar_i(w) \subseteq Ar_i(u)$
Negative introspection of arguments	$w\mathcal{R}_i u \Rightarrow Ar_i(u) \subseteq Ar_i(w)$
General negative introspection of arguments	$w\mathcal{R}_i u \Rightarrow \bigcup_{j\in\mathbb{A}\text{gt}} Ar_j(u) \subseteq Ar_i(w)$

Table 1: Some properties of EA-models

area of 2 only contains a and b. Moreover, agent 1 is also mistaken in his/her understanding of the attacks between arguments, inasmuch as argument d attacks argument b in the actual world w_0, but 1 believes this fact to be false (no attack in w_1).

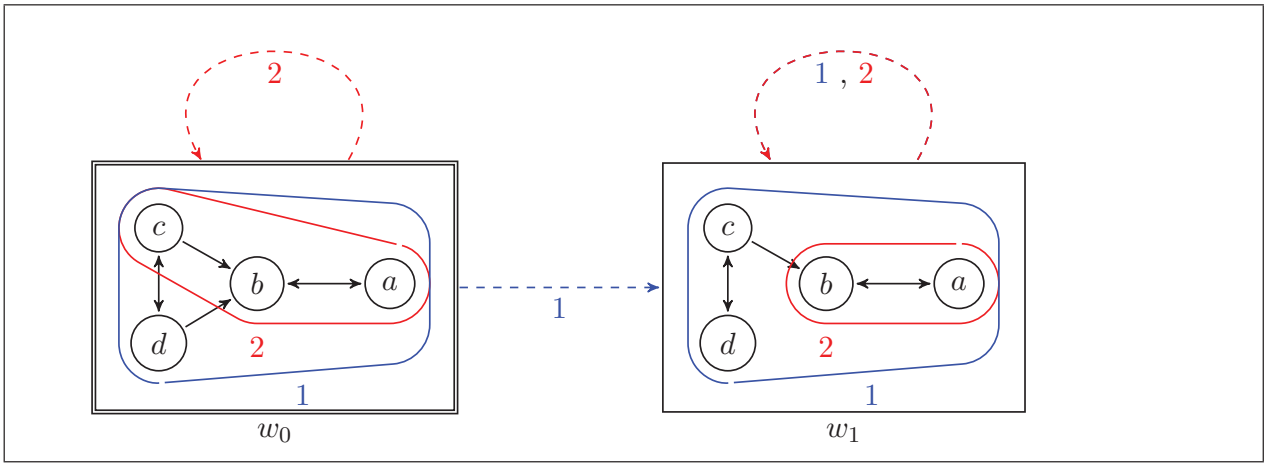


Figure 9: A simple EA-model

Languages Different languages have been proposed to talk about EA-models (or some of their extensions). The common denominator in all of them is the use of doxastic/epistemic modalities to jump from one possible world to another. The main differences among them lie in the resources chosen to talk about the MAF representing each world. Let us briefly comment on them.

Three of these languages were introduced in [86]. The first one, later extended by [81], is just the standard multi-agent doxastic language with a set of distinguished atoms $\{\mathsf{aware}_i(x) \mid i \in \mathbb{A}\text{gt}, x \in \mathbf{A}\}$, which are used to talk about the set of arguments

that each agent is aware of at each world (i.e. to talk about $Ar_i(w)$). The lack of a syntactic device to talk about attacks implicitly assumes that the set of attacks assigned by $\mathcal{D}$ to w (denoted by $att(w)$) is the same for every $w \in S$. This assumption is dropped by the second language introduced by [86]: a two-layer language where one type of formula is used to talk about the MAF assigned to each world and the other one is used as a way of (modally) jumping from one world to the other. Their third language, interpreted over a product-model version of EA-models, is essentially one where the previous two layers are fused together so that argumentative and doxastic modalities can be nested (just as the one we discussed in Section 5.3.1).

If one restricts to finite sets of arguments, as most of the literature does, then propositional languages are sufficient for reasoning about AFs and their semantics [21]. In this case, using modalities to describe a single MAF is an overkill. In line with this, [82], use an enriched version of [86]'s first language to talk about the MAF assigned to each world in EA-models. This enrichment consists of a new set of atoms for describing the attack relation $\{\mathsf{r}_{x,y} \mid x, y \in \mathbf{A}\}$. Note that, with the variables ($\mathsf{aware}_i(x)$ and $\mathsf{r}_{x,y}$), all properties of Table 1 are easily definable (e.g., $\neg\mathsf{aware}_i(x) \rightarrow \Box_i \bigwedge_{j\in\mathbb{A}\mathrm{gt}} \neg\mathsf{aware}_j(x)$ for the last one).[18]

Finally, a third approach is presented in [106, pp. 161-165]. This can be understood as some kind of compromise between the full power of bi-modal languages to describe MAFs [86; 57]) vs. the use of propositional languages [82]. Summing up, the idea is to use yet a new kind of propositional variable $\{\mathsf{in}_x \mid x \in \mathbf{A}\}$ meant to describe an arbitrary set of arguments (e.g., an extension), plus one extra modality that quantifies over valuations changing the truth values of these variables. This new modality (inspired by works that combine dynamic logic and abstract argumentation [38]) allows expressing maximality and minimality checking, and it is therefore expressive enough to capture all standard argumentation semantics for MAFs through polynomially long formulas.

Dynamics of information EA-models were systematically dinamised in [82], where they are combined with event models [13], imported from the field of dynamic epistemic logic [100]. With the resulting framework, one is able to model nuanced forms of epistemic and argumentative dynamics, such as the action of privately searching for a counter-argument in the context of a debate. For presentational purposes, we just introduce here the simplest kind of epistemic-argumentative action, which was first studied in [81]: the public addition or disclosure of an argument.

[18] The main shortcoming of this approach is that with no quantifiers over atoms – or any other equally expressive device – the encoding of some notions (e.g., those requiring maximality checking, as preferred semantics) requires propositional formulas that are exponentially long (on the size of $\mathbf{A}$), and therefore not very appealing from a computational point of view.

Informally, the idea is to model what happens when, within a group discussion, an agent publicly puts forward an argument. Formally:

Definition 5.8 (Public update with an argument)**.** *Given a EA-model* $M = \langle S, \mathcal{R}, \mathcal{D}\rangle$ *and an argument* $a \in Ar$*, the update of* M *by* a *is the model* $M^{a!} = \langle S, \mathcal{R}, \mathcal{D}^{a!}\rangle$ *where* $\mathcal{D}^{a!}$ *only differs from* $\mathcal{D}$ *in the value assigned to the awareness set of each agent at each world:* $Ar_i^{a!}(w) = Ar_i(w) \cup \{a\}$ *for every* $i \in \mathbb{A}\text{gt}$ *and every* $w \in S$*.*

Applications to opponent modelling The main application of EA-models and their dynamics is the systematic analysis of different forms of opponent modelling in abstract argumentation. EA-models represent an expressive and epistemically transparent formalism where other proposals can be translated, so as to fully understand their hidden epistemic assumptions. Let us quickly review some of these reductions.

Incomplete AFs One of the simplest ways for modelling the opponent of an agent within a debate is through the notion of *incomplete argumentation framework* (IAF) [19]. The idea is to provide a compact specification of the uncertain view that the agent has of her opponent's information about a debate. Formally, an IAF is a tuple $\langle Ar, Ar^?, att, att^?\rangle$ where $Ar, Ar^? \subseteq \mathbf{A}$ are two disjoint sets of arguments, respectively representing certain and uncertain arguments; and $att, att^? \subseteq (Ar \cup Ar^?) \times (Ar \cup Ar^?)$ are two disjoint sets of attacks respectively representing certain and uncertain attacks. Perhaps more intuitively, Ar is the set of arguments that the agent believes her opponent to be aware of; $Ar^?$ is the set of arguments such that the agent does not know whether her opponent is aware of; and something analogous for att and $att^?$.

Reasoning about IAFs needs the notion of completion, i.e. a hypothetical removal of uncertainty from the IAF. Formally, a completion of $\langle Ar, Ar^?, att, att^?\rangle$ is any AF $\langle Ar^*, att^*\rangle$ such that $Ar \subseteq Ar^* \subseteq Ar \cup Ar^?$ and $att \cap (Ar^* \times Ar^*) \subseteq att^* \subseteq (att \cup att^?) \cap (Ar^* \times Ar^*)$.[19] As a key for reduction to EA-models, completions can be understood as possible worlds. Indeed, an IAF can be seen as a single-agent EA-model where each possible world represents a completion. If an argument a is present in one completion, then the atom $\mathsf{aware}_i(a)$ is going to be true in its corresponding world of the model, and the same holds for attacks. In this way, reasoning problems over IAFs become model-checking problems in EA-models[82]. Note that the correspondence is not strict: There are EA-models that do not represent the set of completions of

[19]Classical acceptability problems over AFs (sceptical and credulous acceptability) can be reformulated so as to get a new layer of quantification (over completions), obtaining in this way necessary (in all completions) and possible (in at least one completion) variants of classical problems.

any IAF. Hence, a legitimate question: what logic would we get if we only consider EA-models that represent IAFs?

The previous question is answered by [61]. The keystone for this finding is the connection between IAFs and the epistemic logic of visibility studied in [59]. Interestingly, the two central axioms of the resulting logic are:

$$\Box_i\varphi \rightarrow \Diamond_i\varphi$$
$$\Box_i(l_1 \vee \cdots \vee l_n) \rightarrow (\Box_i l_1 \vee \cdots \vee \Box_i l_n)$$

where $l_1 \ldots l_n$ is a sequence of consistent literals from $\{\mathsf{aware}_i(x) \mid x \in \mathbf{A}\} \cup \{\mathsf{r}_{x,y} \mid x, y \in \mathbf{A}\}$. The first axiom is just (D) (see Sect. 2.1 and the introduction to the current section), so it shows that IAFs model an epistemic attitude that is, at least, consistent (which seems quite reasonable for intelligent artificial agents). The second one expresses that the captured epistemic attitudes distribute over disjunctions of consistent literals (literals that capture the status of arguments and attacks). This second property looks more difficult to justify, unless for its efficient computational behaviour.

Control AFs (CAFs) [37] extend IAFs in two directions: uncertainty and dynamics. Regarding uncertainty, CAFs include a new attack relation $att^{\leftrightarrow}$ which is meant to capture attacks whose existence is known by the agent but whose direction is unknown. For example, imagine that a and b are two arguments disclosed, respectively, by two politicians of opposing parties. Imagine, moreover, that the agent knows that her opponent is biased towards one side of the political spectrum, but she is not sure about which side it is. Hence, the agent considers two possible completions (for her opponent): one where a attacks b and one where b attacks a. It is easy to show that no IAF represent that precise set of completions, so the previous kind of scenarios justify the introduction of $att^{\leftrightarrow}$. On the dynamic side, CAFs expand IAFs with an AF $\langle Ar_C, att_C\rangle$ formed by *control arguments* and *control attacks*. Intuitively, these arguments are the ones that the agent knows privately (she knows them and knows her opponent does not know). Then, reasoning over CAFs introduces yet another quantification layer (over subsets of Ar_C), raising so-called *controllability problems*: is there a set of control arguments such that a target argument gets sceptically/credulously accepted in all/at least one completion? These problems too were reduced to EA-model-checking problems in [82]. Such a reduction shows at least three interesting things. First, the axiom of IAFs capturing distribution over disjunctions of consistent literals is dropped because of the introduction of $att^{\leftrightarrow}$. Second, control arguments are in tight connection with the notion of public disclosure (see above). Third, the effects of communicating these arguments (e.g., the resulting perception of control attacks) are strongly idealised: the agent always knows what the effects of her communication act are. The last two points motivate

the study of further forms of representing argument communication in a compact setting.

Recursive opponent models The reductions of incomplete AFs and control AFs to EA models do not exploit all the expressive power of the latter. Indeed, and as far as uncertainty and multi-agency are concerned, incomplete AFs and control AFs can be seen as depth-1 epistemic attitudes about the arguments and attacks that the opponent of the agent owns. In other words, incomplete and control AF only talk about what the agent believes about her opponent's view of the debate. However, as pointed out by [96] concerning strategic argumentation, higher-order epistemic attitudes might as well be relevant in an argumentative context. As an example, imagine that you want to surprise your opponent. Then you might disclose an argument that *you believe your opponent believes that you are not aware of*, which is a depth-2 epistemic attitude. Along these lines, more expressive (and hence more complex) forms of opponent modelling have been studied in the context of strategic argumentation. These opponent models are, in their qualitative version,[20] EA models in disguise. The details of the reduction can be found in [82, Sect. 8.3].

New formalisms In recent years, two different tools for representing qualitative uncertainty about AFs appeared [85; 3]. Although they have not yet been reduced to EA-models, they should be reducible if one looks at the respective complexity classes of the relevant reasoning problems. We believe these reductions to be interesting open problems, so as to carve deeper into the epistemic assumptions underlying these new formalisms and get a more complete picture of how to represent qualitative uncertainty and multi-agency over AFs.

6 Related work

The first works on the combination of formal argumentation and modal logic appeared less than 20 years ago ([24] is the pioneering one, to the best of our knowledge). However, the literature is already quite vast. This section points to further readings on the topic that were left out of this article either because we wanted to keep its extension between reasonable limits or because they were already analysed in [8].

The first line of work, started by [38] and followed by [40; 39; 60; 107], consists in applying the Dynamic Logic of Propositional Assignments (DL-PA) to reason about abstract argumentation formalisms. DL-PA is a lightweight, well-behaved variant of dynamic propositional logic. In the quoted papers, it is used to capture argu-

[20] The are also probabilistic versions of opponent models for argumentation that go beyond the expressivity of EA models.

mentation frameworks, their semantics, their dynamics, and their extensions with qualitative uncertainty. These works share the same general target of those adopting an *argumentative reading of modalities*:[21] using modal logic to reason about argumentation systems. However, instead of interpreting modalities over attack relations and arguments, they are based on propositional encodings of argumentation formalisms [21], and the role of modalities is basically capturing restricted propositional quantification.

A second line of work proposes to use modal languages as the object languages of structured argumentation formalisms (e.g., ASPIC$^+$ [76] or ABA [42]). A particular instance of this idea is the insertion of modal languages into deductive argumentation systems, something that was previously covered in [23]. Going further than deductive reasoning, and hence incorporating non-monotonic inferences, a recent paper [99] has approached the use of deontic modalities in an argumentation system from the ASPIC family.

Finally, there is another interesting research line oriented towards integrating temporal and modal language formulas to represent arguments in the nodes of an argumentation network, as done in [18]. This approach can be seen as an extension of the traditional Dung networks, which depict arguments as atomic entities and study the relationships of attack between them. That way more content can be added to nodes in the network (e.g., proofs in some logic or simply just formulas from a richer language). Argumentation networks have also been applied in modelling so-called *argumentation with many lives* [48], where the network stands for a survival game (and thus the various traditional Dung semantics can be viewed as defining extensions in the form of possible survival group). With many lives available, there can be sets of nodes "living together" (so that members can attack but not able to kill one another). Recent research work in many lives argumentation networks [49] included temporal aspects, modelled through evolutionary temporal logic.

7 Conclusion

Argumentation and modal logic are two important theories and techniques within the field of knowledge representation and reasoning. Although their combination started only a couple of decades ago, the literature of the topic has flourished very rapidly since then. In this article we analysed three different lines of work to combine modal logic and formal argumentation: a) a logic-based framework that combines dynamic logic and argumentation for value-based planning; b) alternating-time temporal logic extended with coalitional argumentation; c) a combined approach for

[21] Those discussed in the introduction and in [8, Sect. 3.2.1], i.e., [53] and subsequent papers.

integrating epistemic logics and argumentation. These three alternatives give clear evidence of the heterogeneity of problems that can be approached by the combination of these two families of formal tools. Moreover, they also show the different relative positions in which modal logic and argumentation can be put together. In a broader sense, they also allow us to think about the possible interplay between classical logics and non-monotonic logics and how they can function respectively when solving particular problems.

Joint uses of formal argumentation and modal logic face numerous challenges in the current state of the art. We have seen how each work focuses on a particular reading of modalities (e.g., argumentative, epistemic, dynamic, or temporal). Although some papers tackle the fundamental question of their combination (e.g., argumentative and epistemic [57]), there are many possibilities which are yet unexplored. For instance, there seems to be strong rationale for motivating the design and study of structured argumentation systems that allow for the integration of deontic and epistemic modalities for reasoning about complex scenarios. Moreover, both strategic logic (Section 4.1) and strategic argumentation (end of Section 5.4) are used for reasoning about agents' strategic behavior in the context of multi-agent systems. If one sees the announcement of an argument as the performance of an action, then it seems natural looking at the transfer of strategic argumentation to ATL-like logics that are interpreted over transition systems. That would allows us to represent and reason about the argumentative abilities of agents in the style of modal logic.

References

[1] Thomas Ågotnes, Wiebe van der Hoek, and Michael Wooldridge. Normative system games. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, pages 1–8, 2007.

[2] Natasha Alechina, Mehdi Dastani, and Brian Logan. Reasoning about normative update. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 20–26. AAAI Press, 2013.

[3] Gianvincenzo Alfano, Sergio Greco, Francesco Parisi, and Irina Trubitsyna. Epistemic abstract argumentation framework: Formal foundations, computation and complexity. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 409–417, 2023.

[4] R. Alur, T. A. Henzinger, and O. Kupferman. Alternating-time Temporal Logic. *Journal of the ACM*, 49:672–713, 2002.

[5] Leila Amgoud. An argumentation-based model for reasoning about coalition structures. In *ArgMAS*, pages 217–228, 2005.

[6] Leila Amgoud. Towards a formal model for task allocation via coalition formation. In *AAMAS*, pages 1185–1186, 2005.

[7] Leila Amgoud and Henri Prade. Formalizing practical reasoning under uncertainty: An argumentation-based approach. In *2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT'07)*, pages 189–195. IEEE, 2007.

[8] Ofer Arieli, AnneMarie Borg, Jesse Heyninck, and Christian Straßer. Logic-based approaches to formal argumentation. In DM Gabbay, Massimiliano Giacomin, Guillermo R Simari, and Matthias Thimm, editors, *Handbook of Formal Argumentation*, volume 3, pages 707–838. College Publications, 2021.

[9] Sergei Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(4):477–513, 2008.

[10] Katie Atkinson and Trevor Bench-Capon. Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, 171(10-15):855–874, 2007.

[11] Katie Atkinson and Trevor Bench-Capon. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20, 2018.

[12] Alexandru Baltag, Nick Bezhanishvili, Aybüke Özgün, and Sonja Smets. Justified belief and the topology of evidence. In Jouko Väänänen, Åsa Hirvonen, and Ruy de Queiroz, editors, *Logic, Language, Information, and Computation*, pages 83–103. Springer, 2016.

[13] Alexandru Baltag and Lawrence S Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[14] Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief change, soft evidence and defeasible knowledge. In *Logic, Language, Information and Computation: 19th International Workshop, WoLLIC 2012, Buenos Aires, Argentina, September 3-6, 2012. Proceedings 19*, pages 168–190. Springer, 2012.

[15] Alexandru Baltag, Bryan Renne, and Sonja Smets. The logic of justified belief, explicit knowledge, and conclusive evidence. *Annals of Pure and Applied Logic*, 165(1):49–81, 2014.

[16] Alexandru Baltag and Sonja Smets. A qualitative theory of dynamic interactive belief revision. In Wiebe van der Hoek, Giacomo Bonanno, and Michael Wooldridge, editors, *Logic and the foundations of game and decision theory (LOFT 7)*, volume 3 of *Texts in Logic and Games*, pages 9–58. Amsterdam University Press, 2008.

[17] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. Abstract argumentation frameworks and their semantics. In Pietro Baroni, Dov M. Gabbay, Massimilino Giacomin, and Leendert van der Torre, editors, *Handbook of formal argumentation*, pages 159–236. College Publications, 2018.

[18] Howard Barringer, Dov M. Gabbay, and John Woods. Modal and temporal argumentation networks. *Argument Comput.*, 3(2-3):203–227, 2012.

[19] Dorothea Baumeister, Matti Järvisalo, Daniel Neugebauer, Andreas Niskanen, and Jörg Rothe. Acceptance in incomplete argumentation frameworks. *Artificial Intelli-*

gence, 295:103470, 2021.

[20] Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. Using argumentation to model agent decision making in economic experiments. *Autonomous Agents and Multi-Agent Systems*, 25(1):183–208, 2012.

[21] Philippe Besnard and Sylvie Doutre. Checking the acceptability of a set of arguments. In James P. Delgrande and Torsten Schaub, editors, *Proceedings of the NMR*, pages 59–64. AAAI Press, 2004.

[22] Philippe Besnard, Alejandro Garcia, Anthony Hunter, Sanjay Modgil, Henry Prakken, Guillermo Simari, and Francesca Toni. Introduction to structured argumentation. *Argument & Computation*, 5(1):1–4, 2014.

[23] Philippe Besnard and Anthony Hunter. A review of argumentation based on deductive arguments. *Handbook of Formal Argumentation*, 1:437–484, 2018.

[24] Guido Boella, Joris Hulstijn, and Leendert W. N. van der Torre. A logic of abstract argumentation. In *ArgMAS*, volume 4049 of *Lecture Notes in Computer Science*, pages 29–41. Springer, 2005.

[25] Thomas Bolander. A gentle introduction to epistemic planning: The DEL approach. In Sujata Ghosh and R. Ramanujam, editors, *Proceedings of the Ninth Workshop on Methods for Modalities, M4M@ICLA 2017, Indian Institute of Technology, Kanpur, India, 8th to 10th January 2017*, volume 243 of *EPTCS*, pages 1–22, 2017.

[26] Thomas Bolander and Mikkel Birkegaard Andersen. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34, 2011.

[27] Nils Bulling, Carlos Iván Chesñevar, and Jürgen Dix. An argumentative approach for modelling coalitions using ATL. In Iyad Rahwan and Pavlos Moraitis, editors, *Argumentation in Multi-Agent Systems, Fifth International Workshop, ArgMAS 2008, Estoril, Portugal, May 12, 2008. Revised Selected and Invited Papers*, volume 5384 of *Lecture Notes in Computer Science*, pages 197–216. Springer, 2008.

[28] Nils Bulling, Jürgen Dix, and Carlos Iván Chesñevar. Modelling coalitions: ATL + argumentation. In Lin Padgham, David C. Parkes, Jörg P. Müller, and Simon Parsons, editors, *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 2*, pages 681–688. IFAAMAS, 2008.

[29] Nils Bulling, Wojciech Jamroga, and Jürgen Dix. Reasoning about temporal properties of rational play. *Ann. Math. Artif. Intell.*, 53(1-4):51–114, 2008.

[30] Alfredo Burrieza and Antonio Yuste-Ginel. Basic beliefs and argument-based beliefs in awareness epistemic logic with structured arguments. In Henry Prakken, Stefano Bistarelli, Francesco Santini, and Carlo Taticchi, editors, *Proceedings of the COMMA 2020*, pages 123–134. IOS Press, 2020.

[31] Alfredo Burrieza and Antonio Yuste-Ginel. An awareness epistemic framework for belief, argumentation and their dynamics. In Joseph Y. Halpern and Andrés Perea, editors, *Proceedings Eighteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, volume 335 of *EPTCS*, pages 69–83, 2021.

[32] Martin Caminada. Rationality postulates: applying argumentation theory for non-monotonic reasoning. *Journal of Applied Logics*, 4(8):2707–2734, 2017.

[33] Martin WA Caminada and Dov M Gabbay. A logical account of formal argumentation. *Studia Logica*, 93(2-3):109–145, 2009.

[34] Brian F Chellas. *Modal logic: an introduction*. Cambridge university press, 1980.

[35] Carlos Iván Chesñevar, Ana Gabriela Maguitman, and Ronald Prescott Loui. Logical models of argument. *ACM Comput. Surv.*, 32(4):337–383, dec 2000.

[36] E. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, 1999.

[37] Yannis Dimopoulos, Jean-Guy Mailly, and Pavlos Moraitis. Control argumentation frameworks. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. AAAI Press, 2018.

[38] Sylvie Doutre, Andreas Herzig, and Laurent Perrussel. A dynamic logic framework for abstract argumentation. In C. Baral, G. De Giacomo, and T. Eiter, editors, *Fourteenth International Conference on the Principles of Knowledge Representation and Reasoning*. AAAI Press, 2014.

[39] Sylvie Doutre, Andreas Herzig, and Laurent Perrussel. Abstract argumentation in dynamic logic: Representation, reasoning and change. In Beishui Liao, Thomas Ågotnes, and Yi N. Wang, editors, *Dynamics, Uncertainty and Reasoning*, pages 153–185. Springer, 2019.

[40] Sylvie Doutre, Faustine Maffre, and Peter McBurney. A dynamic logic framework for abstract argumentation: adding and removing arguments. In Salem Benferhat, Karim Tabia, and Moonis Ali, editors, *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems.*, volume 10351 of *LNCS*, pages 295–305. Springer, 2017.

[41] Phan Minh Dung. On the acceptability of arguments and its fundamental role in non-monotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357, 1995.

[42] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. pages 199–218. Springer, 2009.

[43] Paul E. Dunne and Trevor J. M. Bench-Capon. Two party immediate response disputes: Properties and efficiency. *Artif. Intell.*, 149(2):221–250, 2003.

[44] Paul Égré, Paul Marty, and Bryan Renne. Knowledge, justification, and adequate reasons. *The review of symbolic logic*, 14(3):687–727, 2021.

[45] Ronald Fagin and Joseph Y Halpern. Belief, awareness, and limited reasoning. *Artificial intelligence*, 34(1):39–76, 1987.

[46] Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Vardi. *Reasoning about knowledge*. MIT press, 2004.

[47] Dov Gabbay and Davide Grossi. When are two arguments the same? equivalence in abstract argumentation. In *Johan van Benthem on Logic and Information Dynamics*, pages 677–701. Springer, 2014.

[48] Dov M. Gabbay and Gadi Rozenberg. Introducing abstract argumentation with many

lives. *FLAP*, 7(3):295–336, 2020.

[49] Dov M. Gabbay and Gadi Rozenberg. Evolutionary temporal logic for modelling many-lives argumentation networks. *FLAP*, 10(5):909–966, 2023.

[50] Dov M Gabbay and Valentin B Shehtman. Products of modal logics, part 1. *Logic journal of IGPL*, 6(1):73–146, 1998.

[51] Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963.

[52] Cristian Gratie, Adina Magda Florea, and John-Jules Ch Meyer. Full hybrid μ-calculus, its bisimulation invariance and application to argumentation. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 181–194. Springer, 2012.

[53] Davide Grossi. Doing argumentation theory in modal logic. 2009.

[54] Davide Grossi. Argumentation in the view of modal logic. In Peter McBurney, Iyad Rahwan, and Simon Parsons, editors, *International Workshop on Argumentation in Multi-Agent Systems*, volume 6614 of *LNCS*, pages 190–208. Springer, 2010.

[55] Davide Grossi. On the logic of argumentation theory. In W. van der Hoek, G.A. Kaminka, Y. Lesperance, M. Luck, and S. Sen, editors, *9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2010)*, pages 409–416. IFAAMAS, 2010.

[56] Davide Grossi and Simon Rey. Credulous acceptability, poison games and modal logic. In N. Agmon, M. E. Taylor, E. Elkind, and M. Veloso, editors, *18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019*, pages 1994–1996. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2019.

[57] Davide Grossi and Wiebe van der Hoek. Justified beliefs by justified arguments. In Chitta Baral, Giuseppe De Giacomo, and Thomas Eiter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourteenth International Conference*. AAAI Press, 2014.

[58] Christos Hadjinikolis, Yiannis Siantos, Sanjay Modgil, Elizabeth Black, and Peter McBurney. Opponent modelling in persuasion dialogues. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[59] Andreas Herzig, Emiliano Lorini, and Faustine Maffre. Possible worlds semantics based on observation and communication. In Hans van Ditmarsch and Gabriel Sandu, editors, *Jaakko Hintikka on Knowledge and Game-Theoretical Semantics*, pages 339–362. Springer, 2018.

[60] Andreas Herzig and Antonio Yuste-Ginel. Abstract argumentation with qualitative uncertainty: An analysis in dynamic logic. In Pietro Baroni, Christoph Benzmüller, and Yi N. Wáng, editors, *Logic and Argumentation*, volume 13040 of *LNCS*, pages 190–208. Springer, 2021.

[61] Andreas Herzig and Antonio Yuste-Ginel. On the Epistemic Logic of Incomplete Argumentation Frameworks. In M. Bienvenu, G. Lakemeyer, and E. Erdem, editors, *Proceedings of the 18th International Conference on Principles of Knowledge Repre-*

sentation and Reasoning, pages 681–685, 11 2021.

[62] Kaarlo Jaakko Juhani Hintikka. Knowledge and belief: An introduction to the logic of the two notions. 1962.

[63] W. Jamroga and N. Bulling. A general framework for reasoning about rational agents. In *Proceedings of AAMAS'07*, pages 592–594, Honolulu, Hawaii, USA, 2007. ACM Press. Short paper.

[64] Wojtek Jamroga and Nils Bulling. A logic for reasoning about rational agents. In F. Sadri and K. Satoh, editors, *Proceedings of CLIMA '07*, pages 54–69, Porto, Portugal, 2007. Univesidade Do Porto.

[65] Max Knobbout and Mehdi Dastani. Reasoning under compliance assumptions in normative multiagent systems. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 331–340. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[66] Max Knobbout, Mehdi Dastani, and John-Jules Meyer. A dynamic logic of norm change. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 886–894, 2016.

[67] Max Knobbout, Mehdi Dastani, and John-Jules Ch Meyer. Reasoning about dynamic normative systems. In *European Workshop on Logics in Artificial Intelligence*, pages 628–636. Springer, 2014.

[68] Agi Kurucz, Frank Wolter, Michael Zakharyaschev, and Dov M Gabbay. *Many-dimensional modal logics: theory and applications*. Gulf Professional Publishing, 2003.

[69] Beishui Liao, Michael Anderson, and Susan Leigh Anderson. Representation, justification, and explanation in a value-driven agent: an argumentation-based approach. *AI and Ethics*, 1(1):5–19, 2021.

[70] Beishui Liao, Nir Oren, Leendert van der Torre, and Serena Villata. Prioritized norms in formal argumentation. *Journal of Logic and Computation*, 29(2):215–240, 2019.

[71] Beishui Liao and Leendert van der Torre. Explanation semantics for abstract argumentation. In Henry Prakken, Stefano Bistarelli, Francesco Santini, and Carlo Taticchi, editors, *Computational Models of Argument - Proceedings of COMMA 2020, Perugia, Italy, September 4-11, 2020*, volume 326 of *Frontiers in Artificial Intelligence and Applications*, pages 271–282. IOS Press, 2020.

[72] Jieting Luo, Beishui Liao, and Dov M. Gabbay. Value-based practical reasoning: Modal logic + argumentation. In *COMMA*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, pages 248–259. IOS Press, 2022.

[73] Jieting Luo, Beishui Liao, and Dov M. Gabbay. Value-based practical reasoning: Modal logic + argumentation. In *COMMA*, volume 353 of *Frontiers in Artificial Intelligence and Applications*, pages 248–259. IOS Press, 2022.

[74] Jieting Luo, John-Jules Meyer, and Max Knobbout. A formal framework for reasoning about opportunistic propensity in multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(4):457–479, 2019.

[75] Jean-Guy Mailly. Yes, no, maybe, i don't know: Complexity and application

of abstract argumentation with incomplete knowledge. *Argument & Computation*, 13(3):291–324, 2022.

[76] Sanjay Modgil and Henry Prakken. A general account of argumentation with preferences. *Artificial Intelligence*, 195:361–397, 2013.

[77] Aybüke Özgün. *Evidence in epistemic logic: a topological perspective*. PhD thesis, Université de Lorraine, 2017.

[78] Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

[79] John L Pollock. Defeasible reasoning. *Cognitive science*, 11(4):481–518, 1987.

[80] Henry Prakken. Combining sceptical epistemic reasoning with credulous practical reasoning. *COMMA*, 144:311–322, 2006.

[81] Carlo Proietti and Antonio Yuste-Ginel. Persuasive argumentation and epistemic attitudes. In Luís Soares Barbosa and Alexandru Baltag, editors, *Dynamic Logic. New Trends and Applications*, volume 12005 of *LNCS*, pages 104–123. Springer, 2020.

[82] Carlo Proietti and Antonio Yuste-Ginel. Dynamic epistemic logics for abstract argumentation. *Synthese*, 199(3):8641–8700, 2021.

[83] Iyad Rahwan and Leila Amgoud. An argumentation based approach for practical reasoning. In *Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '06, page 347–354, New York, NY, USA, 2006. Association for Computing Machinery.

[84] Tjitze Rienstra, Matthias Thimm, and Nir Oren. Opponent models with uncertainty for strategic argumentation. In Francesca Rossi, editor, *Twenty-Third International Joint Conference on Artificial Intelligence IJCAI 2013*. AAAI Press, 2013.

[85] Chiaki Sakama and Tran Cao Son. Epistemic argumentation framework: Theory and computation. *Journal of Artificial Intelligence Research*, 69:1103–1126, 2020.

[86] François Schwarzentruber, Srdjan Vesic, and Tjitze Rienstra. Building an epistemic logic for argumentation. In Luis Fariñas del Cerro, Andreas Herzig, and Jérôme Mengin, editors, *Logics in Artificial Intelligence*, volume 7519 of *LNCS*, pages 359–371. Springer, 2012.

[87] C Shi, S Smets, and FR Velázquez-Quesada. Argument-based belief in topological structures. In J Lang, editor, *Proceedings of the Sixteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, EPTCS. Open Publishing Association, 2017.

[88] Chenwei Shi. *Reason to believe*. PhD thesis, University of Amsterdam, 2018.

[89] Chenwei Shi. No false grounds and topology of argumentation. *Journal of Logic and Computation*, 31(4):1079–1101, 2021.

[90] Chenwei Shi, Sonja Smets, and Fernando R Velázquez-Quesada. Beliefs based on evidence and argumentation. In *Proceedings of WoLLIC 2018*, volume 10944 of *LNCS*, pages 289–306. Springer, 2018.

[91] Chenwei Shi, Sonja Smets, and Fernando R Velázquez-Quesada. Logic of justified beliefs based on argumentation. *Erkenntnis*, pages 1–37, 2021.

[92] Chenwei Shi, Sonja Smets, and Fernando R. Velázquez-Quesada. Beliefs supported by

binary arguments. *Journal of Applied Non-Classical Logics*, 28(2-3):165–188, 2018.

[93] Guillermo Ricardo Simari and Ronald Prescott Loui. A mathematical treatment of defeasible reasoning and its implementation. *Artif. Intell.*, 53(2-3):125–157, 1992.

[94] Dan Sperber. Intuitive and reflective beliefs. *Mind & Language*, 12(1):67–83, 1997.

[95] Robert Stalnaker. On logics of knowledge and belief. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 128(1):169–199, 2006.

[96] Matthias Thimm. Strategic argumentation in multi-agent systems. *KI-Künstliche Intelligenz*, 28(3):159–168, 2014.

[97] Johan van Benthem, David Fernández-Duque, and Eric Pacuit. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133, 2014.

[98] Johan van Benthem, David Fernández-Duque, Eric Pacuit, et al. Evidence logic: A new look at neighborhood structures. volume 9 of *Advances in modal logic*, pages 97–118. College Publications, 2012.

[99] Leendert van der Torre. From classical to non-monotonic deontic logic using aspic. In *Logic, Rationality, and Interaction: 7th International Workshop, LORI 2019, Chongqing, China, October 18–21, 2019, Proceedings*, volume 11813, page 71. Springer Nature, 2019.

[100] Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic epistemic logic*. Springer, 2007.

[101] Georg Henrik Von Wright. *An essay in modal logic*. Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Company, 1951.

[102] Yì N Wáng and Xu Li. A logic of knowledge based on abstract arguments. *Journal of Logic and Computation*, 31(8):2004–2027, 2021.

[103] Timothy Williamson. *Knowledge and its Limits*. Oxford University Press on Demand, 2002.

[104] Timothy Williamson. 20 knowledge first epistemology. *The Routledge companion to epistemology*, 2011.

[105] Liuwen Yu, Dongheng Chen, Lisha Qiao, Yiqi Shen, and Leendert van der Torre. A principle-based analysis of abstract agent argumentation semantics. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 18, pages 629–639, 2021.

[106] Antonio Yuste-Ginel. *Arguments to believe and beliefs to argue. Epistemic logics for argumentation and its dynamics*. PhD thesis, 2022.

[107] Antonio Yuste-Ginel and Andreas Herzig. Qualitative uncertainty and dynamics of argumentation through dynamic logic. *Journal of Logic and Computation*, 33(2):370–405, 2023.

Received December 2024