Arguably argumentative: A formal approach to the argumentative theory of reason

No Author Given

No Institute Given

Abstract.

1 Introduction

sjurNeeds rewriting, but this is easy.

2 The social dimension of interaction – the need for new foundations

(sjur: Why all our references from various fields are implicitly in need of some common understanding of rational interaction that is not reducible to individual rationality.)

3 The argumentative theory – an attempt at a foundational theory

(sjur: What Mercier and Sperber's theory is and what it is not. It introduces the idea that agents are utility-maximizing in such a way that they seek to win arguments. This leads to behaviour typically seen as irrational and/or illogical. However, this irrationality might still lead to outcomes of interaction that appear reasonable/rational, sometimes also precisely *because* individual agents are not themselves rational in the classical sense. Hence the theory contains the idea that social rationality is not reducible to individual utility-maximizing (this is trivially true; it does not make sense to talk about a society that tries to "win" arguments, so the standard is obviously different at the social level). However, a weakness of Mercier and Sperber's theory is that they remain vague about what exactly the social rationality consists in. They only appeal to intuitions that appear classical, but do not inform us how these are to be defined when they target outcomes of deliberation as opposed to targeting internal reasoning processes. In some other work they attempt to tackle this by focusing on argument reception processes, giving a descriptive account of how outcomes are influenced by various conditions placed on deliberation. (I have some references I think...). This suggests characterizing social rationality in terms of such conditions. Perhaps even the conditions for deliberation are even more important than the perceived "classically correct" nature of the outcome? We don't now, but we propose formal logic as a means for further study of such questions....)

4 Background on abstract argumentation

Following Dung [**?**] we represent argumentation structures by directed graphs $\mathsf{F} = (S, E)$, such that S is a set of arguments and $E \subseteq S \times S$ encode the attacks between them, i.e., such that if $(x, y) \in E$ then the argument x attacks y. Traditionally, most work in formal argumentation theory has focused on defining and investigating notions of successful sets of arguments, in a setting where the argumentation framework is given and remains fixed. Such notions are typically formalized by an *argumentation semantics*, an operator ε which returns, for any AF F, the set of sets of arguments from $\mathsf{F} = (S, E)$ that are regarded as successful combinations, i.e., such that $\varepsilon(\mathsf{F}) \subseteq 2^S$. Many proposals exists in the literature, we point to [**?**] for a survey and formal comparison of different semantics. While some semantics, such as the grounded and ideal semantics, return a unique set of arguments, the "winners" of the argumentation scenario encoded by F, most semantics return more than one possible collection of arguments that *would* be successful if they were held together. For instance, the admissible semantics, upon which many of the other well-known semantics is built, returns, for each AF F, the following sets of arguments:

$$a(\mathsf{F}) = \{A \subseteq S \mid E^-(A) \subseteq E^+(A) \subseteq S \setminus A\}$$

That is, the admissible sets are those that can defend themselves against attacks (first inclusion), and do not involve any internal conflicts (second inclusion). A strengthening that is widely considered more appropriate (yet incurs some computational costs) is the *preferred* semantics p, which is defined by taking only those admissible sets that are set-theoretically maximal, i.e., such that they are not contained in any other admissible set. In general, an AF admits many preferred sets, and even more admissible ones. Indeed, notice that the empty set is always admissible by the default (the inappropriateness of which provides partial justification for using preferred semantics instead). As a simple example, consider F below.

$\mathsf{F}:$  $a(\mathsf{F}) = \{\emptyset, \{p\}, \{q\}\},\ \mathsf{p}(\mathsf{F}) = \{\{\mathsf{p}\}, \{\mathsf{q}\}\}$

Indeed, it seems hard to say which one of p and q should be regarded as successful in such a scenario. In the absence of any additional information, it seems safest to concede that choosing either one will be a viable option. Alternatively, one may take the view that due to the undetermined nature of the scenario, it should not be permitted to regard either argument as truly successful. This, indeed, is the view taken by unique status semantics, such as the grounded and ideal semantics. However, while such a restrictive view might be appropriate in some circumstance, it seems unsatisfactory for a general theory of argumentation. Surely, in most real-world argumentation situations, it is not tenable for an arbitrator to refrain from making a judgment whenever doing so would involve some degree of discretion on his part.

Since argumentation semantics typically only restrict the choice of successful arguments, without determining it completely, a modal notion of *accep-*

tance arises, usually referred to as *skeptical* acceptance in argumentation parlor, whereby an argument is said to be skeptically accepted by F under ε if $\forall S \in \varepsilon(\mathsf{F}) : p \in S$. The dual notion is called *credulous* acceptance, and obtains just in case $\exists S \in \varepsilon(\mathsf{F}) : p \in S$. Moreover, since the choice among elements of $\varepsilon(\mathsf{F})$ can itself be a contentious issue, and is not one which can be satisfactorily resolved by single-agent argumentation theory, there has been research devoted to giving an account of multi-agent interaction concerning the choice among members of $\varepsilon(\mathsf{F})$, see [**?**,**?**]. While this is interesting, it seems that another aspect of real-world argumentation has an even stronger multi-agent flavor, namely the process by which one arrives at a common AF in the first place. Certainly, two agents, a and b, might disagree about whether to choose p or q in F considered above, but as it stands, such a choice appears arbitrary and, most likely, the two agents would also be willing to admit as much. Arguably, then, the disagreement itself is only superficial. The agents disagree, but they provide no *reason* for their different preferences, and do bot provide any content or structure to substantiate them. This leaves an arbitrator in much the same position as he was in before: he might note the different opinions raised, but he has no basis upon which to inquire into their merits, and so his choice must, eventually, still be an exercise in discretion.

In practice, however, it would have to be expected that if the agents a and b were really committed to their stance, they would not simply accept that F correctly encodes the situation and that the choice is in fact arbitrary. Rather, they would produce *arguments* to back up their position. It might be, for instance, that agent a, who favors p, claims that q is inconsistent for some reason, while agent b, who favors q, makes the same accusation against the argument p. Then, however, we are no longer justified in seeing this as disagreement about which choice to make from $\varepsilon(\mathsf{F})$. Rather, the disagreement concerns the nature of the argumentation structure itself. The two agents, in particular, put forth different *views* on the situation. For instance, in our toy example, we would have to consider the following two AFs, where V_a, V_b encode the views of a and b respectively.

$$V_a: \quad \circlearrowleft p \rightleftarrows q \qquad\qquad V_b: \quad p \rightleftarrows q \circlearrowright \tag{1}$$

Then the question arises: what are we to make of this?

In the following, we address this question, and we approach it from the conceptual starting point that evaluating (higher-order) differences of opinion such as that expressed by V_a, V_b takes place iteratively, through a process of *deliberation*, leading, in a step-by-step fashion, to an aggregated *common* F. Such a process might be instantiated in various ways: it could the agents debating the matter among themselves and reaching some joint decision, or it could be an arbitrator who considers the different views and reasons about them by emulating such a process. Either way, we are not interested in attempting to provide any guidance towards the "correct" outcome, which is hardly possible in general. Rather, we are interested in investigating the modalities that arise when we consider the space of all possible outcomes (where possible will be defined in

due course). Moreover, we are interested in investigating structural questions, asking, for instance, about the importance of the order in which arguments are considered, and the consequences of limiting attention to only a subset of arguments.

We use a dynamic modal logic to facilitate this investigation, and in the next section we define the basic framework and show that model checking is decidable even on infinite AF's, as long as the agent's views remains finitely branching, i.e., as long as no argument is attacked by infinitely many other arguments. We will parameterize our logic by an argumentation semantics, so that it can be applied to any such semantics which satisfies a normality condition. In particular, let $C(\mathsf{F}) = \{C_1^{\mathsf{F}}, \ldots, C_i^{\mathsf{F}}, \ldots\}$ denote the (possibly infinite) set of maximal connected components from F (the set of all maximal subsets of S such that any two arguments in the same set are connected by a sequence of attacks). Then we say that a semantics ε is *normal* if we have, for any $\mathsf{F} = (S, E)$

$$A \in \varepsilon(\mathsf{F}) \Leftrightarrow A = \bigcup_i A_i \text{ for some } A_1, \ldots, A_i, \ldots \text{ s.t. } A_i \in \varepsilon(C_i^{\mathsf{F}}) \text{ for all } i \quad (2)$$

That is, a semantics is normal if the status of an argument depends only on those arguments to which it has some (indirect) relationship through a sequence of attacks. We remark that all argumentation semantics of which we are aware satisfies this requirement, hence we feel justified in dubbing it normality.

5 Deliberative dynamic logic

We assume given a finite non-empty set $\mathcal{A}$ of agents and a countably infinite set Π of arguments.[1] The basic building block of dynamic deliberative logic is provided in the following definition.

Definition 1. *A basis for deliberation is an $\mathcal{A}$-indexed collection of digraphs $\mathcal{B} = (V_a)_{(a \in \mathcal{A})}$, such that for each $a \in \mathcal{A}$, $V_a \subseteq \Pi \times \Pi$.*

Given a basis which encodes agents' view of the arguments, we are interested in the possible ways in which agents can deliberate to reach *agreement* on how arguments are related. That is, we are interested in the set of all AFs that can plausibly be seen as resulting from a *consensus* regarding the status of the arguments in Π. What restrictions is it reasonable to place on a consensus? It seems that while many restrictions might arise from pragmatic considerations, and be implemented by specific protocols for "good" deliberation in specific contexts, there are few restrictions that can be regarded as completely general. For instance, while there is often good reason to think that the position held by the majority will be part of a consensus, it is hardly possible to stipulate an axiomatic restriction on the notion of consensus amounting to the principle of majority rule. Indeed, sometimes deliberation takes place and leads to a single

[1] Possibly "statements" or "positions", depending on the context of application.

dissenting voice convincing all the others, and often, these deliberative processes are far more interesting than those that transpire along more conventional lines. However, it seems reasonable to assume that whenever *all* agents agree on how an argument p is related to an argument q, then this relationship is part of any consensus. This, indeed, is the only restriction we will place on the notion of a consensus; that when the AF F is a consensus for *basis*, it must satisfy the following *faithfulness* requirement.

– *For all $p, q \in \Pi$, if there is no disagreement about p's relationship to q (attack/not attack), then this relationship is part of* F

This leads to the following definition of the set $\Upsilon(\mathcal{B})$, which we will call the set of *complete assents* for $\mathcal{B}$, collecting all AFs that are faithful to $\mathcal{B}$.

$$\Upsilon(\mathcal{B}) = \left\{ \mathsf{F} \subseteq \Pi \times \Pi \;\middle|\; \bigcap_{a \in \mathcal{A}} V_a \subseteq \mathsf{F} \bigcup_{a \in \mathcal{A}} V_a \right\} \tag{3}$$

An element of $\Upsilon(\mathcal{B})$ represents a possible consensus among agents in $\mathcal{A}$, but it is an *idealization* of the notion of assent, since it disregards the fact that in practice, assent tends to be *partial*, since it results from a dynamic process, emerging through *deliberation*. Indeed, as long as the number of arguments is not bounded we can *never* hope to arrive at complete assent via deliberation. We can, however, initiate a process by which we reach agreement on more and more arguments, in the hope that this will approximate some complete assent, or maybe even be *robust*, in the sense that there is *no* deliberative future where the results of current partial agreement end up being undermined. Complete assent, however, arises only in the limit.

When and how deliberation might successfully lead to an approximation of complete assent is a question well suited to investigation with the help of dynamic logic. The dynamic element will be encoded using a notion of a deliberative event – centered on an argument – such that the set of ways in which to relate this arguments to arguments previously considered gives rise to a space of possible deliberative time-lines, each encoding the continued stepwise construction of a joint point of view. This, in turn, will be encoded as a monotonically growing AF $\mathsf{F} = (S, E)$ where $S \subseteq \Pi, E \subseteq S \times S$ and such that faithfulness is observed by all deliberative events. That is, an event consists in adding to F the agents' combined view of p with respect to the set $S \cup \{p\}$. This leads to the following collection of possible events, given a basis $\mathcal{B}$, a partial consensus[2] $\mathsf{F} = (S, E)$ and an argument $p \in \Pi$:

$$\mathcal{U}_{\mathcal{B}}(\mathsf{F}, p) = \left\{ X \;\middle|\; \bigcap_{a \in \mathcal{A}} V_a|_{S \cup \{p\}} \subseteq X \subseteq \bigcup_{a \in \mathcal{A}} V_a|_{S \cup \{p\}} \right\} \tag{4}$$

To provide a semantics for a logical approach to deliberation based on such events, we will use Kripke models.

[2] These "partial consensuses" are sometimes referred to as "contexts" when they are used to describe graphs inductively, as we will do later.

Definition 2 (Deliberative Kripke model). *Given an argumentation semantics ε and a set of views $\mathcal{B}$, the deliberative Kripke models induced by $\mathcal{B}$ and ε is the triple $\mathcal{K}_{(\mathcal{B},\varepsilon)} = (Q_{\mathcal{B}}, \mathsf{R}_{\mathcal{B}}, \mathbf{e}_{\varepsilon})$ such that*

- *$Q_{\mathcal{B}}$, the set of points, is the set of all pairs of the form $q = (q_S, q_E)$ where $q_S \subseteq \Pi$ and*

$$\bigcap_{a \in \mathcal{A}} V_a|_{q_S} \subseteq q_E \subseteq \bigcup_{a \in \mathcal{A}} V_a|_{q_S}$$

  The basis $\mathcal{B}$ together with our definition of an event, given in Equation 4, induces the following function, mapping states to their possible deliberative successors, defined for all $p \in \Pi, q \in Q_{\mathcal{B}}$ as follows

$$succ(p, q) \quad := \quad \{\ (q_S \cup \{p\}, q_E \cup X) \mid X \in \mathcal{U}_{\mathcal{B}}(q, p)\ \}$$

  We also define a lifting, for all states $q \in Q_{\mathcal{B}}$:

$$succ(q) \quad := \quad \{\ q' \mid \exists p \in \Pi : q' \in succ(q, p)\}$$

- *$\mathsf{R}_{\mathcal{B}} : \Pi \cup \{\exists\} \to 2^{Q_{\mathcal{B}} \times Q_{\mathcal{B}}}$ is a map from symbols to relations on $Q_{\mathcal{B}}$ such that*
  - *$\mathsf{R}_{\mathcal{B}}(p) = \{(q, q') \mid q' \in succ(p, q)\}$ for all $p \in \Pi$ and*
  - *$\mathsf{R}_{\mathcal{B}}(\exists) = \{(q, q') \mid q' \in succ(q)\}$,*

- *$\mathbf{e}_{\varepsilon} : Q_{\mathcal{B}} \to 2^{(3^{\Pi})}$ maps states to labellings such that for all $q \in Q_{\mathcal{B}}$ we have $\mathbf{e}_{\varepsilon}(q) = (\pi_1, \pi_0, \pi_{\frac{1}{2}})$ with*

$$\mathbf{e}_{\varepsilon}(q) = \{\pi \mid \pi_1 \in \varepsilon(q), \pi_0 = \{p \in q_S \mid \exists q \in \pi_1 : (q, p) \in q_E\}\}$$

Notice that in the last point, we essentially map q to the sets of extensions prescribed by ε when q is viewed as an AF. We encode this extension as a three-valued labeling, however, following [**?**]. Notice that the default status, attributed to all arguments not in q_S, is $\frac{1}{2}$. The logical language we will use consists in two levels. For the lower level, used to talk about static argumentation, we follow [**?**,**?**] in using Łukasiewicz three-valued logic. Then, for the next level, we use a dynamic modal language which allows us to express consequences of updating with a given argument, and also provides us with existential quantification over arguments, allowing us to express claims like "there is an update such that ϕ". This leads to the language $\mathcal{L}_{\mathrm{DDL}}$ defined by the following BNF's

$$\phi \quad ::= \blacklozenge\alpha \mid \neg\phi \mid \phi \wedge \phi \mid \langle p \rangle \phi \mid \Diamond\phi$$

where $p \in \Pi$ and $\alpha \in \mathcal{L}^{\blacklozenge}$ where $\mathcal{L}^{\blacklozenge}$ is defined by the following grammar:

$$\alpha ::= p \mid \neg\phi \mid \phi \to \phi$$

for $p \in \Pi$.

We also use standard abbreviations such that $\Box\phi = \neg\Diamond\neg\phi$, $[p]\phi = \neg\langle p \rangle\neg\phi$ and $\blacksquare\alpha = \neg\blacklozenge\neg\alpha$. We also consider that standard boolean connectives abbreviated as usual for connectives not occurring inside a $\blacklozenge$-connective and abbreviations for connectives of Łukasiewicz logic in the scope of $\blacklozenge$-connectives.

Next we define truth of formulas on deliberative Kripke models. We begin by giving the valuation of complex formulas from $\mathcal{L}^{\blacklozenge}$, which is simply three-valued Łukasiewicz logic.

Definition 3 (α-satisfaction). *For any three-partitioning $\pi = (\pi_1, \pi_0, \pi_{\frac{1}{2}})$ of Π, we define*

$$\begin{aligned} \overline{\pi}(p) &= x \ \ s.t\ p \in \pi_x \\ \overline{\pi}(\neg\alpha) &= 1 - \overline{\pi}(\alpha) \\ \overline{\pi}(\alpha_1 \to \alpha_2) &= \min\{1, 1 - (\overline{\pi}(\alpha_1) - \overline{\pi}(\alpha_2))\} \end{aligned}$$

Now we can give a semantic interpretation of the full language as follows.

Definition 4 ($\mathcal{L}_{\mathbf{DDL}}$-satisfaction). *Given an argumentation semantics ε and a basis $\mathcal{B}$, truth on $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ is defined inductively as follows, in all points $q \in Q_{\mathcal{B}}$.*

$$\begin{aligned} \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \blacklozenge\alpha &\iff \text{there is } \pi \in \mathbf{e}_\varepsilon(q) \text{ s.t. } \overline{\pi}(\phi) = 1 \\ \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \neg\phi &\iff \text{not } \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \phi \\ \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \phi \wedge \psi &\iff \text{both } \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \phi \text{ and } \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \psi \\ \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \langle\pi\rangle p &\iff \exists (q, q') \in \mathsf{R}_{\mathcal{B}}(p) : \mathcal{K}_{(\mathcal{B},\varepsilon)}, q' \vDash \phi \\ \mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \Diamond\phi &\iff \exists (q, q') \in \mathsf{R}_{\mathcal{B}}(\exists) : \mathcal{K}_{(\mathcal{B},\varepsilon)}, q' \vDash \phi \end{aligned}$$

To illustrate the definition, we return to the example depicted in (1). In Figure 1, we depict this basis together with a fragment of the corresponding Kripke model, in particular the fragment arising from the p-successors of $(\emptyset, \emptyset)$.

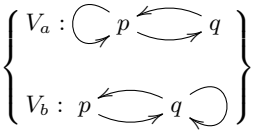
$\mathcal{B} =$ 

Fig. 1. A fragment of the deliberative Kripke model for $\mathcal{B}$.

Let us assume that $\varepsilon = \mathsf{p}$ is the preferred semantics. Then the following list gives some formulas that are true on $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ at the point $(\emptyset, \emptyset)$, and the reader should easily be able to verify them by consulting the above fragment of $\mathcal{K}_{(\mathcal{B},\varepsilon)}$.

$$\begin{aligned} &\langle p\rangle\blacksquare p, \quad \Diamond\blacksquare p, \quad [p]\Diamond\blacksquare q, \\ &\neg[p]\Diamond\blacklozenge p, \ \langle p\rangle\Diamond\blacksquare\neg p, \ \Diamond\Diamond(\blacklozenge p \wedge \blacklozenge q) \end{aligned}$$

We can also record some validities that are easy to verify against Definition 2.

Proposition 1. *The following formulas are all validities of $\mathcal{L}_{\mathrm{DDL}}$, for any $p, q \in \Pi$, $\phi \in \mathcal{L}_{\mathrm{DDL}}$.*

1. $\langle p\rangle\langle q\rangle\phi \leftrightarrow \langle q\rangle\langle p\rangle\phi$
2. $\langle p\rangle[q]\phi \rightarrow [q]\langle p\rangle\phi$
3. $\Diamond\Box\phi \rightarrow \Box\Diamond\phi$
4. $\langle p\rangle\langle p\rangle\phi \rightarrow \langle p\rangle\phi$

We remark that $[q]\langle p\rangle\phi \rightarrow \langle p\rangle[q]\phi$ is *not* valid, as witnessed for instance by the following basis $\mathcal{B}$, for which we have $\mathcal{K}_{(\mathcal{B},p)}, (\emptyset,\emptyset) \models [q]\langle p\rangle\blacksquare p$ but also $\mathcal{K}_{(\mathcal{B},p)}, (\emptyset,\emptyset) \models [p]\langle q\rangle\blacksquare q$ (as the reader may easily verify by considering the corresponding Kripke model).

$$\mathcal{B} = \left\{ \begin{array}{l} V_a : \ p \longrightarrow q \\ V_b : \ p \longleftarrow q \end{array} \right\}$$

Finally, let us notice that as Π is generally infinite, we must expect to encounter infinite bases. This means, in particular, that our Kripke models are often infinite. However, in the next section we show that as long as $\mathcal{B}$ is *finitary*, meaning that no agent $a \in \mathcal{A}$ has a view where an argument is attacked by infinitely many other arguments, we can solve the model-checking problem also on infinite models.

6 Model checking on finitary models

Towards this result, we now introduce some notation and a few abstractions to simplify our further arguments. We will work with labeled trees, in particular, where we take a tree over labels X to be some non-empty, prefix-closed subset of X^* (finite sequences of elements of X). Notice that trees thus defined contain no infinite sequences. This is intentional, since we will "shrink" our models (which may contain infinite sequences of related points), by mapping them to trees. To this end we will use the following structures.

Definition 5. *Given a basis $\mathcal{B}$, we define $\mathcal{I}(\mathcal{B})$, a set of sequences over $\Pi \times 2^\Pi$ labeled by AFs, defined inductively as follows*

Base case: $\epsilon \in \mathcal{I}(\mathcal{B})$ *and is labeled by the AF* $\mathsf{F}(\epsilon) = (S(\epsilon), E(\epsilon))$ *where* $S(\epsilon) = \emptyset = E(\epsilon)$.

Induction step: *If $x \in \mathcal{I}(\mathcal{B})$, then for any $p \in \Pi$ and any partial assent $X = \mathcal{U}_\mathcal{B}(x,p)$, we have $x;(p,X) \in \mathcal{I}(\mathcal{B})$ labeled by the AF $\mathsf{F}(x;(p,X))$ where $S(x;(p,X)) = S(x) \cup \{p\}$ and $E(x;(p,X)) = E(x) \cup X$.*

To adhere to standard naming we use ϵ to denote the empty string. It should not be confused with the argumentation semantics ε. This will also be clear from the context. We next define tree-representations of our Kripke models.

Definition 6. *Let $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ be some model. The* tree representation *of $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ is the set T, together with the representation map $\gamma : Q_\mathcal{B} \rightarrow 2^T$, defined inductively as follows*

Base case *$\epsilon \in T$ is the root with $\gamma((\emptyset,\emptyset)) = \{\epsilon\}$.*

Induction step *For any* $x \in T, q \in \mathcal{K}_{(\mathcal{B},\varepsilon)}$ *with* $x \in \gamma(q)$ *and* $q' \in succ(q)$ *witnessed by* $p \in \Pi$ *and* $X \in \mathcal{U}_{\mathcal{B}}(x,p)$*, we have* $x;(p,X) \in T$ *with* $q' \in \gamma(x;(p,X))$*.*

Notice that the tree-representation is a tree where each node is an element of $\mathcal{I}(\mathcal{B})$. Some single states in $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ will have several representations in a tree. That is, $\gamma(q)$ may not be a singleton. On the other hand, it is easy to see that for every state $q \in \mathcal{K}_{(\mathcal{B},\varepsilon)}$, and every path from $(\emptyset,\emptyset)$ to q, there will be a node $x \in T$ such that $q \in \gamma(x)$.

The main result of our paper is that model checking $\mathcal{L}_{\mathrm{DDL}}$-truth at $(\emptyset,\emptyset)$ is tractable as long as all views are *finitely branching*, i.e., such that for all $a \in \mathcal{A}, p \in \Pi$, p has only finitely many attackers in V_a. Clearly this requires shrinking the models since the modality $\Diamond$ quantifies over an infinite domain whenever Π is infinite. We show, however, that attention can be restricted to arguments from Π that are *relevant* to the formula we are considering. To make the notion of relevance formal, we will need the following measure of complexity of formulas.

Definition 7. *The* white modal depth *of* $\phi \in \mathcal{L}_{\mathrm{DDL}}$ *is* $|\phi|^{\Diamond} \in \mathbb{N}$*, which is defined inductively as follows*

$$
\begin{array}{rcll}
|\alpha|^{\Diamond} & := & 0 & \textit{no white connectives in these formulas} \\
|\blacklozenge\alpha|^{\Diamond} & := & 0 & \\
|\neg\phi|^{\Diamond} & := & |\phi|^{\Diamond} & \textit{depth is deepest nesting of} \\
|\phi \wedge \psi|^{\Diamond} & := & \max\{|\phi|^{\Diamond}, |\psi|^{\Diamond}\} & \textit{white connectives} \\
|\Diamond\phi|^{\Diamond} & := & 1 + |\phi|^{\Diamond} & \\
|\langle p\rangle\phi|^{\Diamond} & := & 1 + |\phi|^{\Diamond} &
\end{array}
$$

We let $\Pi|_{\phi}$ denote the set of arguments occurring in ϕ in sub-formulas from $\mathcal{L}^{\blacklozenge}$. Notice that given a state $q \in Q_{\mathcal{B}}$, the satisfaction of a formula of the form $\phi = \blacklozenge\alpha$ at the AF encoded by q is not dependent on the entire digraph $q = (q_S, q_E)$.

Indeed, this is what motivated our definition of normality for an argumentation semantics, leading to the following simple lemma, which is the first step towards shrinking Kripke structures for the purpose of model checking. Given a model $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ and a state $q \in Q_{\mathcal{B}}$, we let $C(q,\Phi)$ denote the digraph consisting of all connected components from q which contains a symbol from Φ. Then we obtain the following.

Lemma 1. *Given a semantics* ε *and two bases* $\mathcal{B}$ *and* $\mathcal{B}'$*, we have, for any two states* $q \in \mathcal{K}_{(\mathcal{B},\varepsilon)}$ *and* $q' \in \mathcal{K}_{(\mathcal{B}',\varepsilon)}$ *and for any formula* $\phi \in \mathcal{L}_{\mathrm{DDL}}$ *with* $|\phi|^{\Diamond} = 0$*:*

$$\big(C(q,\Pi|_{\phi}) = C(q',\Pi|_{\phi})\big) \Rightarrow \big(\mathcal{K}_{(\mathcal{B},\varepsilon)}, q \vDash \phi \Leftrightarrow \mathcal{K}_{(\mathcal{B}',\varepsilon)}, q' \vDash \phi\big)$$

In order to complete our argument in this section, we will make use of n-bisimulations modulo a set of symbols.

Definition 8. *Given two models (with possibly different bases, but with common set of symbols Π and semantic ε) $K_{\mathcal{B}} = \langle Q_{\mathcal{B}}, R, \varepsilon\rangle$ and $K'_{\mathcal{B}} = \langle Q_{\mathcal{B}'}, R', \varepsilon\rangle$, states $q \in Q_{\mathcal{B}}$ and $q' \in Q'_{\mathcal{B}}$, a natural number n and a set $\Phi \subseteq \Pi$, then we say that q and q' are n-bisimilar modulo Φ (denoted $(K_{\mathcal{B}}, q) \underline{\leftrightarrow}_n^{\Phi} (K_{\mathcal{B}'}, q')$), if, and only if, there are $n+1$ relations relation $Z_n \subseteq Z_{n-1} \subseteq \cdots \subseteq Z_0 \subseteq Q_{\mathcal{B}} \times Q'_{\mathcal{B}}$ such that*

1. *qZ_nq',*
2. *whenever $(v, v') \in Z_0$, then $C(v, \Phi) = C(v', \Phi)$,*
3. *whenever $(v, v') \in Z_{i+1}$ and vRu, then there is a u' s.t. $v'R'u'$ and uZ_iu',*
4. *whenever $(v, v') \in Z_{i+1}$ and $v'R'u'$, then there is a u s.t. vRu and uZ_iu'.*

Let us now also define a particular subset of arguments, the arguments which have at most distance n from some given set of arguments:

Definition 9. *Given a basis $\mathcal{B} = (V_a)_{(a\in\mathcal{A})}$, a subset $\Phi \subseteq \Pi$ and a number n, the n-vicinity of Φ is $D(\mathcal{B}, \Phi, i) \subseteq \Pi$, defined inductively as follows*

$$\begin{aligned} D(\mathcal{B}, \Phi, 0) &= \Phi \\ D(\mathcal{B}, \Phi, n+1) &= D(\mathcal{B}, \Phi, n) \\ &\quad \cup \left\{ p \in \Pi \mid \exists q \in D(\mathcal{B}, \Phi, n) : \{(p,q),(q,p)\} \cap \bigcup_{a\in\mathcal{A}} V_a \neq \emptyset \right\} \end{aligned}$$

Notice that as long as Φ is finite and all agents' views have finite branching, then the set D is also finite. Also notice that an equivalent characterization of the set $D((V_a)_{(a\in\mathcal{A})}, \Phi, i)$ can be given in terms of paths as follows: an argument $p \in \Pi$ is in $D(\mathcal{B}, \Phi, i)$ if, and only if, there is a path $p = x_1x_2\ldots x_n$ in $\bigcup_{a\in\mathcal{A}} V_a$ such that $x_n \in \Phi$ and $n \leq i$ (we consider an argument p equivalently as an empty path at p).

Definition 10. *Given a formula $\phi \in \mathcal{L}_{\mathrm{DDL}}$. Let $(V_a)_{(a\in\mathcal{A})}$ be a possibly infinite basis , we define $\rho_\phi((V_a)_{(a\in\mathcal{A})})$ such that*

- *for every $a \in \mathcal{A}$, $\rho_\phi(V_a) := V_a \cap D(V_a, \Pi|_\phi, |\phi|^\Diamond)$*

Notice that the Kripke model for $\rho(\mathcal{B})$ will have finite branching as long as the argument symbols in the $|\phi|^\Diamond$-vicinity of the argument symbols in ϕ have finite branching in all agents' views. In the following, we will show that for any finitely branching $\mathcal{B}$ and normal ε, we have $\mathcal{K}_{(\mathcal{B},\varepsilon)}, (\emptyset,\emptyset) \models \phi$ if, and only if, $\mathcal{K}_{(\rho_\phi(\mathcal{B}),\varepsilon)}, (\emptyset,\emptyset) \models \phi$.

Theorem 1. *Let $\mathcal{B}$ be an arbitrary basis , and and $\phi \in \mathcal{L}_{\mathrm{DDL}}$.*

$$\left(\mathcal{K}_{(\mathcal{B},\varepsilon)}, (\emptyset,\emptyset)\right) \quad \underline{\leftrightarrow}_{|\phi|^\Diamond}^{\Pi|_\phi} \quad \left(\mathcal{K}_{(\rho_\phi(\mathcal{B}),\varepsilon)}, (\emptyset,\emptyset)\right)$$

Proof. Let $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ be an arbitrary model and let T denote its tree representation, while T' denotes the tree representation of $\mathcal{K}_{(\rho_\phi(\mathcal{B}),\varepsilon)}$.

We take $n = |\phi|^\Diamond$ and let Φ be the atoms occurring in ϕ inside the scope of some $\blacklozenge$-operator. Moreover, for brevity, we denote $D = D(\mathcal{B}, \Phi, n)$.

Definition of $(Z_i)_{(0\leq i\leq |\phi|^\Diamond)}$: We define all the relations Z_i inductively using the tree-representations as follows.

Base case: $(i = 0)$ For all $0 \leq i \leq n$, we let $\epsilon Z_i \epsilon$.

Induction step: $(0 < i \leq n)$ For all $y = x; (v, X) \in T$ and $y' = x'; (v', X') \in T'$, both of length i, with $x(Z_{i+1})x'$. We let, for every $k \leq i$, $y(Z_k)y'$ if, and only if, $v = v'$, *and* $X \cap (D \times D) = X'$.

Notice that if $x(Z_i)x'$, then $S(x) = S(x')$ and $|S(x)| \leq (n - i)$. Moreover, by consulting Definition 6 it is not hard to see that for all $q \in Q_{\mathcal{B}}, q' \in Q_{\rho(\mathcal{B})}$ we have, for all $0 \leq i \leq n$ and all $q \in Q_{\mathcal{B}}, q' \in Q_{\rho(\mathcal{B})}$:

$$\forall x_1, x_2 \in \gamma(q) : \forall x_1', x_2' \in \gamma(q') : x_1(Z_i)x_2 \iff x_1'(Z_i)x_2'$$

This means, in particular, that the following lifting of $(Z_i)_{0\leq i\leq n}$ to models is well-defined, for all $q \in Q_{\mathcal{B}}, q' \in Q_{\rho(\mathcal{B})}$ and all $0 \leq i \leq n$:

$$q(Z_i)q' \iff x(Z_i)x'$$

for some $x \in \gamma(x), x' \in \gamma(q')$.

Next we show that $(Z_i)_{0\leq i\leq n}$ so defined is an n-bisimulation between $\mathcal{K}_{(\mathcal{B},\varepsilon)}$ and $\mathcal{K}_{(\rho(\mathcal{B}),\varepsilon)}$.

$(Z_i)_{0\leq i\leq n}$ witnesses n-bisimulation: We address all the points of the definition of n-bisimulation modulo Φ in order.

1. Clearly, $(\emptyset, \emptyset)Z_n(\emptyset, \emptyset)$. Hence the first condition of the definition is satisfied.
2. Consider any arbitrary states q, q' and let $x = x_1; x_2; \ldots; x_m$ and $x' = x_1'; x_2'; \ldots; x_m'$ be the corresponding nodes from T, T' that witnesses to $q(Z_0)q'$. By definition of Z_0 we have $S(x) = S(x')$, but it is possible that we have $E(x) \neq E(x')$. However, we must have $C(\mathsf{F}(x), \Phi) = C(\mathsf{F}(x'), \Phi)$, and to see this, it is enough to observe that as $m \leq n$, each of x and x' contains at most n nodes. Then, since $\mathsf{F}(x) = q$ and $\mathsf{F}(x') = q'$ are the same on D, and the distance from $\Pi \setminus D$ to Φ is greater than n. That is, any path from an argument in $\Pi \setminus D$ to an argument in $\Phi = \Pi|_\phi$ would be a path consisting of at least $n+1$ nodes. It follows that no element from Φ can be in a connected components containing elements outside of D.
3. Consider now q, q' corresponding to x and x' such that $x(Z_{i+1})x'$. Notice that $(q, r) \in \mathsf{R}_{\mathcal{B}}(\exists)$ if, and only if, there is a (p, X) such that $xR(x; (p, X))$. So all we need to show is that $X \cap (D \times D)$ is in $\mathcal{U}_{\rho_\phi(\mathcal{B})}(x', p)$. Then it will follow that there is a successor to x', namely $(p, X \cap (D \times D))$, with $(x')R'(x'; (p, X \cap (D \times D)))$. This is a straightforward consequence of the Definition 10 of ρ. The argument for the particular sub relations $\mathsf{R}_{\mathcal{B}}(p)$ is analogous.

4. Finally consider q, q' corresponding to x and x' such that $x(Z_{i+1})x'$ for (p, X) such that $x'R(x'; (p, X'))$. Again we need to ensure that there is an $X \in \mathcal{U}_{\mathcal{B}}(x, p)$ such that $X' = X \cap (D \times D)$, and again this follows from the Definition 10 of ρ. The argument for the particular sub relations $\mathsf{R}_{\mathcal{B}}(p)$ is analogous.

Proposition 2. *Let $\phi \in \mathcal{L}_{\mathrm{DDL}}$ and $\mathcal{B}, \mathcal{B}'$ arbitrary bases. If states $q \in \mathcal{K}_{(\mathcal{B},\varepsilon)}$ and $q' \in \mathcal{K}_{(\mathcal{B}',\varepsilon)}$ are $|\phi|^{\Diamond}$-bisimilar modulo $\Pi|_{\phi}$, then $\mathcal{K}_{(\mathcal{B},\varepsilon)}, q \models \phi \Leftrightarrow \mathcal{K}_{(\mathcal{B}',\varepsilon)}, q' \models \phi$. Or, succinctly*

$$\left((\mathcal{K}_{(\mathcal{B},\varepsilon)}, q) \underline{\leftrightarrow}^{\Pi|_{\phi}}_{|\phi|^{\Diamond}} (\mathcal{K}_{(\mathcal{B}',\varepsilon)}, q')\right) \Rightarrow \left(\mathcal{K}_{(\mathcal{B},\varepsilon)}, q \models \phi \Leftrightarrow \mathcal{K}_{(\mathcal{B}',\varepsilon)}, q' \models \phi\right).$$

Proof. The proof is by induction on $|\phi|^{\Diamond}$.

Base case: ($|\phi|^{\Diamond} = 0$) There are no white connectives, and our states, q and q', are clearly 0-bisimilar modulo Φ. It is also easy to see, consulting Definition 2, that the truth of a formula of modal depth 0 is only dependent on the AF q. Then it follows from the fact that ε is assumed to be normal that the truth of ϕ is in fact only dependent on $C(q, \Phi)$. From $q(Z_0)q'$, we obtain $C(q, \Phi) = C(q', \Phi)$ and the claim follows.

Induction step: ($|\phi|^{\Diamond} > 0$) We skip the boolean cases as these are trivial, so let $\phi := \Diamond\psi$ (the case of white connectives with an explicit argument is similar). Suppose $|\phi|^{\Diamond} = i + 1$ and $q(Z_{i+1})q'$. Suppose further that $\mathcal{K}_{(\mathcal{B},\varepsilon)}, q \models \Diamond\psi$. Then there is a successor of q, $v \in succ(q)$ such that $\mathcal{K}_{(\mathcal{B},\varepsilon)}, v \models \psi$. All successors of q will be i-bisimilar to a successor of q' (point 3. of Definition 8). So we have $(\mathcal{K}_{(\mathcal{B},\varepsilon)}, v) \underline{\leftrightarrow}^{\Phi}_{i} (\mathcal{K}_{(\mathcal{B}',\varepsilon)}, v')$. As $|\psi|^{\Diamond} < |\Diamond\psi|^{\Diamond}$ we can apply our induction hypothesis to obtain $\mathcal{K}_{(\mathcal{B}',\varepsilon)}, v' \models \psi$, and $\mathcal{K}_{(\mathcal{B},\varepsilon)}, q' \models \Diamond\psi$ as desired.

7 Conclusion and future work

We have argued for a logical analysis of deliberative processes by way of modal logic, where we avoid making restrictions that may not be generally applicable, and instead focus on logical analysis of the space of possible outcomes. The deliberative dynamic logic (DDL) was put forth as a concrete proposal, and we showed some results on model checking.

We notice that DDL only allows us to study deliberative processes where every step in the process is explicitly mentioned in the formula. That is, while we quantify over the arguments involved and the way in which updates take place, we do not quantify over the *depth* of the update. For instance, a formula like $\Diamond\Box p$ reads that there is a deliberative update such that no matter what update we perform next, we get ϕ. A natural next step is to consider instead a formula $\Diamond\Box^{*}\phi$, with the intended reading that there is an update which not only makes ϕ true, but ensures that it remains true for all possible future *sequences* of updates. Introducing such formulas to the logic, allowing the deliberative modalities to be iterated, is an important challenge for future work. Moreover,

we would also like to consider even more complex temporal operators, such as those of computational tree logic, or even μ-calculus.

Finding finite representations for the deliberative truths that can be expressed in such languages appears to be much more challenging, but we would like to explore the possibility of doing so.

Also, we would like to explore the question of validity for the resulting logics, and the possibility of obtaining some compactness results. Indeed, it seems that if we introduce temporal operators we will be able to express truths on arbitrary points $q \in Q_{\mathcal{B}}$ by corresponding formulas that are true at $(\emptyset, \emptyset)$, thus capturing the way in which complete assent can be faithfully captured by a finite (albeit unbounded) notion of iterated deliberation.

If the history of the human race is anything to go by, it seems clear that we never run out of arguments or controversy. But it might also be that some patterns or structures are decisive enough that they warrant us to conclude that the *truth* has been settled, even if deliberation may go on indefinitely. A further logical inquiry into this and related questions will be investigated in future work.

References