Journal of applied logics (Print) ISSN 2631-9810
Journal of applied logics (Online) ISSN 2631-9829

Journal of Applied Logics

The IfCoLog Journal of Logics and their Applications

Volume 10 • Issue 3 • May 2023

Special Issue:
Advances in Argumentation in AI

Guest Editors
Marcello D'Agostino
Fabio Aurelio D'Asaro
Costanza Larese

Available online at
www.collegepublications.co.uk/journals/ifcolog/
Free open access

Journal of Applied Logics - IfCoLog
Journal of Logics and their Applications

Volume 10, Number 3

May 2023

Disclaimer

Statements of fact and opinion in the articles in Journal of Applied Logics - IfCoLog Journal of Logics and their Applications (JALs-FLAP) are those of the respective authors and contributors and not of the JALs-FLAP. Neither College Publications nor the JALs-FLAP make any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

ISBN 978-1-84890-427-9
ISSN (E) 2631-9829
ISSN (P) 2631-9810

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Editorial Board

Scope and Submissions

This journal considers submission in all areas of pure and applied logic, including:

pure logical systems
proof theory
constructive logic
categorical logic
modal and temporal logic
model theory
recursion theory
type theory
nominal theory
nonclassical logics
nonmonotonic logic
numerical and uncertainty reasoning
logic and AI
foundations of logic programming
belief change/revision
systems of knowledge and belief
logics and semantics of programming
specification and verification
agent theory
databases

dynamic logic
quantum logic
algebraic logic
logic and cognition
probabilistic logic
logic and networks
neuro-logical systems
complexity
argumentation theory
logic and computation
logic and language
logic engineering
knowledge-based systems
automated reasoning
knowledge representation
logic in hardware and VLSI
natural language
concurrent computation
planning

This journal will also consider papers on the application of logic in other subject areas: philosophy, cognitive science, physics etc. provided they have some formal content.

Submissions should be sent to Jane Spurr (jane@janespurr.net) as a pdf file, preferably compiled in LaTeX using the IFCoLog class file.

Contents

ARTICLES

Advancing the Boundaries of Formal Argumentation: Reflections on the AI^3 2021 Special Issue

Marcello D'Agostino
LUCI Group, University of Milan, Italy
`marcello.dagostino@unimi.it`

Fabio Aurelio D'Asaro*
Ethos Group, University of Verona, Italy
`fabioaurelio.dasaro@univr.it`

Costanza Larese
LUCI Group, University of Milan, Italy
`costanza.larese@unimi.it`

Abstract

This article reflects on the Special Issue based on invited papers from the *5th Workshop on Advances in Argumentation in Artificial Intelligence* (*AI^3 2021*), showcasing the latest advancements in the field made by the Italian community on argumentation, as well as other researchers worldwide. This Special Issue highlights the importance of advancing logical-based AI approaches, such as formal argumentation, in the continuously expanding landscape of Artificial Intelligence. Papers in this Special Issue cover a diverse range of topics, including argument game-based proof theories, analysis of legal cases, decomposability in abstract argumentation, meta-argumentation approaches, explanations for model outputs using causal models, representation of natural argumentative discourse, and Paraconsistent Weak Kleene logic-based belief revision. By emphasizing these innovative research contributions, this article underscores the need for continued progress in the field of Formal Argumentation to complement and enhance the ongoing developments in AI.

*Corresponding Author

1 Introduction

The study of *argumentation theory* has deep roots in logic and philosophy, and has recently become a burgeoning field in Artificial Intelligence (AI) as researchers explore methods for formalizing and reasoning with arguments and conflicting information. Argumentation provides procedures for making and explaining decisions and is able to capture diverse kinds of reasoning and dialogue activities in a formal yet intuitive way, enabling the integration of different specific techniques and the development of trustable applications.

With the advent of Dung's abstract argumentation frameworks [8], a foundation for representing conflicts between arguments was established, leading to a wide range of applications and advancements in the AI community. Dung's work inspired the development of several alternative and complementary argumentation frameworks, such as bipolar argumentation frameworks (see, e.g., [6]), which consider both support and attack relations between arguments; value-based argumentation frameworks (see, e.g., [2]), which incorporate the role of values and preferences in the evaluation of arguments; and structured argumentation frameworks, such as Assumption-Based Argumentation (see, e.g., [9]) and Defeasible Logic Programming (see, e.g., [10]), which provide more detailed representations of the internal structure and content of arguments. Some papers included in this Special Issue (see [5, 1, 3, 4, 11]) focus on defining and examining new argumentation frameworks, as well as representing argumentation processes.

Over the past two decades, formal argumentation has developed into a thriving area of AI research. As theoretical models have been established, practical applications have emerged in various fields, including social network dialogues, law, and medicine. In this Special Issue, some papers (see [7, 12]) are driven by practical needs, such as legal argumentation, and explainability in AI.

Given that the study of argumentation is inherently interdisciplinary, the goal of the *Advances in Argumentation in Artificial Intelligence* (*AI^3*) workshop series, co-located with the *International Conference of the Italian Association for Artificial Intelligence* (*AIxIA*), is to stimulate discussions and promote scientific collaboration among researchers not only directly involved in argumentation, but also from research fields indirectly related to argumentation. Cross-fertilization with different fields, including non-monotonic reasoning, logic programming, linguistics, natural language processing, philosophy, and psychology, is essential for updating and extending foundations in Argumentation Theory, as well as tackling a number of open issues currently debated in the area. Interdisciplinary collaborations are necessary to foster the adoption of argumentation as a viable AI paradigm with a wide range of applications.

In this special issue, we bring together extended selected papers from the 5th edition of the AI^3 workshop held in 2021 (see `http://sites.google.com/view/ai3-2021` for the website of the workshop), which showcases state-of-the-art applications and developments in the field. The contributions in this issue highlight recent advances in various types of argumentation frameworks, including alternatives to Dung's abstract argumentation frameworks, innovative algorithms for reasoning with arguments, and real-world use cases demonstrating the practical impact of argumentation techniques. Furthermore, these articles provide valuable insights into the challenges and future directions of argumentation research, helping to shape the ongoing discourse in this exciting and evolving field. In Section 2 we introduce and discuss the contributions to this Special Issue. Some final remarks conclude this editorial in Section 3.

2 Description of the Papers in the Special Issue

We grouped together papers in this Special Issue according to whether they are inspired by theoretical motivations or applicative ones. In particular, the first subgroup focuses on dialectical argument games, argumentation frameworks, the modeling of the burden of persuasion, and modeling or representation of argumentation processes, highlighting the need to refine and advance the theoretical foundations of argumentation in various contexts. The second subgroup emphasizes the application of argumentation theory to real-world cases, legal argumentation, and explainability in AI, demonstrating the practical value and potential impact of argumentation research on diverse domains. By organizing the papers in this way, we aim to showcase the rich interplay between theoretical advancements and practical applications in the field of argumentation, fostering further developments and cross-disciplinary collaboration.

2.1 Theoretical Foundations and Advances in Argumentation

Papers in this subsection explore the theoretical foundations and advances in argumentation. These include novel frameworks and formalisms to better understand and represent argumentative discourse and reasoning, as well as innovative approaches to address specific challenges faced by resource-bounded agents. Among the key topics covered are dialectical argument game proof theories, the decomposability of semantics in abstract argumentation, adpositional argumentation for representing natural argumentative discourse, the introduction of a PWK-style argumentation framework, and the modeling of the burden of persuasion.

"Decomposing Semantics in Abstract Argumentation" by Pietro Baroni, Federico Cerutti and Massimiliano Giacomin [1]: This paper introduces a general model for investigating decomposability in abstract argumentation, which is the possibility of determining the labellings prescribed by a semantics based on evaluations of local functions in sub-frameworks. The main aim is to analyze the range of decomposable semantics with varying degrees of local information and to devise a constructive procedure to identify local functions. The research questions addressed include modeling diverse kinds of information exploited in local computations, determining the range of decomposable semantics under different degrees of local information, determining the local counterpart of an argumentation semantics to guarantee decomposability, and exploiting the model and results to analyze semantics decomposability properties.

The paper establishes a monotone relationship between the degree of information available locally and the set of decomposable semantics. It also investigates the construction of local functions for the computation of local labellings by introducing a general constructive procedure independent of the specific semantics definitions. Two kinds of local functions are identified that enforce decomposability if the semantics and the local information exploited make it possible. Finally, the decomposability properties of stable, grounded, and preferred semantics are analyzed under local information concerning close neighbors.

"Dialectical Argument Game Proof Theories for Classical Logic" by Federico Castagna [5]: The paper introduces argument games for Dialectical Classical Logic Argumentation (*Cl-Arg* for short), an approach that provides dialectical characterizations of Cl-Arg arguments by resource-bounded agents while preserving the rational criteria established by the rationality postulates and practical desiderata. These argument games aim to better approximate bounded non-monotonic reasoning processes.

Dialectical Cl-Arg revolves around the core notion of dialectical defeats, which enable argumentative interactions more aligned with the dialectical reasoning of resource-bounded agents. The study aims to develop argument games for Dialectical Cl-Arg that address the following main aspects of argumentation by resource-bounded agents: (i) demonstrating the inconsistencies of an opponent's argument by assuming its premises, (ii) handling finite subsets of the arguments of the AFs, (iii) reducing resource consumption while still satisfying the rationality postulates and practical desiderata by employing dialectical means. The author developed dialectical argument games for the admissible, preferred, and grounded semantics of Dialectical Cl-Arg, discovering interesting properties that differentiate these games from standard argument games. Dialectical games have specific relevance conditions that characterize their protocols, unique winning strategies, and conflict-freeness of the set of arguments moved by the proponent in the winning strategy. Conflict-freeness is particularly important as it provides various efficiency improvements for the games, such as preventing the proponent from playing self-defeating arguments, playing arguments already moved by the opponent, and playing arguments that defeat or are defeated by other arguments already moved by the proponent. Additionally, the paper suggests that efficiency improvement can be obtained by forbidding the opponent from repeating arguments that have already been defeated in the dialectical admissible/preferred game unless they have also been defended or indirectly defended by other arguments.

"The logic of the arguer. Representing natural argumentative discourse in Adpositional Argumentation" by Marco Benini, Federico Gobbo and Jean H.M. Wagemans [3]: This paper presents Adpositional Argumentation, a framework for representing natural argumentative discourse at various levels of abstraction, ranging from linguistic to pragmatic aspects. The framework's granularity allows analysts to study the unfolding of an arguer's logic throughout the discourse without imposing any specific interpretation.

Natural argumentative discourse is defined as a piece of natural language used to convince an audience of the acceptability of a particular point of view. The authors recognize that the lack of interaction between argumentation theory and computational argumentation has limited the development of tools and models for natural argumentative discourse. They propose Adpositional Argumentation to bridge this gap, offering a formalism that is uniform across multiple levels of abstraction.

The authors argue that the logic of the arguer is dynamic and unfolds throughout the discourse. By providing a detailed and unambiguous representation, Adpositional Argumentation can help analysts gain insights into the logic of the arguer and improve their understanding of the argumentative discourse. This framework lays the foundation for further research in areas such as inquiring strategies, representation of complex argumentation, and the dynamics of attacking and defending an argument in dialogues.

"A PWK-style Argumentation Framework and Expansion" by Massimiliano Carrara, Filippo Mancini and Wei Zhu [4]: This paper explores argumentation as an epistemic process performed by an agent to extend and revise beliefs and gain knowledge, focusing on the possibility of suspending the claim under evaluation. The authors propose to distinguish between two kinds of suspensions: critical and non-critical. Non-critical suspension occurs when an agent neither believes nor disbelieves certain information and can still form a judgment or continue processing an argument. Critical suspension, on the other hand, occurs when an agent gains irrelevant, off-topic, or even malicious information, which should be filtered and set apart from the argumentation process.

The paper introduces a Paraconsistent Weak Kleene logic (*PWK* for short) based belief revision theory, which uses the notion of topic to distinguish between the two kinds of suspensions. PWK logic includes a non-standard truth value, u, which is interpreted as "off-topic". This helps to account for critical and non-critical suspensions in argumentation.

The authors develop a PWK-style argumentation framework that extends the abstract argumentation framework and enables the distinction between critical and non-critical suspensions. They also present a PWK belief revision model, which serves as an expansion of the classical AGM belief revision model with two kinds of suspension.

"Burden of persuasion: a meta-argumentation approach" by Giuseppe Pisano, Roberta Calegari, Andrea Omicini and Giovanni Sartor [11]: This paper presents a burden of persuasion meta-argumentation model, which interprets the burden of persuasion as a set of meta-arguments. It separates the model into two levels: an object level, which deals with standard arguments, and a meta-level, which addresses the burden of persuasion. Bimodal graphs are used to define the interaction between these two levels. The proposed framework includes three main components: object-level argumentation, meta-level argumentation, and bimodal graphs.

The paper extends previous work by introducing a novel technological reification of the model that supports the burden inversion mechanism. It also positions the contribution against the state of the art and discusses related work, highlighting strengths and limitations compared to other approaches. The model is able to handle various nuances of burdens, such as reasoning over the concept of the burden itself, resulting in a comprehensive, interoperable framework that is open to further extensions. Additionally, the model effectively deals with the inversion of the burden.

2.2 Practical Applications and Real-world Implications of Argumentation

Papers in this subsection present innovative methodologies and frameworks in the field of applied argumentation, including the analysis of legal judgments and generating explanations for the outputs of machine learning classifiers using causal models and argumentation.

"A Formal Argumentation Exercise on the Karadžic Trial Judgment" by Federico Cerutti and Yvonne Mcdermott [7]: This paper presents the methodology and results of applying argumentation theory to map evidence and arguments regarding whether Radovan Karadžić, President of the Serb Republic, possessed the *mens rea* (i.e., knowledge of wrongdoing) for genocide in Srebrenica. The analysis results were submitted to the Mechanism for International Criminal Tribunals as an amicus curiae brief.

Using the argumentation-based techniques available in the `CISpaces.org` tool, the authors manually analyzed a subset of the judgment to highlight three reasoning lines that lead to the conclusion that Karadžić in fact possessed the requisite *mens rea*. Two of these reasoning lines might merit further discussion, and the last one relies on a single witness.

The main contribution of the paper is to show that the proposed methodology can be used to identify the strengths and weaknesses of a case. This can be useful for the plaintiff, defendant, judges, and jurors as it helps clarify which elements are proven beyond reasonable doubt and which ones are not. This is currently a live issue in international criminal law, with debates regarding whether each piece of evidence should be evaluated on its own merits in light of other evidence on the record or whether Trial Chambers should base their decisions on the accumulation of all evidence without needing to link factual and legal findings to the final decisions. Although the Appeals Chamber denied the admissibility of the application, the interest triggered in the international criminal law community suggests potential for future work in this area.

"Explaining Classifiers' Outputs with Causal Models and Argumentation" by Antonio Rago, Fabrizio Russo, Emanuele Albini, Francesca Toni and Pietro Baroni [12]: This paper introduces a novel approach to generate argumentation frameworks from causal models to forge explanations for the outputs of AI models, specifically machine learning classifiers. The methodology proposed involves reinterpreting properties of argumentation framework semantics as explanation moulds, characterizing argumentative relations. The authors focus on relation-based explanations, as they claim different users may need different forms of explanations based on their cognitive abilities, background, and goals.

The main contributions of the paper include proposing a new concept for defining relation-based explanations for causal models by inverting properties of argumentation semantics, defining a novel form of reinforcement explanation for causal models, and demonstrating the deployment of reinforcement explanations with two machine-learning models from which causal models are drawn. Moreover, an empirical evaluation shows promising preliminary results and indicates directions for future work.

The authors demonstrate their methodology by reinterpreting the property of bivariate reinforcement in bipolar Argumentation Frameworks, showing how extracted bipolar Argumentation Frameworks may be used as counterfactual explanations for the outputs of causal models. They then evaluate their method empirically, comparing it to a popular approach from the literature, and show advantages in highlighting specific relationships between feature and classification variables and generating counterfactual explanations with respect to a commonly used metric.

3 Conclusion

This Special Issue brings together a collection of seven papers from the 5th edition of the Workshop on Advances in Argumentation in Artificial Intelligence. Articles in this Special Issue explore various aspects of argumentation theory, from dialectical reasoning in classical logic, to applying argumentation in real-world legal cases, to investigating decomposability and burden of persuasion, and to generating explanations for machine learning classifiers. The contributions in this issue also emphasize the importance of understanding and modeling natural argumentative discourse and the development of new frameworks to handle the complexity of argumentation as an epistemic process.

These articles illustrate the variety of applications and the interdisciplinary nature of argumentation research, spanning artificial intelligence, computer science, logic, linguistics, philosophy, and law. They showcase innovative methodologies, novel frameworks, and empirical evaluations that advance our understanding of argumentation theory and its practical applications. Moreover, they highlight the necessity of bridging the gap between theoretical and computational aspects of argumentation to develop more accurate and efficient models that capture the complexities of real-world reasoning processes.

As argumentation theory continues to evolve, future research will likely focus on improving existing methodologies, expanding their applications to new domains, and refining the understanding of the intricate dynamics that underlie argumentation. The articles in this Special Issue challenge researchers to further advance the field of argumentation theory and its practical applications.

References

[1] Pietro Baroni, Federico Cerutti, and Massimiliano Giacomin. Decomposing semantics in abstract argumentation. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

[2] Trevor Bench-Capon. Value based argumentation frameworks. *arXiv preprint cs/0207059*, 2002.

[3] Marco Benini, Federico Gobbo, and Jean H.M. Wagemans. The logic of the arguer. representing natural argumentative discourse in adpositional argumentation. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

[4] Massimiliano Carrara, Filippo Mancini, and Wei Zhu. A PWK-style argumentation framework and expansion. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

[5] Federico Castagna. Dialectical argument game proof theories for classical logic. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

[6] Claudette Cayrol and Marie-Christine Lagasquie-Schiex. On the acceptability of arguments in bipolar argumentation frameworks. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005, Barcelona, Spain, July 6-8, 2005. Proceedings 8*, pages 378–389. Springer, 2005.

[7] Federico Cerutti and Yvonne Mcdermott. A formal argumentation exercise on the karadžic trial judgment. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

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

[9] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. *Argumentation in artificial intelligence*, pages 199–218, 2009.

[10] Alejandro J García and Guillermo R Simari. Defeasible logic programming: An argumentative approach. *Theory and practice of logic programming*, 4(1-2):95–138, 2004.

[11] Giuseppe Pisano, Roberta Calegari, Andrea Omicini, and Giovanni Sartor. Burden of persuasion: a meta-argumentation approach. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

[12] Antonio Rago, Fabrizio Russo, Emanuele Albini, Francesca Toni, and Pietro Baroni. Explaining classifiers' outputs with causal models and argumentation. *IfCoLog Journal of Logics and their Applications*, 10(3), 2023.

Received 9 April 2023

Dialectical Argument Game Proof Theories for Classical Logic

Federico Castagna
School of Computer Science, University of Lincoln, U.K.
fcastagna@lincoln.ac.uk

Abstract

Argument game-based proof theories provide procedural structures capable of determining the status of an argument. Given an argumentation framework, argument games identify the membership of an argument in a specific extension simulating a dispute between two opposing contenders. The semantics intended to be captured dictate the rules of the played game, which serve to describe how the players can achieve victory. Dialectical Classical logic Argumentation (Dialectical Cl-Arg) is a recent approach that provides real-world dialectical characterisations of Cl-Arg arguments by resource-bounded agents while preserving the rational criteria established by the rationality postulates and practical desiderata. This paper combines both subjects and introduces argument games for Dialectical Cl-Arg, highlighting the properties and benefits enjoyed by these games in comparison with the standard ones. The result will be a proof theory better equipped to approximate real-world non-monotonic single-agent reasoning processes.

1 Introduction

Since Aristotle's Organon [1, 33] and its considerable influence on the history of Western thought, rich scholarly literature has been investigating the intertwined notions of arguments, reasoning, and logic. For example, Walton claimed that *"logic is the evaluation of reasoning in arguments"* [35], whereas Mercier and Sperber emphasised the argumentative characterisation of reasoning:

> *"Reasoning is generally seen as a means to improve knowledge and make better decisions. However, much evidence shows that reasoning often*

The author would like to thank Peter McBurney and Marcello D'Agostino for the invaluable help and comments provided to previous drafts of the current paper.

> *leads to epistemic distortions and poor decisions. This suggests that the function of reasoning should be rethought. Our hypothesis is that the function of reasoning is argumentative. It is to devise and evaluate arguments intended to persuade."* [23]

Trying to consolidate possessed information by formulating reasons (via arguments) that challenge or defend them is an ordinary procedure in which humans engage. This process is not only common but even necessary: how could it be possible, otherwise, to decide what to believe or trust without being misled by a non-reliable source of information? This 'scaffolding' (as defined in [24]) role of dialogues and arguments can be seen in social and lone thinking practices where the reasoner(s) evaluates the possessed information by constructing counter-arguments that assess their acceptability. Thanks to its important role, argumentation has been developed as a rich, interdisciplinary area of research spanning Philosophy, Linguistics, Psychology and Artificial Intelligence. Able to characterize a promising paradigm for modelling reasoning in the presence of conflict and uncertainty, formal-logical accounts of the argumentation theory have come to be increasingly central as a core study within Artificial Intelligence. According to such a theory, in order to determine if a piece of information is acceptable, it will suffice to prove that the argument (in which the considered information is embedded) is justified under specific semantics. A way of doing this is to show the membership of the argument in a winning strategy of an argument game (as described, for example, in [25, 34] and [9]). Indeed, argument game-based proof theories provide procedural structures capable of determining the status of an argument according to the semantics intended to be captured.

Dung's abstract argumentation framework (AF) [17] has been considered the formalism from which stemmed most of the subsequent studies in this fruitful research field. Nevertheless, although a plethora of works has successfully shown various additions and instantiations of Dung's abstract AF and achieved different goals, none of these approaches managed to provide a full rational account for real-world resource-bounded agents. Undoubtedly, the introduction of the rationality postulates [6, 7], as well as desiderata for practical applications [20], have allowed eschewing the arising of counter-intuitive results in AFs instantiations. However, such requirements demand a consumption of resources that typically far exceed the availability of real-world agents.

1.1 Contribution

The main contribution of this research paper is the development of argument games for Dialectical Classical Logic Argumentation (Dialectical Cl-Arg [15]), a recent approach that provides real-world dialectical characterisations of AFs by resource-bounded agents. This approach satisfies the practical desiderata and the rationality postulates (under minimal requirements) and revolves around the core notion of *dialectical defeats*. Such defeats enable argumentative interactions more aligned with the dialectical reasoning of real-world resource-bounded agents. Thus, their presence requires the implementation of *dialectical argument game* proof theories capable of conveying the same idea as single-agent reasoning processes.

1.2 Paper Overview

The paper is organized as follows. Section 2 outlines an overview of the main definitions of Dung's argumentation framework, the standard argument games and Dialectical Cl-Arg. Section 3 provides the first contributions by establishing the general formal background that characterises the dialectical argument games. The other contributions occur in Sections 4 and 5, where (a) the protocol of the dialectical admissible game (which also yields the credulous preferred game) and (b) the protocol of the dialectical grounded game are given along with (c) their respective soundness and completeness results. The specific properties enjoyed by dialectical games in comparison with the standard ones are illustrated in Section 6, whereas Section 7 introduces potential efficiency improvements that may be embedded in the developed protocols. Section 8 presents the related works and some promising research paths that might be investigated in the future. Finally, Section 9 draws the conclusions and summarizes the paper findings.

2 Background

Argumentation has been developed as a theory able to characterize the essence of non-monotonic reasoning via the dialectical interplay of arguments. According to Dung's seminal paper [17], an Argumentation Framework (AF) is composed of a set of arguments 'AR' and a binary relation called *'attacks'*, which denotes conflicts existing between arguments in AR, i.e., AF = $\langle$AR, *attacks*$\rangle$. Various semantics have also been presented and each of them specifies the status of (*sceptically* or *credulously*) justified (i.e., acceptable) arguments. Several works stemmed from [17], some of which introduced different ways of structuring arguments and instantiating Dung's abstract AF [18, 31, 27]. For example, Classical Logic Argumentation (Cl-arg) [21, 2] is one such instantiation that builds AFs using classical logic as its underlying language.

2.1 Dialectical Classical Logic Argumentation

Unlike the standard formalisation of Cl-Arg, real-world agents behave pragmatically and do not need to: (i) always construct every argument defined by a base, (ii) enforce consistency and subset minimality checks on their arguments (nor do they have enough computational power to do these checks, given their limited resources). Dialectical Cl-Arg provides a formalisation of real-world modes of dialectical reasoning from resource-bounded agents whilst satisfying both the rationality postulates [6, 7] and practical desiderata [20].

Definition 1. [Dialectical Arguments] [15] $X = (\Delta, \Gamma, \alpha)$ *is a* dialectical argument *defined by a base $\mathcal{B}$ of classical wff, if* $(\Delta \cup \Gamma) \subseteq \mathcal{B}$*,* $\Delta \cap \Gamma = \emptyset$*, and* $\Delta \cup \Gamma \vdash_c \alpha$*. If* $\alpha = \curlywedge$ *then X is said to be a* falsum *argument. If* $\Gamma = \emptyset$ *then X is said to be* unconditional*; else X is* conditional*. Finally, if* $\Delta = \emptyset$ *then X is said to be* unassailable*.*

Δ*,* Γ *and* α *are respectively referred to as the* premises ($\mathit{Prem}(X)$)*,* suppositions ($\mathit{Supp}(X)$) *and* conclusion ($\mathit{Con}(X)$) *of* $X = (\Delta, \Gamma, \alpha)$*. Also, the union of premises and suppositions of X can be referred to as the* assumptions ($\mathit{Assumptions}(X)$) *of the argument.*

Attacks and defeats for Dialectical Cl-Arg work differently than their respective counterparts for Classical Logic Argumentation (Cl-Arg). The reason is the presence of suppositions embedded in the internal structure of the arguments. Intuitively, it is common practice for interlocutors in dialogues to differentiate between their own arguments' premises, regarded as true, and their opponents' premises that they want to challenge: "by considering what I deem to be valid and supposing what you have committed to, I can show your premises inconsistency". This motivates such an epistemic distinction between information considered true (i.e., $\mathit{Prem}(X)$, the *premises* of an argument X) and opponents' information supposed true (i.e., $\mathit{Supp}(X)$, the *supposition* of an argument X) which proves useful also in solving the so-called 'foreign commitment problem'[1].

Definition 2. [Attacks and Defeats][15] *Let AR be a set of dialectical arguments defined by a base $\mathcal{B}$. The attack relation 'attacks'* $\subseteq AR \times AR$ *is defined as follows. For any* $X = (\Delta, \Gamma, \alpha)$*,* $Y = (\Pi, \Sigma, \beta) \in AR$*: attacks*$(X, Y)$ *iff:*

- *if* $\alpha \neq \curlywedge$ *then* $\overline{\alpha} \in \Pi$ *(X attacks Y on $\overline{\alpha}$, equivalently on* $Y' = (\{\overline{\alpha}\}, \emptyset, \overline{\alpha})$*);*

[1] As extensively explained in [8], the foreign commitment problem is the issue that arises in dialogical applications when agents are forced to commit to the premises of their interlocutors in order to challenge their arguments.

- *if* $\alpha = \curlywedge$ *(*X *attacks* Y *on any* $\phi \in \Gamma \cap \Pi$*, equivalently on any* $Y' = (\{\phi\}, \emptyset, \phi)$*).*

Let $\prec$ *be a strict partial ordering over AR. Then, for every* X*,* Y *such that attacks*(X, Y)*, defeats*(X, Y) *iff exactly one of the following holds:*

- *either* X *is an argument of the form* $(\emptyset, \Gamma, \curlywedge)$*;*
- *else,* $\exists \psi \in$ *Prem*(Y) *such that attacks*(X, Y) *on* ψ*, and* $X \not\prec (\{\psi\}, \emptyset, \psi)$*.*

$X \Rightarrow Y$ *will stand for "defeats*(X, Y)*", and* $X \not\Rightarrow Y$ *will stand for "*$\neg$*defeats*(X, Y)*".*

The description of Dialectical Cl-Arg formalism provided in [15] accounts only for *undermine* attacks and the ensuing defeats based upon this type of conflict. Undermines are those kinds of attacks that occur when the conclusion of the attacking argument targets the premises of the challenged argument. Nevertheless, the literature (e.g., [29, 32]) presents *undercuts* and *rebuttals* as additional categories of conflicts. The first denotes arguments arguing against the defeasible inference rule used to derive the attackee's conclusion, whereas the second depicts a disagreement towards the attackee's defeasible conclusion. However, none of these conflicts can be transposed in Dialectical Cl-Arg since no defeasible rules (but only the classical entailment $\vdash_c$) are employed in the construction of the arguments.

The strict partial ordering of Definition 2 refers to the Elitist Preference Ordering. In addition, the authors of [15] show that such preference is also *'redundance-coherent'* in the sense that arguments are not strengthened when redundantly weakening with syntactically disjoint assumptions[2]. This is an important property that ensures the satisfaction of the non-contamination (i.e., Non-Interference and Crash Resistance) rationality postulates for Dialectical Cl-Arg.

Definition 3. [Elitist Preference Ordering]
Let X, Y *be dialectical classical logic arguments defined by a base* $\mathcal{B}$*, and* $\leq$ *a partial preordering over* $\mathcal{B}$*. Then:*

(i) $X \prec Y$ *iff* $\exists \alpha \in$ *Assumptions*(X) *such that* $\forall \beta \in$ *Assumptions*(Y)*,* $\alpha < \beta$*.*

(ii) $\prec$ *is redundance-coherent iff:* $\forall X, X', Y$ *such that* $X = (\Gamma, \emptyset, \alpha)$*,* $X' = (\Delta \cup \Gamma, \emptyset, \alpha)$*, and* $\Delta \parallel \Gamma \cup \{\alpha\}$*: if* $X \prec Y$ *then* $X' \prec Y$*.*

[2] Here 'weakening' denotes that a logical entailment from, say, Δ continues to be valid when adding some Γ to Δ. Also, we consider 'syntactically disjoint' (denoted by using '$\parallel$') two sets of formulae that do not have symbols in common.

Cl-Arg assumes instantiation of an AF by all arguments defined by a base $\mathcal{B}$ of classical wff, a task that proves to be unfeasible for agents with limited resources. As such, dialectical arguments (Definition 1) along with the described defeat relation (Definition 2) allow us to introduce a *dialectical AF* as an argumentation framework $\langle$AR, *defeats*$\rangle$ where AR is any subset of the dialectical arguments defined by a base $\mathcal{B}$.

$A_1 = (\{a\}, \emptyset, a)$	$B_1 = (\{b\}, \emptyset, b)$
$F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$	$G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
$F_2 = (\{b\}, \{\neg a \vee \neg b\}, \neg a)$	$G_2 = (\{a\}, \{\neg a \vee \neg b\}, \neg b)$
$F_3 = (\{\neg a \vee \neg b\}, \{b\}, \neg a)$	$G_3 = (\{\neg a \vee \neg b\}, \{a\}, \neg b)$
$N_1 = (\{a \supset b\}, \{\neg b\}, \neg a)$	$N_2 = (\{a \supset b, \neg b\}, \emptyset, \neg a)$
$N_3 = (\{a \supset b, a\}, \emptyset, b)$	$O_1 = (\{\neg(a \supset b)\}, \emptyset, \neg(a \supset b))$
$L_1 = (\{\neg b\}, \emptyset, \neg b)$	$X_3 = (\{b\}, \{\neg b\}, \curlywedge)$
$C_1 = (\{\neg a \vee \neg b\}, \emptyset, \neg a \vee \neg b)$	$H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
$X_1 = (\emptyset, \{a, b, \neg a \vee \neg b\}, \curlywedge)$	$X_2 = (\{a, b, \neg a \vee \neg b\}, \emptyset, \curlywedge)$

Table 1: Example of dialectical arguments defined by a base $\mathcal{B} = \{a,\ b,\ \neg a \vee \neg b,\ \neg b,\ a \supset b,\ \neg(a \supset b)\}$.

Defeats and dialectical defeats for dialectical AFs present an important difference: the reference to a set $\mathcal{S}$ of arguments. The general idea is that, when challenging the acceptability of an argument with respect to a set $\mathcal{S}$, the defeating argument can also suppose premises from all the arguments in $\mathcal{S}$. Whereas, the argument that defends $\mathcal{S}$ can only suppose the premises of the defeating argument. This new kind of defeat compelled the authors of [15] to adjust the standard semantics accordingly.

Definition 4. [Dialectical defeats and semantics for dialectical AFs][15] *Let $\langle AR,\ defeats\rangle$ be a dialectical AF, $\mathcal{S} \subseteq AR$ and X, $Y \in AR$. Then:*

1) X dialectically defeats Y with respect to $\mathcal{S}$, *denoted* $X \Rightarrow_{\mathcal{S}} Y$, *if defeats*$(X, Y)$ *and Supp*$(X) \subseteq$ *Prem*$(\mathcal{S} \cup \{Y\})$.

2) $\mathcal{S}$ *is* conflict-free *if* $\forall Z,\ Y \in \mathcal{S},\ Z \not\Rightarrow_{\mathcal{S}} Y$.

3) Y *is* acceptable *with respect to* $\mathcal{S}$ *if* $\forall X$ *such that* $X \Rightarrow_{\mathcal{S}} Y$, $\exists Z \in \mathcal{S}$ *such that* $Z \Rightarrow_{\{X\}} X$.

4) *Let* $\mathcal{S}$ *be conflict-free. Then* $\mathcal{S}$ *is: an* admissible *extension iff* $X \in \mathcal{S}$ *implies* X *is acceptable with respect to* $\mathcal{S}$*; a* complete *extension iff* $\mathcal{S}$ *is admissible and*

if X is acceptable with respect to $\mathcal{S}$ then $X \in \mathcal{S}$; a preferred *extension iff it is a set inclusion maximal complete extension; the* grounded *extension iff it is the set inclusion minimal complete extension.*

The following example depicts a scenario that clarifies the role of dialectical defeats while also providing a comparison between Dialectical Cl-Arg and Cl-Arg arguments. Since rigorous Cl-Arg formal definitions can be found in [2, 21, 15], for simplicity, Example 1 will consider such arguments as being identical to Dialectical Cl-Arg arguments devoided of suppositions.

Example 1. *Consider Figure 1. Let A_1, $B_1 \in \mathcal{S}$ be the dialectical arguments introduced in Table 1, and let $Z_1 = (\{a \supset \neg b\}, \{a\}, \neg b)$ be a dialectical argument that defeats B_1 with respect to $\mathcal{S}$, i.e., $Z_1 \Rightarrow_{\mathcal{S}} B_1$. Notice that such defeat occurs only due to the presence of the formula $a \in Prem(A_1)$. The supposition of the formula a by the dialectical argument Z_1 (i.e., $Supp(Z_1) \subseteq Prem(\mathcal{S} \cup \{B_1\})$ allows concluding $\neg b$, hence defeating argument B_1. However, $Z_0 = (\{a \supset \neg b\}, a \supset \neg b)$, the Cl-Arg argument that has the same premises as Z_1, is not capable of moving the same defeat to B_1. Indeed, the absence of the formula a among the premises prevents Z_0 from classically entailing the conclusion $\neg b$, hence precluding the defeat of argument B_1. This example shows how, by supposing formulae (from single arguments or sets), additional attacks and defeats may arise for Dialectical Cl-Arg arguments in comparison with Cl-Arg arguments.*

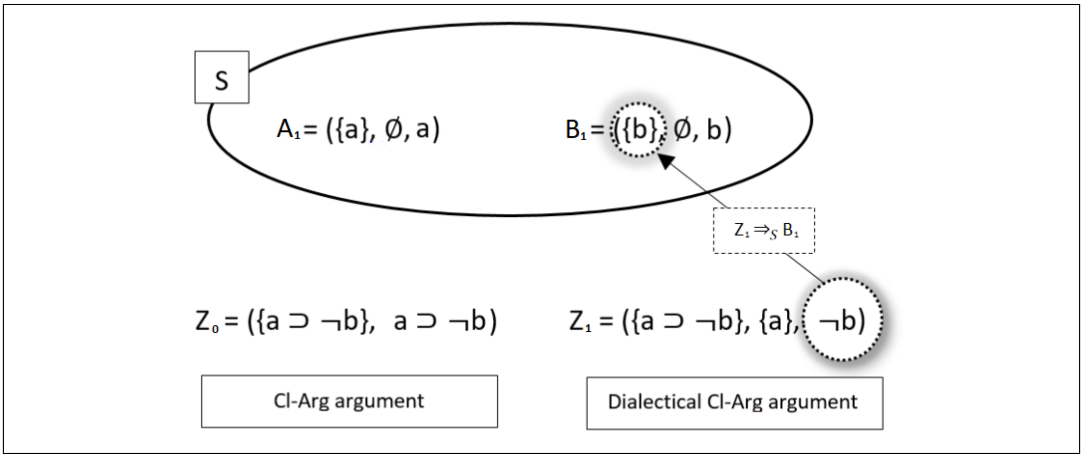


Figure 1: An example of differences between Cl-Arg and Dialectical Cl-Arg.

The conclusions of an extension in Dialectical Cl-Arg may derive from conditional arguments that only suppose the truth of the premises without any commitment. As such, we should account for a more restrictive definition of conclusions. That is to say, once the extensions are defined, we detach only the conclusions of *unconditional* arguments all of whose assumptions are premises presumed true.

Definition 5. [Conclusions of an Extension in Dialectical Cl-Arg] *Let E be an extension of a dialectical AF. Then $C(E) = \{\phi \mid (\Delta, \emptyset, \phi) \in E\}$.*

Dialectical AFs enjoy some specific properties, as explained in [15]. Here we are going to outline five of them (P1, P2, P3, P4, P4′), which will be used later in the next sections.

Proposition 1. *Given a dialectical $AF = \langle AR,\ defeats\rangle$:*

($P1$) *$\forall X \in AR$: $\alpha \in Prem(X)$ implies that $(\{\alpha\}, \emptyset, \alpha) \in AR$ (where $(\{\alpha\}, \emptyset, \alpha)$ is denoted as the* 'elementary argument' *of X defined by α);*

($P2$) *$\forall X \in AR$: if $X' \in [X]$, that is to say, if X' is the* logically equivalent argument *of X (i.e., the only difference between X and X' is the different distribution of premises and supposition), then $X' \in AR$;*

($P3$) *If $(\Delta, \emptyset, \alpha) \in AR$ and $(\Gamma, \emptyset, \overline{\alpha}) \in AR$, then either $(\Delta, \emptyset, \curlywedge) \in AR$ or $(\Gamma, \emptyset, \curlywedge) \in AR$ or $(\Delta \cup \Gamma, \emptyset, \curlywedge) \in AR$;*

($P4$) *If $(\Gamma, \emptyset, \alpha) \in AR$, $\Delta \subseteq \Gamma$, $\Delta \neq \emptyset$ and $\Delta \parallel \Gamma \setminus \Delta \cup \{\alpha\}$, then either $(\Delta, \emptyset, \curlywedge) \in AR$ or $(\Gamma \setminus \Delta, \emptyset, \alpha) \in AR$;*

($P4'$) *If $(\Gamma, \emptyset, \alpha) \in AR$, $\Delta \subseteq \Gamma$, $\Delta \neq \emptyset$ and $\Delta \parallel \Gamma \setminus \Delta \cup \{\alpha\}$, then $(\Delta, \emptyset, \curlywedge) \in AR$.*

We can now refer to ⟨AR, *defeats*⟩ as a *partially instantiated dialectical AF* (*pdAF*) if AR corresponds to any subset of the dialectical arguments defined by a base $\mathcal{B}$ such that AR satisfies $P1$, $P2$, $P3$ and $P4$.

A *non-redundant pdAF* is, instead, a pdAF such that AR satisfies $P1$, $P2$, $P3$, $P4'$ and there are no redundantly contaminated arguments[3].

[3] A redundantly contaminated argument is an argument that employs redundant assumptions, that is to say, a subset of the assumptions is unnecessary for drawing the argument conclusion. This may occur due to the fact that Dialectical Cl-Arg drops subset minimality checks. To avoid violation of the non-contamination postulates, the adopted preference relation has to be 'redundance-coherent'. Indeed, this is the case of the Elitist preference of Definition 3.

2.1.1 Rationality Postulates for Dialectical Cl-Arg

The rationality postulates are specific properties whose satisfaction ensures that any concrete instantiations of an argumentation framework fulfil some rational criteria. Dialectical Cl-arg satisfies the rationality postulates and does so by requiring that the AF enjoys $P1$-$P4$. This would impose minimally restrictive assumptions[4] as to the arguments that agents should be able to construct, thus providing a rational account of arguments more suited for the limited availability of resources that characterises real-world agents. A detailed report of the postulates, along with lemmas, theorems and respective proofs of their validity, is given in [15].

Theorem 1. **[Sub-argument Closure]** *Let E be a complete extension of a dialectical $AF = \langle AR, \mathit{defeats}\rangle$ such that AR satisfies P1. Let $X \in E$. Then if $\alpha \in \mathit{Prem}(X)$ then $(\{\alpha\}, \emptyset, \alpha) \in E$. That is to say, all the elementary arguments associated with $\mathit{Prem}(X)$ are in E.*

Theorem 2. **[Direct Consistency]** *Let E be an admissible extension of a dialectical $AF = \langle AR, \mathit{defeats}\rangle$. If AR satisfies P1, P2 and P3, then $\forall \alpha, \beta \in C(E)$, $\alpha \neq \curlywedge$ and $\beta \neq \overline{\alpha}$. That is to say, no conflicting or unconditional falsum arguments are in E.*

Theorem 3. **[Premise Consistency]** *Let $\langle AR, \mathit{defeats}\rangle$ be a dialectical AF such that AR satisfies P2. If for some $\Delta \subseteq \mathit{Prem}(E)$: $(\Delta, \emptyset, \curlywedge) \in AR$, then E cannot be an admissible extension of $\langle AR, \mathit{defeats}\rangle$.*

Closure under Strict Rules for Dialectical Cl-Arg slightly differs from its standard version. That is caused by the limited availability of resources that characterises real-world agents. Indeed, although it may be the case that $C(E) \vdash_c \alpha$, it may not be that there exists an $X \in E$ such that X concludes α, given that agents are not logically omniscient and do not construct all arguments from a base. Hence, the following version of the postulate:

Theorem 4. **[Closure under Strict Rules]** *Let E be a complete extension of a dialectical $AF = \langle AR, \mathit{defeats}\rangle$, where AR satisfies P1. Let $E' \subseteq E$ and $C(E') \vdash_c \alpha$. If there exists an $X = (\Delta, \emptyset, \alpha) \in AR$ such that $\Delta = \mathit{Prem}(E')$, then $X \in E$.*

Non-contamination postulates provide means for eschewing different kinds of contaminations that may negatively affect the dialectical AFs. In particular, the satisfaction of *Non-Interference* ensures that no syntactically disjoint bases $\mathcal{B}$ (i.e., bases that do not share predicate or function symbols) influence each other's argumentation defined inferences. On the other hand, *Crash Resistance* guarantees

[4] Especially the satisfaction of $P1$-$P3$.

that no set of formulae yields the same outcome when merged with a syntactically disjoint set of formulae.

Theorem 5. [Non-Interference] *Non-interference is satisfied by (non-redundant) pdAFs.*

Theorem 6. [Crash Resistance] *Crash Resistance is satisfied if there does not exist a contaminating base $\mathcal{B}$ for pdAFs and non-redundant pdAFs.*

2.1.2 Dung's Fundamental Lemma and Monotonicity of the Characteristic Function for Dialectical Cl-Arg

Among the most important key results of Dung's seminal paper [17] are the *fundamental lemma* and the *monotonicity of the AF's characteristic function $\mathcal{F}_{AF}$* (that yields the constructive definition of the grounded extension via its iterations). However, unlike Dung's standard AFs, these properties cannot be straightforwardly shown, since when determining the acceptability of X w.r.t. E, the defeats on X are not independent of the set E under consideration. For dialectical AFs, the defeats on X w.r.t. E may be a subset of the defeats on X w.r.t. $E' \supset E$ (due to the additional premises committed to in E'). To avoid this issue, the authors of [15] have devised specific 'epistemically maximal' sets of arguments by means of whose it is possible to show the desired properties.

Definition 6. [Epistemically maximal sets] *Let $\langle AR, defeats\rangle$ be a dialectical AF. Then $E \subseteq AR$ is* epistemically maximal (em) *iff:*

$$\textit{If } X = (\Delta, \Gamma, \alpha) \in E,\ \Gamma' \subseteq (\Gamma \cap \textit{Prem}(E)),\ \textit{then } X' = (\Delta \cup \Gamma', \Gamma \setminus \Gamma', \alpha) \in E \qquad (\bullet)$$

The function $Cl_{em} : 2^{AR} \to 2^{AR}$ maps any E to its epistemically maximal set. As such, $Cl_{em}(E)$ denotes the smallest superset of E that is closed under condition $(\bullet)$.

Notice that adding all arguments up to some i to a set E, and then closing, yields the same result as adding each argument one by one and closing prior to each subsequent addition [15]. It is now possible to prove a variant of the fundamental lemma that involves *em* sets:

Lemma 1. [Fundamental Lemma for Dialectical Cl-Arg][15] *Let X, X' be acceptable w.r.t. an admissible extension E of a dialectical $AF = \langle AR,\ defeats\rangle$. Then:*

(1) *$Cl_{em}(E \cup \{X\})$ is admissible, and*

(2) *X' is acceptable w.r.t. $Cl_{em}(E \cup \{X\})$*

Lemma 1 entails:

Proposition 2. *Every admissible extension of a dialectical AF is a subset of a preferred extension.*

Proposition 2 guarantees that it suffices to show that an argument X is in an admissible extension, in order to prove that X is credulously justified under the preferred semantics (exactly as Dung's standard AFs).

Finally, by employing a variant of the framework characteristic function, i.e., $\mathcal{F}_p$, whose domain is composed of sets E that are *em* admissible and that returns $Cl_{em}(\mathcal{F}(E))$, we can also show the constructive definition of the grounded extension. Indeed, starting with the empty set and iteratively applying $\mathcal{F}_p$, the monotonically increasing sequence approximates, and in the case of a *finitary* dialectical AF, it constructs, the least fixed point of $\mathcal{F}_p$, i.e., the grounded extension:

Proposition 3. [15] *Let $\langle AR,\ defeats \rangle$ be a dialectical AF, and $F^0 = \emptyset$, $F^{i+1} = \mathcal{F}_p(F^i)$. Let E be the grounded extension of $\langle AR,\ defeats \rangle$. Then:*

1. $E \subseteq \bigcup_{i=0}^{\infty}(F^i)$.

2. *If $\langle AR,\ defeats \rangle$ is* finitary*, i.e., $\forall X \in AR$, the set $\{Y \mid defeats(Y, X)\}$ is finite, then $E = \bigcup_{i=0}^{\infty}(F^i)$.*

In the remainder of the paper, we are going to see how harnessing the properties and formalism thus far introduced will shape the dialectical characterisation of standard argument games.

2.2 Standard Argument Games

Before moving forward, let us now review the fundamental notions of the standard argument games as described in [25]. Notice that these definitions have been modified to accommodate dialectical AFs (which is a fair straightforward adaptation). However, recall that the main contributions of this paper concern the development of argument games for Dialectical Cl-Arg that involves *dialectical defeats* (Definition 4): this entails a non-trivial modification of the standard games.

In a nutshell, an argument game is played by two players, PRO (for *proponent*) and OPP (for *opponent*), each of which is referred to as the other's 'counterpart'. PRO starts the game by moving an initial argument X that it wants to test. After that, both players take turns in moving arguments against their counterpart's moves. This generates disputes:

Definition 7. [Dispute] *A sequence of moves in which each player moves against its counterpart's argument is referred to as a* dispute*. Formally,* $d = X$—Y—Z— $\cdots$ *is a dispute, and* X—Y *denotes a player moving argument* Y *against an argument* X *played by its counterpart (similarly,* Y—Z*). A* sub-dispute d' *of a dispute* d *is any sub-sequence of* d *that starts with the same initial argument. For example, if* $d = X$—Y—Z*, then* $d' = X$—Y *would be a sub-dispute of* d*.*

Notice that, to avoid ambiguity, each argument of a dispute will be labelled with either P or O (that stands for either one of the two players, PRO or OPP). Hence, $d =$ (P)X—(O)Y—(P)Z is a dispute where PRO moves the argument X, followed by Y played by OPP and countered by another move from PRO, Z.

We can now introduce the notion of the (unique) dispute tree, which represents the 'playing field' of the standard argument games. In other words, the dispute tree is the data structure that contains all the potential moves (and sequences of moves) available to the players.

Definition 8. [Dispute Tree] *Let* $AF = \langle AR, defeats\rangle$ *be a finite dialectical argumentation framework, and let* $\mathrm{A} \in AR$*. The* dispute tree *induced by* A *in the AF is the (upside-down) tree* $\mathcal{T}$ *of arguments, such that* $\mathcal{T}$*'s root node is* A*, every branch of the tree (from root to leaf) is a different dispute, and* $\forall X, Y \in AR$*:* X *is a child of* Y *in* $\mathcal{T}$ *iff* $defeats(X, Y)$*.*

From here on, we are going to write PRO($*$) and OPP($*$) to denote the sets of all PRO and OPP arguments in $*$, where $*$ can be replaced with d, $\mathcal{T}$ or any other tree that will be introduced in the remainder of the paper. Also, LAST(d) will identify the last argument played in a dispute d.

An argument game is said to be won by the proponent only if it has a winning strategy. That is to say, only if it can successfully defend the argument it wants to test (i.e., the root of $\mathcal{T}$) against any possible arguments moved by the opponent. PRO loses otherwise. In other words, this may be interpreted as a formalisation of the simple principle already emphasised by Dung: *"The one who has the last word laughs best"* [17].

Definition 9. [Winning Strategy] *Let* $\mathcal{T}$ *be the dispute tree induced by* A *in a finite dialectical* $AF = \langle AR, defeats\rangle$*. Let also* d *be a dispute in* $\mathcal{T}$*. Then, a* winning strategy $\mathcal{T}'$ *for* A *is the dispute tree* $\mathcal{T}$ *pruned in a way such that:*

(9.1) *The set* $\mathcal{T}'_D$ *of disputes in* $\mathcal{T}'$ *is a non-empty finite set such that each dispute* $d \in \mathcal{T}'_D$ *is finite and is won by PRO (i.e.,* $LAST(d) \in PRO(\mathcal{T})$*);*

(9.2) $\forall d \in \mathcal{T}'_D$, *$\forall d'$ such that d' is some sub-dispute of d, LAST$(d') = X$ and $X \in$ PRO$(\mathcal{T})$, then $\forall Y \in$ OPP$(\mathcal{T})$ such that $Y \Rightarrow X$, there is a $d'' \in \mathcal{T}'_D$ such that d'— Y is a sub-dispute of d''.*

Informally, the previous definition states that a winning strategy is the dispute tree $\mathcal{T}$ pruned in a way such that (9.1) $\mathcal{T}'_D$ is a non-empty finite set, its disputes are finite, end with a PRO argument and (9.2) are such that OPP has moved exhaustively (i.e., all the moves that OPP could have played, had been played) and also PRO has countered every defeating argument moved by OPP.

3 Developing Dialectical Argument Games

In the following sections, we are going to develop argument games for Dialectical Cl-Arg that accommodate the dialectical defeats and semantics introduced in Definition 4. The resulting proof theory will present some specific features that will distinguish it from the standard argument games, although the general structure remains similar. Intuitively, winning a dialectical game for an argument A means having a 'dialectical procedure' (depending on the semantics that the proof theory is meant to capture) for defending the information contained in A, hence showing the admissibility of the encoded data.

The main difference between a dispute tree $\mathcal{T}$ and a *dialectical dispute tree* $\mathcal{D}$ can be identified with the additional reference to a subset $\mathcal{S} \subseteq$ PRO$(\mathcal{T})$. That is to say, $\mathcal{S}$ represents a candidate admissible set of PRO arguments such that PRO commits to their premises. Recall once again that, when challenging the acceptability of an argument with respect to a set $\mathcal{S}$, the defeating argument can suppose premises from all the arguments in $\mathcal{S}$. Whereas, the argument that defends $\mathcal{S}$ can only suppose the premises of the defeating argument. Another important difference between standard and dialectical games is that the latter handles *partially instantiated dialectical AFs* (*pdAFs*)[5]. As a consequence, each dialectical game enjoys specific properties that encapsulate the dialectical uses of arguments by real-world resource-bounded agents, thus succeeding in better approximating a process capable of bridging formal (proof-theoretical) and informal (real-world exchange of arguments) single-agent reasoning.

We can now formally introduce the (unique) dialectical dispute tree induced by A wrt a set $\mathcal{S}$:

Definition 10. **[Dialectical Dispute Tree]** *Let $\mathcal{T}$ be the dispute tree induced by* A *in a finite pdAF = $\langle AR, defeats\rangle$. Let also $\mathcal{S} \subseteq$ PRO$(\mathcal{T})$. Then, the* dialectical

[5] Refer to Proposition 1.

dispute tree $\mathcal{D}$ *induced by* A *with respect to* $\mathcal{S}$ *is the dispute tree* $\mathcal{T}$ *pruned in a way such that* $\forall X, Y \in AR$*:* X *is a child of* Y *in* $\mathcal{D}$ *iff* $defeats(X, Y)$ *and:*

1. *If* $X \in PRO(\mathcal{D})$ *and* $Y \in OPP(\mathcal{D})$*, then* $X \Rightarrow_{\{Y\}} Y$*, i.e.* X *defeats* Y *and* $Supp(X) \subseteq Prem(Y)$*;*

2. *If* $X \in OPP(\mathcal{D})$ *and* $Y \in PRO(\mathcal{D})$*, then* $X \Rightarrow_{\mathcal{S}} Y$*, i.e.* X *defeats* Y *with respect to* $\mathcal{S}$ *and* $Supp(X) \subseteq Prem(\mathcal{S} \cup \{Y\})$*.*

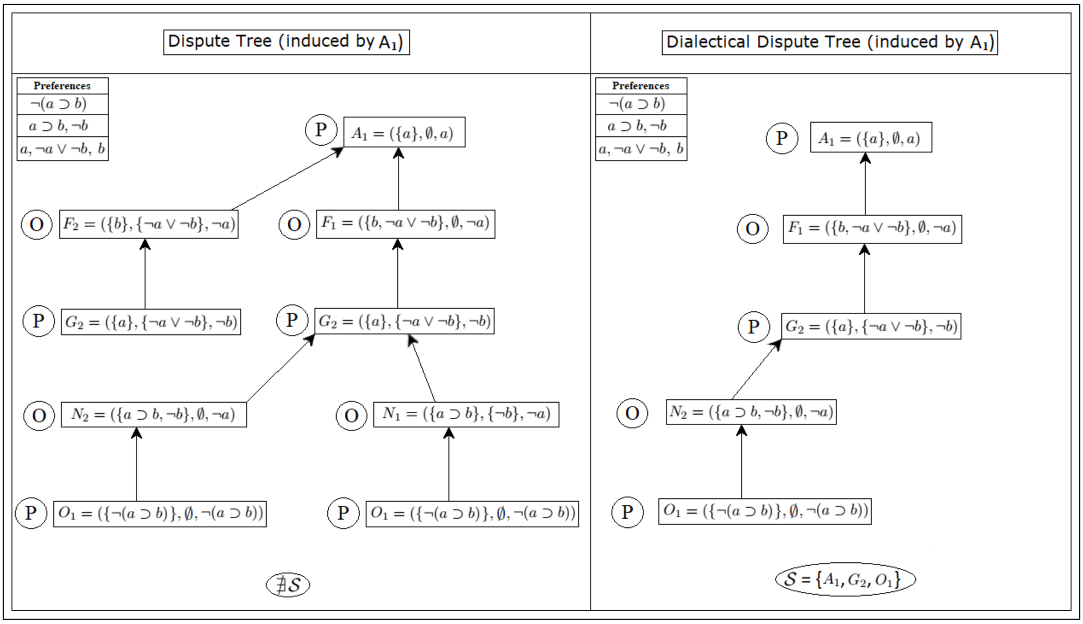

Dispute Tree (induced by A₁)
Dialectical Dispute Tree (induced by A₁)
Preferences
¬(a ⊃ b)
a ⊃ b, ¬b
a, ¬a ∨ ¬b, b
P
O
A₁ = ({a}, ∅, a)
F₂ = ({b}, {¬a ∨ ¬b}, ¬a)
F₁ = ({b, ¬a ∨ ¬b}, ∅, ¬a)
G₂ = ({a}, {¬a ∨ ¬b}, ¬b)
N₂ = ({a ⊃ b, ¬b}, ∅, ¬a)
N₁ = ({a ⊃ b}, {¬b}, ¬a)
O₁ = ({¬(a ⊃ b)}, ∅, ¬(a ⊃ b))
∄S
S = {A₁, G₂, O₁}

Figure 2: The (incomplete) dispute tree $\mathcal{T}$ (on the left) induced by A_1 in a finite pdAF = ⟨AR, *defeats*⟩ and the corresponding (incomplete) dialectical dispute tree $\mathcal{D}$ (on the right) induced by A_1 wrt $\mathcal{S} = \{A_1, G_2, O_1\}$ in the same pdAF = ⟨AR, *defeats*⟩.

The ‘playing field’ of the dialectical argument games (i.e., the data structure on the basis of which the games are played) is still depicted by the dispute tree $\mathcal{T}$. Indeed, the relationship existing between the dispute tree $\mathcal{T}$ induced by A in a finite pdAF and the dialectical dispute tree $\mathcal{D}$ induced by A wrt $\mathcal{S}$ is such that $\mathcal{D}$ is ‘contained’ in $\mathcal{T}$ (since $\mathcal{D}$ is a pruned version of $\mathcal{T}$), as shown in the following example.

Example 2. *Figure 2 presents the (incomplete) dispute tree* $\mathcal{T}$ *induced by* A_1 *in a finite pdAF* $= \langle AR, defeats\rangle$ *and the corresponding (incomplete) dialectical dispute* *tree $\mathcal{D}$ induced by A_1 wrt $\mathcal{S} = \{A_1, G_2, O_1\}$ in the same pdAF. Both trees are incomplete since the purpose of the example is just to show the relationship existing between them. For the same reason, we also avoid listing all the arguments of the pdAF.*

Observe that, unlike $\mathcal{T}$, where no set is taken into consideration, the defeats in $\mathcal{D}$ are parametrized to the set $\mathcal{S}$. This implies that, when defeating PRO's arguments, OPP can only suppose the premises of the arguments in the set $\mathcal{S}$ (besides the premises of the targeted argument). No such restrictions exist for $\mathcal{T}$. Notice that, even if we keep extending both trees, dispute $d = (P)A_1 — (O)F_2 — (P)G_2$ will never be part of $\mathcal{D}$. This is because, according to Definition 10 (which also emphasizes how dialectical defeats work), PRO can move G_2 only if $Supp(G_2) \subseteq Prem(F_2)$. However, this is never going to be the case since the formula $\neg a \vee \neg b \notin Prem(F_2)$. Therefore, even if the two trees were identical in every other branch, the absence of dispute d will still make $\mathcal{D}$ 'contained' in $\mathcal{T}$.

Dialectical argument games share with the standard argument games the notion of a winning strategy: in order to win the game for an argument A, PRO must have a winning strategy for it. It will lose otherwise. However, the two definitions slightly differ since a dialectical winning strategy has to take into account the set $\mathcal{S}$ targeted by the dialectical defeats:

Definition 11. [Dialectical Winning strategy] *Let $\mathcal{D}$ be the dialectical dispute tree induced by* A *wrt $\mathcal{S}$ in a finite pdAF = $\langle AR, defeats\rangle$ and let d be a dispute in $\mathcal{D}$. Then, a* dialectical winning strategy *$\mathcal{W}$ for* A *corresponds to the dialectical dispute tree $\mathcal{D}$ pruned in a way such that:*

(11.1) *The set $\mathcal{W}_D$ of disputes in $\mathcal{D}$ is a non-empty finite set such that each dispute $d \in \mathcal{W}_D$ is finite and is won by PRO (i.e., $LAST(d) \in PRO(\mathcal{D})$);*

(11.2) *$\forall d \in \mathcal{W}_D$, $\forall d'$ such that d' is some sub-dispute of d, $LAST(d') = X$ and $X \in PRO(\mathcal{D})$, then $\forall Y \in OPP(\mathcal{D})$ such that $Y \Rightarrow_{\mathcal{S}} X$, there is a $d'' \in \mathcal{W}_D$ such that $d' — Y$ is a sub-dispute of d''.*

Similarly to Definition 9, the previous definition states that a dialectical winning strategy corresponds to the dialectical dispute tree $\mathcal{D}$ pruned in a way such that (11.1) $\mathcal{W}_D$ is a non-empty finite set, its disputes are finite, end with a PRO argument and are such that (11.2) OPP has moved exhaustively and also PRO has countered each defeating argument moved by OPP. The difference is in the dialectical defeats: the nodes are no more connected by means of the defeats relations among arguments, but through dialectical defeats among arguments that target the set $\mathcal{S}$.

We now have all the elements needed to formally introduce the protocol of the dialectical admissible/preferred game. Similar to a list of instructions, this protocol determines the legal moves that can be performed by the players. The game unfolds as a result of the legal arguments played and terminates when there are no more valid moves available. When this happens, the status of the root of the tree is evaluated. The presence of a winning strategy for such an argument assigns the victory to PRO. Strictly speaking, OPP never wins: its purpose is to counter each argument moved by the proponent in order to assist it in testing the admissibility of the root argument (indeed, argument games are formalisations of single-agent reasoning processes). Nevertheless, OPP can still prevent PRO's victory by invalidating its winning strategy.

3.1 Progressively Constructing Dialectical Dispute Trees

When we play a Φ-dialectical game we are increasingly building, starting from the root A and following the legal moves licensed by the protocol Φ, a dialectical dispute tree denoted as Φ-$\mathcal{D}^n$. Each node of such a tree corresponds to an argument progressively played by either PRO or OPP that is labelled with a positive integer i (with $1 \leq i \leq n$). These additional labels allow identifying the order in which the arguments have been played, hence, also determining the current stage (i.e., the nth-stage) of the Φ-dialectical game. Recall that the dispute tree $\mathcal{T}$ induced by A represents the playing field of the games, and every Φ-dialectical game for A is contained within its data structure (i.e., Φ-$\mathcal{D}^n$ is a 'pruned-version' of $\mathcal{T}$). Moreover, being a dialectical dispute tree, even Φ-$\mathcal{D}^n$ is constructed wrt a set $\mathcal{S} \subseteq \text{PRO}(\mathcal{T})$, however, such $\mathcal{S}$ can gradually increase with each new move made by PRO during the game. Indeed, $\mathcal{S}$ is composed of the same arguments moved by PRO in Φ-$\mathcal{W}^n$ (i.e., a dialectical winning strategy for A of Φ-$\mathcal{D}^n$), which can be extended while the game proceeds[6]. As it will be shown, observe also that $\mathcal{S}$ is still a different set than PRO(Φ-$\mathcal{W}^n$), meaning that it will modify its members according to the changes in PRO(Φ-$\mathcal{W}^n$), but it will never be empty even if there is no winning strategy Φ-$\mathcal{W}^n$.

In order to formally describe a Φ-dialectical game, we first need to define a *partial dialectical dispute tree* $\mathcal{D}^n$ which will stand as a potential 'game template' deprived of a protocol:

Definition 12. [Partial dialectical dispute tree] *A partial dialectical dispute tree*

[6] Although the set $\mathcal{S}$ can increase the number of its members while the game goes on, it can never exceed the size of PRO($\mathcal{T}$). Indeed, keep in mind that every Φ-dialectical game for A is contained in the dispute tree $\mathcal{T}$ induced by A (since $\mathcal{T}$ corresponds to the playing field of the game).

$\mathcal{D}^n$ induced by A *wrt $\mathcal{S} \subseteq PRO(\mathcal{T})$ (with $\mathcal{S} \neq \emptyset$) in a finite pdAF $= \langle AR, defeats\rangle$ is the (upside-down) tree that starts from the argument* A*, and it is progressively built up to the nth-move by one of the two players, such that each node of the tree is labelled with a positive integer i (for $1 \leq i \leq n$). Moreover, every branch of the tree (from root to leaf) constitutes a different dispute. Also $\forall X, Y \in AR$: X is a child of Y in $\mathcal{D}^n$ iff defeats(X, Y) and:*

1. *If $X \in PRO(\mathcal{D}^n)$ and $Y \in OPP(\mathcal{D}^n)$, then $X \Rightarrow_{\{Y\}} Y$, i.e. X defeats Y and Supp$(X) \subseteq$ Prem(Y);*

2. *If $X \in OPP(\mathcal{D}^n)$ and $Y \in PRO(\mathcal{D}^n)$, then $X \Rightarrow_{\mathcal{S}} Y$, i.e. X defeats Y with respect to a set $\mathcal{S}$ and Supp$(X) \subseteq$ Prem$(\mathcal{S} \cup \{Y\})$.*

Finally, $\mathcal{W}^n$ will denote a dialectical winning strategy for A *of $\mathcal{D}^n$ as per Definition 11 (substituting $\mathcal{D}$ with $\mathcal{D}^n$).*

Every stage of a Φ-dialectical game can then be identified with a specific dialectical dispute tree Φ-$\mathcal{D}^n$, i.e., a *partial dialectical dispute tree* of Definition 12 where each of its nodes also fulfils the legal move requirements according to the protocol Φ. Consider that every such stage of the game is not unique: playing the same game multiple times does not necessarily hold the same Φ-$\mathcal{D}^n$ at identical stages n. They can indeed differ depending on the way in which the legal arguments have been deployed by the players. As we are going to see, this notion is essential for a proper account of the dialectical defeats in the game protocol[7].

3.2 Disqualified Defeats

It is interesting to notice that, during a Φ-dialectical game, a dialectical defeat that occurred in an early stage of the game might not take place in a more advanced phase of the same game. This can be caused by an update of the current $\mathcal{S}$, the set parametrized by OPP for performing dialectical defeats. We denote this anomaly as 'disqualified defeats'.

Definition 13. [Disqualified dialectical defeats] *Let Φ-$\mathcal{D}^n$ be the dialectical dispute tree of a Φ-dialectical game built up to the nth-move where X and Y denote*

[7] Observe that it is possible for one (or more, depending on the protocol) dialectical winning strategy Φ-$\mathcal{W}^n$ for A of Φ-$\mathcal{D}^n$ to exist, although there is no dialectical winning strategy $\mathcal{W}$ for A of $\mathcal{D}$. This can happen, for example, when $\mathcal{D}$ is composed only by infinite disputes (recall that we need finite disputes to have winning strategies, as stated by Definition 11.1), whilst Φ-$\mathcal{D}^n$ is composed by finite disputes, due to the restrictions imposed by the protocol Φ. In this situation, it is possible to identify in Φ-$\mathcal{D}^n$ a winning strategy Φ-$\mathcal{W}^n$. Such an example is illustrated in Figure 3(b).

arguments played respectively by OPP and PRO in Φ-$\mathcal{D}^n$. Let also $X \Rightarrow_{\mathcal{S}} Y$ by supposing $\alpha \in Prem(\mathcal{S})$. If, after the game goes on, we will reach a stage Φ-$\mathcal{D}^{n+k}$ (for $k > 0$) where $\alpha \notin Prem(\mathcal{S})$, then the defeat moved by X against Y will be invalidated and will be denoted as 'disqualified'. As such, X and all the arguments following it in the same dispute will be (temporarily) pruned from the tree.

Consider indeed that the status of disqualified defeats might be temporary and be updated again in a further stage of the game (when these defeats will become valid once more). Definition 13 entails the following proposition:

Proposition 4. *Let Φ-$\mathcal{D}^n$ be the dialectical dispute tree of a Φ-dialectical game built up to the nth-move:*

(*I*) *If the nth-move is an argument X played by OPP, then moving X cannot disqualify the dialectical defeat that X performs against a PRO argument.*

(*II*) *The presence of OPP arguments whose defeats have been disqualified will not affect the dialectical winning strategy.*

Proof.

(*I*) Since X is the last argument (legally) played in Φ-$\mathcal{D}^n$, it trivially does not comply with Definition 13.

(*II*) Even if the dialectical defeats moved by OPP arguments have been disqualified (hence are no more a threat for PRO), the requirements of the dialectical winning strategy have not changed. That is to say, every dispute of Φ-$\mathcal{W}^n$ must terminate with a PRO argument (Definition (11.1)).

□

Notice that every dialectical game protocol Φ takes into account disqualified defeats, which are then also contemplated by the dialectical dispute tree Φ-$\mathcal{D}^n$ (and dialectical winning strategy Φ-$\mathcal{W}^n$).

4 Dialectical Admissible/Preferred Games

We can now formally introduce the protocol for the dialectical admissible/preferred game. As already stated, during each dialectical argument game, the players have to comply with a protocol Φ that identifies the legal moves allowed.

Definition 14. [Dialectical Admissible Game legal moves] *Let $\mathcal{D}^n$ and $\mathcal{W}^n$ be defined as in Definition 12, let d be a dispute of $\mathcal{D}^n$ and d' be a sub-dispute of d. Let also $(PL_n)X$ (for $n > 0$) denote the argument X played by either one of the two players (P or O) as the (last) nth-move. Then Φ_P identifies legal moves in the following way:*

(14.0) *PRO moves the first argument.*

(14.1) *If $(PL_n)X$ and $n = 2k$ (for $k > 0$), then the next move n+1, say Y, is by PRO and it is such that:*

(*a*) *$Y \Rightarrow_{\{Z\}} Z$, where $Z \in$ OPP($\mathcal{D}^n$);*

(*b*) *There exists a $\mathcal{W}^{n+1}$ for A of $\mathcal{D}^{n+1}$.*

(14.2) *If $(PL_n)X$ and $n = 2k + 1$ (for $k \geq 0$), then the next move n+1, say Y, is by OPP and it is such that:*

(*a*) *$Y \Rightarrow_{\mathcal{S}} Z$, where $Z \in \mathcal{S}$ and $\mathcal{S}$:= PRO($\mathcal{W}^n$)[8];*

(*b*) *If $d = d'$—Z, then $Y \notin$ OPP(d');*

(*c*) *For each $d = d'$—J—$\cdots$, where $J \in$ OPP($\mathcal{D}^n$) and its defeat has been disqualified, then LAST(d) = LAST(d') until next OPP's turn.*

A Φ_P-dialectical game is said to be terminated when, during its turn, the corresponding player runs out of the legal moves identified by (14.1(*a*-*b*)) or (14.2(*a*-*b*)) of the protocol Φ_P. PRO wins only if it has a winning strategy once the game terminates. It loses otherwise.

The previous protocol can be informally summarised as follows. PRO starts the game by playing the first argument [(14.0)] and, after that, OPP will make its move. Then, the two players alternate in playing only one argument at a time to reply to one of their counterpart's arguments. Observe that when $\mathcal{S}$ is initialized in the game and, subsequently, every time its arguments are updated by the changes in PRO($\mathcal{W}^n$) [(14.2(a))], it is always the beginning of OPP's turn. This means that the condition for which $\mathcal{S} \neq \emptyset$ is continuously respected[9].

[8] The symbol ':=' denotes a variable initialization rather than an equivalence relation. That is to say, at the beginning of each OPP's turn, the content of $\mathcal{S}$ is initialized to the current PRO($\mathcal{W}^n$), i.e, the arguments member of $\mathcal{S}$ are the same as PRO($\mathcal{W}^n$). This operation overwrites the previous contents of $\mathcal{S}$.

[9] That is because a situation in which $\mathcal{S}$ = PRO($\mathcal{W}^n$) = $\emptyset$ never occurs at the beginning of OPP's turn.

Notice that the established protocol allows *backtracking* to other arguments. That is to say, when PRO moves it can either target the last argument played by OPP or another argument moved by OPP in the dialectical dispute tree generated thus far (i.e., an argument member of the set OPP($\mathcal{D}^n$)) [(14.1(a))]. Similarly, when OPP moves it can either target the last argument played by PRO or another argument moved by PRO in the current dialectical winning strategy (i.e., an argument member of the set PRO($\mathcal{W}^n$)) [(14.2(a))]. The *relevance conditions* [(14.1(b)) for PRO; (14.2(a)) for OPP] ensure that: after PRO has made its move, there will be a winning strategy $\mathcal{W}^{n+1}$, hence providing the victory to PRO; after OPP has moved, instead, the previous winning strategy will cease to exist, thus preventing PRO from winning. That is to say, PRO will be forced to generate a dialectical winning strategy during each of its turns, while OPP will have to invalidate such a winning strategy during every one of its turns. Backtracking and relevance conditions are strictly connected. Although it is possible for a player to defeat an argument other than the one previously posited by its counterpart, such a move needs to comply with the protocol relevance conditions. This combination ensures that both participants exhaustively account for every option available, otherwise restricted around the last argument played (which may be unassailable, hence preventing further move against it). Indeed, given the goal of changing the winning status at the end of their respective turns, PRO and OPP may choose which argument to defeat, thus leaving for a later moment the other (if still available) alternatives.

The restriction (14.2(b)) on the moves played by OPP is necessary (as also shown in the standard games of [25, 34] and [9]). Indeed, allowing OPP to repeat its arguments, since OPP is required to move exhaustively, could imply the generation of infinite disputes. To see why let us suppose that $(\mathrm{PL}_n)X$ (for $n > 1$) identifies an argument X played by either one of the two players (denoted as P or O) as its nth move in a Φ-dialectical game. Then, there could be an infinite dispute d like the following:

$$d = (P_1)A\text{—}\cdots\text{—}(O_n)Y\text{—}(P_{n+1})Z\text{—}(O_{n+2})Y\text{—}(P_{n+3})Z\text{—}(O_{n+4})Y\text{—}\cdots$$

Intuitively, since Z is capable of defending itself by defeating Y, there is no need to further extend the dispute by repeating the same arguments: this is because Z has already shown its acceptability wrt PRO($\mathcal{W}^{n+1}$). Therefore, the only way for avoiding infinite disputes (and infinite dialectical admissible/preferred games) is to prevent OPP from repeating its arguments in the same disputes.

Finally, (14.2(c)) ensures that the disqualified defeats (Definition 13) are taken into account throughout the game. That is to say, whenever a dialectical defeat moved by an argument J is disqualified, the protocol guarantees the pruning of J and all the arguments that follow in the same dispute, until the next turn of OPP, when a new check for disqualified defeats will occur.

Remark 1. *Similarly to the standard argument games presented in [25], the protocol of the dialectical admissible games is identical to the protocol of the dialectical credulous preferred games. Indeed, it suffices to show the membership of an argument* A *in an admissible extension to show also that* A *is credulously justified under the preferred semantics as well. That is because every admissible extension of a dialectical AF is a subset of a preferred extension. This is a consequence of the Fundamental Lemma (Lemma 1) and its entailed property (Proposition 2).*

4.1 Soundness and Completeness

As it has been defined, the admissible/preferred game satisfies the properties of soundness and completeness. This proves the equivalence existing between the victory of the Φ_P-dialectical game for an argument A and the membership of the same A to an admissible/preferred extension of the corresponding finite pdAF.

Theorem 7. *Let Φ_P-$\mathcal{D}^n$ identifies a terminated Φ_P-dialectical game for* A*.*
Then, there exists a dialectical winning strategy Φ_P-$\mathcal{W}^n$ for A*, such that the set PRO(Φ_P-$\mathcal{W}^n$) of arguments moved by PRO in Φ_P-$\mathcal{W}^n$ is conflict-free, iff* A *is included in an admissible extension* Adm *of the pdAF.*

Proof.

Soundness. We have to prove that if A is a member of the conflict-free set PRO(Φ_P-$\mathcal{W}^n$), then $A \in \mathsf{Adm}$. To simplify the notation, let E = PRO(Φ_P-$\mathcal{W}^n$). Assume that A is a member of the conflict-free set E, then:

- By Definition 11.2, the existence of the winning strategy implies that: each argument played by OPP against arguments moved by PRO in the winning strategy has been successfully countered by PRO. That is to say, $\forall X \in$ E, if $\exists Y \in$ AR such that $Y \Rightarrow_E X$, then $\exists Z \in$ E, such that $Z \Rightarrow_{\{Y\}} Y$, ensuring in this way that X is acceptable wrt E.
- Recall that the set of disputes of Φ_P-$\mathcal{W}^n$ is finite and composed of finite disputes (by Definition 11.1). As such, E is composed of a finite number of arguments.

We have thus shown that E is a finite, conflict-free set and every argument in E is acceptable wrt it. Therefore, E corresponds to an admissible extension, hence, if A is a member of the conflict-free set PRO(Φ_P-$\mathcal{W}^n$), then $A \in \mathsf{Adm}$.

Completeness. We show that if $A \in \mathsf{Adm}$, then A is a member of the conflict-free set PRO(Φ_P-$\mathcal{W}^n$). We are going to do this by constructing a dialectical winning strategy Φ_P-$\mathcal{W}^n$ for A.

– Assume that $A \in \mathsf{Adm}$. Since the pdAF is finite, then it is also finitary, meaning that every argument in Adm has a finite number of defeaters. Then we can build a dialectical winning strategy Φ_P-$\mathcal{W}^n$ for A if PRO starts the game with A and, for each argument Y dialectically defeating A and moved by OPP, PRO chooses one argument X from Adm (even A itself) such that $\mathrm{X}\Rightarrow_{\{Y\}}\mathrm{Y}$. Notice that the generation of infinite disputes is prevented by the admissible/preferred protocol (Definition 14.2(b)). This procedure can be repeated for every argument Z dialectically defeating X, and so on, until OPP runs out of legal moves according to the protocol Φ_P (which will happen for sure since A is a member of an admissible set).

The result will be a dialectical winning strategy Φ_P-$\mathcal{W}^n$ for A, hence, A is a member of the conflict-free set PRO(Φ_P-$\mathcal{W}^n$). We have thus shown that, if $A \in \mathsf{Adm}$, then A is a member of the conflict-free set PRO(Φ_P-$\mathcal{W}^n$).

□

5 Dialectical Grounded Games

The dialectical grounded game protocol Φ_G enjoys the same notations and definitions introduced thus far, but presents also important differences compared to the dialectical admissible/preferred game. Indeed, the protocol should be designed such that, when the game terminates and PRO is the winner, the set PRO(Φ_G-$\mathcal{W}^n$) of arguments moved by PRO in a dialectical winning strategy Φ_G-$\mathcal{W}^n$ is a subset of the grounded extension Grd of the pdAF. In this way, by iterating the framework characteristic function $\mathcal{F}$ from PRO(Φ_G-$\mathcal{W}^n$), we are able to obtain the grounded extension Grd. However, recall that it is the monotonicity of the function, in the case of a finitary pdAF[10], that ensures the construction of the least fixed point of $\mathcal{F}$ which corresponds to the grounded extension.

In Dialectical Cl-Arg [15] the monotonicity of $\mathcal{F}$ holds only under the domain of *epistemically maximal* (*em*) admissible sets of arguments (described in Definition 6). Then, to get the grounded extension via the iteration of $\mathcal{F}$ from the set PRO(Φ_G-$\mathcal{W}^n$), we will need PRO(Φ_G-$\mathcal{W}^n$) to be *em*. Otherwise, we might have to face a situation in which argument A, whose membership in Grd we wanted to

[10] Being finitary, it can be shown that $\mathcal{F}$ is also $\omega-$continuous (as explained in [17] for standard AFs and in [15] for pdAFs).

test via the dialectical grounded game, is not acceptable wrt Grd, although $A \in$ PRO(Φ_G-$\mathcal{W}^n$). To address this issue, we are going to adapt the protocol Φ_G accordingly.

Definition 15. [Dialectical Grounded Game legal moves] *Let $\mathcal{D}^n$ and $\mathcal{W}^n$ be characterized as in Definition 12, let d be a dispute of $\mathcal{D}^n$ and d' be a sub-dispute of d. Let also $(PL_n)X$ (for $n > 0$) denote the argument X played by either one of the two players (P or O) as the (last) nth-move. Then Φ_G identifies legal moves in the following way:*

(15.0) *PRO moves the first argument.*

(15.1) *If $(PL_n)X$ and $n = 2k$ (for $k > 0$), then the next move n+1, say Y, is by PRO and it is such that:*

- (a) *$Y \Rightarrow_{\{Z\}} Z$, where $Z \in$ OPP($\mathcal{D}^n$);*
- (b) *There exists a $\mathcal{W}^{n+1}$ for A of $\mathcal{D}^{n+1}$;*
- (c) *If $d = d'$—Z, then $Y \notin$ PRO(d').*

(15.2) *If $(PL_n)X$ and $n = 2k+1$ (for $k \geq 0$), then the next move n+1, say Y, is by OPP and it is such that:*

- (a) *$Y \Rightarrow_{\mathcal{S}} Z$, where $Z \in \mathcal{S}$ and $\mathcal{S} :=$ PRO($\mathcal{W}^n$).*
- (b) *For each $d = d'$—J—$\cdots$, where $J \in$ OPP($\mathcal{D}^n$) and its defeat has been disqualified, then LAST(d) = LAST(d') until next OPP's turn.*

(15.3) *If, at the beginning of its turn, OPP cannot perform the move described by (15.2(a)), then apply function Cl_{em} (Definition 6) on PRO($\mathcal{W}^n$).*

Notice that a Φ_G-dialectical game is said to be terminated when, during its turn, at least one player runs out of the legal moves identified by (15.1(a-c)) or (15.2(a)) of the protocol Φ_G. PRO wins only if it has a winning strategy once the game terminates. It loses otherwise.

As per Definition 14, the previous protocol can be informally summarised as follows. PRO starts the game by playing the first argument [(15.0)] and after that OPP will make its move. Then, the two players alternate in playing only one argument at a time to reply to one of their counterpart's arguments. Observe that when $\mathcal{S}$ is initialized in the game and, subsequently, every time its arguments are updated by the changes in PRO($\mathcal{W}^n$) [(15.2(a))], it is always the beginning of OPP's turn. This means that the condition for which $\mathcal{S} \neq \emptyset$ is continuously respected.

Notice also that the established protocol allows *backtracking* to other arguments. That is to say, when PRO moves it can either target the last argument played by OPP or another argument moved by OPP in the dialectical dispute tree generated thus far (i.e., an argument member of the set OPP($\mathcal{D}^n$)) [(15.1(a))]. Similarly, when OPP moves it can either target the last argument played by PRO or another argument moved by PRO in the current dialectical winning strategy (i.e., an argument member of the set PRO($\mathcal{W}^n$)) [(15.2(a))]. The *relevance conditions* [(15.1(b)) for PRO; (15.2(a)) for OPP] ensure that: after PRO has made its move, there will be a winning strategy $\mathcal{W}^{n+1}$, hence providing the victory to PRO; after OPP has moved, instead, the previous winning strategy will cease to exist, thus preventing PRO from winning. That is to say, PRO will be forced to generate a dialectical winning strategy during each of its turns, while OPP will have to invalidate such a winning strategy during every one of its turns. Observe also that backtracking and relevance conditions are strictly correlated (similarly to Definition 14).

The restriction (15.1(c)) emphasises the additional burden of proof entailed by the membership to the grounded extension. This is intuitively captured by the idea that in defending an argument X's membership to the grounded extension Grd, PRO must 'appeal to' some argument other than X itself. This is reflected in the game by the fact that PRO cannot repeat the arguments it has already moved in the same disputes.

Moreover, (15.2(b)) ensures that the disqualified defeats (Definition 13) are taken into account throughout the game. That is to say, whenever a dialectical defeat moved by an argument J is disqualified, the protocol guarantees the pruning of J and all the arguments that follow in the same dispute, until the next turn of OPP, when a new check for disqualified defeats will occur.

Finally, in light of the previously underlined epistemically maximal requirement, an additional one-time move has been included. Recall that adding all arguments up to some i to a set E, and then *em* closing, yields the same result as adding each argument one by one and closing prior to each subsequent addition. As such, once the game is terminated in favour of PRO and immediately before PRO is declared the winner, it suffices to apply function Cl_{em} (Definition 6) over the resulting set PRO($\mathcal{W}^n$) rendering it *em*, therefore, a subset of the grounded extension of the pdAF.

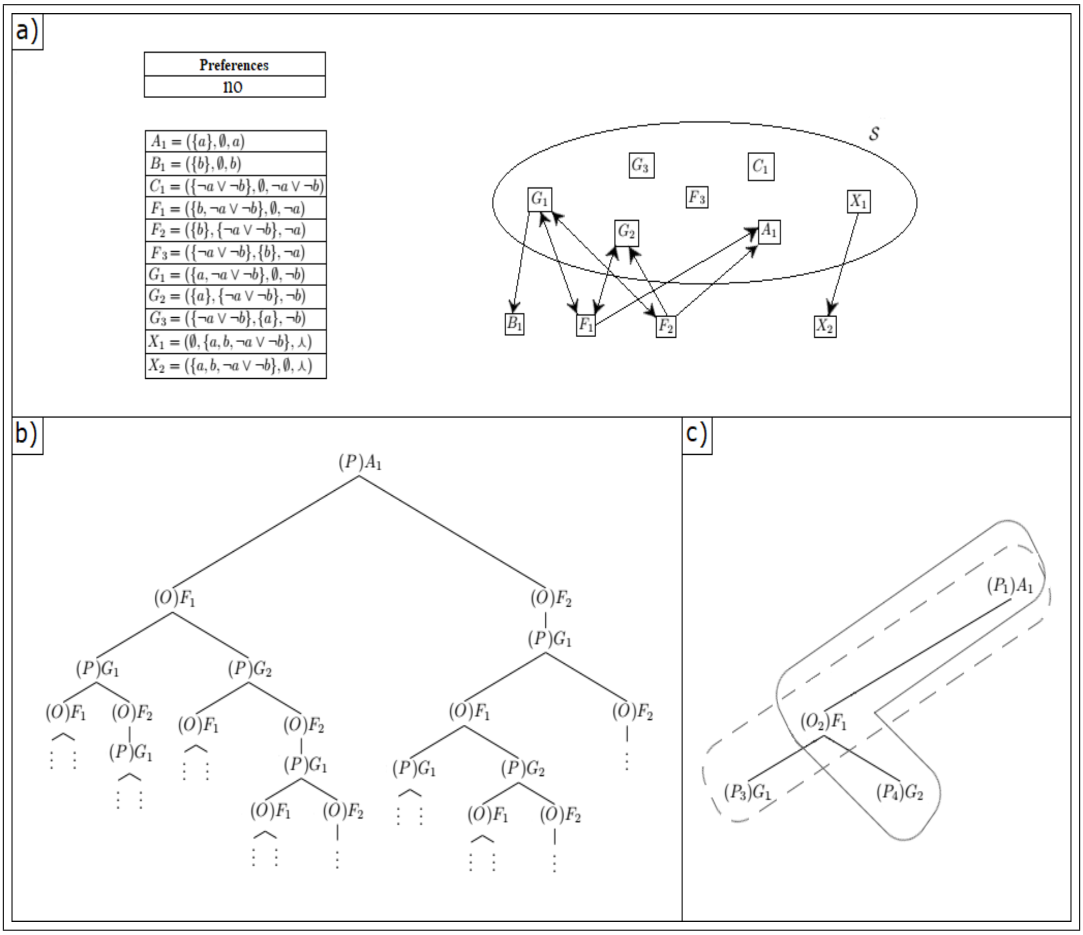

a)
Preferences
no
$A_1 = (\{a\}, \emptyset, a)$
$B_1 = (\{b\}, \emptyset, b)$
$C_1 = (\{\neg a \vee \neg b\}, \emptyset, \neg a \vee \neg b)$
$F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$
$F_2 = (\{b\}, \{\neg a \vee \neg b\}, \neg a)$
$F_3 = (\{\neg a \vee \neg b\}, \{b\}, \neg a)$
$G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
$G_2 = (\{a\}, \{\neg a \vee \neg b\}, \neg b)$
$G_3 = (\{\neg a \vee \neg b\}, \{a\}, \neg b)$
$X_1 = (\emptyset, \{a, b, \neg a \vee \neg b\}, \lambda)$
$X_2 = (\{a, b, \neg a \vee \neg b\}, \emptyset, \lambda)$
$\mathcal{S}$
G_3
C_1
G_1
F_3
X_1
G_2
A_1
B_1
F_1
F_2
X_2
b)
$(P)A_1$
$(O)F_1$
$(O)F_2$
$(P)G_1$
$(P)G_2$
$(O)F_1$
$(O)F_2$
c)
$(P_1)A_1$
$(O_2)F_1$
$(P_3)G_1$
$(P_4)G_2$

Figure 3: Figure a) illustrates a pdAF with a list of its arguments and the set $\mathcal{S}$ that is parametrized by the dialectical defeats. Consider also that X_2 is defeated by all the arguments of the pdAF, except A_1, B_1, and C_1 (the arrows that should have highlighted such defeats have been omitted to avoid unnecessary graphical confusion). Figure b) displays the dialectical dispute tree $\mathcal{D}$ induced by A_1 wrt $\mathcal{S}$ in the pdAF of Figure a). Notice that $\mathcal{D}$ is composed of infinite disputes (the vertical dots represent the endless continuation of the disputes), as such, it does not have a winning strategy. A dialectical dispute tree Φ-$\mathcal{D}^n$, with $n = 4$, is depicted in Figure c) and corresponds to a Φ-dialectical game played up to the nth-move. Observe that the number of each move (next to the label P or O) represents the order in which the arguments have been played in the game. In this example, we are assuming a protocol Φ that licenses legal moves where PRO can play more than one argument per turn, therefore, Φ-$\mathcal{D}^n$ has two winning strategies (both of which are encircled in the figure).

5.1 Soundness and Completeness

In the following proofs, we are going to employ the framework characteristic function $\mathcal{F}_p$, which iterates over admissible epistemically maximal extensions:

Definition 16. *Let $\langle AR, defeats\rangle$ be a pdAF and AR_p the set of all the em admissible subsets of AR. Then $\mathcal{F}_p : AR_p \mapsto AR_p$, where $\mathcal{F}_p(E) = Cl_{em}(\mathcal{F}(E))$.*

We can now show that the dialectical grounded game satisfies the properties of soundness and completeness.

Theorem 8. *Let Φ_G-$\mathcal{D}^n$ identifies a terminated Φ_G-dialectical game for* A*. Then, there exists a dialectical winning strategy Φ_G-$\mathcal{W}^n$ for* A*, such that the em closure $Cl_{em}(PRO(\Phi_G\text{-}\mathcal{W}^n))$ of the set of arguments moved by PRO in Φ_G-$\mathcal{W}^n$ is conflict-free, iff* A *is included in the grounded extension* Grd *of the pdAF.*

To simplify the notation, let us abbreviate $Cl_{em}(\mathrm{PRO}(\Phi_G\text{-}\mathcal{W}^n))$ in Cl_{em}.

Proof.

Soundness. We have to prove that if A is a member of the conflict-free set Cl_{em}, then $A \in \mathsf{Grd}$. Hence, assuming that A is a member of the conflict-free set Cl_{em}:

- Clearly, all of Φ_G-$\mathcal{W}^n$ leaves, say X_i, are in $\mathcal{F}_p(E_0)$ since they have no defeaters and are then acceptable wrt $\emptyset$. Now, consider that in every branch of Φ_G-$\mathcal{W}^n$, the arguments defended[11] by each X_i are acceptable with respect to $\mathcal{F}_p(E_0)$ and so are in $\mathcal{F}_p(E_1)$. This process can be repeated until, say, $\mathcal{F}_p(E_i)$ when the root A of Φ_G-$\mathcal{W}^n$ is reached. Since $Cl_{em} \subseteq \mathcal{F}_p(E_i)$, and further iterations of $\mathcal{F}_p(E_i)$ will yield the generation of the least fixed point Grd, then A will be a member of Grd.

This suffices to show that if A is a member of the conflict-free set Cl_{em}, then $A \in \mathsf{Grd}$.

Completeness. We have to prove that if $A \in \mathsf{Grd}$, then A is a member of the conflict-free set Cl_{em}. Employing the acceptable arguments in the characteristic function $\mathcal{F}_p$ we are going to show that we can build a Φ_G-winning strategy for A.

[11] Recall that an argument X defends an argument Z iff: when $\exists Y \in AR$ such that Y defeats Z, then X defeats Y.

- Assume that $A \in \mathsf{Grd}$. Since the pdAF is finite, it is also finitary, hence we know that there is a least number i such that $A \in \mathcal{F}_p(E_i)$. Then we will have a dialectical winning strategy Φ_G-$\mathcal{W}^n$ for A if PRO starts the game with A and: for each argument Y dialectically defeating A and moved by OPP, PRO chooses one argument X from $\mathcal{F}_p(E_{i-1})$ such that X$\Rightarrow_{\{Y\}}$Y. This procedure can be iterated for every argument Z dialectically defeating X, and so on, until PRO can choose an argument from $\mathcal{F}_p(E_0)$. $\mathcal{F}_p(E_0)$ has no defeaters and, as such, OPP cannot play any legal move (licensed by the protocol Φ_G) against it. Finally, the grounded game protocol will also ensure the epistemically maximality of the set of arguments moved by PRO in Φ_G-$\mathcal{W}^n$ (15.3).

The result yields a dialectical winning strategy Φ_G-$\mathcal{W}^n$ for A, such that A is a member of the conflict-free set Cl_{em}. We have thus shown that, if $A \in \mathsf{Grd}$, then A is a member of the conflict-free set Cl_{em}.

□

6 Main Features of Dialectical Argument Games

Dialectical argument games hold specific features that differentiate them from the standard argument games of [25, 9, 34] and depend upon their protocols and the properties possessed by each pdAF (especially P1, P2 and P3). Although, for convenience, we are going to outline these features using the dialectical admissible/preferred game (Definition 14), notice that the choice of the protocol is irrelevant.

6.1 Feature 1 (F1)

(F1) The set of all the arguments moved by PRO in a dialectical winning strategy (i.e., PRO(Φ_P-$\mathcal{W}^n$)), is always conflict-free.

Every pdAF = $\langle$AR, *defeats*$\rangle$ prevents any conflicts existing between arguments in a set E $\subseteq$ AR if each argument in E is acceptable with respect to it. Since this has already been formally proven and shown[12], here we will try to explain it through an example. Notice also the rationale underpinning *F1*: due to their limited resources, it would be unrealistic to demand that real-world agents actually perform conflict-free checks on every set E of arguments.

[12] **Lemma 17** of [15] states that: *Let $E \subseteq AR$ such that every argument in E is acceptable w.r.t. E, and AR satisfies $P1, P2$ and $P3$. Then E is conflict-free.*
The proof can be found in the same paper.

Example 3. *Consider a pdAF that includes the arguments listed in Table 1 and such that all the arguments composing the set PRO(Φ_P-$\mathcal{W}^n$) are acceptable wrt it. To simplify the notation, let E = PRO(Φ_P-$\mathcal{W}^n$).*
Among the arguments of E, suppose that there are two conflicting arguments as $G_2 = (\{a\}, \{\neg a \vee \neg b\}, \neg b)$ and $F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$: we are going to show how this will lead to a contradiction. Due to property P1, $A_1 = (\{a\}, \emptyset, a) \in AR$. Hence, by property P3, $X_2 = (\{a, b, \neg a \vee \neg b\}, \emptyset, \curlywedge) \in AR$ and by property P2, $X_1 = (\emptyset, \{a, b, \neg a \vee \neg b\}, \curlywedge) \in AR$. However, if this is the case, $X_1 \Rightarrow_E G_2$ (and, similarly, $X_1 \Rightarrow_E F_1$). Since X_1 is unassailable, $\nexists Z \in E$ such that $Z \Rightarrow_{\{X_1\}} X_1$ and this will contradict the assumption that all the arguments members of E are acceptable wrt to it. Therefore, since all the arguments that compose the set PRO(Φ_P-$\mathcal{W}^n$) are acceptable wrt it, PRO(Φ_P-$\mathcal{W}^n$) must be conflict-free.

6.2 Feature 2 (F2)

(F2) The relevance conditions, i.e., the conditions of the protocol that compel both players to change the outcome of the game at the end of every turn, are essential to the unfolding of the dialectical argument games. This also justifies why the set $\mathcal{S}$ cannot be initialized with any set other than PRO(Φ_P-$\mathcal{W}^n$).

The relevance conditions (14.1(b) and 14.2(a) of Definition 14) can be summarised as the conditions that force the two players to change the outcome of the game at the end of every turn[13]. These requirements are fundamental for real-world agents that reason with limited availability of resources. Indeed, it would be illogical to allow such players to move arguments useless for the result of the game: this would simply mean wasting valuable resources[14]. Moreover, the relevance conditions clarify why the set $\mathcal{S}$, referenced in the admissible/preferred protocol, corresponds to the current set of arguments moved by PRO in Φ_P-$\mathcal{W}^n$, that is to say, PRO(Φ_P-$\mathcal{W}^n$). This, in turn, allows avoiding a specific issue that could permanently prevent the victory of PRO, as the following example will show.

Example 4. *The examples of Figures 4, 5 and 6 depict a dialectical admissible game played using the arguments of Table 1, where $F_1 \not\prec (\{a\}, \emptyset, a)$, $\forall T \in \{G_1, L_1\}$, $T \not\prec (\{b\}, \emptyset, b)$, $\forall V \in \{N_3, X_3\}$, $V \not\prec (\{\neg b\}, \emptyset, \neg b)$, while $H_1 \not\prec (\{\neg a \vee \neg b\}, \emptyset, \neg a \vee$*

[13] The research presented in [28] introduces a series of relevant properties for dialogue protocols. Property *R1* seems quite similar to our relevance conditions, although our study concerns argument game proof theories rather than dialogues.

[14] Notice that we are dealing with pdAFs, and so, small subsets of the respective overall set of arguments of the considered framework. As such, positing only relevant arguments is not going to be particularly expensive for agents' resources.

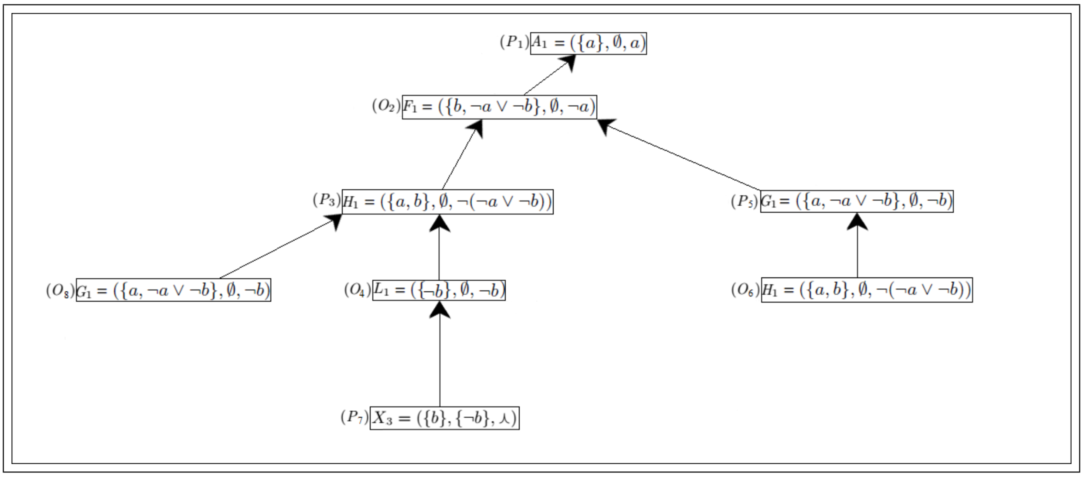

(P_1) $A_1 = (\{a\}, \emptyset, a)$
(O_2) $F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$
(P_3) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_5) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_8) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_4) $L_1 = (\{\neg b\}, \emptyset, \neg b)$
(O_6) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_7) $X_3 = (\{b\}, \{\neg b\}, ⅄)$

Figure 4: The Figure illustrates a dialectical dispute tree Φ_P-$\mathcal{D}^n$, hence generated following the protocol for the dialectical admissible/preferred games. Notice that the arrows indicate the defeats between the arguments. Starting with the root argument A_1, the other arguments are played according to the order highlighted by the numbers near their labels (P or O). The last player to move is OPP, which moves G_1. Since $G_1 \Rightarrow_{\mathcal{S}} H_1$ (where $\mathcal{S} :=$ PRO(Φ_P-$\mathcal{W}^{n\text{-}1}$), i.e., $\mathcal{S} = \{A_1, H_1, X_3\}$) and $G_1 \not\prec (\{b\}, \emptyset, b)$, this ensures OPP invalidates the winning strategy Φ_P-$\mathcal{W}^{n\text{-}1}$. Hence, there is no winning strategy in Φ_P-$\mathcal{D}^n$.

$\neg b$). Starting with the root A_1, the order in which the arguments are played is outlined in the brackets, next to the labels PRO and OPP. The dialectical dispute tree Φ_P-$\mathcal{D}^n$ (Figure 4) has been generated following the protocol for the dialectical admissible/preferred games, however, its extension into Φ-$\mathcal{D}^{n+1}$ (Figure 5) does not take into account PRO's relevance condition (14.1(b) of Definition 14). This immediately raises an issue: without the relevance condition, we could have to face a situation in which PRO is still losing even after its turn has ended (Figure 5). In this circumstance, during the next turn of OPP, there will be no winning strategy, hence no set of arguments moved by PRO in Φ_P-$\mathcal{W}^{n+1}$ (i.e., the set PRO(Φ_P-$\mathcal{W}^{n+1}$)), that can be targeted as $\mathcal{S}$. Suppose, for the sake of the example, that the protocol of the game allows searching for another set $\mathcal{S}$. What could then be the set $\mathcal{S}$ parametrised by the dialectical defeats moved by OPP? Without PRO(Φ_P-$\mathcal{W}^{n+1}$) the only reasonable alternative is to consider a different set $\mathcal{S}$ initialized in a way such that $\mathcal{S} \subseteq$ PRO(Φ_P-$\mathcal{D}^{n+1}$). Nevertheless, notice that if OPP is allowed to suppose the premises of arguments in a non-conflict-free set $\mathcal{S}$, then OPP would have enough resources for playing an unassailable argument (as X_1). As shown in Figure 6, H_1, G_1 $\in \mathcal{S}$, and $B_1 \in AR$ by property P1 of the pdAF. By P3, $X_2 \in AR$, while by property

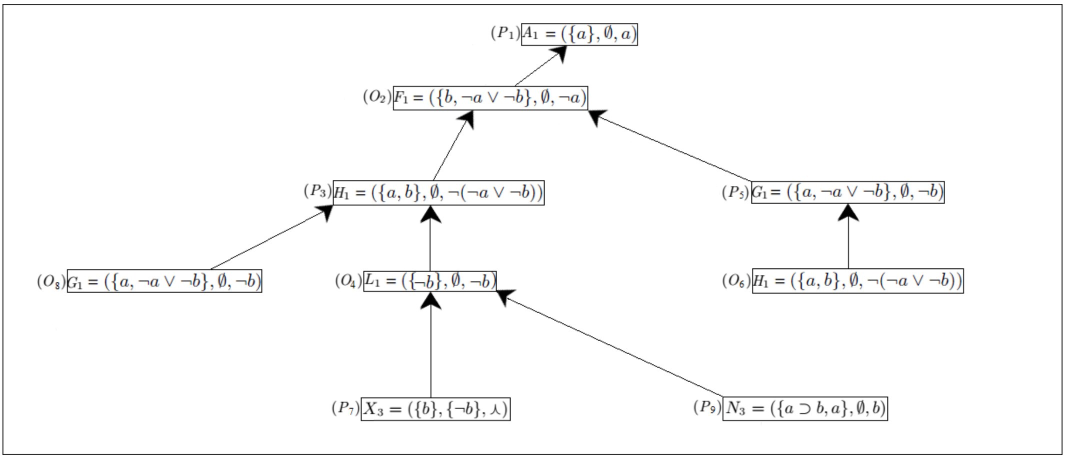

(P_1) $A_1 = (\{a\}, \emptyset, a)$
(O_2) $F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$
(P_3) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_5) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_8) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_4) $L_1 = (\{\neg b\}, \emptyset, \neg b)$
(O_6) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_7) $X_3 = (\{b\}, \{\neg b\}, \curlywedge)$
(P_9) $N_3 = (\{a \supset b, a\}, \emptyset, b)$

Figure 5: The Figure illustrates the extension of the dialectical dispute tree Φ_P-$\mathcal{D}^n$ into Φ_P-$\mathcal{D}^{n+1}$ due to argument N_3 played by PRO. As we can see, if PRO's relevance condition is dropped, then PRO is free to move any argument and not only the ones that will reinstate the winning strategy. $N_3 \Rightarrow_{\{L_1\}} L_1$ and $N_3 \not\rightharpoonup (\{\neg b\}, \emptyset, \neg b)$. However, this implies that, even after PRO moves, there is no winning strategy in Φ_P-$\mathcal{D}^{n+1}$ (because the argument G_1 played by OPP has not yet been defeated).

P2, also $X_1 \in AR$ (since X_1 is the logically equivalent argument of X_2). Argument X_1 constitutes the problem: it defeats A_1 and has empty premises, which implies it cannot be defeated. This means that, by playing X_1, OPP will change the final outcome of the game invalidating any other possible attempt from PRO of reinstating the winning strategy. However, this happened in the example because there was no set PRO(Φ_P-$\mathcal{W}^{n+1}$) and OPP had to suppose the premises of the arguments members of a different set $\mathcal{S} \subseteq$ PRO(Φ_P-$\mathcal{D}^{n+1}$) which was not conflict-free. In other words, unassailable arguments as X_1 can be moved only when (i) arguments that defeat each other or (ii) unconditional arguments with conflicting conclusions are in $\mathcal{S}$. Moving such arguments will immediately trigger property P3 of the pdAF, which will highlight the inconsistency of their premises, while property P2 will ensure the generation of the corresponding unassailable argument.

Nevertheless, without requiring a resource-consuming conflict-free check on every $\mathcal{S} \subseteq$ PRO(Φ_P-$\mathcal{D}^{n+1}$), how would it be possible to ensure the conflict-freeness of the set $\mathcal{S}$? The only set of arguments moved by PRO which satisfies this condition (without requiring a conflict-free check) in a dialectical argument admissible game is the set PRO(Φ_P-$\mathcal{W}^{n+1}$), thanks to property F1. Therefore, $\mathcal{S}$ has to be initialized to PRO(Φ_P-$\mathcal{W}^{n+1}$).

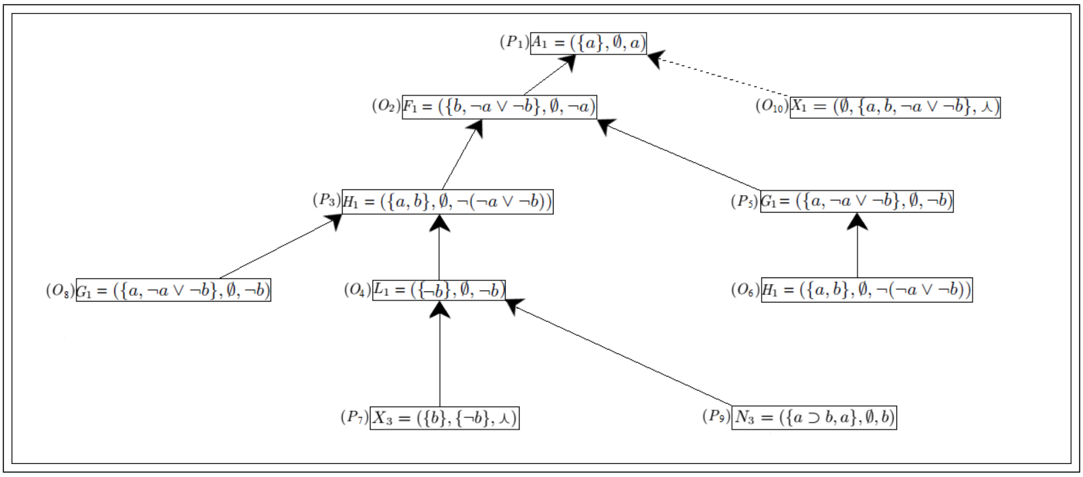

(P_1) $A_1 = (\{a\}, \emptyset, a)$
(O_2) $F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$
(O_{10}) $X_1 = (\emptyset, \{a, b, \neg a \vee \neg b\}, \curlywedge)$
(P_3) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_5) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_8) $G_1 = (\{a, \neg a \vee \neg b\}, \emptyset, \neg b)$
(O_4) $L_1 = (\{\neg b\}, \emptyset, \neg b)$
(O_6) $H_1 = (\{a, b\}, \emptyset, \neg(\neg a \vee \neg b))$
(P_7) $X_3 = (\{b\}, \{\neg b\}, \curlywedge)$
(P_9) $N_3 = (\{a \supset b, a\}, \emptyset, b)$

Figure 6: The Figure illustrates the extension of the dialectical dispute tree Φ_P-$\mathcal{D}^{n+1}$ into Φ_P-$\mathcal{D}^{n+2}$ due to argument X_1 played by OPP. It is possible to move X_1 because there is no winning strategy in Φ_P-$\mathcal{W}^{n+1}$, hence there is no set PRO(Φ_P-$\mathcal{W}^{n+1}$): this forces OPP to target the premises of a different set $\mathcal{S}$, initialized in a way such that $\mathcal{S} \subseteq$ PRO(Φ_P-$\mathcal{D}^{n+1}$) (in the case of the example, $\mathcal{S} :=$ PRO(Φ_P-$\mathcal{D}^{n+1}$), i.e., $\mathcal{S} = \{A_1, H_1, G_1, X_3, N_3\}$). The danger of arguments such as X_1 lies in their unassailability and the fact that they always succeed as defeats (underlined by the dashed arrow in the picture and explained in Definition 2). That is to say, the final outcome of the game can then be changed if $\mathcal{S} \neq$ PRO(Φ_P-$\mathcal{W}^{n+1}$) because it can allow OPP to move arguments as X_1 against the root of the tree (preventing PRO from reinstating any other possible winning strategy).

The implication of what has been shown in Example 4 is that the relevance conditions need to be part of the protocols of any dialectical argument game. Indeed, if this is not the case, we could have to face a situation in which PRO is still losing even after its turn has ended. In this circumstance, during the next turn of OPP, there will be no set PRO(Φ_P-$\mathcal{W}^n$) that can be used to initialise $\mathcal{S}$. Hence, once again, the issue outlined in Example 4 could arise and change the final outcome of the game by permanently invalidating PRO's winning strategy. This then means that $\mathcal{S} :=$ PRO(Φ_P-$\mathcal{W}^n$) and cannot be otherwise.

6.3 Feature 3 (F3)

Before the introduction of the third feature (F3) enjoyed by the dialectical admissible/preferred argument games, we need to formally define the *uniqueness* of the dialectical winning strategy regardless of the employed protocol.

Definition 17. [Uniqueness of the dialectical winning strategy] *Let $\mathcal{D}^n$ and* *let $\mathcal{W}^n$ be defined as in Definition 12. Then $\mathcal{W}^n$ is said to enjoy the uniqueness property if there is no other dialectical winning strategy for* A *wrt $\mathcal{S}$ simultaneously present in $\mathcal{D}^n$.*

Let us consider a dialectical dispute tree $\mathcal{D}^n$ identical (although without the implementation of a specific game protocol) to the one in Figure 3(c). This tree has two winning strategies, say $\mathcal{W}_1^n$ and $\mathcal{W}_2^n$, each of which is composed of a single dispute. That is to say: $d_1 = (P_1)A_1$—$(O_2)F_1$—$(P_3)G_1$ and $d_2 = (P_1)A_1$—$(O_2)F_1$—$(P_4)G_2$, such that $\mathcal{W}_1^n$ is composed of d_1, while $\mathcal{W}_2^n$ is composed of d_2. Obviously, $\mathcal{D}^n$ does not enjoy the uniqueness property. Indeed, both G_1 and G_2 defeat the same argument F_1, whereas only one of such defeats is actually needed. This implies that it suffices that either $\mathcal{W}_1^n$ or $\mathcal{W}_2^n$ is present for PRO to win (at least temporarily) the game. For the final outcome of the game, it is pointless to have both winning strategies simultaneously. It is also resource-consuming, meaning that it does not comply well with the Dialectical Cl-Arg purpose of capturing resource-bounded real-world agents' reasoning.

(F3) Any dialectical winning strategy Φ_P-$\mathcal{W}^n$ enjoys the uniqueness property.

Uniqueness is a property enforced on a dialectical winning strategy Φ_P-$\mathcal{W}^n$ by the protocol of the dialectical admissible/preferred argument game. Uniqueness is certainly a desirable property since it allows for shorter and simpler games. This ensures a quicker evaluation of the status of the dialectical dispute tree root.

The following Lemma shows that the protocol of the dialectical admissible/preferred game ensures the uniqueness of Φ_P-$\mathcal{W}^n$.

Lemma 2. *Let Φ_P-$\mathcal{D}^n$ identifies a Φ_P-dialectical game for* A*. Then, there exists only one dialectical winning strategy Φ_P-$\mathcal{W}^n$ for* A *wrt $\mathcal{S}$ that is simultaneously present in Φ_P-$\mathcal{D}^n$.*

Proof. Since the protocol of the admissible/preferred game forces the players to move only one argument per turn, the only other way to have multiple winning strategies simultaneously is by having different arguments moved by PRO (in different turns) that defeat the same argument played by OPP. We are going to show how this case cannot occur under the Φ_P protocol.

Let d_1 be a dispute in Φ_P-$\mathcal{W}^n$ and d' a sub-dispute of d_1. Let also $d_1 = d'$—$(O_{n-i})Y$ —$(P_{n-i+1})X$, for $n - i > 1$. As usual, the index near the player labels denotes the order in which the moves have been played. Suppose now that the last (*nth*) argument moved is an argument $Z \neq X$ from PRO that dialectically defeats Y and generates $d_2 = d'$—$(O_{n-i})Y$—$(P_n)Z$, which is another dispute in Φ_P-$\mathcal{D}^n$ and d' is a sub-dispute of d_2 as well, then it is easy to see that PRO has played against the protocol Φ_P. That is because:

- If PRO defeats an argument without affecting the existing game status it will violate its relevance condition (Definition 14.1(b)).

Playing argument Z will then be prevented by PRO's relevance condition, ensuring in this way the uniqueness of the dialectical winning strategy Φ_P-$\mathcal{W}^n$. □

7 Efficiency Improvements

The protocols thus far developed can benefit from a range of efficiency improvements. They follow from the properties of the dialectical games and Dialectical Cl-Arg in general, which means that they will preserve the already proven soundness and completeness results. In particular, we can obtain shorter games thanks to (I1), which allows us to avoid meaningless repetitions of defeated arguments from OPP. Moreover, (I2) and (I3) show how, due to the features enjoyed by the dialectical games and without additional restrictions on the legal moves available to the players (unlike in [25]), it is possible to obtain other specific efficiency improvements. In the next section, these enhancements will be examined and, when required, also formalised and integrated into the protocols of the dialectical games.

7.1 List of Efficiency Improvements for Dialectical Games

In the admissible/preferred dialectical game, OPP is forbidden to repeat any arguments (*and not just in a dispute*) *which have already been defeated, and not defended or indirectly defended by another argument, in the game.*[15]

Let us assume that OPP's argument Y has been defeated, and not defended, by an argument X moved by PRO in a dispute d. If now OPP repeats Y in a different dispute, then PRO can simply repeat X defeating Y once again.

Example 5. *For instance, let Φ_P-$\mathcal{D}^n$ be a dialectical dispute tree and let d be a dispute in Φ_P-$\mathcal{D}^n$. Suppose also that X is an argument moved by PRO in d, while Y is an argument played by OPP in d such that $X \Rightarrow_{\{Y\}} Y$. Then, if the game goes on* (*up to $n + k$ moves, for $k > 1$*)*, whenever Y will 'appear' in a different dispute,*

[15] According to the recursive definition of indirect defence, an argument X *indirectly defends* an argument A if: *i*) X *defends* A; *ii*) X *defends* Z, and Z *indirectly defends* A.

PRO can simply play X again. As such, playing argument Y proves to be just a waste of resources.

We can now formalise this idea by substituting condition (14.2(b)) from the protocol Φ_P (Definition 14) with the following constraint (I1). The purpose of forbidding such moves is to avoid extending the game by adding useless sequences of arguments to it:

Definition 18 (Improved legal move)**.** *The following additional constraint for OPP* (*where OPP's argument Y is the next move played in the game*) *substitutes* (14.2(b)) *from the protocol* Φ_P*:*

(I1) *If* $\exists J \in$ *OPP*(Φ-$\mathcal{D}^n$) *such that J is defeated and not defended* (*neither directly nor indirectly defended*) *by another argument, then* $Y \neq J$.

The soundness and completeness results of the dialectical games will not be affected by restriction (I1), as the following lemma will prove:

Lemma 3. *Let* Φ_P-$\mathcal{D}^n$ *identifies a terminated* Φ_P*-dialectical game for* A*. Then, there exists a dialectical winning strategy* Φ_P-$\mathcal{W}_1^n$ *for* A*, iff there exists a dialectical winning strategy* Φ_P-$\mathcal{W}_2^n$ *for* A *constructed using a protocol that employs* (I1)*.*

Proof.

[→] If there exists a dialectical winning strategy Φ_P-$\mathcal{W}_2^n$, then there also trivially exists a dialectical winning strategy Φ_P-$\mathcal{W}_1^n$. Indeed, if OPP cannot repeat its defeated (and not defended) arguments (I1), it cannot as well repeat its arguments in the same disputes ((14.2(b)) of Definition 14). That is to say, Φ_P-$\mathcal{W}_2^n$ follows every requirement established by protocol Φ_P.

[←] We are going to show that every dialectical winning strategy Φ_P-$\mathcal{W}_1^n$ can be transformed into a dialectical winning strategy Φ_P-$\mathcal{W}_2^n$. Suppose that there is a dispute d in Φ_P-$\mathcal{W}_1^n$ in which it appears the sequence J—X of arguments such that J is moved by OPP, X is moved by PRO and $X \Rightarrow_{\{J\}} J$. We also know that J is not defended (or indirectly defended) because, being a dispute in the winning strategy, d terminates with a PRO argument. Notice that, since J is an OPP argument moved in a dispute, it must be preceded by a PRO argument. Hence, if we now remove every other J—X—$\cdots$ sequence (including whatever follows after X) from the dialectical winning strategy, we will not affect PRO's victory and we will generate a new dialectical winning strategy, i.e., Φ_P-$\mathcal{W}_2^n$.

□

The following improvements are similar to the ones already introduced in [25], with an important difference. Unlike the standard games, dialectical games do not need to enforce specific restrictions on their protocols in order to benefit from these efficiency enhancements: they are ensured by the properties enjoyed by any dialectical game.

($I2$) *PRO does not move self-defeating arguments (i.e., arguments which defeat themselves).*

Whenever a self-defeating argument, say X, is played by PRO, PRO violates property *F1.* Indeed, even if X reinstates a dialectical winning strategy Φ-$\mathcal{W}^n$, the same X will also conflict with an argument member of PRO(Φ-$\mathcal{W}^n$), i.e., X itself.

($I3$) *PRO does not play an argument that defeats (or is defeated by) an argument in PRO(Φ-$\mathcal{W}^n$).*

That is to say, PRO does not move arguments that conflict with the arguments it has already moved in the winning strategy. Indeed, if PRO plays an argument X defeated by (a member of) PRO(Φ-$\mathcal{W}^n$) or that defeats an argument member of PRO(Φ-$\mathcal{W}^n$), the resulting winning strategy will not be conflict-free. This will then violate property *F1.*

Example 6. *Consider the dialectical dispute tree of Figure* 4 *and assume that PRO decides to counter its opponent's last move by playing argument* $F_1 = (\{b, \neg a \vee \neg b\}, \emptyset, \neg a)$ *such that* $F_1 \Rightarrow_{\{G_1\}} G_1$ *on* $(\{a\}, \emptyset, a)$*. However, since* F_1 *defeats, hence conflicts, with* $H_1 \in$ *PRO*(Φ-$\mathcal{W}^n$) (F_1 *is also dialectically defeated by* H_1) *this move will violate property* F1 (*the situation will then be similar to the one described in Example 3*).

Remark 2. *Notice that* (*I3*) *also subsumes the fact that* PRO does not move an argument X in a dispute d if such an argument has already been played by OPP in d. *Indeed, playing argument X will reinstate the dialectical winning strategy* Φ-$\mathcal{W}^n$. *However, at the same time, X is an argument moved by OPP (hence X complies with OPP's relevance condition). As such, playing X will imply defeating once again an argument in PRO*(Φ-$\mathcal{W}^n$)*, violating property F1*[16].

[16] It is interesting to observe that this is not generally the case if PRO repeats (i) an OPP argument or (ii) an already defeated PRO argument, say X, in a different dispute of the dialectical dispute tree. That is because it might be that the opponent cannot suppose anymore the same premises that (ii) allowed it to defeat X the first time or (i) allowed it to defeat an argument in PRO(Φ-$\mathcal{W}^n$). For example, assume that an argument Y moved by OPP dialectically defeated X

As shown, (I2) and (I3) follow directly from the property *F1*, which is enjoyed by any dialectical game. As such, no modifications to the game protocols are needed, meaning that the soundness and completeness results will be preserved.

8 Related and Future Work

Initially introduced in [14], the dialectical approach of Dialectical Cl-Arg has been subsequently examined from different perspectives. For example, the investigation concerning argumentative characterisations of Brewka's Preferred Subtheories (PS) [3] showed that, compared with the standard approach, the grounded semantics applied to Dialectical Cl-Arg more closely approximates sceptical inference in PS [16]. In addition, the research presented in [26] provides a full rational account of structured (ASPIC$^+$) arguments under resource bounds by adapting the approach of Dialectical Cl-Arg.

Extending further the study commenced in [10] and continued in [11], we plan to increase the range of dialectical argument game protocols investigating the stable [9], semi-stable [4] and ideal semantics [19, 5]. Similarly to the work presented in [25], we could also consider adapting the standard 3-values labelling approach (where each label represents the *IN, OUT*, and *UNDEC* status of an argument with respect to the examined semantics) and devise algorithmic procedures for the enumeration of specific extensions. Starting from the preliminary study proposed in [12], the design of fully-fledged algorithms would also help in additionally assessing the soundness and completeness properties of the dialectical argument games. Finally, another research direction that will be pursued involves generalising the developed dialectical argument games to dialogues, following the guidelines of the already existing literature in the field (mainly [22, 30, 13]). This would have the interesting consequence of allowing to move from non-monotonic single-agent inference to distributed non-monotonic reasoning.

9 Conclusion

The main aspects of the real-world uses of argumentation by resource-bounded agents include: (i) showing the inconsistencies of an opponent's argument by supposing the premises of its arguments; (ii) handling only finite subsets of the arguments of the AFs; (iii) reducing the consumption of resources by employing dialectical

drawing its suppositions α from Prem(PRO(Φ-$\mathcal{W}^n$)). However, after the game goes on, it might be that $\alpha \notin$ Prem(PRO(Φ-$\mathcal{W}^{n+k}$)). Then Y cannot dialectically defeat X anymore (i.e., Y defeat against X is disqualified), therefore X is now a perfectly viable move for PRO.

means (while still satisfying the rationality postulates and practical desiderata) [15]. These features would constitute the hallmarks of an argument game based on Dialectical Cl-Arg, thus capable of better approximating non-monotonic single-agent real-world reasoning processes than the standard argument games. In this paper, we have achieved some important results. We have developed argument game proof theories (denoted as *dialectical argument games*) for the admissible, preferred and grounded semantics of Dialectical Cl-Arg. Incorporating dialectical defeats in the standard structure of the argument games proved to be a non-trivial process which yielded the discovery of interesting properties that differentiate dialectical games from the standard ones. That is to say, dialectical games enjoy (a) specific relevance conditions that characterise their protocols and yield (b) the uniqueness of their winning strategies, whilst property *F1* ensures (c) the conflict-freeness of the set of arguments moved by the proponent in the winning strategy. The last is of particular importance since it provides the games with a various range of efficiency improvements. Without the need to perform any additional checks or to enforce additional restrictions in the protocols (unlike in [25]), *F1* allows each dialectical game to prevent the proponent from: playing self-defeating arguments; playing arguments already moved by the opponent (in the same dispute); and playing arguments that defeat (or are defeated by) other arguments already moved by the proponent. Finally, another efficiency improvement can be obtained if the opponent is forbidden to repeat arguments that have already been defeated in the dialectical admissible/preferred game, such that none of them has also been defended or indirectly defended by other arguments.

References

[1] Aristotle. *The Organon.* Jazzybee Verlag, 2015.

[2] Philippe Besnard and Anthony Hunter. *Elements of argumentation*, volume 47. MIT press Cambridge, 2008.

[3] Gerhard Brewka. Preferred subtheories: An extended logical framework for default reasoning. In *IJCAI*, volume 89, pages 1043–1048, 1989.

[4] Martin Caminada. An algorithm for computing semi-stable semantics. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 222–234. Springer, 2007.

[5] Martin Caminada. A labelling approach for ideal and stage semantics. *Argument and Computation*, 2(1):1–21, 2011.

[6] Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5-6):286–310, 2007.

[7] Martin Caminada, Walter Carnielli, and Paul Dunne. Semi-stable semantics. *Journal of Logic and Computation*, 22(5):1207–1254, 2012.

[8] Martin Caminada, Sanjay Modgil, and Nir Oren. Preferences and unrestricted rebut. *Computational Models of Argument*, 2014.

[9] Martin Caminada and Yining Wu. An argument game for stable semantics. *Logic Journal of IGPL*, 17(1):77–90, 2009.

[10] Federico Castagna. Argument games for Dialectical Classical Logic Argumentation. *Online Handbook of Argumentation for AI*, pages 2–6, 2020.

[11] Federico Castagna. A dialectical characterisation of argument game proof theories for Classical Logic Argumentation. *International Workshop on Advances in Argumentation in Artificial Intelligence (AI^3)*, 2021.

[12] Federico Castagna. Labelling procedures for Dialectical Classical Logic Argumentation. *Online Handbook of Argumentation for AI*, pages 7–11, 2021.

[13] Eva Cogan, Simon Parsons, and Peter McBurney. New types of inter-agent dialogues. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 154–168. Springer, 2005.

[14] Marcello D'Agostino and Sanjay Modgil. A rational account of classical logic argumentation for real-world agents. In *ECAI*, pages 141–149, 2016.

[15] Marcello D'Agostino and Sanjay Modgil. Classical logic, argument and dialectic. *Artificial Intelligence*, 262:15–51, 2018.

[16] Marcello D'Agostino and Sanjay Modgil. A study of argumentative characterisations of preferred subtheories. In *IJCAI*, pages 1788–1794, 2018.

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

[18] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. Assumption-based argumentation. In *Argumentation in artificial intelligence*, pages 199–218. Springer, 2009.

[19] Phan Minh Dung, Paolo Mancarella, and Francesca Toni. Computing ideal sceptical argumentation. *Artificial Intelligence*, 171(10-15):642–674, 2007.

[20] Phan Minh Dung, Francesca Toni, and Paolo Mancarella. Some design guidelines for practical argumentation systems. In *Computational Models of Argument*, pages 183–194. IOS Press, 2010.

[21] Nikos Gorogiannis and Anthony Hunter. Instantiating abstract argumentation with classical logic arguments: Postulates and properties. *Artificial Intelligence*, 175(9-10):1479–1497, 2011.

[22] Peter McBurney and Simon Parsons. Dialogue games for agent argumentation. In *Argumentation in artificial intelligence*, pages 261–280. Springer, 2009.

[23] Hugo Mercier and Dan Sperber. Why do humans reason? Arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

[24] Sanjay Modgil. Dialogical scaffolding for human and artificial agent reasoning. In *AIC*, pages 58–71, 2017.

[25] Sanjay Modgil and Martin Caminada. Proof theories and algorithms for abstract argumentation frameworks. *Argumentation in artificial intelligence*, 105129, 2009.

[26] Sanjay Modgil and Marcello D'Agostino. A fully rational account of structured argumentation under resource bounds. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Main track.*, pages 1841–1847. International Joint Conferences on Artificial Intelligence (IJCAI-20), 2020.

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

[28] Simon Parsons, Peter McBurney, Elizabeth Sklar, and Michael Wooldridge. On the relevance of utterances in formal inter-agent dialogues. In *International Workshop on Argumentation in Multi-Agent Systems*, pages 47–62. Springer, 2007.

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

[30] Henry Prakken. Coherence and flexibility in dialogue games for argumentation. *Journal of logic and computation*, 15(6):1009–1040, 2005.

[31] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[32] Henry Prakken, Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, Leendert van der Torre, et al. *Historical overview of formal argumentation*, volume 1. College Publications, 2018.

[33] Robin Smith. Aristotle's Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Winter 2022 edition, 2022.

[34] Gerard A.W. Vreeswik and Henry Prakken. Credulous and sceptical argument games for preferred semantics. In *European Workshop on Logics in Artificial Intelligence*, pages 239–253. Springer, 2000.

[35] Douglas N. Walton. What is reasoning? What is an argument? *The journal of Philosophy*, 87(8):399–419, 1990.

Received 31 May 2022

A Formal Argumentation Exercise on the Karadžić Trial Judgment

Federico Cerutti
University of Brescia, Italy
federico.cerutti@unibs.it

Yvonne McDermott
Swansea University
yvonne.mcdermottrees@swansea.ac.uk

Abstract

We present the methodology and the results of an application of argumentation theory to map the evidence and arguments as to whether Radovan Karadžić, President of the Serb Republic, possessed the requisite *mens rea* — the knowledge of wrongdoing that constitutes part of a crime — for genocide in Srebrenica. To evaluate the strengths and weaknesses of Trial Chamber's findings in the publicly available judgment, we used argumentation-based techniques available in the CISpaces.org tool. The results of our analysis were submitted to the Appeals Chamber in the same case as an *amicus curiæ* brief, to assist the Appeals Chamber in its consideration of whether the Trial Chamber erred in finding that Karadžić possessed the requisite *mens rea*.

1 Introduction

In this paper — which is an extended version of [9] — we present the methodology and the results of an application of argumentation theory to map the evidence and arguments as to whether Karadžić possessed *mens rea*[1] for genocide in relation to the Srebrenica mass killing. The results of our analysis were submitted to the Mechanism for International Criminal Tribunals as an *amicus curiæ*[2] brief [19] pursuant to Rule

[1] *Mens rea*: the intention or knowledge of wrongdoing that constitutes part of a crime. For the crime of genocide, it must be shown that the perpetrator intended to destroy, in whole or in part, a national, ethnic, racial, or religious group.

[2] *Amicus curiæ*: a non-party in a lawsuit who argues or presents information relevant to the lawsuit.

83 of the MICT Rules of Procedure and Evidence. We based our analysis only on the judgment of Prosecutor v. Radovan Karadžić [15].[3]

On 24th March 2016, Radovan Karadžić was convicted for genocide in Srebrenica by the International Criminal Tribunal for the former Yugoslavia (ICTY). As reported in [15], at least 5,115 men were killed by members of the Bosnian Serb Forces in July 1995 in Srebrenica (Section 3).

The Trial Chamber's finding that the accused possessed the *mens rea* — *i.e.*, the intention and knowledge of wrongdoing that constitutes part of a crime — for genocide in relation to the Srebrenica joint criminal enterprise (JCE) was the subject of academic critique at the time of the Trial judgment, *e.g.*, [26].

Using the argumentation-based techniques available in the CISpaces.org[4] tool [7], reviewed in Section 2, we manually analysed a sub-set of the 2615 pages of [15] to highlight the three reasoning lines that are present in the judgment and that lead to the conclusion that Karadžić possessed the requisite *mens rea.* Of those, two of them might merit further discussions, and the last one relies on a single witness.

Our main contribution is to show that the methodology we propose in Section 4 can be used to show the weakness and strengths of a case — cf. Section 6. This can be of use for the plaintiff, the defendant, but also judges and jurors, as it helps clarifying which elements are proven beyond any reasonable doubt, and which ones are not. This is currently a live issue in international criminal law: one of the authors of this paper argues that "each piece of evidence should be evaluated on its own merits, in light of the other evidence on the record, to determine whether a point has been proven beyond reasonable doubt," [18] as also supported by several judgments. The opposite is often argued, namely that the Trial Chambers should find their decision on the basis of the the accumulation of all the evidence in the case, but without the need to link factual and legal findings to the final decisions.

The submission of our *amicus curiæ* triggered reactions from the academic community interested in international criminal justice, practitioners at the United Nations courts of law, and media. We critically analyse our research and comment on its impact and related work in Section 7.

[3] In the following, we will heavily rely upon the judgment [15] as the only source of information for our analysis and paper.

[4] Although the project's name is CISpaces.org, it is still a research-grade prototype not yet stable enough to be released to the general public, hence **it is not accessible** at `https://cispaces.org`. However, the source code **is available** at `https://github.com/cispaces` and a best-effort service is provided at `https://tiresia.unibs.it/cispaces/`.

2 Background

For this analysis, we used the tools available from the CISpaces project [28] and then further developed in its CISpaces.org version, introduced in [7], that rely on argumentation schemes and computational models of argumentation.

A fundamental concept in computational models of argumentation is the one of defeasible inference *rule*,[5] where a statement (*antecedent*) becomes a (*prima facie*) reason to believe another statement (*consequent*). For instance, "Mary, a witness, says that John committed the fraud" (antecedent) can be seen as a *prima facie* reason to believe that "John committed the fraud" (consequent).

Rules provide the building blocks for the notion of *argument*, that — borrowing from the ASPIC literature [24] — is iterative in the chaining of rules. Statements that are tentatively assumed to hold provide the base case for such an iteration, and thus they are defined as arguments having the statement itself both as premise and as conclusion, where *premises* and *conclusion* are two attributes of the notion of argument. The premises of arguments constructed using this base case also take the name of *ordinary premises* in our approach. Iteratively, an argument requires the existence of a rule whose antecedents are the conclusions of other arguments (*sub-arguments*), and, as a consequent, a statement that becomes the conclusion of this new argument, while its premises are the union of all the premises of its subarguments.

A statement is the contrary of another one when they cannot be both true, albeit they can both be false. A flexible way of using such a notion of *contrariness* [24] is by allowing for a statement to be the contrary of another one. By requiring the vice versa, the two statements would become *contradictory.* We will make use of such a flexibility in the following of our analysis.

The notion of *contrariness* between statements leads to the concept of *defeat* between arguments: an argument defeats another argument if the former *rebuts* or *undermines* the latter. When the conclusion of an argument contradicts the conclusion of another argument, it is the case that the first *rebuts* the second, as well as all the other arguments that have such a second argument as sub-argument. If, instead, the conclusion of an argument contradicts one of the ordinary premises of another one, then the former *undermines* the latter.

Given a set of arguments and defeats between them, we need criteria to assess which arguments collectively survive the defeats and thus can provide a reasonable viewpoint (or *extension*) based on the statements and the rules that we were considering. Such criteria usually consider *conflict-freeness*, *i.e.*, the absence of defeats

[5] We will not make use of strict rules in this work.

within the viewpoint; *admissibility*, *i.e.*, if an argument in the viewpoint is defeated by a second argument, the latter must in turn be defeated by a third argument also in the viewpoint; and *maximality*, *i.e.*, a viewpoint cannot be a strict subset of another viewpoint. Multiple viewpoints can exist for the same set of arguments and the defeats between them: two equally reliable witnesses, each providing one reason for contradictory conclusions, lead to the situation that each of the two arguments per se is a reasonable viewpoint, hence there are two of them. In this case, if an argument belongs to at least one viewpoint, it is said to be *credulously accepted*. If, instead, an argument belongs to all the viewpoints, it is said to be *sceptically accepted*. In the following we will be making use of this notion of sceptically acceptance in connection with the principle of *proof beyond a reasonable doubt*.

CISpaces.org provides a convenient visual language and an effective Human-Machine Interface for argumentation mapping. It builds on top of the Argument Interchange Format AIF [11] that specifies a graph structure composed of two types of nodes connected by links. The nodes can be either information (in the following identified by squared boxes) or scheme nodes (in the following identified by round boxes). Information nodes define the antecedents and consequents that we will be making use in the generation of arguments. Scheme nodes can be either rule of inference applications or conflict applications. A rule of inference application provides the connection between antecedents and a consequent: if one or more information nodes are linked to an inference node, and the latter is in turn linked to another information node, we will interpret this sub-graph as an inference rule. Conflict nodes, instead, express the contrariness relationship between two inference nodes: once again, links here are directed too.

2.1 Argumentation Schemes

Argumentation schemes [34] are abstract reasoning patterns commonly used in everyday conversational argumentation, legal, scientific argumentation, etc. Schemes have been derived from empirical studies of human arguments and debate. Each scheme has a set of critical questions that represent standard ways of critically probing into an argument to find aspects of it that are open to criticism. For instance, the following is the scheme for arguments from *evidence to hypothesis* [34]:

Major Premise: If A (a hypothesis) is true, then B (a proposition reporting an event) will be observed to be true.

Minor Premise: B has been observed to be true in a given instance.

Conclusion: Therefore, A is true.

CQ1: Is it the case that if A is true, then B is true?

CQ2: Has B been observed to be true?

CQ3: Could there be some reason why B is true, other than its being true because of A being true?

The other argumentation schemes used in this analysis are: the *abductive argumentation scheme*; the *argumentation from cause to effect*; the *argumentation from witness testimony*; and the *argumentation from (popular) opinion* [34].

An abductive argument aims at identifying a chain of inferences to fill in the gaps in the line of reasoning towards a given conclusion. It often involves identifying reasonable causes for a given outcome. It can be criticised on the basis of discussing alternative causes or on the actual explanatory power of the identified probable cause.

Connected to the previous scheme, an argument from cause to effect link two phenomena, A and B, in a possible causal link, hence stating that if A occurs, then B will (might) occur. This is also the main element for critique, namely how strong is such a causal generalisation?

Moving towards schemes widely used in trials, witness testimony is a strong argument when there is no direct access to the facts. In this case, to evaluate it one needs to rely upon comparison to other available evidence and evaluation of its consistency, both internal and external. It is worth mentioning that for this work we did not have access to the original set of testimonies as they are not available *verbatim* in the judgement.

An appeal to (popular) opinion may refer to just a majority in a reference group. In general, the argument from popular opinion may be undermined under three aspects: the actual agreement of the majority with the proposition; the weakness of the argument itself when used to prove the truth of a proposition; and the link with the true opinion.

2.2 Charting Arguments, Mapping into ASPIC+, and Evaluating them

CISpaces.org [28, 7] enables a user to draw a directed graph representing an argument map, which can then be compiled into an ASPIC+ theory for automatic reasoning. In particular, an argument map is a directed graph ($\mathsf{WDG} = \langle N, \rightsquigarrow \rangle$) based on the AIF format [11], thus with two distinct types of nodes: information nodes (or I-nodes) and scheme nodes (or S-nodes). S-nodes can be either rule of inference application (RA-nodes), or conflict application (CA-nodes), respectively represented as *Pro* and *Con* nodes. Pro links can be further labelled with the argumentation scheme they instantiate. In a WDG, nodes are connected by edges whose semantics are implicitly defined by their use [11].

Similarly to [24, 17], a WDG can be mapped into an ASPIC+ system [20]. Assume a logical language $\mathcal{L}$, and a set of *strict* or *defeasible* inference rules — resp. $\varphi_1, \ldots, \varphi_n \longrightarrow \varphi$ and $\varphi_1, \ldots, \varphi_n \Longrightarrow \varphi$. A strict rule inference always holds — *i.e.*, if the *antecedents* $\varphi_1, \ldots, \varphi_n$ hold, the *consequent* φ holds as well — while a defeasible inference "usually" holds.

An *argumentation system* is as tuple $AS = \langle \mathcal{L}, \mathcal{R}, ^-, \nu \rangle$ where:

- $^- : \mathcal{L} \mapsto 2^{\mathcal{L}}$ is a contrariness function s.t. if $\varphi \in \overline{\psi}$ and: $\psi \notin \overline{\varphi}$, then φ is a *contrary* of ψ; $\psi \in \overline{\varphi}$, then φ is a *contradictory* of ψ ($\varphi = -\psi$);
- $\mathcal{R} = \mathcal{R}_d \cup \mathcal{R}_s$ is a set of strict ($\mathcal{R}_s$) and defeasible ($\mathcal{R}_d$) inference rules such that $\mathcal{R}_d \cap \mathcal{R}_s = \emptyset$;
- $\nu : \mathcal{R}_d \mapsto \mathcal{L}$, is a partial function.[6]

A *knowledge base* $\mathcal{K}$ in an AS is a set of *axioms* $\mathcal{K}_n$ that cannot be attacked, and *ordinary premises* $\mathcal{K}_p$ that can be attacked, *i.e.*, $\mathcal{K}_n \cup \mathcal{K}_p = \mathcal{K} \subseteq \mathcal{L}$.

Building upon the notion of an argumentation system and of a knowledge base, an *argumentation theory* is a pair $AT = \langle AS, \mathcal{K} \rangle$.

To map a WDG into an ASPIC+ system, let us assume that:

- $P \subseteq N$ is the set of I-nodes, where each I-node in the graph is written p_i;
- ℓ_{type}, with $type = \{Pro, Con\}$, refers to a S-node;
- $[p_1, \ldots, p_n \rightsquigarrow \ell_{Pro} \rightsquigarrow p_\phi]$ indicates an inference rule, where $p_1, \ldots, p_q$ are parent nodes of the S-node ℓ_{Pro}, and p_ϕ is a child of ℓ_{Pro};
- conflict schemes can be either $[p_1 \rightsquigarrow \ell_{Con} \rightsquigarrow p_2]$ or $[p_1, \ldots, p_n \rightsquigarrow \ell_{Con} \rightsquigarrow p_\phi]$.

For this work, we make use of a subset of the ASPIC+ system: in particular, we will use neither strict rules nor preferences.

Given a $\mathsf{WDG} = \langle N, \rightsquigarrow \rangle$, its corresponding ASPIC+ system $AS = \langle \mathcal{L}, \mathcal{R}, ^-, \nu \rangle$ is such that:

- $\forall p \in P \subseteq N$, $p \in \mathcal{L}$;

[6] Informally, $\nu(r)$ is a wff in $\mathcal{L}$ which says that the defeasible rule r is applicable. However, we will not make use of this feature in the following.

- $\mathcal{R} = \mathcal{R}_s \cup \mathcal{R}_d$ with $\mathcal{R}_s = \emptyset$ and $\forall [p_1, \ldots, p_n \rightsquigarrow \ell_{Pro} \rightsquigarrow p_\phi]$, $p_1, \ldots, p_n \Rightarrow p_\phi \in \mathcal{R}_d$;
- $\forall [p_1 \rightsquigarrow \ell_{Con} \rightsquigarrow p_2]$, $p_1 \in \overline{p_2}$;
- $\forall [p_1, \ldots, p_n \rightsquigarrow \ell_{Con} \rightsquigarrow p_\phi]$, is mapped as $p_1, \ldots, p_n \Rightarrow p_h \in \mathcal{R}_d$ and $p_h \in \overline{p_\phi}$;

and the knowledge base $\mathcal{K}_n \cup \mathcal{K}_p = \mathcal{K} \subseteq \mathcal{L}$ is such that, given $[p_1, \ldots, p_n \rightsquigarrow \ell_{Pro} \rightsquigarrow p_\phi]$, $\forall p_i \in \{p_1, \ldots, p_n\}$, if p_i is not a conclusion of any inference rule $\nexists [\ell_{Pro} \rightsquigarrow p_i] \in \rightsquigarrow$, $p_i \in \mathcal{K}_p$. In addition, assume $\mathsf{WDG}' = \langle N', \rightsquigarrow' \rangle$ a subset of WDG— *i.e.*, such that $N' \subseteq N$ and $\rightsquigarrow' \subseteq \rightsquigarrow$—containing only a single cycle of inference schemes— *i.e.*, analogous to the case $p_i \Rightarrow p_i$—then $\forall p_i \in P' \subset N'$, if $[\ell_{Pro} \rightsquigarrow p_i]$, $[p_i \rightsquigarrow \ell_{Pro}] \in \rightsquigarrow'$, then $p_i \in \mathcal{K}_p$ is an ordinary premise.

Following [20], an *argument* **a** on the basis of a $AT = \langle AS, \mathcal{K} \rangle$, $AS = \langle \mathcal{L}, \mathcal{R}, ^{-}, \nu \rangle$ is:

1. φ if $\varphi \in \mathcal{K}$ with: $\texttt{Prem}(\mathbf{a}) = \{\varphi\}$; $\texttt{Conc}(\mathbf{a}) = \varphi$; $\texttt{Sub}(\mathbf{a}) = \{\varphi\}$; $\texttt{Rules}(\mathbf{a}) = \texttt{DefRules}(\mathbf{a}) = \emptyset$; $\texttt{TopRule}(\mathbf{a}) =$ undefined.
2. $\mathbf{a}_1, \ldots, \mathbf{a}_n \longrightarrow / \Longrightarrow \psi$ if $\mathbf{a}_1, \ldots, \mathbf{a}_n$, with $n \geq 0$, are arguments such that there exists a strict/defeasible rule $r = \texttt{Conc}(\mathbf{a}_1), \ldots, \texttt{Conc}(\mathbf{a}_n) \longrightarrow / \Longrightarrow \psi \in \mathcal{R}_s/\mathcal{R}_d$. $\texttt{Prem}(\mathbf{a}) = \bigcup_{i=1}^{n} \texttt{Prem}(\mathbf{a}_i)$; $\texttt{Conc}(\mathbf{a}) = \psi$; $\texttt{Sub}(\mathbf{a}) = \bigcup_{i=1}^{n} \texttt{Sub}(\mathbf{a}_i) \cup \{\mathbf{a}\}$; $\texttt{Rules}(\mathbf{a}) = \bigcup_{i=1}^{n} \texttt{Rules}(\mathbf{a}_i) \cup \{r\}$; $\texttt{DefRules}(\mathbf{a}) = \{d \mid d \in \texttt{Rules}(\mathbf{a}) \cap \mathcal{R}_d\}$; $\texttt{TopRule}(\mathbf{a}) = r$

An argument can be attacked in its premises (*undermining*) or its conclusion (*rebuttal*). Since we will not use the preference ordering between arguments, we will omit it from the definition. Similarly for the notion of *undercut* on the inference rule (cf. [20]).

Given **a** and **b** arguments, **a** *defeats* **b** iff **a** *successfully rebuts* or *successfully undermines* **b**, where: **a** *successfully rebuts* **b** (on **b**′) iff $\texttt{Conc}(\mathbf{a}) \notin \overline{\varphi}$ for some $\mathbf{b}' \in \texttt{Sub}(\mathbf{b})$ of the form $\mathbf{b}''_1, \ldots, \mathbf{b}''_n \Longrightarrow -\varphi$;**a** *successfully undermines* **b** (on φ) iff $\texttt{Conc}(\mathbf{a}) \notin \overline{\varphi}$, and $\varphi \in \texttt{Prem}(\mathbf{b}) \cap \mathcal{K}_p$.

An *argumentation framework* (*AF*) [13] is a pair $\Delta = \langle \mathcal{A}, \rightarrow \rangle$ where $\mathcal{A}$ is a set of arguments[7] and $\rightarrow \subseteq \mathcal{A} \times \mathcal{A}$ is an attack relation. We denote with $\mathbf{a}_2 \rightarrow \mathbf{a}_1$ when $\langle \mathbf{a}_2, \mathbf{a}_1 \rangle \in \rightarrow$.

An *AF* $\langle \mathcal{A}, \rightarrow \rangle$ is the *abstract argumentation framework defined by* $AT = \langle AS, \mathcal{K} \rangle$, $AS = \langle \mathcal{L}, \mathcal{R}, ^{-}, \nu \rangle$ if $\mathcal{A}$ is the set of all finite arguments constructed from $\mathcal{K}$ (as above); and $\rightarrow$ is the defeat relation on $\mathcal{A}$.

[7] In this paper we consider only *finite* sets of arguments: see [3] for a discussion on infinite sets of arguments.

Given an AF $\Delta = \langle \mathcal{A}, \rightarrow \rangle$: a set $S \subseteq \mathcal{A}$ is a *conflict-free* set of Δ if $\nexists\ \mathbf{a}_1, \mathbf{a}_2 \in S$ s.t. $\mathbf{a}_1 \rightarrow \mathbf{a}_2$; an argument $\mathbf{a}_1 \in \mathcal{A}$ is *acceptable* with respect to a set $S \subseteq \mathcal{A}$ of Δ if $\forall \mathbf{a}_2 \in \mathcal{A}$ s.t. $\mathbf{a}_2 \rightarrow \mathbf{a}_1$, $\exists\ \mathbf{a}_3 \in S$ s.t. $\mathbf{a}_3 \rightarrow \mathbf{a}_2$; a set $S \subseteq \mathcal{A}$ is an *admissible* set of Δ if S is a conflict-free set of Δ and every element of S is acceptable with respect to S.

A set of argument $S \subseteq \mathcal{A}$ is a *preferred extension* if and only if it is a maximal (with respect to set inclusion) admissible set.

An argument is *skeptically accepted* with regards to preferred semantics if and only if it belongs to *each* preferred extension. Checking this is a problem that lies at the second level of the polynomial hierarchy [14], hence the need — in general — for efficient implementations [10].

3 Karadžić and Srebrenica

What follows is a short historical summary of the events that lead to the Srebrenica massacre as reported in [15]. Figure 1 summarises the timeline of the most relevant events for our analysis starting from 6 July 1995.

The Socialist Republic of Bosnia and Herzegovina (SRBiH) was one of the six republics that once constituted the Socialist Federal Republic of Yugoslavia (SFRY): unlike the other republics, it possessed no single majority ethnic grouping. One of its political parties, the Serbian Democratice Party or SDS — led by Radovan Karadžić, campaigned to establish separate Serbian institutions. Following a plebiscite held on 9 and 10 November 1991, an autonomous Serb Republic (Republika Srpska) was proclaimed in 1992.

Among other key personnel within the Serb Republic, Radovan Karadžić served as President and Supreme Commander of the Bosnian Serb Army (VRS). Tomislav Kovač was the Assistant Minister of the Ministry of Intern (MUP), and the acting Ministry from September 1993 until January 1994. Ratko Mladić served as Commander of Main Staff, the highest operative body of the VRS. His assistant commander for Security Administration was Ljubiša Beara, with duties of management of the main staff of the Military Police, as well as co-ordinating with the bodies of the Ministry of the Interior. Momir Nikolić was Chief of the Security and Intelligence Organ, which was responsible for issues of security in the corps composing the VRS, including the arrest and detention of prisoners of war and other persons.

When in 1992 the population of Republic of Bosnia and Herzegovina voted for independence from the Socialist Federal Republic of Yugoslavia in a referendum, forces of the Serb Republic attacked different parts of the Republic of Bosnia and Herzegovina, whose state administration effectively ceased to function having lost

6th	Shelling of Srebrenica began
11th, Afternoon	Srebrenica has fallen Karadžić appoints Deronjić as Civilian Commissioner for Srebrenica
11th, Night	A column of Bosnian Muslim men tried to escape by walking in a northwesterly direction towards the safe haven of Tuzla
12th, Morning	Shelling of the column began
12th, Afternoon	Large numbers of the members of the columns surrendered
13th, Morning	Groups of detainees from the column marched towards the Kravica Warehouse
13th, 1630h–Night Kravica Warehouse	One of the Bosnian Muslim detainees took away the rifle of a soldier and shot him dead: other soldiers started shooting at the detainees in response. Others shot at the detainees with machine guns and automatic rifles. Hand-grenades were thrown in the warehouse through the windows. By nightfall, between 755 and 1,016 Bosnian Muslim men were killed.
13th, 1700h–1840h Pale	Karadžić had an hour-long conversation on the phone during which he was briefed by General Maladić that Srebrenica "[wa]s done."
13th, 2010h	Intercepted call between Deronjić and Karadžić through an intermediary **:** Deronjić, the President is asking how many thousands? **D:** About two for the time being. [. . .] **:** Deronjić, the President says: "All the goods must be placed inside the warehouses before twelve tomorrow." **D:** Right. **:** Deronjić, not in the warehouses over there, but somewhere else.
13th, around 2100h Bratunac SDS Office	Deronjić ordered to bury the detainees that had been killed at the Kravica Warehouse in a bauxite mine in Milići.
14th Just after midnight	Momir Nikolić drove Beara to the Bratunac SDS office, where Beara met with Deronjić and Vasić. Beara and Deronjić argued about where the Bosnian Muslim men were to be executed, as Beara insisted that he had instructions from his "boss" that the detainees were to remain in Bratunac, and Deronjić countered that the Accused had instructed him that all detainees in Bratunac should be transferred to Zvornik. Eventually, Beara and Deronjić agreed that the detainees would indeed be transferred to Zvornik. Detainees began to be transferred to the first of four detention sites in Zvornik.
14th, 1240h–1310h	Karadžić met with Deronjić alone.
14th, afternoon, after 1310h	Karadžić and Deronjić met with Srebrenica represetatives for about four hours.
14th, 2245h–2310h	Kovač met with Karadžić after touring Srebrenica, and the Bratunac and Zvornik areas on 13 and 14 July.
15th, 0035h–0125h	Bajagić — who has a substantive knowledge of the events in Srebrenica being the technical service procurement clerk — met with Karadžić.
16th	By the end of 16th July 1995, at least 3365 Bosnian Muslims men were killed.

Figure 1: Timeline of some of the most relevant events related to the Srebrenica mass killing. All dates refer to the month of July 1995.

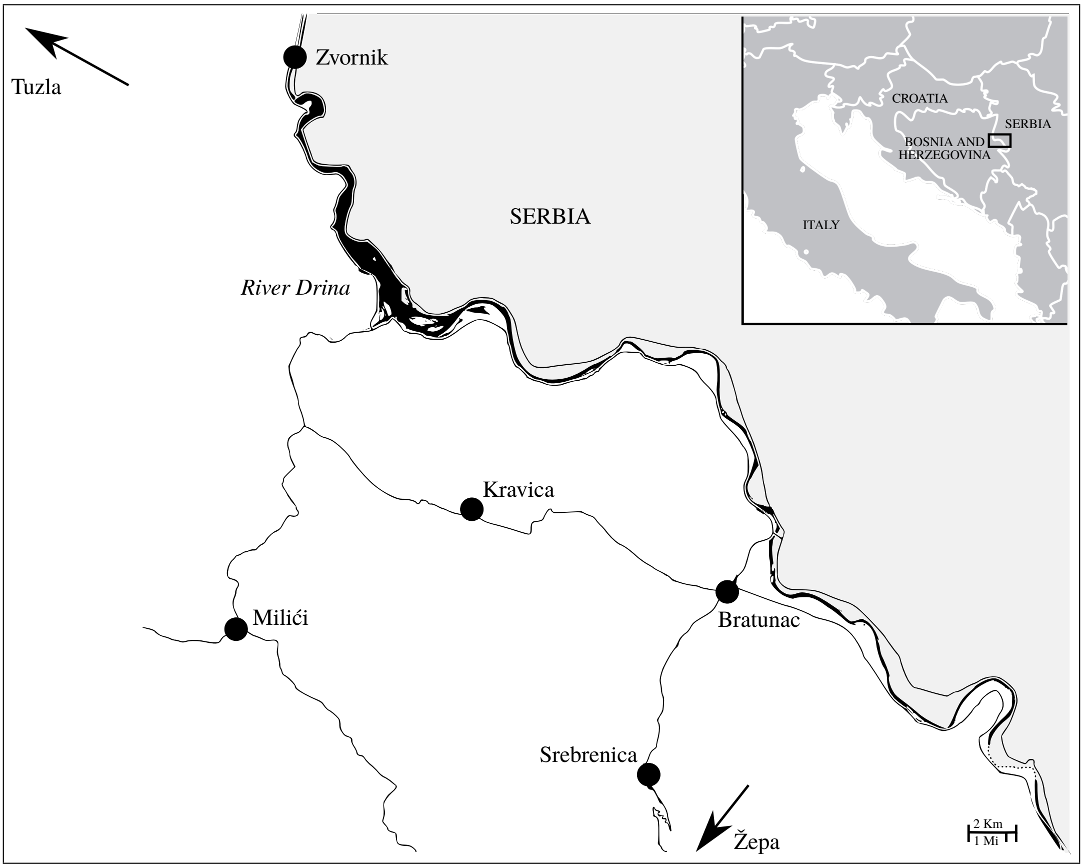


Figure 2: Relevant locations next to the Drina river. In white in the main chart the Socialist Republic of Bosnia and Herzegovina.

control over the entire territory. The Assembly of the Serb Republic adopted the strategic goal to eliminate the border with Serbia: Srebrenica — a town with a majority of Bosnian Muslims — was close to that border (Figure 2).

In late June 1995, Karadžić gave a combat assignment that led to an offensive against Srebrenica and ultimately to the killing of at least 5,115 Bosnian Muslim men.

4 Methodology

The goal of our *amicus curiae* brief [19] was to identify the precise factual and inferential bases for the Trial Chamber's findings of Karadžić's genocidal intent in the Trial Chamber judgment, and to elucidate the forms of reasoning that led to these conclusions. We limited our analysis to the reasoning process that can be fathomed from the Trial Chamber's judgment. As such, we did not analyse issues such as the reliability of witnesses or evidence since they are the purview of the Trial Chamber alone, and also because the entire set of evidence used by the Trial Chamber is not publicly available.

In the present case, Karadžić's *mens rea* is an element of the offence of genocide in Srebrenica, as genocide requires each member of the joint criminal enterprise to be knowledgeable of the *dolus specialis* of the principal perpetrator. The *material facts* upon which proof of *mens rea* hinged were the Trial Chamber judgment's findings on: (1) Karadžić's knowledge of the expansion of the plan to remove Bosnian Muslims from Srebrenica to include the killing of men and boys, hence Karadžić sharing the intent to destroy the Bosnian Muslims in Srebrenica; and (2) his active involvement in the killings.

Following [21], we manually and in full agreement identified the arguments — and their general argumentation schemes when possible — that the Trial Chamber put forward in [15] related to the two hypotheses (*Was Karadžić aware of the intent to kill the detainees?* and *Was Karadžić actively involved in the oversight of the killing of the men and boys?*, *cf.*, Section 5), together with (1) those instances of schemes for which not all critical questions have been satisfactorily addressed; (2) and particular facts that seem missing but necessary to expose the entire line of reasoning, labelling them with **Unstated**. In those cases, we did not include an analysis of critical questions for the inferences based on such unstated pieces of information, as it would be a detour from the purpose of this work.

It is worth noticing that the text proved resistant to attempts of automatic analysis. This is also evident in the graphical charting of our analysis (Figure 3), where we consider pieces of information spanning more than 210 pages (Para 5312 fn. 18025 [15, p. 2203] to Para 5808 [15, p. 2413]), in addition to historical information scattered around the entire document. For instance, the information that Srebrenica has fallen on 11 July 1995 has been presented in Para 5033 [15, p. 2079], 331 pages before being used in an argument to support the hypothesis.

5 Results

5.1 Was Karadžić aware of the intent to kill the detainees?

Figure 3 depicts our understanding of the reasoning lines that the Trial Chamber describes in its judgment [15] in support of the hypothesis that Karadžić was aware of the intent to kill the detainees. This is also the conclusion of the skeptically accepted

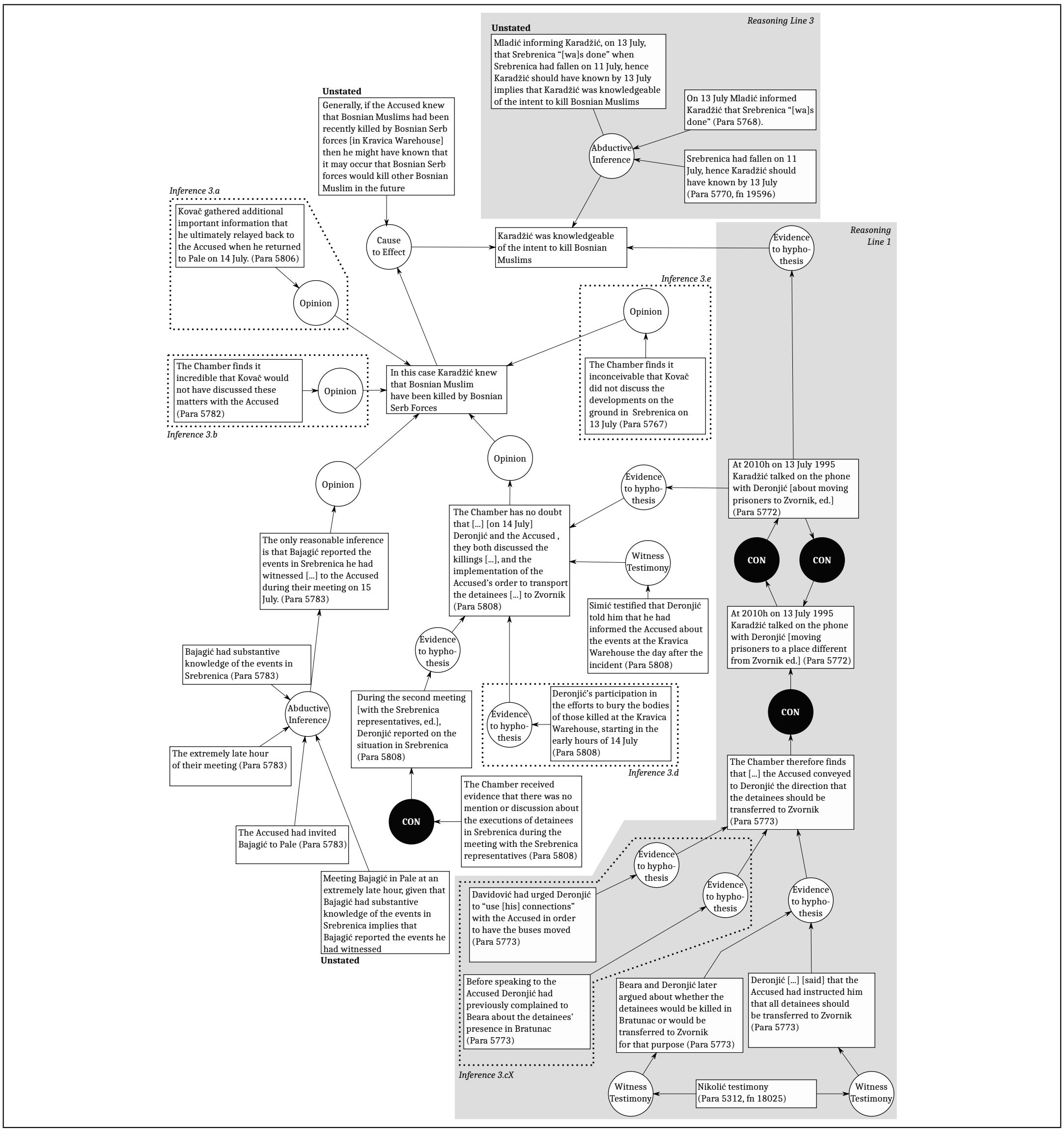

Reasoning Line 3
Unstated
Mladić informing Karadžić, on 13 July, that Srebrenica "[wa]s done" when Srebrenica had fallen on 11 July, hence Karadžić should have known by 13 July implies that Karadžić was knowledgeable of the intent to kill Bosnian Muslims
On 13 July Mladić informed Karadžić that Srebrenica "[wa]s done" (Para 5768).
Abductive Inference
Srebrenica had fallen on 11 July, hence Karadžić should have known by 13 July (Para 5770, fn 19596)
Unstated
Generally, if the Accused knew that Bosnian Muslims had been recently killed by Bosnian Serb forces [in Kravica Warehouse] then he might have known that it may occur that Bosnian Serb forces would kill other Bosnian Muslim in the future
Inference 3.a
Kovač gathered additional important information that he ultimately relayed back to the Accused when he returned to Pale on 14 July. (Para 5806)
Cause to Effect
Karadžić was knowledgeable of the intent to kill Bosnian Muslims
Reasoning Line 1
Evidence to hypo-thesis
Opinion
Inference 3.e
Opinion
The Chamber finds it incredible that Kovač would not have discussed these matters with the Accused (Para 5782)
Opinion
In this case Karadžić knew that Bosnian Muslim have been killed by Bosnian Serb Forces
The Chamber finds it inconceivable that Kovač did not discuss the developments on the ground in Srebrenica on 13 July (Para 5767)
Inference 3.b
Opinion
At 2010h on 13 July 1995 Karadžić talked on the phone with Deronjić [about moving prisoners to Zvornik, ed.] (Para 5772)
Opinion
Evidence to hypo-thesis
The Chamber has no doubt that [...] [on 14 July] Deronjić and the Accused , they both discussed the killings [...], and the implementation of the Accused's order to transport the detainees [...] to Zvornik (Para 5808)
The only reasonable inference is that Bajagić reported the events in Srebrenica he had witnessed [...] to the Accused during their meeting on 15 July. (Para 5783)
CON
CON
Witness Testimony
Simić testified that Deronjić told him that he had informed the Accused about the events at the Kravica Warehouse the day after the incident (Para 5808)
At 2010h on 13 July 1995 Karadžić talked on the phone with Deronjić [moving prisoners to a place different from Zvornik ed.] (Para 5772)
Evidence to hypo-thesis
Bajagić had substantive knowledge of the events in Srebrenica (Para 5783)
Abductive Inference
During the second meeting [with the Srebrenica representatives, ed.], Deronjić reported on the situation in Srebrenica (Para 5808)
Evidence to hypo-thesis
Deronjić's participation in the efforts to bury the bodies of those killed at the Kravica Warehouse, starting in the early hours of 14 July (Para 5808)
CON
The extremely late hour of their meeting (Para 5783)
Inference 3.d
The Chamber therefore finds that [...] the Accused conveyed to Deronjić the direction that the detainees should be transferred to Zvornik (Para 5773)
The Chamber received evidence that there was no mention or discussion about the executions of detainees in Srebrenica during the meeting with the Srebrenica representatives (Para 5808)
CON
The Accused had invited Bajagić to Pale (Para 5783)
Evidence to hypo-thesis
Meeting Bajagić in Pale at an extremely late hour, given that Bajagić had substantive knowledge of the events in Srebrenica implies that Bajagić reported the events he had witnessed
Unstated
Davidović had urged Deronjić to "use [his] connections" with the Accused in order to have the buses moved (Para 5773)
Evidence to hypo-thesis
Evidence to hypo-thesis
Before speaking to the Accused Deronjić had previously complained to Beara about the detainees' presence in Bratunac (Para 5773)
Beara and Deronjić later argued about whether the detainees would be killed in Bratunac or would be transferred to Zvornik for that purpose (Para 5773)
Deronjić [...] [said] that the Accused had instructed him that all detainees should be transferred to Zvornik (Para 5773)
Inference 3.cX
Witness Testimony
Nikolić testimony (Para 5312, fn 18025)
Witness Testimony

Figure 3: Analysis of arguments in [15] in favour of the hypothesis that Karadžić was knowledgeable of the intent of killing Bosnian Muslims men. Each "Para" reference refers to a paragraph of [15]. Names and events are introduced in Section 3, except for Milorad Davidović, who was a senior official in the MUP and, later on, a witness. Squared boxes are claims; white circles are *Pro* nodes, labelled with the argumentation schemes they refer to; while black circles are *Con* nodes. Dotted areas identify inferences for which there are critical questions that were not explicitly addressed in [15]. Three reasoning lines are highlighted as they are referred to in Section 5.1.

arguments with regards to preferred semantics (cf. Section 2), quite unsurprisingly given the scarce number of conflicts: this is expected since the judgment does not record each exchange of arguments between the defence and the prosecution.

There are three main lines of reasoning in favour of this conclusion. The first one is based upon Nikolić's testimony that he overheard Deronjić saying that the accused had instructed Deronjić that all detainees should be transferred to Zvornik, cf. Figure 1, 14th July 1995, just after midnight. This testimony gives reasons to the chamber to refute the alternative explanation — highlighted by the defence and reported in the judgment — that Karadžić was referring to a place different from Zvornik in the intercepted conversation with Deronjić, cf. Figure 1, 13th July 1995, 2010h. In this line of reasoning, the Chamber decided also to link additional pieces of information (*Inference 3.cX* of Figure 3), as supporting the conclusion that Karadžić ordered that detainees should be transferred to Zvornik, such as a complaint to Beara by Deronjić about the presence of detainees in Bratunac. However, for those facts, it appeared that the Chamber did not consider some relevant critical question, *e.g.*, *Is there any other reasonable explanation for why Deronjić had previously complained to Beara about the detainees' presence in Bratunac, other than it being true because Karadžić conveyed to Deronjić the direction that the detainees should be transferred to Zvornik?* Despite those additional pieces of information (*Inference 3.cX* of Figure 3), this line of reasoning does not rely on unstated findings or pieces of information for which critical questions have not explicitly been answered. It will be recalled that we methodologically chose not to assess the reliability of Nikolić's testimony as we did not have access to the entire trial records and besides, credibility and reliability are adjudged on a number of factors, including the witness's demeanour and/or evasiveness in the witness box, which would be difficult to determine from a transcript of proceedings [12].

A second line of reasoning justifying the hypothesis is based on Simić's testimony that Deronjić told him that he had informed Karadžić about the events at the Kravica Warehouse the day after the incident, in conjunction with the unstated assumption that if the accused knew that Bosnian Muslims had been recently killed by Bosnian Serb forces (in Kravica Warehouse), then he might have been known that it may occur that Bosnian Serb forces would kill other Bosnian Muslims in the future. It could be questioned whether all relevant critical questions find an answer in the judgment [15], with regard to *Inferences 3.a, 3.b, 3.d, 3.e* of Figure 3. For instance, *what evidence supported the finding that Kovač relayed back additional important information to Karadžić when he returned to Pale on 14 July?* (*Inference 3.a*, Figure 3); or *what evidence supported the finding that Kovač discussed these matters with Karadžić?* (*Inference 3.b*, Figure 3).

A third line of reasoning is based on an abductive inference with the unstated premise that Mladić informed Karadžić, on 13 July, that Srebrenica "[wa]s done." The Trial Chamber appears to have concluded that, given that Srebrenica had fallen on 11 July, Karadžić would have known this by 13 July. From that unstated inference, it drew a further inference that the conversation implied that Karadžić knew of the intent to kill the Bosnian Muslims of Srebrenica.

5.2 Was Karadžić actively involved in the oversight of the killing of the men and boys?

Figure 4 depicts our understanding of the Chamber's line of reasoning in concluding that the accused was actively involved in the oversight of the killing of the men and boys after the 13 July conversation. This is also the conclusion of the skeptically accepted arguments with regard to preferred semantics.

However, for each inference line supporting this conclusion, either necessary premises are unstated (hence left to the reader to assume), or at least one relevant critical question is not explicitly answered, namely:

- Regarding *Inference 4.a*: *Was there any other reasonable explanation for the statements that 'several thousand fighters did manage to get through' and 'we were not able to encircle the enemy and destroy them', other than that they were an illustration of regret that the corridor had been opened on 16 July?*

- Regarding *Inference 4.b*:

    1. *Was it established as true that the accused received the request for access from international organisations?*
    2. *Was there any other reasonable explanation for why international organisations were not granted access to Srebrenica, other than this being true because the accused was actively involved in the oversight of the killings after the 13 July conversation?*

- Regarding *Inference 4.c*: *Was there any other reasonable explanation for why, in late July and early August 1995, the accused promoted and praised Mladić, Živanović, and Krstić, other than it being true because the accused was actively involved in the oversight of the killings after the 13 July conversation?*

- Regarding *Inference 4.d*: *Was there any other reasonable explanation for why no investigations were ever carried out, other than it being true because the accused was actively involved in the oversight of the killings after the 13 July conversation?*

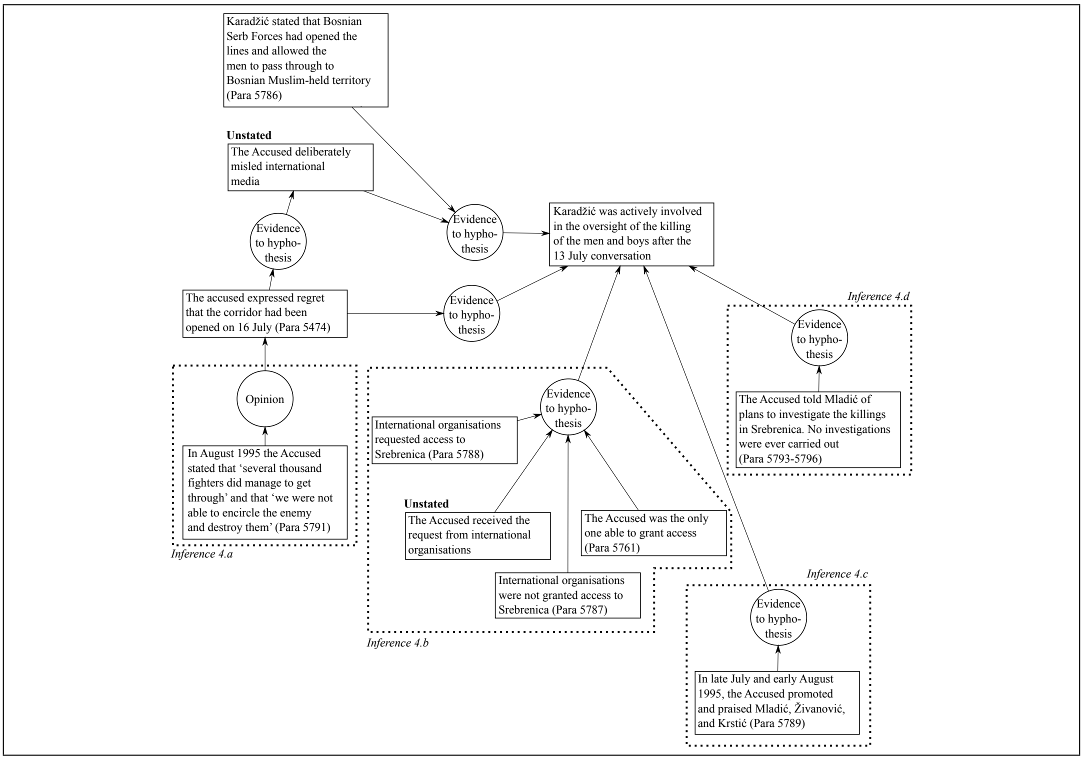

Karadžić stated that Bosnian Serb Forces had opened the lines and allowed the men to pass through to Bosnian Muslim-held territory (Para 5786)
Unstated
The Accused deliberately misled international media
Evidence to hypho- thesis
Evidence to hypho- thesis
Karadžić was actively involved in the oversight of the killing of the men and boys after the 13 July conversation
The accused expressed regret that the corridor had been opened on 16 July (Para 5474)
Evidence to hypho- thesis
Inference 4.d
Evidence to hypho- thesis
Opinion
Evidence to hypho- thesis
International organisations requested access to Srebrenica (Para 5788)
The Accused told Mladić of plans to investigate the killings in Srebrenica. No investigations were ever carried out (Para 5793-5796)
In August 1995 the Accused stated that 'several thousand fighters did manage to get through' and that 'we were not able to encircle the enemy and destroy them' (Para 5791)
Unstated
The Accused received the request from international organisations
The Accused was the only one able to grant access (Para 5761)
Inference 4.a
International organisations were not granted access to Srebrenica (Para 5787)
Inference 4.c
Evidence to hypho- thesis
Inference 4.b
In late July and early August 1995, the Accused promoted and praised Mladić, Živanović, and Krstić (Para 5789)

Figure 4: Analysis of arguments in [15] in favour of the hypothesis that Karadžić was actively involved in the oversight of the killing of the men and boys after the 13 July conversation. Each "Para" reference refers to a paragraph of [15]. Names and events are introduced in Section 3, except for Milorad Davidović, who was a senior official in the MUP and later on a witness. Squared boxes are claims; white circles are *Pro* nodes, labelled with the argumentation schemes they refer to; while black circles are *Con* nodes. Dotted areas identify inferences for which there are critical questions that were not explicitly addressed in [15]. Three reasoning lines are highlighted as they are referred to in Section 5.2.

6 Discussion

On 21 February 2018, we sent to the United Nations Mechanism for International Criminal Tribunals (MICT) a request for leave to make submissions as *amicus curiæ* pursuant to Rule 83 of the MICT Rules of Procedure and Evidence [19].

Given the overall results (cf. Section 5), we believed that our analysis was, on balance, probably more helpful to the prosecution than to the defence in the appeal, insofar as it illustrated that, while some inferential steps could have been explicated in greater detail, the Trial Chamber's reasoning was broadly sound.

This is clearly not the first attempt to apply formal argumentation to judicial findings. In [31], Verheij introduces the notion of *automated argument assistance* which is in spirit very close to our work here, as it explicitly aims at drafting and testing of court pleadings. Walton [35] provides an extensive account of argumentation in legal systems, and argues for a *new method* for legal argumentation which, among others, includes the use of "argument diagramming to map out the network of inference in a given case" [35, p. 323]. For completeness of discussion, authoritative colleagues criticise the use of argument diagramming, notably van Gelder in [30]. In reflecting on his experience, he notices that argument diagramming might not serve well the purpose of deliberation, possibly because deliberation is a dialectical activity rich in nuances. However, he did not consider deliberation activities where the incentives for a proper epistemic investigation are significant, such as writing a judgment for an international criminal case.

The work by Walton on legal argumentation [35] and in general on argumentation schemes — summarised in [34] — motivated researchers in deriving computational models, thus building on the tradition initiated by Verheij [31]. Bex et.al. [6] expanded on the idea of using argumentation schemes for providing a formal account of reasoning, and subsequently in [5] they also considered the advantages of merging it with storytelling. The latter also takes into consideration the different positions of the plaintiff and the defendant, which is also the case of [23] — where a formal dialogue system is used as a formalisation tool — and [22], where ASPIC+ is used for formalising legal case-based reasoning. In contrast to previous approaches, we considered explicitly the role of argumentative semantics using skeptical acceptance according to preferred semantics as a proxy for the *beyond any reasonable doubt* standard of proof. This is clearly questionable, but it looks a reasonable approximation as it is a rather conservative choice, although it might be a little difficult to explain to non-experts. It, however, builds upon the assumption that all the other reasonable alternatives have been explored and correctly mapped. Further analysis using other semantics are already planned for future work, as well as a deeper comparison with the ANGELIC methodology [1], in particular after the recent paper [2] showing a correspondence with ASPIC+.

We also feel that there is very little we can add to van Gelder's observations in [29], where he analyses some legal arguments. His comments strongly resonate with us, as we also experienced "little use of verbal indicators of logical structure, and often use obscure or vague indicators" [29]. We also encountered incomplete arguments, with text scattered across the document, and possibly serving multiple purposes.[8] Although far from providing an off-the-shelf support tool, in retrospect our analysis would have significantly benefit from (1) entity-relations extractors, *e.g.*, [27] and (2) topic modelling system, *e.g.*, [36], which together might transform a static PDF document into a database that can be queried.

In contrast to previous approaches, we considered a case under discussion at ICTY offering the results of our analysis as an *amicus curiæ* brief to the Appeal Chambers. It unfortunately denied admissibility of our application on 28 March 2018, observing that "the issues regarding whether Karadžić possessed the *mens rea* for genocide in relation to the Srebrenica JCE were extensively litigated before the Trial Chamber and have been fully briefed by Karadžić and the Prosecution on appeal." The Appeals Chamber also seems to criticise the fact that "the *Amicus Curiæ* Observations seek to guide the Appeals Chamber's analysis of the Trial Judgment without consideration of or access to the entire record that is relevant to the Trial Chamber's conclusions." This however would raise the question: what is the purpose of having a 2615 page judgment, if the judgment does not actually fully reflect the grounds for the conclusion? Finally, confirmation that our analysis was trustworthy comes from the Appeal Chamber Judgment [16] that in its section D.2 provides a summary of the Trial's Chamber Judgment regarding whether Karadžić was knowledgeable of the intent to kill Bosnian Muslims which is almost entirely present in our resulting argumentation network depicted in Figure 3.

7 Conclusion

In this paper, we presented the methodology and the results of an application of argumentation theory to map the evidence and arguments as to whether Karadžić possessed *mens rea* for genocide in Srebrenica based on [15]. As discussed in Section 5, we summarised the results of our analysis testing whether Karadžić was knowledgeable of the intent — of General Mladić and others — to kill Bosnian Muslims. This hypothesis is supported within the Trial Chamber's judgment [15] by three lines of reasoning, two of which might merit further discussion, and the last one relying on a single witness.

Although at first sight this paper seems to be similar to other attempts to analyse legal reasoning with formal argumentation, *e.g.*, [23, 22], it differs from them substantially as we did not try to capture the debate component, hence distinguishing between Prosecutor and Defence claims. Instead, our analysis is closer to works

[8] A reviewer of an earlier version of this paper commented that some of the instances of argument from opinion in Figure 3 seem more of instance of argument from ignorance. This is something that only a judge mindful of the purpose of their prose could clarify when writing the document.

aimed at analysing arguments in a single document, like, for instance, [21] that analyses the role of argumentation in written financial communications.

Although the Appeals Chamber denied the admissibility of our application, the interest that applying formal argumentation theories triggered in the international criminal law community suggests that there is scope for future work in this area. We cannot claim that the methodology used in this analysis is beyond critique, but we can claim that it can help creating a better judgment that fully reflects the grounds for the overall conclusion.

This is the long-term aspiration of the ongoing research underpinning this paper, and we are fully aware that this will require to provide answers and innovative proposals both from a technical perspective as well as from the legal one. From a technical perspective, for instance, we still lack appropriate methodologies for adequately transforming statements of natural language into formal logic — a problem most students of logic encounter without being presented with satisfactory solutions, cf. among others [4] — thus inevitably exposing the subjectivity of each formalisation. In addition, following [8], we will also work in the direction of assessing the quality and the strengths of different argumentation reasoning lines, by taking into consideration quantitative measurements of uncertainty and trust, thus enriching the community proposals looking at probabilistic elements in legal reasoning, *e.g.*, [25, 33, 32].

Acknowledgments

This research formed part of a project entitled Developing a Tool to Support Judicial Reasoning in International Criminal Trials, funded in part by Cherish-DE, the United Kingdom's Digital Economy Crucible (http://cherish-de.uk).

This research was partially sponsored by the Italian Ministry of Research through a Rita Levi-Montalcini Personal Fellowship (D.M. n. 285, 29/03/2019) and by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

References

[1] Latifa Al-Abdulkarim, Katie Atkinson, and Trevor J. M. Bench-Capon. A methodology for designing systems to reason with legal cases using abstract dialectical frameworks. *Artificial Intelligence and Law*, 24(1):1–49, 2016.

[2] Katie Atkinson and Trevor J. M. Bench-Capon. Relating the ANGELIC methodology and ASPIC+. In *COMMA 2018*, pages 109–116, 2018.

[3] Pietro Baroni, Federico Cerutti, Paul E. Dunne, and Massimiliano Giacomin. Automata for Infinite Argumentation Structures. *Artificial Intelligence*, 203(0):104–150, 2013.

[4] Michael Baumgartner and Timm Lampert. Adequate formalization. *Synthese*, 164(1):93–115, 2008.

[5] F. Bex, S. van den Braak, H. van Oostendorp, H. Prakken, B. Verheij, and G. Vreeswijk. Sense-making software for crime investigation: how to combine stories and arguments? *Law, Probability and Risk*, 6(1-4):145–168, oct 2007.

[6] Floris Bex, Henry Prakken, Chris Reed, and Douglas Walton. Towards a Formal Account of Reasoning about Evidence: Argumentation Schemes and Generalisations. *Artificial Intelligence and Law*, 11(2/3):125–165, 2003.

[7] Federico Cerutti, Timothy J. Norman, Alice Toniolo, and Stuart E. Middleton. Cispaces.org: from fact extraction to report generation. In *COMMA 2018*, pages 269 – 280, 2018.

[8] Federico Cerutti and Gavin Pearson. Supporting scientific enquiry with uncertain sources. In *21st International Conference on Information Fusion*, pages 404–411, 2018.

[9] Federico Cerutti and Yvonne McDermott Rees. Did karadzic possess the mens rea for genocide in srebrenica? In Marcello D'Agostino, Fabio Aurelio D'Asaro, and Costanza Larese, editors, *Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021*, volume 3086 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021.

[10] Federico Cerutti, Mauro Vallati, and Massimiliano Giacomin. An Efficient Java-Based Solver for Abstract Argumentation Frameworks: jArgSemSAT. *International Journal on Artificial Intelligence Tools*, 26(02):1750002, 2017.

[11] Carlos Iván Chesnevar, Jarred McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo R. Simari, Matthew South, Gerard A. W. Vreeswijk, and Steven Willmot. Towards an argument interchange format. *The Knowledge Engineering Review*, 21(04):293, 2006.

[12] Gabrielė Chlevickaitė, Barbora Holá, and Catrien Bijleveld. Judicial witness assessments at the icty, ictr and icc: Is there 'standard practice'in international criminal justice? *Journal of International Criminal Justice*, 18(1):185–210, 2020.

[13] Phan Minh Dung. On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.

[14] Paul E. Dunne and Michael Wooldridge. Complexity of abstract argumentation. In

Argumentation in AI, chapter 5, pages 85–104. Springer-Verlag, 2009.

[15] ICTY. Prosecutor v. Karadžić. Judgement. `http://www.icty.org/x/cases/karadzic/tjug/en/160324_judgement.pdf`, 2016.

[16] ICTY. Prosecutor v. Karadžić: Appeal Judgement. MICT-13-55-0660/2, 2019.

[17] Mathilde Janier, John Lawrence, and Chris Reed. OVA+: An argument analysis interface. In *COMMA 2014*, pages 463–564, 2014.

[18] Yvonne McDermott. Inferential reasoning and proof in international criminal trials: The potentials of wigmorean analysis. *Journal of International Criminal Justice*, 13(3):507–533, 2015.

[19] Yvonne McDermott Rees and Federico Cerutti. Request for leave to make submissions as amicus curiae. `http://jrad.unmict.org/webdrawer/webdrawer.dll/webdrawer/rec/240941/view/`, 2018.

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

[21] Rudi Palmieri, Andrea Rocci, and Nadzeya Kudrautsava. Argumentation in earnings conference calls. corporate standpoints and analysts' challenges. *Studies in Communication Sciences*, 15(1):120 – 132, 2015.

[22] H. Prakken, A. Wyner, T. Bench-Capon, and K. Atkinson. A formalization of argumentation schemes for legal case-based reasoning in ASPIC+. *Journal of Logic and Computation*, 25(5):1141–1166, 2015.

[23] Henry Prakken. Formalising ordinary legal disputes: a case study. *Artificial Intelligence and Law*, 16(4):333–359, 2008.

[24] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument & Computation*, 1(2):93–124, June 2010.

[25] Henry Prakken. Argument schemes for discussing Bayesian modellings of complex criminal cases. In *JURIX 2017*, pages 69–78, 2017.

[26] Milena Sterio. The Karadžić Genocide Conviction: Inferences, Intent, and the Necessity to Redefine Genocide. *Emory International Law Review*, jan 2017.

[27] Mihai Surdeanu, David McClosky, Mason R. Smith, Andrey Gusev, and Christopher D. Manning. Customizing an information extraction system to a new domain. In *ACL2011*, pages 2–10, 2011.

[28] Alice Toniolo, Timothy J. Norman, Anthony Etuk, Federico Cerutti, Robin Wentao Ouyang, Mani Srivastava, Nir Oren, Timothy Dropps, John A. Allen, and Paul Sullivan. Agent Support to Reasoning with Different Types of Evidence in Intelligence Analysis. In *AAMAS 2015*, pages 781–789, 2015.

[29] Tim van Gelder. Why are legal arguments to hard to follow? `https://timvangelder.com/2009/08/13/why-are-legal-arguments-so-hard-to-follow/` (on 23 Aug 2018), 2009.

[30] Tim van Gelder. Cultivating Deliberation for Democracy. *Journal of Public Deliberation*, 8(1), apr 2012.

[31] Bart Verheij and Bart. Automated argument assistance for lawyers. In *ICAIL '99*,

pages 43–52, New York, New York, USA, 1999.

[32] Bart Verheij, Floris Bex, Sjoerd T. Timmer, Charlotte S. Vlek, John-Jules Ch. Meyer, Silja Renooij, and Henry Prakken. Arguments, scenarios and probabilities: connections between three normative frameworks for evidential reasoning. *Law, Probability and Risk*, 15(1):35–70, mar 2016.

[33] Charlotte S. Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. A method for explaining Bayesian networks for legal evidence with scenarios. *Artificial Intelligence and Law*, 24(3):285–324, sep 2016.

[34] D Walton, C Reed, and F Macagno. *Argumentation schemes.* Cambridge University Press, NY, 2008.

[35] Douglas N Walton. *Legal argumentation and evidence.* Pennsylvania State University Press, 2002.

[36] Hai Zhuge and Lei He. Automatic maintenance of category hierarchy. *Future Generation Computer Systems*, 67:1 – 12, 2017.

Received 15 June 2022

Decomposing Semantics in Abstract Argumentation

Pietro Baroni
University of Brescia, Italy
pietro.baroni@unibs.it

Federico Cerutti
University of Brescia, Italy
federico.cerutti@unibs.it

Massimiliano Giacomin
University of Brescia, Italy
massimiliano.giacomin@unibs.it

Abstract

The paper introduces a general model for the investigation on decomposability in abstract argumentation, i.e. the possibility of determining the labellings prescribed by a semantics based on evaluations of local functions in subframeworks. By exploiting this model, the paper shows the range of decomposable semantics with varying degrees of local information. A constructive procedure for identifying local functions is then devised, encompassing two kinds of local functions, both of them able to enforce decomposability whenever the semantics is decomposable. As an example of application, the decomposability properties of stable, grounded and preferred semantics are analyzed when local information concerning close neighbors is available.

1 Introduction

Dung's model provides an abstract account of argumentation where arguments are simply represented as nodes of a directed graph, called *argumentation framework*, and where the graph's edges represent the binary attacks between them [17]. This

The authors are indebted to the anonymous reviewers for their helpful comments. This work has been partially supported by the GNCS-INdAM project CUP_E55F22000270001.

formalism is able to capture several approaches in nonmonotonic reasoning and structured argumentation. Its importance lies in the formal methods, called *argumentation semantics*, used to assess the acceptability of sets of arguments and then to determine their justification status, thus providing a basis for evaluating the status of the relevant conclusions in structured instances of the abstract model. This is necessary since conflicts between arguments prevent them from being accepted altogether, and a formal method is needed to solve the conflict [1].

While in the original definitions of argumentation semantics an argumentation framework is considered as a monolithic structure and arguments are evaluated at a global level, in recent years attention has been devoted to semantics definition in a modular fashion, i.e. determining the semantics outcome based on local evaluations in subframeworks [4, 21, 10]. Several motivations underlie this research direction. First, a local approach can save computation time [9, 15] possibly applying parallel computation techniques [16] or exploiting incremental computation in a dynamic context [22]. Second,various equivalence relations [8, 18, 24] heavily rely on modules and can also help summarizing (possibly complex) argumentation frameworks [5]. Furthermore, this research issue is a starting point to tackle the problem of combining different argumentation semantics, i.e. regarding a global argumentation framework as composed of a set of interacting parts each associated with a (possibly) different semantics [23, 19], e.g. to model a multi-agent context or to integrate different kinds of reasoning [7].

In a previous paper [5], the modular definition of argumentation semantics has been investigated without any restriction on how an argumentation framework is partitioned into subframeworks. In particular, the property of *decomposability* of argumentation semantics has been introduced concerning the correspondences between semantics outcome at global and local level. A semantics **S** is decomposable if, given a partition of an argumentation framework into a set of subframeworks, the outcomes produced by **S** can be obtained as a combination of the outcomes produced by a local counterpart of **S** applied separately on each subframework, and vice versa.

While, to the best of our knowledge, the framework proposed in [5] has been the first one to support decomposability analysis at this level of generality, it assumes that the local computation in each subframework can have access only to a specific kind of information about the outcome of the computations in the outside components. In particular, the available information is relatively limited, including the set of outside attackers, the labels assigned to them by the computations in other subframeworks and the unidirectional attacks from such arguments to inner arguments. It turns out that, among the most common semantics proposed in the literature, full decomposability with respect to every arbitrary partition is satisfied by some semantics. In contrast, others require the partition to be based on the strongly connected components of the argumentation framework. A few semantics then lack the decomposability property even under this restriction.

An interesting issue is whether exploiting further information would be useful and lead more semantics to be decomposable. For instance, we may consider attacks from inner arguments to outside arguments, or we may consider a larger set of outside arguments, such as the attackers of attackers, and so on.

This paper aims at providing a model for the investigation of the above issue at a general level, as well as general results that do not rely on the specific argumentation semantics definitions. More specifically, this paper addresses the following research questions:

1. How to model in general the diverse kinds of information that can be exploited in local computations, and the relevant functions representing the local counterparts of argumentation semantics;

2. Determining the range of semantics that are decomposable under different degrees of local information exploited, investigating in particular the extreme cases of null and complete information, respectively;

3. How to determine, in general, the local counterpart of an argumentation semantics to guarantee decomposability;

4. How to exploit the introduced model and the relevant results to analyze semantics decomposability properties.

After some background provided in Section 2, the first question is dealt with in Section 3, by introducing the notions of local information function and local function. On this basis, a generalized notion of decomposability w.r.t. [5] is provided. Section 4 is devoted to the second question, by studying how the set of decomposable semantics depend on the partial order between local information functions. The third question is tackled in general terms in Section 5, by introducing a constructive procedure which does not rely on specific semantics definitions, and thus can be applied to all semantics. In particular, the procedure is based on the selection of argumentation frameworks, where the output of the local function can be determined by applying the semantics at hand. This procedure is shown to be general enough to encompass two kinds of local functions, both of them enforcing decomposability if the semantics and the local information exploited make it possible. Section 6 exploits the procedure to devise the canonical local function for any semantics, which enforces decomposability if the semantics is decomposable, while Section 7 identifies an alternative ‘light’ local function, which achieves the same result under some constraints concerning in particular the local information available. An example of application is provided in Section 8, where the decomposability properties of stable, grounded and preferred semantics are analyzed under local information concerning close neighbors, i.e. the direct attackers and attacked arguments of the subframework. Finally, Section 9 concludes the paper with some discussion and perspectives for further work.

The paper integrates and extends two previous contributions [20, 6] by the same authors. In particular, the technical contents have been enhanced by including some proofs previously omitted and expanding other ones. In order to illustrate and clarify the concepts and definition presented in the paper several examples have been added. Moreover, Section 8 provides a novel contribution showing the application of the results of the previous sections for the study of the decomposability properties of three well-known semantics under some specific types of local information.

2 Preliminaries

We follow the traditional definition of argumentation framework introduced by Dung [17] and define its restriction to a subset of arguments.

Definition 1. *An* argumentation framework *is a pair $AF = (\mathcal{A}, att)$ in which $\mathcal{A}$ is a finite*[1] *set of arguments and $att \subseteq \mathcal{A} \times \mathcal{A}$. Given two arguments α and β such that $(\alpha, \beta) \in att$, we say that α* attacks *β or, equivalently, that α is an* attacker *of β. An argument α which attacks itself, i.e. such that $(\alpha, \alpha) \in att$, is called* self-attacking*. Given a set $Args \subseteq \mathcal{A}$, the restriction of AF to $Args$, denoted as $AF{\downarrow_{Args}}$, is the argumentation framework $(Args, att \cap (Args \times Args))$. The (infinite) set of all possible argumentation frameworks is denoted as SAF.*

We will also need two relations and an operator between argumentation frameworks.

Definition 2. *Given two argumentation frameworks $AF_1 = (\mathcal{A}_1, att_1)$ and $AF_2 = (\mathcal{A}_2, att_2)$:*

- $AF_1 \subseteq AF_2$ *iff* $\mathcal{A}_1 \subseteq \mathcal{A}_2$ *and* $att_1 \subseteq att_2$
- $AF_1 \sqsubseteq AF_2$ *iff* $\mathcal{A}_1 \subseteq \mathcal{A}_2$ *and* $AF_2{\downarrow_{\mathcal{A}_1}} = AF_1$
- $AF_2 \setminus AF_1 \triangleq AF_2{\downarrow_{\mathcal{A}_2 \setminus \mathcal{A}_1}}$

[1] We restrict to the finite case in this paper, while in the more general original definition the set of arguments may be infinite.

The relation $\subseteq$ extends set inclusion to argumentation frameworks, while $AF_1 \sqsubseteq AF_2$ holds if AF_1 is a subframework[2] of AF_2. In this case, $AF_2 \setminus AF_1$ returns the subframework of AF_2 involving the arguments outside AF_1.

It can be easily proved that $\subseteq$ and $\sqsubseteq$ between argumentation frameworks are partial orders.

Proposition 1. *The relations $\subseteq$ and $\sqsubseteq$ between argumentation frameworks are reflexive, antisymmetric and transitive.*

Proof. The proof that $\subseteq$ and $\sqsubseteq$ are reflexive is immediate from the relevant definitions. The fact that $\subseteq$ is antisymmetric directly follows from the fact that the set-inclusion relation $\subseteq$ is antisymmetric, and since $\sqsubseteq$ is stricter than $\subseteq$ it is antisymmetric in turn. As to transitivity, the proof for $\subseteq$ is immediate taking into account that the set-inclusion relation $\subseteq$ is transitive. As to $\sqsubseteq$, if $AF_1 \sqsubseteq AF_2$ and $AF_2 \sqsubseteq AF_3$ then by transitivity of $\subseteq$ and the fact that $\sqsubseteq$ is stricter than $\subseteq$ it holds that $AF_1 \subseteq AF_3$, and in particular $\mathcal{A}_1 \subseteq \mathcal{A}_3$. Since $AF_3{\downarrow}_{\mathcal{A}_2} = AF_2$ and $AF_2{\downarrow}_{\mathcal{A}_1} = AF_1$, we have that $AF_3{\downarrow}_{\mathcal{A}_1} = AF_1$, thus $AF_1 \sqsubseteq AF_3$. □

In this paper we adopt the labelling-based approach to the definition of argumentation semantics [2].

A labelling assigns to each argument of an argumentation framework a label taken from a predefined set Λ. We adopt the most common choice for Λ, i.e. $\{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$, where the label `in` means that the argument is accepted, the label `out` means that the argument is rejected, and the label `undec` means that the status of the argument is undecided. For technical reasons, we define labellings both for argumentation frameworks and for arbitrary sets of arguments.

Definition 3. *Given a set of arguments Args, a* labelling *of Args is a total function $Lab : Args \rightarrow \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\}$. The set of all* labellings *of Args is denoted as $\mathfrak{L}_{Args}$. Given an argumentation framework $AF = (\mathcal{A}, att)$, a* labelling *of AF is a labelling of $\mathcal{A}$. The set of all* labellings *of AF is denoted as $\mathfrak{L}(AF)$. For a labelling Lab of Args, the restriction of Lab to a set of arguments $Args' \subseteq Args$, denoted as $Lab{\downarrow}_{Args'}$, is defined as $Lab \cap (Args' \times \{\mathtt{in}, \mathtt{out}, \mathtt{undec}\})$. We extend this notation to sets of labellings, i.e. given a set of a labellings $\mathfrak{L} \subseteq \mathfrak{L}_{Args}$, $\mathfrak{L}{\downarrow}_{Args'} \triangleq \{Lab{\downarrow}_{Args'} \mid Lab \in \mathfrak{L}\}$. Moreover, if $Lab \in \mathfrak{L}(AF)$ and $AF' \subseteq AF$, where $AF' = (\mathcal{A}', att')$, $Lab{\downarrow}_{AF'}$ will denote $Lab{\downarrow}_{\mathcal{A}'}$.*

Labellings can be partially ordered on the basis of the commitment relation [2, 14].

[2] It is immediate to see that $\sqsubseteq$ is stricter than $\subseteq$, i.e. $AF_1 \sqsubseteq AF_2$ entails $AF_1 \subseteq AF_2$.

Definition 4. *Given two labellings $Lab_1, Lab_2 \in \mathfrak{L}_{Args}$, we say that Lab_1 is less or equally committed than Lab_1 (or, equivalently, that Lab_2 is more or equally committed than Lab_1), written $Lab_1 \sqsubseteq Lab_2$, iff $\forall \alpha \in Args\ Lab_1(\alpha) = \texttt{in} \to Lab_2(\alpha) = \texttt{in}$ and $Lab_1(\alpha) = \texttt{out} \to Lab_2(\alpha) = \texttt{out}$.*

A labelling-based semantics prescribes a set of labellings for each argumentation framework.

Definition 5. *Given an argumentation framework $AF = (\mathcal{A}, att)$, a labelling-based semantics* **S** *associates with AF a subset of $\mathfrak{L}(AF)$, denoted as $\mathbf{L_S}(AF)$.*

As shown in [11, 2], for the semantics considered in this paper there is a direct correspondence with the "traditional" extension-based approach [17].

In general, a semantics encompasses a set of alternative labellings for a single argumentation framework. If a semantics **S** is defined in such a way that such a set is always non empty, i.e. $\forall AF, \mathbf{L_S}(AF) \neq \emptyset$, then **S** is said to be *universally defined.* Moreover, a semantics may be defined so that a unique labelling is always prescribed, i.e. for every argumentation framework AF, $|\mathbf{L_S}(AF)| = 1$. In this case the semantics is said to be *single-status*, while in the general case it is said to be *multiple-status.*

Many semantics exist, corresponding to specific criteria to identify labellings. In this respect, we consider as a basic requirement for a semantics **S** to satisfy *conflict-freeness*, i.e. $\forall AF \in SAF$ and $\forall Lab \in \mathbf{L_S}(AF)$, Lab is conflict-free according to the following definition, taken from [12].

Definition 6. *Let Lab be a labelling of an argumentation framework $AF = (\mathcal{A}, att)$. Lab is* conflict-free *if for each $\alpha \in \mathcal{A}$ it holds that*

- *if α is labelled* `in` *then it does not have an attacker that is labelled* `in`
- *if α is labelled* `out` *then it has at least an attacker that is labelled* `in`

Most semantics enforce a stricter condition, prescribing admissible labellings as defined in the following definition [13].

Definition 7. *Let Lab be a labelling of an argumentation framework $AF = (\mathcal{A}, att)$. Lab is* admissible *if for each $\alpha \in \mathcal{A}$ it holds that*

- *if α is labelled* `in` *then all of its attackers (if any) are labelled* `out`
- *if α is labelled* `out` *then it has at least an attacker that is labelled* `in`

As an extreme case of semantics corresponding to the most skeptical one, which has a theoretical interest rather than a practical one, we consider the semantics **UND**, a single-status semantics which assigns to all arguments the label `undec`.

Definition 8. *The semantics* **UND** *is defined as follows:* $\forall AF = (\mathcal{A}, att) \in SAF$, $\mathbf{L}_{\mathbf{UND}}(AF) = \{Lab\}$*, where* $\forall \alpha \in \mathcal{A}, Lab(\alpha) = \texttt{undec}$.

Traditional semantics select labellings among the complete ones, defined as follows.

Definition 9. *Let Lab be a labelling of an argumentation framework* $AF = (\mathcal{A}, att)$*. Lab is* complete *if for each* $\alpha \in \mathcal{A}$ *it holds that*

- α *is labelled* `in` *if and only if all of its attackers (if any) are labelled* `out`
- α *is labelled* `out` *if and only if it has an attacker that is labelled* `in`

It is easy to see that in a complete labelling an argument is `undec` if and only if it has an attacker labelled `undec` and no attacker labelled `in`.

In this paper we will consider *stable*, *grounded* and *preferred* semantics, denoted as **ST**, **GR** and **PR**, respectively.

Definition 10. *Let Lab be a labelling of an argumentation framework* $AF = (\mathcal{A}, att)$*:*

- *Lab is* stable*, i.e.* $Lab \in \mathbf{L}_{\mathbf{ST}}(AF)$*, if it is complete and there are no arguments labelled* `undec`
- *Lab is* grounded*, i.e.* $Lab \in \mathbf{L}_{\mathbf{GR}}(AF)$*, if it is complete and minimizes the set of arguments labelled* `in` *(or, equivalently, maximizes the set of arguments labelled* `undec`*) among complete labellings*
- *Lab is* preferred*, i.e.* $Lab \in \mathbf{L}_{\mathbf{PR}}(AF)$*, if it is complete and maximizes the set of arguments labelled* `in` *among complete labellings*

The uniqueness and existence of the grounded labelling has been proved in [17]. Accordingly, grounded semantics is single-status, while the other semantics are multiple-status. Moreover, all semantics but stable are universally defined (a counterexample for stable semantics is e.g. an argumentation framework including a self-attacking argument only).

Preferred semantics can be equivalently defined by referring to admissible labellings, as proved in [14].

Proposition 2. *Let Lab be a labelling of an argumentation framework* $AF = (\mathcal{A}, att)$*.* $Lab \in \mathbf{L}_{\mathbf{PR}}(AF)$ *iff Lab is a maximal (w.r.t.* $\sqsubseteq$*) admissible labelling.*

The set of labellings prescribed by the semantics can be used to assess argument justification. Various notions of justification can be considered in this respect. The most common one considers an argument *skeptically justified* in an argumentation framework AF according to a semantics $\mathbf{S}$ if it is assigned the label `in` by all labellings of $\mathbf{L_S}(AF)$.

3 A general model for decomposability

The proposed model for the analysis of decomposability of argumentation semantics is articulated in two layers. The first layer deals with the representation, in a general way, of the information locally used for the computation of labellings in subframeworks. The second layer focuses on the modelling of this computation through the notion of the local function. In order to help the reader to follow the structure of the model, Table 1 lists the main definitions and notations introduced in the two layers.

3.1 Modelling local information

Given a subframework of the global argumentation framework, the information needed for the local computation of the labellings in this subframework necessarily includes the topology of the subframework itself, i.e. the set of arguments and the relevant attack relation. On the other hand, some information from the outside is also needed, e.g. some arguments in the subframework might be attacked from external arguments which might be assigned different labellings. Accordingly, in general information from the outside comprises two parts:

- some knowledge of the topology of the neighboring part of the subframework;
- the labelling assigned to this neighboring part by the local computations on external subframeworks, in order to extend it with a local labelling of the subframework.

As to the first point, the topological information specifically available depends on the kind of information known and/or exploited for the local computation. For instance, one might decide to consider external attackers with the unidirectional attacks from them, or one might also contemplate the external nodes attacked by the subframework, or the attackers of the attackers might also be considered, and so on. To model all these possibilities we introduce the notion of *local information function*, which takes in input a “global” argumentation framework AF^* and one of its subframeworks AF, and returns as output the portion of AF^* which can be taken

Modelling local information
Local information function: $\mathcal{LI}_{AF^*}(AF) \in SAF$ (Def. 11) - Specific local information functions: $m\mathcal{LI}$, $M\mathcal{LI}$, $inp\mathcal{LI}$, $Binp\mathcal{LI}$, $out\mathcal{LI}$, $Bout\mathcal{LI}$, $inpout\mathcal{LI}$ and $inp - k - \mathcal{LI}$ (Defs 12, 14, 16) - Partial order between local information functions: $\preceq$ (Def. 17)
Argumentation framework with input (AF, AF', Lab) (Def. 18) - derived from $\mathcal{LI}$ in AF^*: $(AF, AF', \mathit{Lab}) \in AF^{\mathit{inp}}_{\mathcal{LI},AF^*}$ (Def. 19) - derived from $\mathcal{LI}$: $(AF, AF', \mathit{Lab}) \in AF^{\mathit{inp}}_{\mathcal{LI}}$ (Def. 19) - realized from $\mathcal{LI}$ in AF^* under **S**: $(AF, AF', \mathit{Lab}) \in RAF^{\mathit{inp}}_{\mathcal{LI},AF^*,\mathbf{S}}$ (Def. 20) - realized from $\mathcal{LI}$ under **S**: $(AF, AF', \mathit{Lab}) \in RAF^{\mathit{inp}}_{\mathcal{LI},\mathbf{S}}$ (Def. 20)
Local functions and decomposability
Local function F for $\mathcal{LI}$: $F(AF, AF', \mathit{Lab}) \in 2^{\mathfrak{L}(AF)}$ (Def. 21) - enforcing decomposability of **S** under $\mathcal{LI}$ (Def. 22) - enforcing top-down | bottom-up decomposability of **S** under $\mathcal{LI}$ (Def. 23)
(Fully) decomposable semantics **S** under $\mathcal{LI}$ (Def. 22)

Table 1: Main definitions and notations introduced in Section 3.

into account to compute the labellings of AF (note that this portion extends AF). Some constraints are also introduced concerning the role of AF^* (see the relevant explanation later).

Definition 11. *A local information function is a function*

$$\mathcal{LI} : \{(AF^*, AF) \mid AF^*, AF \in SAF \wedge AF \sqsubseteq AF^*\} \rightarrow SAF$$

such that $\forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$

- $AF \sqsubseteq \mathcal{LI}(AF^*, AF)$ *and* $\mathcal{LI}(AF^*, AF) \subseteq AF^*$
- *if* $AF^* \subseteq AF^{**}$ *then either* $\mathcal{LI}(AF^{**}, AF) = \mathcal{LI}(AF^*, AF)$ *or it is not the case that* $\mathcal{LI}(AF^{**}, AF) \subseteq AF^*$

For ease of notation, in the following $\mathcal{LI}(AF^*, AF)$ will be denoted as $\mathcal{LI}_{AF^*}(AF)$.

Some explanation on the constraints introduced in the above definition is in order.

As to the first item, $AF \sqsubseteq \mathcal{LI}(AF^*, AF)$ signifies that the local subframework must be known to compute the appropriate labellings, and thus is part of the available information. The other condition $\mathcal{LI}(AF^*, AF) \subseteq AF^*$ expresses that the neighboring part of AF returned by the function is taken from AF^*. Here the use of $\subseteq$ rather than $\sqsubseteq$ gives more freedom in the choice of the local information, since it makes it possible to neglect some attacks that otherwise should be taken into account (e.g. one might consider external attackers with the relevant attacks directed towards AF but neglect the attacks directed from AF to such attackers).

The second item concerns the role of AF^*, which must be used only to identify the neighboring part of the subframework available locally. However, in principle there might be some further information hidden in the way the output of the function, say AF', is selected depending on AF^*, e.g. subtle dependencies could be introduced where part of the external topology might be artificially excluded to take into account the topology of $AF^* \setminus AF'$. To avoid this possibility, the constraint requires that if AF^* is enlarged, then either AF' does not change, or the additional elements of the enlarged global framework play an explicit role, i.e. some appear in the novel output of the local information function.

It should be noted that, according to Definition 11, for any local information function $\mathcal{LI}$ and any argumentation framework AF it holds that $\mathcal{LI}_{AF}(AF) = AF$, i.e. obviously there is no external information w.r.t. the whole argumentation framework.

Definition 11 encompasses various local information functions corresponding to different criteria to select the local information taken into account.

As two extreme cases, we introduce the local information functions[3] $m\mathcal{LI}$, called the *minimum local information function*, and $M\mathcal{LI}$, called the *maximum local information function*. The first function models the case where no external information is available, i.e. $m\mathcal{LI}$ returns as output just the subframework where local labellings are computed. The second function models the case where all external topological information is available, i.e. $M\mathcal{LI}$ returns as output the whole global argumentation framework.

Definition 12. *$m\mathcal{LI}$ is the local information function such that $\forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$, $m\mathcal{LI}_{AF^*}(AF) = AF$. $M\mathcal{LI}$ is the local information function such that $\forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$, $M\mathcal{LI}_{AF^*}(AF) = AF^*$.*

[3] The proof that the second item of Definition 11 is satisfied will be given later (see Proposition 4).

Example 1. *Considering the argumentation frameworks AF^* and AF depicted in Figure 1, we have that $m\mathcal{LI}_{AF^*}(AF) = AF$, i.e. the function returns the sub-framework AF without external information, and $M\mathcal{LI}_{AF^*}(AF) = AF^*$, i.e. the function returns the whole global framework modelling complete knowledge of the external topology.*

There are plenty of other local information functions between the two extreme cases described above, and in the following we introduce some of them just for the sake of the example. In order to make their definitions easier, we first introduce some notations.

Definition 13. *Given an argumentation framework $AF = (\mathcal{A}, att)$ and a set of arguments $Args \subseteq \mathcal{A}$:*

- $Args^{inp}_{AF} = \{\alpha \in \mathcal{A} \setminus Args \mid \exists \beta \in Args, (\alpha, \beta) \in att\}$
- $Args^{att-inp}_{AF} = att \cap (Args^{inp}_{AF} \times Args)$
- $Args^{Batt-inp}_{AF} = Args^{att-inp}_{AF} \cup (att \cap (Args \times Args^{inp}_{AF}))$
- $Args^{out}_{AF} = \{\alpha \in \mathcal{A} \setminus Args \mid \exists \beta \in Args, (\beta, \alpha) \in att\}$
- $Args^{att-out}_{AF} = att \cap (Args \times Args^{out}_{AF})$
- $Args^{Batt-out}_{AF} = Args^{att-out}_{AF} \cup (att \cap (Args^{out}_{AF} \times Args))$

In words, $Args^{inp}_{AF}$ is the set of the arguments attacking $Args$ from the outside, $Args^{att-inp}_{AF}$ includes the attacks from $Args^{inp}_{AF}$ to $Args$ (but not vice versa), $Args^{Batt-inp}_{AF}$ includes the attacks from $Args^{inp}_{AF}$ to $Args$ and vice versa. $Args^{out}_{AF}$ includes the outside arguments attacked by $Args$, $Args^{att-out}_{AF}$ includes the attacks from $Args$ to $Args^{out}_{AF}$, while $Args^{Batt-out}_{AF}$ also includes the existing reverse attacks.

Example 2. *Consider the argumentation framework $AF^* = (\{\alpha, \beta, \gamma_1, \gamma_2, \gamma_3, \delta_1, \delta_2, \eta\}, \{(\delta_1, \gamma_1), (\delta_1, \gamma_2), (\gamma_1, \gamma_2), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2), (\alpha, \beta), (\beta, \alpha), (\beta, \gamma_3), (\gamma_3, \delta_2), (\delta_2, \eta)\})$ shown in Figure 1 and the set $Args = \{\alpha, \beta\}$. It turns out that:*

- $Args^{inp}_{AF^*} = \{\gamma_1, \gamma_2\}$
- $Args^{att-inp}_{AF^*} = \{(\gamma_1, \alpha), (\gamma_2, \alpha)\}$
- $Args^{Batt-inp}_{AF^*} = \{(\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2)\}$
- $Args^{out}_{AF^*} = \{\gamma_2, \gamma_3\}$

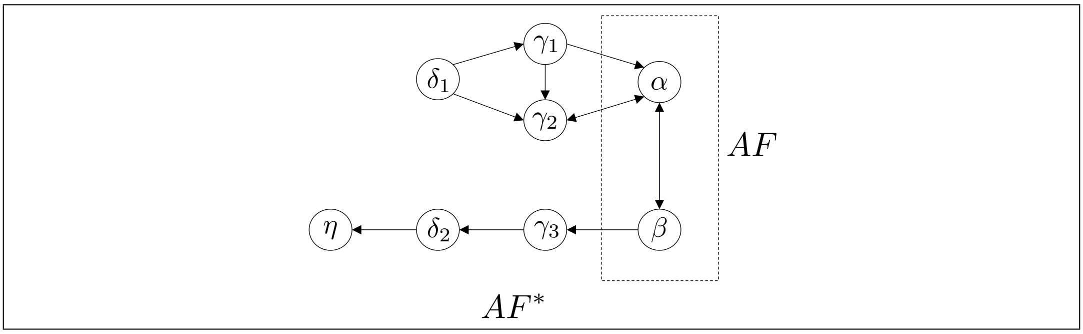


Figure 1: Argumentation frameworks AF^* and AF, with $AF \sqsubseteq AF^*$.

- $Args_{AF^*}^{att-out} = \{(\alpha, \gamma_2), (\beta, \gamma_3)\}$
- $Args_{AF^*}^{Batt-out} = \{(\alpha, \gamma_2), (\gamma_2, \alpha), (\beta, \gamma_3)\}$

The following five local functions only involve close neighbors, i.e. direct attackers of the subframework and/or arguments directly attacked by the subframework.

Definition 14. *The following functions from* $\{(AF^*, AF) \mid AF^*, AF \in SAF \wedge AF \sqsubseteq AF^*\}$ *to* SAF *are defined:*

- $inp\mathcal{LI}_{AF^*}(AF) = (\mathcal{A} \cup \mathcal{A}_{AF^*}^{inp}, att \cup \mathcal{A}_{AF^*}^{att-inp})$
- $Binp\mathcal{LI}_{AF^*}(AF) = (\mathcal{A} \cup \mathcal{A}_{AF^*}^{inp}, att \cup \mathcal{A}_{AF^*}^{Batt-inp})$
- $out\mathcal{LI}_{AF^*}(AF) = (\mathcal{A} \cup \mathcal{A}_{AF^*}^{out}, att \cup \mathcal{A}_{AF^*}^{att-out})$
- $Bout\mathcal{LI}_{AF^*}(AF) = (\mathcal{A} \cup \mathcal{A}_{AF^*}^{out}, att \cup \mathcal{A}_{AF^*}^{Batt-out})$
- $inpout\mathcal{LI}_{AF^*}(AF) = (\mathcal{A} \cup \mathcal{A}_{AF^*}^{inp} \cup \mathcal{A}_{AF^*}^{out}, att \cup \mathcal{A}_{AF^*}^{att-inp} \cup \mathcal{A}_{AF^*}^{att-out})$

where $AF = (\mathcal{A}, att)$.

In words, $inp\mathcal{LI}$ selects as external information the set of outside attackers and the unidirectional attacks from them to AF, $Binp\mathcal{LI}$ is similar but considers both possible directions for the attacks. The functions $out\mathcal{LI}$ and $Bout\mathcal{LI}$ are the counterparts of $inp\mathcal{LI}$ and $Binp\mathcal{LI}$, respectively, that involve attacked arguments instead of attackers. In particular, $out\mathcal{LI}$ selects as external information the set of outside attacked arguments and the unidirectional attacks from AF to them, $Bout\mathcal{LI}$ is similar but considers both possible directions for the attacks. Finally, $inpout\mathcal{LI}$ models complete information about close neighbors, i.e. attackers and attacked arguments with both directions for the attacks, since it holds that $inpout\mathcal{LI}_{AF^*}(AF) = Binp\mathcal{LI}_{AF^*}(AF) \cup Bout\mathcal{LI}_{AF^*}(AF)$.

One may also consider a larger neighboring part w.r.t. direct attackers and attacked arguments. For instance, besides the direct attackers, the local information may involve also their attackers, the attackers of their attackers, and so on until a level k.

Definition 15. *Given an argumentation framework $AF = (\mathcal{A}, att)$, a path in AF of length n from α_0 to α_n is a sequence of arguments $\alpha_0, \ldots, \alpha_n$ such that $(\alpha_i, \alpha_{i+1}) \in att$ for each $i \in \{0, \ldots, n-1\}$. We indicate that a path of length n exists from α_0 to α_n as $p^n_{AF}(\alpha_0, \alpha_n)$. Given a set of arguments $Args \subseteq AF$ and an integer $k > 0$, $Args^{inp-k}_{AF} \equiv \{\alpha \in \mathcal{A} \setminus Args \mid \exists \beta \in \mathcal{A}, p^n_{AF}(\alpha, \beta), n \leq k\}$.*

The following function considers all the ancestors of the arguments in AF (w.r.t. the attack relation) of distance less that or equal to a constant k as well as all involved attacks.

Definition 16. *$inp - k - \mathcal{LI}$ is the function from $\{(AF^*, AF) \mid AF^*, AF \in SAF \wedge AF \sqsubseteq AF^*\}$ to SAF such that $inp - k - \mathcal{LI}_{AF^*}(AF) \equiv AF^*{\downarrow}_{(\mathcal{A} \cup \mathcal{A}^{inp-k}_{AF^*})}$*

Example 3. *Referring again to Example 2, consider the argumentation frameworks AF^* and $AF = AF^*{\downarrow}_{\{\alpha,\beta\}}$, depicted in Figure 1. It turns out that:*

- $inp\mathcal{LI}_{AF^*}(AF) = (\{\alpha, \beta, \gamma_1, \gamma_2\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha)\})$
- $Binp\mathcal{LI}_{AF^*}(AF) = (\{\alpha, \beta, \gamma_1, \gamma_2\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2)\})$
- $out\mathcal{LI}_{AF^*}(AF) = (\{\alpha, \beta, \gamma_2, \gamma_3\}, \{(\alpha, \beta), (\beta, \alpha), (\alpha, \gamma_2), (\beta, \gamma_3)\})$
- $Bout\mathcal{LI}_{AF^*}(AF) = (\{\alpha, \beta, \gamma_2, \gamma_3\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2), (\beta, \gamma_3)\})$
- $inpout\mathcal{LI}_{AF^*}(AF) =$ $(\{\alpha, \beta, \gamma_1, \gamma_2, \gamma_3\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2), (\beta, \gamma_3)\})$
- $inp - 1 - \mathcal{LI}_{AF^*}(AF) = AF^*{\downarrow}_{\{\gamma_1, \gamma_2, \alpha, \beta\}} =$ $(\{\alpha, \beta, \gamma_1, \gamma_2\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2), (\gamma_1, \gamma_2)\})$
- $inp - 2 - \mathcal{LI}_{AF^*}(AF) = AF^*{\downarrow}_{\{\delta_1, \gamma_1, \gamma_2, \alpha, \beta\}} = (\{\alpha, \beta, \gamma_1, \gamma_2, \delta_1\},$ $\{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \gamma_2), (\gamma_1, \gamma_2), (\delta_1, \gamma_1), (\delta_1, \gamma_2)\})$

To show that the above functions are actually local information functions, we have to prove that they satisfy the constraints of Definition 11. The following proposition introduces sufficient conditions that might be easier to verify w.r.t. those of Definition 11. In particular, the constraint concerning the role of the global framework AF^* (second item in Definition 11) holds if a function is monotone w.r.t. AF^* and its output does not change if AF^* is replaced with the same argumentation framework returned as output.

Proposition 3. *Let $\mathcal{LI}$ be a function from $\{(AF^*, AF) \mid AF^*, AF \in SAF \wedge AF \sqsubseteq AF^*\}$ to SAF. If $\forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$ the following conditions are satisfied*

- $AF \sqsubseteq \mathcal{LI}_{AF^*}(AF)$ *and* $\mathcal{LI}_{AF^*}(AF) \subseteq AF^*$
- *for every $AF^{**} \in SAF$ such that $AF^* \subseteq AF^{**}$, it holds that $\mathcal{LI}_{AF^*}(AF) \subseteq \mathcal{LI}_{AF^{**}}(AF)$*
- $\mathcal{LI}_{\mathcal{LI}_{AF^*}(AF)}(AF) = \mathcal{LI}_{AF^*}(AF)$

then $\mathcal{LI}$ is a local information function.

Proof. Referring to Definition 11, only the second item has to be proved since the first one holds by the first hypothesis. To this purpose, we show that for every $AF, AF^*, AF^{**} \in SAF$ such that $AF^* \subseteq AF^{**}$, if $\mathcal{LI}_{AF^{**}}(AF) \subseteq AF^*$ then $\mathcal{LI}_{AF^{**}}(AF) = \mathcal{LI}_{AF^*}(AF)$.

If $\mathcal{LI}_{AF^{**}}(AF) \subseteq AF^*$, by the second hypothesis (monotony w.r.t. the global framework) $\mathcal{LI}_{\mathcal{LI}_{AF^{**}}(AF)}(AF) \subseteq \mathcal{LI}_{AF^*}(AF)$. According to the third hypothesis $\mathcal{LI}_{\mathcal{LI}_{AF^{**}}(AF)}(AF) = \mathcal{LI}_{AF^{**}}(AF)$, thus $\mathcal{LI}_{AF^{**}}(AF) \subseteq \mathcal{LI}_{AF^*}(AF)$. Since $AF^* \subseteq AF^{**}$, the second hypothesis yields $\mathcal{LI}_{AF^*}(AF) \subseteq \mathcal{LI}_{AF^{**}}(AF)$. Thus by antisymmetry of $\subseteq$ we get $\mathcal{LI}_{AF^{**}}(AF) = \mathcal{LI}_{AF^*}(AF)$. □

We can then show that the functions introduced above are local information functions.

Proposition 4. *$m\mathcal{LI}$, $M\mathcal{LI}$, $inp\mathcal{LI}$, $Binp\mathcal{LI}$, $out\mathcal{LI}$, $Bout\mathcal{LI}$, $inpout\mathcal{LI}$ and $inp - k - \mathcal{LI}$ are local information functions.*

Proof. For all of the functions the proof is based on Proposition 3.

As to $m\mathcal{LI}$, if $AF \sqsubseteq AF^*$ then it is immediate to see that $AF \sqsubseteq m\mathcal{LI}_{AF^*}(AF)$ and $m\mathcal{LI}_{AF^*}(AF) \subseteq AF^*$, since $m\mathcal{LI}_{AF^*}(AF) = AF$. Also the second required constraint that $m\mathcal{LI}_{AF^*}(AF) \subseteq m\mathcal{LI}_{AF^{**}}(AF)$ trivially holds, since $m\mathcal{LI}_{AF^*}(AF) = m\mathcal{LI}_{AF^{**}}(AF) = AF$. Finally, as to the third constraint $m\mathcal{LI}_{m\mathcal{LI}_{AF^*}(AF)}(AF) = m\mathcal{LI}_{AF}(AF) = AF = m\mathcal{LI}_{AF^*}(AF)$.

As to $M\mathcal{LI}$, if $AF \sqsubseteq AF^*$ then $AF \sqsubseteq M\mathcal{LI}_{AF^*}(AF)$ and $M\mathcal{LI}_{AF^*}(AF) \subseteq AF^*$ trivially hold, since $M\mathcal{LI}_{AF^*}(AF) = AF^*$. The second constraint holds since

$\mathcal{MLI}_{AF^*}(AF) \subseteq \mathcal{MLI}_{AF^{**}}(AF)$ equates to $AF^* \subseteq AF^{**}$. As to the third constraint, by the definition of $\mathcal{MLI}$ we directly have that $\mathcal{MLI}_{\mathcal{MLI}_{AF^*}(AF)}(AF) = \mathcal{MLI}_{AF^*}(AF)$.

As to the other functions, by inspection of their definitions it is easy to see that for each $\mathcal{LI} \in \{inp\mathcal{LI}, Binp\mathcal{LI}, out\mathcal{LI}, Bout\mathcal{LI}, inpout\mathcal{LI}, inp-k-\mathcal{LI}\}$ $\mathcal{LI}_{AF^*}(AF)$ is obtained by (possibly) adding to AF elements (arguments and attacks) from AF^* that are external to AF. Thus the first item of Proposition 3 is verified. It is also easy to see that all elements of $\mathcal{LI}_{AF^*}(AF)$ are maintained in the output obtained with AF^* enlarged, thus also the second item holds. As to the third item, let $\mathcal{LI}_{AF^*}(AF) = AF'$. According to the definitions of the functions, each element included in $\mathcal{LI}_{AF^*}(AF)$ is still an element of $\mathcal{LI}_{AF'}(AF)$, thus also the last item of Proposition 3 is verified. □

Local information functions can be partially ordered based on the amount of information returned as output. For instance, the local information function $inp\mathcal{LI}$, that returns as output the outside attackers of AF and the relevant unidirectional attacks, is less informative than $inpout\mathcal{LI}$ that also includes outside attacked nodes and both directions of attacks.

Definition 17. *Given two local information functions $\mathcal{LI}_1$ and $\mathcal{LI}_2$, $\mathcal{LI}_1 \preceq \mathcal{LI}_2$ iff $\forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$ it holds that $\mathcal{LI}_{1AF^*}(AF) \subseteq \mathcal{LI}_{2AF^*}(AF)$.*

In words, if $\mathcal{LI}_1 \preceq \mathcal{LI}_2$ then $\mathcal{LI}_1$ always returns an argumentation framework which is contained in that returned by $\mathcal{LI}_2$. It is easy to see that $inp\mathcal{LI} \preceq Binp\mathcal{LI}$, $Binp\mathcal{LI} \preceq inpout\mathcal{LI}$, $out\mathcal{LI} \preceq Bout\mathcal{LI}$, $Bout\mathcal{LI} \preceq inpout\mathcal{LI}$, and $Binp\mathcal{LI} \preceq inp-k-\mathcal{LI}$ with $k \geq 1$.

$\preceq$ is a partial order with a least and a greatest element.

Proposition 5. *$\preceq$ is reflexive, transitive and antisymmetric. $m\mathcal{LI}$ and $\mathcal{MLI}$ are the least and greatest element, respectively, w.r.t. $\preceq$ of the set of local information functions.*

Proof. The proof that $\preceq$ is a partial order is immediate taking into account that by Proposition 1 the relation $\subseteq$ between argumentation frameworks is a partial order.

By definition of local information function (see Definition 11) $\forall\mathcal{LI}, \forall AF^*, AF \in SAF : AF \sqsubseteq AF^*$, it holds that $AF \sqsubseteq \mathcal{LI}_{AF^*}(AF)$ and $\mathcal{LI}_{AF^*}(AF) \subseteq AF^*$. Since $AF \sqsubseteq \mathcal{LI}_{AF^*}(AF)$ entails $AF \subseteq \mathcal{LI}_{AF^*}(AF)$, it holds that $\forall\mathcal{LI}$, $m\mathcal{LI} \preceq \mathcal{LI}$ and $\mathcal{LI} \preceq \mathcal{MLI}$. □

While local information functions model the identification criterion of available topological information for all possible subframeworks of all global argumentation frameworks, the information available for a specific subframework of a given global framework is represented by an *argumentation framework with input*, which besides topological information includes the labelling externally assigned to the neighboring part of the subframework (see the second point at the beginning of the section). This notion is introduced in the next definition.

Definition 18. *An* argumentation framework with input *is a tuple* (AF, AF', Lab) *where* $AF, AF' \in SAF$ *such that* $AF \sqsubseteq AF'$*, and* $Lab \in \mathfrak{L}(AF' \setminus AF)$*.*

Intuitively, AF plays the role of a subframework, while AF' and Lab are the elements affecting the computation of the labellings of AF. In particular, AF' represents the portion of the global argumentation framework which is taken into account, including AF itself, while Lab is the labelling assigned to the relevant arguments outside AF, i.e. those belonging to $AF' \setminus AF$.

The relationships between the notions of local information function and argumentation framework with input are described in Definition 19 and Definition 20.

Definition 19. *An argumentation framework with input* (AF, AF', Lab) *is* derived from *a local information function* $\mathcal{LI}$ *in* AF^**, written* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI},AF^*}$*, if* $AF' = \mathcal{LI}_{AF^*}(AF)$*.*

(AF, AF', Lab) *is derived from* $\mathcal{LI}$*, written* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$*, if* $\exists AF^*$ *such that* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI},AF^*}$*.*

Intuitively, given a subframework AF of AF^*, one can derive in AF^* an argumentation framework with input by applying a local information function to AF and AF^*, obtaining $(AF, \mathcal{LI}_{AF^*}(AF), Lab)$. The second part of the definition removes the reference to a specific global argumentation framework AF^*, by defining an argumentation framework with input as derived from $\mathcal{LI}$ if there is AF^* where it can be derived from $\mathcal{LI}$.

While in the notions introduced above the labelling component of argumentation frameworks with input is not constrained, the notion of *realizability* introduced in the following definition requires the labelling component to correspond to a labelling enforced by the semantics.

Definition 20. *An argumentation framework with input* (AF, AF', Lab) *is* realized *from a local information function* $\mathcal{LI}$ *in an argumentation framework* AF^* *under a semantics* $\mathbf{S}$*, written* $(AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},AF^*,\mathbf{S}}$*, if* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI},AF^*}$ *and* $\exists Lab^* \in \mathbf{L_S}(AF^*)$ *such that* $Lab^* {\downarrow}_{AF' \setminus AF} = Lab$*.*

(AF, AF', Lab) *is realized from a local information function* $\mathcal{LI}$ *under a semantics* $\mathbf{S}$*, written* $(AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},\mathbf{S}}$*, if* $\exists AF^*$ *such that* $(AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},AF^*,\mathbf{S}}$*.*

Example 4. *Referring to Example 3, let*

$$AF' = inp\mathcal{LI}_{AF^*}(AF) = (\{\alpha, \beta, \gamma_1, \gamma_2\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha)\})$$

We have that

$$(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\}) \in AF^{inp}_{inp\mathcal{LI}, AF^*}$$

thus it also holds

$$(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\}) \in AF^{inp}_{inp\mathcal{LI}}$$

i.e. the argumentation framework with input $(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\})$ *is derived from inp$\mathcal{LI}$. Note that these relations hold independently of the labelling of* γ_1 *and* γ_2 *in the argumentation framework with input.*

Under most semantics $\mathbf{S}$*, including the stable, grounded and preferred semantics, it also holds*

$$(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\}) \in RAF^{inp}_{inp\mathcal{LI}, \mathbf{S}}$$

since e.g. letting $AF^* = (\{\gamma_1, \gamma_2, \alpha, \beta\}, \{(\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \beta), (\beta, \alpha)\})$*, it holds that* $(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\}) \in RAF^{inp}_{\mathcal{LI}, AF^*, \mathbf{S}}$.

However, if we consider $inp - 1 - \mathcal{LI}$ *instead of inp$\mathcal{LI}$, so as to include also attack between external attackers, it would not be the case that*

$$(AF, AF', \{(\gamma_1, \mathtt{in}), (\gamma_2, \mathtt{in})\}) \in RAF^{inp}_{inp-1-\mathcal{LI}, \mathbf{S}}$$

since all semantics satisfying conflict-freeness (including $\mathbf{ST}$*,* $\mathbf{GR}$ *and* $\mathbf{PR}$*) prohibit conflicting arguments (*γ_1 *and* γ_2 *in this case) from being all labelled* `in`*.*

3.2 The notions of local function and decomposability

We are now able to define the notion of decomposability of an argumentation semantics[4].

The first step is defining a local function, which represents a local counterpart of the notion of semantics. While a semantics takes as input an argumentation framework and returns a set of its labellings, a local function takes as input an argumentation framework with input and produces as output a set of labellings for the inner local argumentation framework. It makes sense to define a local function with reference to a local information function, since only the argumentation frameworks with input derived from the adopted local information function can play a role (see Definition 19).

[4] This notion generalizes to the setting devised above the homonymous notion introduced in [5].

Definition 21. *A local function F for a local information function $\mathcal{LI}$ assigns to any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ a (possibly empty) set of labellings of AF, i.e. $F(AF, AF', Lab) \in 2^{\mathfrak{L}(AF)}$.*

A semantics $\mathbf{S}$ is *decomposable* (also called fully decomposable) if the labellings prescribed on an argumentation framework AF correspond to the possible combinations of *compatible* labellings obtained by applying a local function F in the subframeworks that partition the global framework.

Definition 22. *A local function F for a local information function $\mathcal{LI}$* enforces decomposability *of a semantics $\mathbf{S}$ under $\mathcal{LI}$ iff for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$, the following condition holds*

$$\begin{aligned}\mathbf{L_S}(AF) = \{&L_{P1} \cup \ldots \cup L_{Pn} \mid \\ &L_{Pi} \in F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})\}\end{aligned}$$

A semantics $\mathbf{S}$ is decomposable (or equivalently fully decomposable) under $\mathcal{LI}$ iff there is a local function F which enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$.

In the above definition, each subframework enriched with the locally available external information is modelled by the argumentation framework with input $(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})$. The first component is the subframework of AF on the partition element P_i. The second component is the available topological information including the neighboring part. The third component is the labelling assigned to the available arguments outside the subframework $AF{\downarrow_{P_i}}$, i.e. those included in the set $\mathcal{LI}_{AF}(AF{\downarrow_{P_i}}) \setminus AF{\downarrow_{P_i}}$. Compatibility refers to the fact that any labelling of a subframework is used by F to compute other labellings in other subframeworks. More specifically, each local labelling L_{Pi} depends on the other ones since the labelling component taken as input by F is obtained from the labellings L_{Pj} (with $j \neq i$) computed in external subframeworks.

A relation holds between the labellings prescribed by a semantics and a local function enforcing decomposability.

Proposition 6. *Let $\mathbf{S}$ be a fully decomposable semantics under a local information function $\mathcal{LI}$, and let F be a local function which enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$. Then, for any argumentation framework AF, $\mathbf{L_S}(AF) = F(AF, AF, \emptyset)$.*

Proof. Letting $AF \equiv (\mathcal{A}, att)$, consider the partition of $\mathcal{A}$ including a single element, i.e. $\mathcal{A}$. According to Definition 22 and taking into account that $AF{\downarrow_{\mathcal{A}}} = AF$

and that $\mathcal{LI}_{AF}(AF) = AF$ we have $\mathbf{L_S}(AF) = \{L_{P1} \mid L_{P1} \in F(AF, AF, \emptyset)\}$, i.e. $\mathbf{L_S}(AF) = F(AF, AF, \emptyset)$. □

Decomposability can be split into two partial decomposability properties.

Definition 23. *A local function F for a local information function $\mathcal{LI}$* enforces top-down decomposability *of a semantics* **S** *under $\mathcal{LI}$ iff for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$, it holds that*

$$\begin{aligned}\mathbf{L_S}(AF) \subseteq \{&L_{P1} \cup \ldots \cup L_{Pn} \mid \\ &L_{Pi} \in F(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i}) \setminus AF{\downarrow}_{P_i}})\}\end{aligned}$$

A local function F for a local information function $\mathcal{LI}$ enforces bottom-up decomposability *of a semantics* **S** *under $\mathcal{LI}$ iff for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$, it holds that*

$$\begin{aligned}\mathbf{L_S}(AF) \supseteq \{&L_{P1} \cup \ldots \cup L_{Pn} \mid \\ &L_{Pi} \in F(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i}) \setminus AF{\downarrow}_{P_i}})\}\end{aligned}$$

In words, F enforces top-down decomposability if the procedure to compute labellings by means of F is complete, i.e. all labellings prescribed by **S** for AF are obtained by applying F to the subframeworks corresponding to the partition and combining the relevant labellings. On the other hand, F enforces bottom-up decomposability if the procedure is sound, i.e. all combinations of local labellings obtained by F give rise to global labellings that are valid according to **S**.

Note that the local function returning for any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ all possible labellings of AF trivially enforces top-down decomposability of any semantics **S**, and the local function always returning the empty set trivially enforces bottom-up decomposability. Thus it would not make much sense to introduce the notions of top-down and bottom-up decomposable semantics.

The following proposition is immediate and thus its proof is omitted.

Proposition 7. *A local function F for $\mathcal{LI}$ enforces decomposability of a semantics* **S** *under $\mathcal{LI}$ iff it enforces both top-down and bottom-up decomposability of* **S** *under $\mathcal{LI}$.*

An immediate consequence of the above proposition is that a semantics is decomposable under $\mathcal{LI}$ iff there is a local function F for $\mathcal{LI}$ which enforces both top-down and bottom-up decomposability of **S** under $\mathcal{LI}$.

4 On the power of local information functions

Intuitively, the more local information is available, the easier it is to determine the global labellings from local computation. Therefore, we expect a more expressive local information function to foster the correct identification of the global labellings, yielding a larger set of decomposable semantics. The next proposition shows that this is the case.

Proposition 8. *If a semantics* $\mathbf{S}$ *is decomposable under* $\mathcal{LI}$*, then for any* $\mathcal{LI}'$ *such that* $\mathcal{LI} \preceq \mathcal{LI}'$*,* $\mathbf{S}$ *is decomposable under* $\mathcal{LI}'$*.*

Proof. By the hypothesis, there is a local function F for $\mathcal{LI}$ such that for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$

$$\begin{aligned} \mathbf{L_S}(AF) = \{ & L_{P1} \cup \ldots \cup L_{Pn} \mid \\ & L_{Pi} \in F(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\setminus AF{\downarrow}_{P_i}})\} \end{aligned} \tag{1}$$

Let us define the local information function F' for $\mathcal{LI}'$ such that for $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}'}$, $F'(AF, AF', Lab) \equiv F(AF, \mathcal{LI}_{AF'}(AF), Lab{\downarrow}_{\mathcal{LI}_{AF'}(AF)\setminus AF})$.

We prove that F' enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}'$, i.e. for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$

$$\begin{aligned} \mathbf{L_S}(AF) = \{ & L_{P1} \cup \ldots \cup L_{Pn} \mid \\ & L_{Pi} \in F'(AF{\downarrow}_{P_i}, \mathcal{LI}'_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}'_{AF}(AF{\downarrow}_{P_i})\setminus AF{\downarrow}_{P_i}})\} \end{aligned}$$

This directly derives from (1) if

$$\begin{aligned} & F'(AF{\downarrow}_{P_i}, \mathcal{LI}'_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}'_{AF}(AF{\downarrow}_{P_i})\setminus AF{\downarrow}_{P_i}}) = \\ & F(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\setminus AF{\downarrow}_{P_i}}) \end{aligned}$$

In order to prove this condition, for the sake of clarity we introduce the following substitutions:

$$\begin{aligned} AF &\to AF^* \\ AF{\downarrow}_{P_i} &\to AF \\ \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}) &\to AF' \\ \mathcal{LI}'_{AF}(AF{\downarrow}_{P_i}) &\to AF'' \\ (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}) &\to Lab \end{aligned}$$

Under these substitutions, it is easy to see that the following conditions hold:

$$AF' = \mathcal{LI}_{AF^*}(AF) \tag{2}$$

$$AF'' = \mathcal{LI}'_{AF^*}(AF) \tag{3}$$

$$AF'' \subseteq AF^* \tag{4}$$

where the last condition is due to the definition of local information function referring to $\mathcal{LI}'$.

Taking into account the substitutions above, the thesis becomes

$$F'(AF, AF'', Lab{\downarrow}_{AF''\setminus AF}) = F(AF, AF', Lab{\downarrow}_{AF'\setminus AF}) \tag{5}$$

According to the definition of F', the first term can be expressed as

$$F(AF, \mathcal{LI}_{AF''}(AF), (Lab{\downarrow}_{AF''\setminus AF}){\downarrow}_{\mathcal{LI}_{AF''}(AF)\setminus AF})$$

Since by definition of local information function $\mathcal{LI}_{AF''}(AF) \subseteq AF''$, it holds that $(Lab{\downarrow}_{AF''\setminus AF}){\downarrow}_{\mathcal{LI}_{AF''}(AF)\setminus AF} = Lab{\downarrow}_{\mathcal{LI}_{AF''}(AF)\setminus AF}$ and thus the same term can be expressed as

$$F(AF, \mathcal{LI}_{AF''}(AF), Lab{\downarrow}_{\mathcal{LI}_{AF''}(AF)\setminus AF})$$

Now, since by (4) $AF'' \subseteq AF^*$, by definition of local information function (in particular, the second constraint of Definition 11) either $\mathcal{LI}_{AF''}(AF) = \mathcal{LI}_{AF^*}(AF)$ or it is not the case that $\mathcal{LI}_{AF^*}(AF) \subseteq AF''$. On the other hand, by the hypothesis that $\mathcal{LI} \preceq \mathcal{LI}'$, $\mathcal{LI}_{AF^*}(AF) \subseteq \mathcal{LI}'_{AF^*}(AF)$ which by (3) yields $\mathcal{LI}_{AF^*}(AF) \subseteq AF''$. Thus the first option holds, yielding the following expression for the term:

$$F(AF, \mathcal{LI}_{AF^*}(AF), Lab{\downarrow}_{\mathcal{LI}_{AF^*}(AF)\setminus AF})$$

which by (2) is equivalent to $F(AF, AF', Lab{\downarrow}_{AF'\setminus AF})$, i.e. the second term of 5, and we are done. □

Note that the constraints introduced in Definition 11 are crucial in the above proof.

In short, the partial order $\preceq$ between local information functions has a direct impact on the capability of capturing the global labellings through local computations. It is then interesting to determine the sets of semantics that are decomposable under the minimum and maximum (w.r.t. $\preceq$) local information functions, i.e. $m\mathcal{LI}$ and $M\mathcal{LI}$, respectively.

Proposition 9. *There are only four semantics satisfying conflict-freeness that are decomposable under $m\mathcal{LI}$:*

- *The semantics* $\mathbf{UND}$
- *The semantics* $\mathbf{S}$ *such that* $\forall AF \in SAF$, $\mathbf{L_S}(AF) = \emptyset$
- *The semantics* $\mathbf{S}$ *such that* $\mathbf{L_S}(AF) = \emptyset$ *if there is a self-attacking argument in* AF, $\mathbf{L_S}(AF) = \mathbf{L_{UND}}(AF)$ *otherwise*
- *The semantics* $\mathbf{S}$ *such that* $\mathbf{L_S}(AF) = \emptyset$ *if there is an argument which is not self-attacking in* AF, $\mathbf{L_S}(AF) = \mathbf{L_{UND}}(AF)$ *otherwise.*

Among these semantics, only $\mathbf{UND}$ *is universally defined.*

Proof. First, to show that the four semantics are fully decomposable, we note that all argumentation frameworks with input in $AF^{inp}_{m\mathcal{LI}}$ have the form $(AF, AF, \emptyset)$. We then select for each semantics $\mathbf{S}$ the local function F such that $F(AF, AF, \emptyset) = \mathbf{L_S}(AF)$ as defined above. It is then easy to see that F enforces decomposability of $\mathbf{S}$ under $m\mathcal{LI}$. In particular, this is immediate for $\mathbf{UND}$ (where the local function always assigns the label `undec` to all arguments) and for the second semantics (where the local function never returns any labelling). As to the third semantics, given an $AF = (\mathcal{A}, att) \in SAF$ we distinguish two cases. If $\exists \alpha \in \mathcal{A}$ such that α is self-attacking, then by definition $\mathbf{L_S}(AF) = \emptyset$. Given a partition $L_{P1} \cup \ldots \cup L_{Pn}$ of $\mathcal{A}$, there must be a partition element L_{Pj} such that $\alpha \in L_{Pj}$, thus $F(AF{\downarrow}_{L_{Pj}}, AF{\downarrow}_{L_{Pj}}, \emptyset) = \emptyset$. As a consequence[5], it holds that

$$\{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in F(AF{\downarrow}_{L_{Pi}}, AF{\downarrow}_{L_{Pi}}, \emptyset)\} = \emptyset$$

[5] It should be noted that in $AF{\downarrow}_{L_{Pj}}$ the local function does not return a set including the empty labelling (i.e. $\{\emptyset\}$), rather it returns the empty set (i.e. $\emptyset$) which prevents from obtaining a combination of local labellings, i.e. there is no $L_{P1} \cup \ldots \cup L_{Pn}$ such that $L_{Pi} \in F(AF{\downarrow}_{L_{Pi}}, AF{\downarrow}_{L_{Pi}}, \emptyset)$.

In the other case, there is no self-attacking argument in $\mathcal{A}$, thus by definition $\mathbf{L_S}(AF) = \mathbf{L_{UND}}(AF)$. For any partition element the local function always assigns the label `undec` to all arguments, thus the combination of local labellings coincides with the labelling of $\mathbf{L_{UND}}(AF)$. Similar considerations apply to the fourth semantics.

Now, it is immediate to verify that all of the four semantics satisfy conflict-freeness, and among them only **UND** is universally defined.

To show that there are no other semantics satisfying conflict-freeness that are decomposable under $m\mathcal{LI}$, for any $AF = (\mathcal{A}, att) \in SAF$, consider the partition $\mathcal{P} = \{\{\alpha\} \mid \alpha \in \mathcal{A}\}$, i.e. consisting of the sets each of them including a single argument. If $\mathbf{S}$ is decomposable under $m\mathcal{LI}$, according to Definition 22 we must have, letting $\mathcal{A} = \{\alpha_1, \ldots, \alpha_n\}$,

$$\mathbf{L_S}(AF) = \{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in F(AF{\downarrow_{\{\alpha_i\}}}, AF{\downarrow_{\{\alpha_i\}}}, \emptyset)\} \tag{6}$$

Note that given an argument α there are only two possibilities for $AF{\downarrow_{\{\alpha\}}}$, i.e. $AF_1 = (\{\alpha\}, \emptyset)$ if α is not self-attacking and $AF_2 = (\{\alpha\}, \{(\alpha, \alpha\})$ otherwise. Let us then evaluate the possible outcomes for $F(AF_1, AF_1, \emptyset)$ and $F(AF_2, AF_2, \emptyset)$. First, the labelling $\{(\alpha, \mathtt{out})\}$ can be ruled out for both $F(AF_1, AF_1, \emptyset)$ and $F(AF_2, AF_2, \emptyset)$ by considering the condition (6) applied to AF_1 and AF_2, since the resulting labelling $\{(\alpha, \mathtt{out})\}$ would violate the second condition of Definition 6. Also the labelling $\{(\alpha, \mathtt{in})\}$ can be ruled out. In particular, as to $F(AF_2, AF_2, \emptyset)$ it is again sufficient to consider the condition (6) applied to AF_2, since the resulting labelling $\{(\alpha, \mathtt{in})\}$ would violate the first condition of Definition 6. As to $F(AF_1, AF_1, \emptyset)$, in the argumentation framework $AF = (\{\alpha_1, \alpha_2\}, \{(\alpha_1, \alpha_2)\})$ the condition (6) would prescribe (possibly among others) the labelling $\{(\alpha_1, \mathtt{in}), (\alpha_2, \mathtt{in})\}$, violating the first condition of Definition 6. As a consequence, only four cases are possible, and according to (6) they correspond to the four semantics above in the relevant order, i.e.

- $F(AF_1, AF_1, \emptyset) = \{\{(\alpha, \mathtt{undec})\}\}$ and $F(AF_2, AF_2, \emptyset) = \{\{(\alpha, \mathtt{undec})\}\}$
- $F(AF_1, AF_1, \emptyset) = \emptyset$ and $F(AF_2, AF_2, \emptyset) = \emptyset$
- $F(AF_1, AF_1, \emptyset) = \{\{(\alpha, \mathtt{undec})\}\}$ and $F(AF_2, AF_2, \emptyset) = \emptyset$
- $F(AF_1, AF_1, \emptyset) = \emptyset$ and $F(AF_2, AF_2, \emptyset) = \{\{(\alpha, \mathtt{undec})\}\}$

□

Proposition 10. *Every semantics* **S** *is decomposable under* $M\mathcal{LI}$.

Proof. For a semantics $\mathbf{S}$, we consider the local function F for $\mathcal{MLI}$ defined as $F(AF, AF', Lab) \equiv \{Lab'{\downarrow}_{\mathcal{A}} \mid Lab' \in \mathbf{L_S}(AF') \wedge Lab'{\downarrow}_{AF' \setminus AF} = Lab\}$, where $\mathcal{A}$ denotes the set of arguments of AF.

We have to prove that for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$, it holds that $\mathbf{L_S}(AF) = \{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in F(AF{\downarrow}_{P_i}, \mathcal{MLI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{\mathcal{MLI}_{AF}(AF{\downarrow}_{P_i}) \setminus AF{\downarrow}_{P_i}})\} = \{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in F(AF{\downarrow}_{P_i}, AF, (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{AF \setminus AF{\downarrow}_{P_i}})\}$, where by the definition of F and taking into account that $AF = (\mathcal{A}, att)$ we have that

$$F(AF{\downarrow}_{P_i}, AF, (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{AF \setminus AF{\downarrow}_{P_i}}) \\ = \{Lab{\downarrow}_{P_i} \mid Lab \in \mathbf{L_S}(AF) \wedge Lab{\downarrow}_{\mathcal{A} \setminus P_i} = (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{\mathcal{A} \setminus P_i}\} \quad (7)$$

Let us first consider a labelling $Lab \in \mathbf{L_S}(AF)$. Since $\mathcal{P}$ is a partition of $\mathcal{A}$, it obviously holds that $Lab = L_{P1} \cup \ldots \cup L_{Pn}$ with $L_{Pi} = Lab{\downarrow}_{P_i}$ and $Lab{\downarrow}_{\mathcal{A} \setminus P_i} = (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{\mathcal{A} \setminus P_i}$. According to (7) we then have that

$$L_{Pi} \in F(AF{\downarrow}_{P_i}, AF, (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{AF \setminus AF{\downarrow}_{P_i}})$$

as desired.

Let us then consider a collection of labellings L_{Pi} for $i = 1 \ldots n$ such that $L_{Pi} \in F(AF{\downarrow}_{P_i}, AF, (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{AF \setminus AF{\downarrow}_{P_i}})$, and let us prove that $L_{P1} \cup \ldots \cup L_{Pn} \in \mathbf{L_S}(AF)$. According to (7), for each L_{Pi} there is a labelling $Lab^i \in \mathbf{L_S}(AF)$ such that $L_{Pi} = Lab^i{\downarrow}_{P_i}$ and $Lab^i{\downarrow}_{\mathcal{A} \setminus P_i} = (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{\mathcal{A} \setminus P_i}$. We show in the following that, according to the last condition,

$$Lab^1 = \ldots = Lab^n = L_{P1} \cup \ldots \cup L_{Pn} \quad (8)$$

entailing the desired conclusion that $L_{P1} \cup \ldots \cup L_{Pn} \in \mathbf{L_S}(AF)$. To this purpose, consider a labelling Lab^i and a partition element P_j, where $i, j \in \{1 \ldots n\}$. If $i = j$, then $L_{Pi} = Lab^i{\downarrow}_{P_i}$. If $i \neq j$, $Lab^i{\downarrow}_{\mathcal{A} \setminus P_i} = (\bigcup_{j=1\ldots n, j \neq i} L_{Pj}){\downarrow}_{\mathcal{A} \setminus P_i}$ entails that $Lab^i{\downarrow}_{P_j} = L_{Pj}$. Summing up, for any Lab^i and for any P_j we have that Lab^i coincides with L_{Pj} in P_j, i.e. (8) holds. □

Summing up, if complete information on the global argumentation framework is available to the local computations, then all semantics become decomposable. If no external information is available, decomposable semantics are only those that are maximally undecided (i.e. those leaving all arguments undecided). This seems to be perfectly reasonable behavior, confirming the suitability of our model and the adopted definition of decomposability.

A constructive procedure for local functions
Standard argumentation framework function f_{ST} for $\mathcal{LI}$: ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab) \subseteq \{AF^* \mid (AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},AF^*,\mathbf{S}}\}$ (Def. 24)
Local function generated by f_{ST} for $\mathbf{S}$ and $\mathcal{LI}$: $F_{f_{ST},\mathbf{S},\mathcal{LI}}$ (Def. 25)
Canonical local function
Canonical local function $F^{\mathcal{LI}}_{\mathbf{S}}$ of $\mathbf{S}$ associated to $\mathcal{LI}$: $F^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$ (Def. 27)
Reduced canonical local functions
AF^* representing $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ under $\mathbf{S}$ and $\mathcal{LI}$: $AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$ (Def. 28)
Pair (AF^i, Lab) derived from $\mathcal{LI}$: $(AF^i, Lab) \in P_{\mathcal{LI}}$ (Def. 29)
$(AF^i, Lab) \in P_{\mathcal{LI}}$ representable under $\mathbf{S}$ and $\mathcal{LI}$: $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$ (Def. 30)
(AF^i, Lab) realized under $\mathbf{S}$: $(AF^i, Lab) \in P^{real}_{\mathbf{S}}$ (Def. 31)
Semantics $\mathbf{S}$ representable | weakly representable w.r.t. $\mathcal{LI}$ (Def. 32)
Reduced canonical local function of $\mathbf{S}$ w.r.t. $\mathcal{LI}$: $RF^{\mathcal{LI}}_{\mathbf{S}}$ (Def. 33)

Table 2: Main definitions and notations introduced in Sections 5, 6 and 7.

5 A constructive procedure for local functions

Once the general model has been designed, the next issue is to identify a local function for any argumentation semantics $\mathbf{S}$ and local information function $\mathcal{LI}$. This issue is addressed in Sections 5, 6 and 7, and Table 2 lists the relevant main definitions and notations.

In order to provide a guidance to the identification of a local function which is valid independently of the specific semantics definitions, we aim at identifying an expression of the local function which is parametric w.r.t. the semantics, and thus does not rely on the properties of a specific semantics.

The expression of the local function is based on the following considerations. First, given an argumentation framework with input $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$, the only way to determine the set of labellings returned as output by the local function on the basis of the semantics $\mathbf{S}$ (given as a parameter) is to apply $\mathbf{S}$ to a set of argumentation frameworks. Since the set of labellings returned by the local function is contained in $\mathfrak{L}(AF)$, each of these argumentation frameworks AF^* must have AF as a subframework, i.e. $AF \sqsubseteq AF^*$, and the returned labellings are obtained by restricting (some of) the labellings in $\mathbf{L}_{\mathbf{S}}(AF^*)$ to AF. Moreover, taking into account the role of AF' and Lab, the argumentation with input (AF, AF', Lab) has to be realized in AF^* from $\mathcal{LI}$, and only the labellings $Lab^* \in \mathbf{L_S}(AF^*)$ compatible with Lab (i.e. such that $Lab^*\downarrow_{AF'\setminus AF} = Lab$) should be taken into account.

In order to model all possible selections of argumentation frameworks for any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$, we introduce the notion of *standard argumentation framework function*, which associates to any argumentation framework with input derived from $\mathcal{LI}$ a (possibly empty) set of argumentation frameworks in which this argumentation framework with input is realized.

Definition 24. *Given a local information function $\mathcal{LI}$, a* standard argumentation framework function *f_{ST} for $\mathcal{LI}$ is a (possibly partial) function which associates to any pair including a semantics $\mathbf{S}$ and an argumentation framework with input $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$, a set of argumentation frameworks, denoted as ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$, such that ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab) \subseteq \{AF^* \mid (AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},AF^*,\mathbf{S}}\}$. A standard argumentation framework function for $\mathcal{LI}$ is* finite *if, $\forall$ $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$, ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ is finite. It is* unitary *if, $\forall$ $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$, either ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ includes a single framework or it is empty.*

Note that if $(AF, AF', Lab) \notin RAF^{inp}_{\mathcal{LI},\mathbf{S}}$ then ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ is not defined, i.e. returns the empty set.

Intuitively, the aim of ${f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ is to provide a set of argumentation frameworks 'representing' all argumentation frameworks where (AF, AF', Lab) can be realized, meaning that such a set is sufficient to construct the output of a local function F. In particular, given a standard argumentation framework function f_{ST} for $\mathcal{LI}$, for any semantics a corresponding local function for $\mathcal{LI}$ can be generated as in the following definition.

Definition 25. *Given a semantics $\mathbf{S}$ and a standard argumentation framework function f_{ST} for a local information function $\mathcal{LI}$, the local function generated by f_{ST} for $\mathbf{S}$ and $\mathcal{LI}$, denoted as $F_{f_{ST},\mathbf{S},\mathcal{LI}}$, is the local function for $\mathcal{LI}$ such that for any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$*

$$F_{f_{ST},\mathbf{S},\mathcal{LI}}(AF, AF', Lab) = \bigcup_{AF^* \in {f_{ST}}^{\mathbf{S},\mathcal{LI}}(AF,AF',Lab)} \{Lab^*\downarrow_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*), \\ Lab^*\downarrow_{AF'\setminus AF} = Lab\}$$

It is easy to see that a monotonic relation between standard argumentation framework functions and generated local functions holds.

Proposition 11. *Given two standard argumentation framework functions f^1_{ST} and f^2_{ST} for $\mathcal{LI}$ and a semantics* $\mathbf{S}$*, if* ${f^1_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab) \subseteq {f^2_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ *then* $F_{f^1_{ST},\mathbf{S},\mathcal{LI}}(AF, AF', Lab) \subseteq F_{f^2_{ST},\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$*.*

Proof. The result easily follows from Definitions 24 and 25. □

We now establish two requirements for a standard argumentation framework function.

First, constructing a local function on the basis of a standard argumentation framework function is easier if the latter is finite. Luckily, since we deal with finite argumentation frameworks, for any generated local function there is always a finite standard argumentation framework function which generates it.

Proposition 12. *Given a standard argumentation framework function f^1_{ST} for a local information function $\mathcal{LI}$ and given a semantics* $\mathbf{S}$*, there exists a finite standard argumentation framework function f^2_{ST} for $\mathcal{LI}$ which generates $F_{f^1_{ST},\mathbf{S},\mathcal{LI}}$.*

Proof. We construct f^2_{ST} generating $F_{f^1_{ST},\mathbf{S},\mathcal{LI}}$ as follows. According to Definition 25, for any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ the output of $F_{f^1_{ST},\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ can be expressed as

$$\bigcup_{AF^* \in {f^1_{ST}}^{\mathbf{S},\mathcal{LI}}(AF,AF',Lab)} \{Lab^*{\downarrow}_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*) \land Lab^*{\downarrow}_{AF' \setminus AF} = Lab\}$$

Since the number of possible labellings of AF, i.e. the cardinality of $\mathfrak{L}(AF)$, is 3^n where n is the number of arguments in AF, obviously the number of distinct labellings $Lab^*{\downarrow}_{AF}$ in the set above is finite as well. Thus there is a finite set of argumentation frameworks, that we let as ${f^2_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$, such that

$$F_{f^1_{ST},\mathbf{S},\mathcal{LI}}(AF, AF', Lab) = \bigcup_{AF^* \in {f^2_{ST}}^{\mathbf{S},\mathcal{LI}}(AF,AF',Lab)} \{Lab^*{\downarrow}_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*) \land Lab^*{\downarrow}_{AF' \setminus AF} = Lab\}$$

This corresponds[6] to our desired f^2_{ST} (see Definition 25). □

Let us now turn to the second requirement. Since by definition a decomposable semantics $\mathbf{S}$ under a local information function $\mathcal{LI}$ admits a (possibly singleton) set of local functions that enforce decomposability of $\mathbf{S}$ under $\mathcal{LI}$, failing to capture at least one of them would not be acceptable for the above construction mechanism. This is expressed by the following definition.

[6] Since we do not specify how the set of argumentation frameworks in ${f^2_{ST}}^{\mathbf{S},\mathcal{LI}}(AF, AF', Lab)$ is selected, the possibility of constructing the function f^2_{ST} relies on the axiom of choice.

Definition 26. *A standard argumentation framework function f_{ST} for $\mathcal{LI}$ is* adequate *if, for every decomposable semantics* $\mathbf{S}$ *under $\mathcal{LI}$, $F_{f_{ST},\mathbf{S},\mathcal{LI}}$ enforces decomposability of* $\mathbf{S}$ *under $\mathcal{LI}$.*

An adequate standard argumentation framework function f_{ST} is pivotal for investigating the decomposability property of a semantics $\mathbf{S}$, since it allows one to select without loss of generality the local function in the condition of Definition 22. In particular, by Definition 26, with f_{ST} adequate, $F_{f_{ST},\mathbf{S},\mathcal{LI}}$ enforces decomposability of $\mathbf{S}$ if the latter is fully decomposable under a local information function $\mathcal{LI}$. Then, the proof that $\mathbf{S}$ is fully decomposable under $\mathcal{LI}$ can focus on the condition of Definition 22 with $F = F_{f_{ST},\mathbf{S},\mathcal{LI}}$, and conversely to show that a semantics is not decomposable it is sufficient to identify an argumentation framework and a partition of its arguments where the same condition is not satisfied by $F_{f_{ST},\mathbf{S},\mathcal{LI}}$.

A significant question is then whether Definitions 24 and 25 or, more generally, the assumptions underlying them, are general enough to capture useful local functions, i.e. whether there is (at least) one adequate standard argumentation framework function. In the next sections we provide a positive answer to this question.

6 The canonical local function

In this section we consider a particular choice for a standard argumentation framework function, motivated by the fact that, as shown in the following proposition, any local function enforcing decomposability includes in its output, for any AF^* such that $(AF, AF', Lab) \in RAF^{inp}_{\mathcal{LI},AF^*,\mathbf{S}}$, the restriction of the labellings of AF^* to the subframework AF.

Proposition 13. *Consider a fully decomposable semantics* $\mathbf{S}$ *under $\mathcal{LI}$, and let $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ be an argumentation framework with input derived from $\mathcal{LI}$. Let AF^* be an argumentation framework such that $AF' = \mathcal{LI}_{AF^*}(AF)$, and $Lab^* \in$ $\mathbf{L}_{\mathbf{S}}(AF^*)$ be a labelling of AF^* such that $Lab^*{\downarrow}_{AF' \setminus AF} = Lab$. Then, for any local function F which enforces decomposability of* $\mathbf{S}$ *under $\mathcal{LI}$, $Lab^*{\downarrow}_{AF} \in F(AF, AF', Lab)$.*

Proof. Taking into account that F enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$, let us apply the condition of Definition 22 to AF^* with the partition $\mathcal{P}$ of its arguments corresponding to the two subframeworks AF and $AF^* \setminus AF$. According to this condition, we have in particular that $Lab^* = Lab_1 \cup Lab_2$ with $Lab_1 = Lab^*{\downarrow}_{AF^* \setminus AF}$, $Lab_2 = Lab^*{\downarrow}_{AF}$ and $Lab_2 \in F(AF, \mathcal{LI}_{AF^*}(AF), Lab_1{\downarrow}_{\mathcal{LI}_{AF^*}(AF) \setminus AF})$. Since by the hypothesis $AF' = \mathcal{LI}_{AF^*}(AF)$, $Lab_1{\downarrow}_{\mathcal{LI}_{AF^*}(AF) \setminus AF} = Lab_1{\downarrow}_{AF' \setminus AF}$, and since $Lab_1 \in$

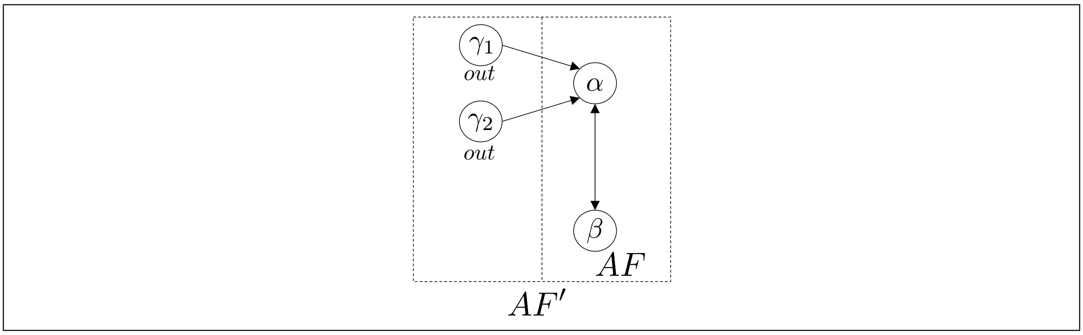


Figure 2: An argumentation framework with input.

$\mathfrak{L}(AF^* \setminus AF)$ the term is in turn equal to $Lab^*{\downarrow}_{AF' \setminus AF}$ which by the hypothesis is Lab. Remembering that $Lab_2 = Lab^*{\downarrow}_{AF}$, we get $Lab^*{\downarrow}_{AF} \in F(AF, AF', Lab)$. □

We should note that the reverse of the above proposition does not hold, i.e. F may require additional labellings w.r.t. those mentioned in the proposition. A labelling included in $F(AF, AF', Lab)$ may not play a role in forming the labellings of AF^* due to the compatibility conditions, but it may be required in a different argumentation framework. This suggests adopting the following definition of the *canonical local function*, which includes all possible labellings that play a role in some argumentation framework.

Definition 27. *Given a semantics* $\mathbf{S}$ *and a local information function* $\mathcal{LI}$*, the canonical local function* $F_{\mathbf{S}}^{\mathcal{LI}}$ *of* $\mathbf{S}$ *associated to* $\mathcal{LI}$ *is defined as follows. For any* $(AF, AF', Lab) \in AF_{\mathcal{LI}}^{inp}$*,*

$$F_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab) = \bigcup_{AF^* | (AF, AF', Lab) \in RAF_{\mathcal{LI}, AF^*, \mathbf{S}}^{inp}} \{Lab^*{\downarrow}_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*) \wedge Lab^*{\downarrow}_{AF' \setminus AF} = Lab\}$$

It is easy to see that the canonical local function of a semantics $\mathbf{S}$ associated to $\mathcal{LI}$ is the local function generated by the maximal standard argumentation framework function for $\mathbf{S}$ and $\mathcal{LI}$, i.e. returning as output *all* argumentation frameworks AF^* such that $(AF, AF', Lab) \in RAF_{\mathcal{LI}, AF^*, \mathbf{S}}^{inp}$ (see Definition 24).

Example 5. *Consider* (AF, AF', Lab) *with* $AF = (\{\alpha, \beta\}, \{(\alpha, \beta), (\beta, \alpha)\})$*,* $AF' = (\{\alpha, \beta, \gamma_1, \gamma_2\}, \{(\alpha, \beta), (\beta, \alpha), (\gamma_1, \alpha), (\gamma_2, \alpha)\})$ *and* $Lab = \{(\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out})\}$*. We provide a pictorial representation in Figure 2 It is easy to see that* $(AF, AF', Lab) \in$

$AF^{inp}_{inp\mathcal{LI}}$, *since e.g. in Example 3 it holds that* $AF' = inp\mathcal{LI}_{AF^*}(AF)$. *We determine* $F^{inp\mathcal{LI}}_{\mathbf{PR}}(AF, AF', Lab)$.

To this purpose, we consider all $AF^* \in SAF$ *such that* $AF' = inp\mathcal{LI}_{AF^*}(AF)$ *and* $\exists Lab^* \in \mathbf{L_{PR}}(AF^*)$ *with* $Lab^*\downarrow_{\{\gamma_1,\gamma_2\}} = \{(\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out})\}$.

First, since any $Lab^* \in \mathbf{L_{PR}}(AF^*)$ *is a complete labelling by definition, the relevant constraints specified in Definition 9 leaves only three possibilities for the labels assigned to* α *and* β*:*

1. $Lab^*(\alpha) = \mathtt{undec}$ *and* $Lab^*(\beta) = \mathtt{undec}$

2. $Lab^*(\alpha) = \mathtt{in}$ *and* $Lab^*(\beta) = \mathtt{out}$

3. $Lab^*(\alpha) = \mathtt{out}$ *and* $Lab^*(\beta) = \mathtt{in}$

In fact, taking into account that there are no further attacks towards α *and* β *besides those in* AF', *according to Definition 9 if* α *is labelled* `in` *then* β *must be labelled* `out`, *if* α *is labelled* `out` *then* β *must be labelled* `in`, *if* α *is labelled* `undec` *then* β *cannot be labelled* `out` *or* `in`, *thus it must be labelled* `undec`.

On the other hand, the first possibility is ruled out by maximality of preferred labellings. We can reason by contradiction, considering $Lab^* \in \mathbf{L_{PR}}(AF^*)$ *such that* $Lab^*(\alpha) = \mathtt{undec}$ *and* $Lab^*(\beta) = \mathtt{undec}$. *Let* $Lab^{*\prime}$ *be the labelling such that* $Lab^{*\prime}(\alpha) = \mathtt{in}$, $Lab^{*\prime}(\beta) = \mathtt{out}$, *and for all other arguments the labels assigned by* $Lab^{*\prime}$ *coincide with those assigned by* Lab^*. *Taking into account that* Lab^* *is preferred and that the attacks towards* α *(and* β*) are identified as in* AF', *it is easy to verify that* $Lab^{*\prime}$ *is admissible (see Definition 7). However,* $Lab^* \sqsubseteq Lab^{*\prime}$ *while the reverse does not hold, contradicting the fact that* Lab^* *is a maximal admissible labelling (see Proposition 2).*

As to the other two possibilities, let us refer again to Example 3 and Figure 1. There are two preferred labellings in AF^*, *i.e.*

$$\{(\delta_1, \mathtt{in}), (\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out}), (\alpha, \mathtt{in}), (\beta, \mathtt{out}), (\gamma_3, \mathtt{in}), (\delta_2, \mathtt{out}), (\eta, \mathtt{in})\}$$

and

$$\{(\delta_1, \mathtt{in}), (\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out}), (\alpha, \mathtt{out}), (\beta, \mathtt{in}), (\gamma_3, \mathtt{out}), (\delta_2, \mathtt{in}), (\eta, \mathtt{out})\}$$

thus $\{(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}$ *and* $\{(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}$ *belong to* $F^{inp\mathcal{LI}}_{\mathbf{PR}}(AF, AF', Lab)$.

Summing up, it holds that

$$F^{inp\mathcal{LI}}_{\mathbf{PR}}(AF, AF', Lab) = \{\{(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}, \{(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}\}$$

Due to the choice of considering all possible labellings compliant with Definitions 24 and 25, the canonical local function of any semantics $\mathbf{S}$ associated to any local information function $\mathcal{LI}$ enforces top-down decomposability of $\mathbf{S}$ under $\mathcal{LI}$, as shown in the following proposition.

Proposition 14. *For any semantics* $\mathbf{S}$ *and local information function* $\mathcal{LI}$*, the canonical local function* $F_{\mathbf{S}}^{\mathcal{LI}}$ *enforces top-down decomposability of* $\mathbf{S}$ *under* $\mathcal{LI}$*.*

Proof. According to Definition 23, we have to prove that for every $AF = (\mathcal{A}, att)$, for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$ and for any labelling $Lab \in \mathbf{L_S}(AF)$, it holds that

$$\begin{aligned} Lab = L_{P1} \cup \ldots \cup L_{Pn} \mid \\ & L_{Pi} \in F_{\mathbf{S}}^{\mathcal{LI}}(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}}) \end{aligned}$$

For any $i \in \{1, \ldots n\}$, let $L_{Pi} \equiv Lab{\downarrow}_{P_i}$. It holds that $Lab = L_{P1} \cup \ldots \cup L_{Pn}$, thus, for any i, $(\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}} = Lab{\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}}$. As a consequence, we have to prove that for any $i \in \{1, \ldots n\}$

$$Lab{\downarrow}_{P_i} \in F_{\mathbf{S}}^{\mathcal{LI}}(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), Lab{\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}})$$

According to the definition of canonical local function (see Definition 27), this amounts to prove that there is an argumentation framework AF^* and a labelling $Lab^* \in \mathbf{L_S}(AF^*)$ such that $\mathcal{LI}_{AF^*}(AF{\downarrow}_{P_i}) = \mathcal{LI}_{AF}(AF{\downarrow}_{P_i})$, $Lab^*{\downarrow}_{AF{\downarrow}_{P_i}} = Lab{\downarrow}_{P_i}$ and $Lab^*{\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}} = Lab{\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\backslash AF{\downarrow}_{P_i}}$. It is easy to see that all these conditions are satisfied by selecting $AF^* = AF$ and $Lab^* = Lab$. In particular, $Lab \in \mathbf{L_S}(AF)$ holds by assumption, $\mathcal{LI}_{AF}(AF{\downarrow}_{P_i}) = \mathcal{LI}_{AF^*}(AF{\downarrow}_{P_i})$ is trivially satisfied, the third condition holds since $Lab{\downarrow}_{AF{\downarrow}_{P_i}} = Lab{\downarrow}_{P_i}$, and finally the last condition trivially holds since $Lab^* = Lab$. □

While top-down decomposability holds for all semantics, i.e. the output of the canonical local function is sufficient to cover all global labellings, the following proposition shows that the output of the canonical local function is necessary to enforce decomposability whenever this is possible, i.e. if the semantics is fully decomposable.

Proposition 15. *Let* $\mathbf{S}$ *be a decomposable semantics under* $\mathcal{LI}$ *and let* F *be a local function which enforces decomposability of* $\mathbf{S}$ *under* $\mathcal{LI}$*. Then,* $\forall(AF, AF', Lab) \in AF_{\mathcal{LI}}^{inp}$*,* $F_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab) \subseteq F(AF, AF', Lab)$*.*

Proof. The proof is an immediate consequence of Proposition 13. □

The reverse of this proposition does not hold since a local function enforcing decomposability can prescribe for a subframework spurious labellings that are not compatible with those of the other subframeworks and thus do not alter the set of labellings obtained by joining the results of local computations.

The above results are sufficient to show that the canonical local function enforces decomposability of all decomposable semantics.

Proposition 16. *If a semantics* $\mathbf{S}$ *is fully decomposable under a local information function* $\mathcal{LI}$*, then* $F_{\mathbf{S}}^{\mathcal{LI}}$ *enforces decomposability of* $\mathbf{S}$ *under* $\mathcal{LI}$*.*

Proof. By the hypothesis there is a local function F for $\mathcal{LI}$ such that for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$

$$\begin{aligned}\mathbf{L}_{\mathbf{S}}(AF) = \{&L_{P1} \cup \ldots \cup L_{Pn} \mid \\ &L_{Pi} \in F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})\}\end{aligned} \tag{9}$$

and we have to prove that for every $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$

$$\begin{aligned}\mathbf{L}_{\mathbf{S}}(AF) = \{&L_{P1} \cup \ldots \cup L_{Pn} \mid \\ &L_{Pi} \in F_{\mathbf{S}}^{\mathcal{LI}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})\}\end{aligned}$$

First, let us consider $Lab \equiv L_{P1} \cup \ldots \cup L_{Pn}$ such that for every $i \in \{1, \ldots, n\}$ $L_{Pi} \in F_{\mathbf{S}}^{\mathcal{LI}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})$. By Proposition 15

$$\begin{aligned}&F_{\mathbf{S}}^{\mathcal{LI}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}) \\ &\qquad\subseteq F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})\end{aligned}$$

Thus by (9) it holds that $Lab \in \mathbf{L}_{\mathbf{S}}(AF)$.

The reverse direction amounts to show that $F_{\mathbf{S}}^{\mathcal{LI}}$ enforces top-down decomposability of $\mathbf{S}$ under $\mathcal{LI}$, and follows from Proposition 14. □

According to Proposition 15 and Proposition 16, the canonical local function of a decomposable semantics $\mathbf{S}$ associated to $\mathcal{LI}$ is the minimal (w.r.t. $\subseteq$) local function enforcing decomposability.

7 Reduced canonical local functions

As mentioned in the previous section, Proposition 13 identifies an argumentation framework AF^* and a relevant set of labellings that are necessary to enforce decomposability. On the other hand, in general a single argumentation framework is not sufficient, i.e. for a given argumentation framework with input different argumentation frameworks may have to be identified in order to determine the whole set of labellings returned as output by the canonical local function.

A single argumentation framework is sufficient, however, if some conditions are verified. These conditions, expressed in the following definition, depend both on the semantics and the local information function.

Definition 28. *Let* $\mathbf{S}$ *be a semantics, let* $\mathcal{LI}$ *be a local information function and* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ *an argumentation framework with input derived from* $\mathcal{LI}$. *An argumentation framework* AF^* represents $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ *under* $\mathbf{S}$ *and* $\mathcal{LI}$, *written* $AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$, *if* $AF' = \mathcal{LI}_{AF^*}(AF)$, $\exists Lab'_1 \in \mathbf{L_S}(AF^* \setminus AF)$ *with* $Lab'_1{\downarrow}_{AF' \setminus AF} = Lab$, *and* $\mathcal{LI}_{AF^*}(AF^* \setminus AF) = AF^* \setminus AF$.

Some comments on the conditions of Definition 28 are in order. In particular, given AF^* such that $AF' = \mathcal{LI}_{AF^*}(AF)$ the key condition is $\mathcal{LI}_{AF^*}(AF^* \setminus AF) = AF^* \setminus AF$, which corresponds to a kind of unidirectional local information function, i.e. while AF is influenced by (part of) $AF^* \setminus AF$ and the relevant labelling, the reverse does not hold. Thus, the role of $AF^* \setminus AF$ is to enforce the labelling *Lab* in $AF' \setminus AF$ independently of AF. On the other hand, this is possible if a labelling compatible with *Lab* is prescribed by the semantics, i.e. $\exists Lab'_1 \in \mathbf{L_S}(AF^* \setminus AF)$ with $Lab'_1{\downarrow}_{AF' \setminus AF} = Lab$.

Example 6. *Referring to the argumentation framework with input* $(AF, AF', Lab) \in AF^{inp}_{inp\mathcal{LI}}$ *introduced in Example 5 (see Figure 2), consider the argumentation framework depicted in Figure 3*

$$AF^* = (\{i, \gamma_1, \gamma_2, \alpha, \beta\}, \{(i, \gamma_1), (i, \gamma 2), (\gamma_1, \alpha), (\gamma_2, \alpha), (\alpha, \beta), (\beta, \alpha\})$$

It turns out that AF^* *represents* (AF, AF', Lab) *under* $\mathbf{S}$ *and* $inp\mathcal{LI}$, *i.e.* $AF^* \in REP^{inp\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$, *for any semantics* $\mathbf{S} \in \{\mathbf{ST}, \mathbf{GR}, \mathbf{PR}\}$.

In particular, $AF' = inp\mathcal{LI}_{AF^*}(AF)$.

Moreover, $AF^* \setminus AF = (\{i, \gamma_1, \gamma_2\}, \{(i, \gamma_1), (i, \gamma 2)\})$ *and* $\mathbf{S}$ *prescribes a unique labelling for* $AF^* \setminus AF$, *i.e.* $\{(i, \texttt{in}), (\gamma_1, \texttt{out}), (\gamma_2, \texttt{out})\}$ *coinciding with Lab in* $\{\gamma_1, \gamma_2\}$, *namely the arguments of* $AF' \setminus AF$.

Finally, since there are no external attackers of $AF^* \setminus AF$, *the key condition* $\mathcal{LI}_{AF^*}(AF^* \setminus AF) = AF^* \setminus AF$ *holds, namely* $AF^* \setminus AF$ *is independent of* AF.

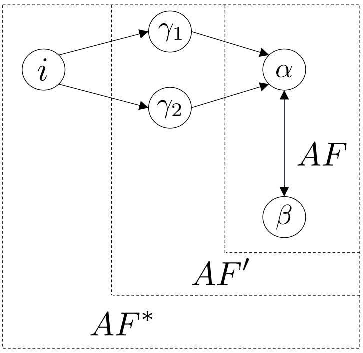


Figure 3: An argumentation framework representing an argumentation framework with input.

The next proposition shows that the reverse of Proposition 13 holds if the conditions of Definition 28 are satisfied, i.e. a single argumentation framework is sufficient to capture the entire output of the local function if it represents the argumentation framework with input.

Proposition 17. *Consider a fully decomposable semantics* $\mathbf{S}$ *under* $\mathcal{LI}$*, and let* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ *be an argumentation framework with input derived from* $\mathcal{LI}$*. Let* AF^* *be an argumentation framework such that* $AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$*. Then, for any local function* F *which enforces decomposability of* $\mathbf{S}$ *under* $\mathcal{LI}$*,* $F(AF, AF', Lab) = \{Lab^*{\downarrow}_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*) \wedge Lab^*{\downarrow}_{AF' \setminus AF} = Lab\}$*.*

Proof. Let us first consider a labelling $Lab^* \in \mathbf{L_S}(AF^*)$ such that $Lab^*{\downarrow}_{AF' \setminus AF} = Lab$. It is easy to see that all the hypotheses of Proposition 13 are satisfied, thus $Lab^*{\downarrow}_{AF} \in F(AF, AF', Lab)$.

As to the reverse direction of the proof, let us first consider the partition of AF^* identified by the subframeworks $AF^* \setminus AF$ and AF. Since by the hypothesis F enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$, according to Definition 22 and taking into account that $AF^*{\downarrow}_{AF^* \setminus AF} = AF^* \setminus AF$ we have that

$$\begin{aligned}\mathbf{L_S}(AF^*) = \{&Lab_1 \cup Lab_2 \mid \\ &Lab_1 \in F(AF^* \setminus AF, \mathcal{LI}_{AF^*}(AF^* \setminus AF), Lab_2{\downarrow}_{\mathcal{LI}_{AF^*}(AF^* \setminus AF) \setminus (AF^* \setminus AF)}), \\ &\qquad Lab_2 \in F(AF, \mathcal{LI}_{AF^*}(AF), Lab_1{\downarrow}_{\mathcal{LI}_{AF^*}(AF) \setminus AF})\}\end{aligned}$$

Taking into account that $\mathcal{LI}_{AF^*}(AF^* \setminus AF) = AF^* \setminus AF$ and $AF' = \mathcal{LI}_{AF^*}(AF)$, it follows that

$$\mathbf{L_S}(AF^*) = \{Lab_1 \cup Lab_2 \mid Lab_1 \in F(AF^* \setminus AF, AF^* \setminus AF, \emptyset), \\ Lab_2 \in F(AF, AF', Lab_1{\downarrow}_{AF' \setminus AF})\} \quad (10)$$

Let us then consider a labelling $Lab_2 \in F(AF, AF', Lab)$. We have to prove that $\exists Lab^* \in \mathbf{L_S}(AF^*)$ such that $Lab^*{\downarrow}_{AF} = Lab_2$ and $Lab^*{\downarrow}_{AF' \setminus AF} = Lab$.

By the hypothesis that $AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$, $\exists Lab_1 \in \mathbf{L_S}(AF^* \setminus AF)$ with $Lab_1{\downarrow}_{AF' \setminus AF} = Lab$. Let us now identify the labelling Lab^* we are looking for as $Lab_1 \cup Lab_2$. It obviously holds that $Lab^*{\downarrow}_{AF} = Lab_2$ and $Lab^*{\downarrow}_{AF' \setminus AF} = Lab_1{\downarrow}_{AF' \setminus AF} = Lab$. Thus, it remains to be proved that $Lab^* \in \mathbf{L_S}(AF^*)$.

Since $Lab_1 \in \mathbf{L_S}(AF^* \setminus AF)$, by Proposition 6 $Lab_1 \in F(AF^* \setminus AF, AF^* \setminus AF, \emptyset)$. Moreover, since $Lab_2 \in F(AF, AF', Lab)$ and $Lab_1{\downarrow}_{AF' \setminus AF} = Lab$, it holds that $Lab_2 \in F(AF, AF', Lab_1{\downarrow}_{AF' \setminus AF})$. Now, by (10) we have $Lab_1 \cup Lab_2 \in \mathbf{L_S}(AF^*)$, i.e. $Lab^* \in \mathbf{L_S}(AF^*)$. □

Example 7. *Continuing Example 6, it turns out that*

$$\mathbf{L_{ST}}(AF^*) = \mathbf{L_{PR}}(AF^*) = \\ = \{\{(i, \mathtt{in}), (\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out})(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}, \\ \{(i, \mathtt{in}), (\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out})(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}\}$$

and

$$\mathbf{L_{GR}}(AF^*) = \{\{(i, \mathtt{in}), (\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{out})(\alpha, \mathtt{undec}), (\beta, \mathtt{undec})\}\}$$

According to Proposition 17, if $\mathbf{S} \in \{\mathbf{ST}, \mathbf{PR}\}$ *then any local function* F *which enforces decomposability of* $\mathbf{S}$ *under inp*$\mathcal{LI}$ *must satisfy*

$$F(AF, AF', Lab) = \{\{(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}, \{(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}\}$$

and any local function F *which enforces decomposability of* $\mathbf{GR}$ *under inp*$\mathcal{LI}$ *must satisfy*

$$F(AF, AF', Lab) = \{\{(\alpha, \mathtt{undec}), (\beta, \mathtt{undec})\}\}$$

Note that it is not guaranteed that $\mathbf{ST}$*,* $\mathbf{PR}$ *or* $\mathbf{GR}$ *is decomposable, i.e.* F *may not exist.*

In order to exploit Proposition 17 to identify a local function generated by a unitary standard argumentation framework function, we need a number of preliminary definitions.

First, for a given argumentation framework with input (AF, AF', Lab) we need to focus on the pair $(AF' \setminus AF, Lab)$, playing for AF the role of the 'input pair' affecting the computation of labellings. Accordingly, we introduce the following definition of a pair derived from a local information function $\mathcal{LI}$.

Definition 29. *Given a local information function $\mathcal{LI}$, a pair (AF^i, Lab), where $AF^i \in SAF$ and $Lab \in \mathfrak{L}(AF^i)$, is derived from $\mathcal{LI}$, written $(AF^i, Lab) \in P_{\mathcal{LI}}$, if $\exists (AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ such that $AF' \setminus AF = AF^i$.*

A pair is *representable* if whenever it appears in an argumentation framework with input, the latter can be represented by an argumentation framework.

Definition 30. *Given a semantics* **S** *and a local information function $\mathcal{LI}$, a pair $(AF^i, Lab) \in P_{\mathcal{LI}}$ is* representable *under* **S** *and $\mathcal{LI}$, denoted as $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$, if for every $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ such that $AF' \setminus AF = AF^i$, we have that $\exists AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$.*

Similarly to the case of a realized argumentation framework with input (see Definition 20), we introduce the notion of realizability of a pair under a semantics.

Definition 31. *Given a semantics* **S***, a pair (AF^i, Lab) is* realized *under* **S***, written $(AF^i, Lab) \in P^{real}_{\mathbf{S}}$, if $\exists AF^* \in SAF$ such that $AF^i \subseteq AF^*$ and $\exists Lab^* \in \mathbf{L_S}(AF^*)$ such that $Lab^*\downarrow_{AF^i} = Lab$.*

In words, there must be an argumentation framework AF^* where $AF^i \subseteq AF^*$, i.e. AF^i appears as a potential external information for a subframework of AF^*, and the semantics enforces the labelling Lab in AF^i. As shown below, if a pair is representable under **S** and $\mathcal{LI}$ then it is also realized under **S**.

Proposition 18. *Given a semantics* **S***, a local information function $\mathcal{LI}$ and a pair $(AF^i, Lab) \in P_{\mathcal{LI}}$, if $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$, then $(AF^i, Lab) \in P^{real}_{\mathbf{S}}$.*

Proof. Since $(AF^i, Lab) \in P_{\mathcal{LI}}$, by Definition 29 $\exists (AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ such that $AF' \setminus AF = AF^i$. Since $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$, by Definition 30 $\exists AF^{**} \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$. Taking into account Definitions 28 and 11, we have in particular $AF \sqsubseteq AF^{**}$, $AF' \subseteq AF^{**}$, and $\exists Lab'_1 \in \mathbf{L_S}(AF^{**} \setminus AF)$ with $Lab'_1\downarrow_{AF' \setminus AF} = Lab$. Letting $AF^* = AF^{**} \setminus AF$ and $Lab^* = Lab'_1$, it must be the case that $AF' \setminus AF \subseteq AF^*$, i.e. $AF^i \subseteq AF^*$, and $\exists Lab^* \in \mathbf{L_S}(AF^*)$ with $Lab^*\downarrow_{AF' \setminus AF} = Lab^*\downarrow_{AF^i} = Lab$. According to Definition 31, $(AF^i, Lab) \in P^{real}_{\mathbf{S}}$. □

On the basis of Proposition 17, if all pairs are representable (and thus realized) then it is possible to construct a local function by means of a unitary standard argumentation framework function. However, this requirement may be impossible to achieve just because of pairs that are not realized under the semantics (and thus cannot be representable as shown by Proposition 18). Then, a weaker requirement is that realized pairs are representable. We introduce accordingly the following definition.

Definition 32. *A semantics* $\mathbf{S}$ *is* representable *w.r.t. a local information function* $\mathcal{LI}$ *if for every* $(AF^i, Lab) \in P_{\mathcal{LI}}$*, it holds that* $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$*, i.e. every pair derived from* $\mathcal{LI}$ *is representable under* $\mathbf{S}$ *and* $\mathcal{LI}$*. A semantics* $\mathbf{S}$ *is* weakly representable *w.r.t.* $\mathcal{LI}$ *if for every* $(AF^i, Lab) \in P_{\mathcal{LI}}$ *such that* $(AF^i, Lab) \in P^{real}_{\mathbf{S}}$*, it holds that* $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$*, i.e. every realized pair under* $\mathbf{S}$ *is representable under* $\mathbf{S}$ *and* $\mathcal{LI}$*.*

It is obvious that a representable semantics w.r.t. $\mathcal{LI}$ is also weakly representable w.r.t. $\mathcal{LI}$.

Example 8. *Consider the local information function* $inp\mathcal{LI}$*.*

Under most semantics $\mathbf{S}$*, in particular if* $\mathbf{S} \in \{\mathbf{GR}, \mathbf{PR}\}$*, for any* $(AF^i, Lab) \in P_{inp\mathcal{LI}}$ *it turns out that* $(AF^i, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$*, thus* $\mathbf{S}$ *is representable w.r.t.* $inp\mathcal{LI}$*. In order to show this, given* $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ *such that* $AF' \setminus AF = AF^i$*, letting* $AF = (\mathcal{A}, att)$ *and* $AF' = (\mathcal{A}', att')$ *we consider the argumentation framework*

$$AF^* = (\mathcal{A}' \cup \{i, u\}, att' \cup \{(u, u)\} \cup \\ \{(i, \alpha) \mid \alpha \in \mathcal{A}' \setminus \mathcal{A} \wedge Lab(\alpha) = \mathtt{out}\} \cup \\ \{(u, \alpha) \mid \alpha \in \mathcal{A}' \setminus \mathcal{A} \wedge Lab(\alpha) = \mathtt{undec}\})$$

In words, AF^* *is obtained by extending* AF' *with two additional arguments* i *and* u*, where* i *is unattacked and* u *is self-attacking and not attacked by other arguments, and by including in the attack relation for any* α *labelled* `out` *by* Lab *the tuple* (i, α)*, and for any* α *labelled* `undec` *by* Lab *the tuple* (u, α) *(see Example 6 for an instance of this construction). It turns out that* $AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$*, as required by Definition 30. In particular, as required by Definition 28,* $AF' = inp\mathcal{LI}_{AF^*}(AF)$ *and* $inp\mathcal{LI}_{AF^*}(AF^* \setminus AF) = AF^* \setminus AF$*. Moreover, the labelling* $Lab'_1 = \{(i, \mathtt{in}), (u, \mathtt{undec})\} \cup Lab$ *of* $AF^* \setminus AF$ *is complete, since it satisfies the constraints specified in Definition 9 (note in particular that, according to the definition of* $inp\mathcal{LI}$*, for any pair* $(AF^i, Lab) \in P_{inp\mathcal{LI}}$ *the attack relation of* AF^i *is empty, i.e. all the arguments in* AF^i *do not receive attacks in* AF^i*). According to Definition 9,* Lab'_1 *is also the unique complete labelling, and thus it is grounded and preferred. In particular, according to Definition 9 (first condition)* i *must be labelled* `in` *by* Lab'_1 *because it is unattacked, and* u *must be labelled* `undec` *since both* `in` *and* `out` *would* *violate the conditions in Definition 9. Since any argument labelled* `out` *by Lab is attacked by i, according to Definition 9 (second condition) it is labelled* `out` *by* Lab'_1*. Since any argument labelled* `in` *by Lab is unattacked in* $AF^* \setminus AF$*, according to Definition 9 (first condition) it is labelled* `in` *by* Lab'_1*. Since any argument labelled* `undec` *by Lab is attacked by u only, according to Definition 9 it is labelled* `undec`*, which is the only label satisfying the conditions. Summing up,* $Lab'_1 \in \mathbf{L_S}(AF^* \setminus AF)$ *and* $Lab'_1\downarrow_{AF' \setminus AF} = Lab$*, as required by Definition 28.*

For the stable semantics we distinguish two cases for the pair $(AF^i, Lab) \in P_{inp\mathcal{LI}}$*. If* $\exists \alpha : Lab(\alpha) =$ `undec` *then* (AF^i, Lab) *is not realized under* $\mathbf{ST}$*, since by definition there is no labelling with* `undec`*-labelled arguments. In the other case, i.e. if Lab does not assign* `undec` *to any argument, then we consider the argumentation framework* AF^* *as above but without the argument u. It can then be verified as above that* $AF^* \in REP_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab)$*. Thus, all realizable pairs are representable, i.e. the stable semantics is weakly representable w.r.t. inp*$\mathcal{LI}$*.*

We are now in a position to introduce the notion of reduced canonical local function. Basically, for any argumentation framework with input (AF, AF', Lab) with a corresponding pair $(AF' \setminus AF, Lab)$ which is realizable, an argumentation framework AF^* is selected that represents (AF, AF', Lab), and the output labellings are identified as in Proposition 17. If instead the pair is not realizable, the function returns an empty set of labellings.

Definition 33. *Given a local information function* $\mathcal{LI}$ *and a weakly representable semantics* $\mathbf{S}$ *w.r.t.* $\mathcal{LI}$*, a reduced canonical local function of* $\mathbf{S}$ *w.r.t.* $\mathcal{LI}$ *is a local function* $RF_{\mathbf{S}}^{\mathcal{LI}}$ *such that for any* $(AF, AF', Lab) \in AF_{\mathcal{LI}}^{inp}$

$$RF_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab) = \begin{cases} \{Lab^*\downarrow_{AF} \mid Lab^* \in \mathbf{L_S}(AF^*) \wedge Lab^*\downarrow_{AF' \setminus AF} = Lab\} \\ \qquad\qquad\qquad\qquad \textit{if } (AF' \setminus AF, Lab) \in P_{\mathbf{S}}^{real} \\ \emptyset \textit{ otherwise} \end{cases}$$

where AF^* *is an argumentation framework such that* $AF^* \in REP_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab)$ *selected to represent* $(AF, AF', Lab) \in AF_{\mathcal{LI}}^{inp}$*.*

Definition 33 is well defined, as shown in the following proposition.

Proposition 19. *Let* $\mathcal{LI}$ *be a local information function and* $\mathbf{S}$ *a weakly representable semantics w.r.t.* $\mathcal{LI}$*. For any* $(AF, AF', Lab) \in AF_{\mathcal{LI}}^{inp}$*, if* $(AF' \setminus AF, Lab) \in P_{\mathbf{S}}^{real}$ *then* $\exists AF^* \in REP_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab)$*, i.e. the selection of an argumentation framework* AF^* *is possible.*

Proof. Since $\mathbf{S}$ is weakly representable, if $(AF' \setminus AF, Lab) \in P_{\mathbf{S}}^{real}$ then $(AF' \setminus AF, Lab) \in P_{\mathbf{S},\mathcal{LI}}^{rep}$. According to Definition 30 it then holds the conclusion, i.e. $\exists AF^* \in REP_{\mathbf{S}}^{\mathcal{LI}}(AF, AF', Lab)$. □

If in particular the semantics is representable, then $(AF' \setminus AF, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$, which by Proposition 18 entails that $(AF' \setminus AF, Lab) \in P^{real}_{\mathbf{S}}$, i.e. the local information function is always defined by the first item in Definition 33.

The suitability of a reduced canonical local function is confirmed by the following propositions.

First, if a semantics is weakly representable w.r.t. a local information function $\mathcal{LI}$, then any reduced canonical local function enforces its decomposability under $\mathcal{LI}$ if this is possible, i.e. if the semantics is decomposable under $\mathcal{LI}$.

Proposition 20. *Let $\mathcal{LI}$ be a local information function and* **S** *a weakly representable semantics w.r.t. $\mathcal{LI}$. If* **S** *is fully decomposable under $\mathcal{LI}$, a reduced canonical local function $RF^{\mathcal{LI}}_{\mathbf{S}}$ of* **S** *w.r.t. $\mathcal{LI}$ enforces decomposability of* **S** *under $\mathcal{LI}$.*

Proof. Since **S** is fully decomposable under $\mathcal{LI}$, there is a local function F for $\mathcal{LI}$ such that for every argumentation framework $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$

$$\mathbf{L_S}(AF) = \{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), \\ (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})\} \quad (11)$$

and we have to prove that for every $AF = (\mathcal{A}, att)$ and for every partition $\mathcal{P} = \{P_1, \ldots, P_n\}$ of $\mathcal{A}$

$$\mathbf{L_S}(AF) = \{L_{P1} \cup \ldots \cup L_{Pn} \mid L_{Pi} \in RF^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), \\ (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})\}$$

Let us first consider $Lab \in \mathbf{L_S}(AF)$. By condition (11), we have that $Lab = L_{P1} \cup \ldots \cup L_{Pn}$ with

$$L_{Pi} \in F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}}).$$

Taking into account that $Lab \in \mathbf{L_S}(AF)$, obviously for any i the pair $(\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}, (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})$ is realized under **S**, thus by Proposition 19 there is an argumentation framework AF^* selected for $RF^{\mathcal{LI}}_{\mathbf{S}}$ to represent $(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\ldots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF{\downarrow_{P_i}})\setminus AF{\downarrow_{P_i}}}})$. We can then apply

Proposition 17, obtaining

$$F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}) =$$
$$\{Lab^*{\downarrow_{AF\downarrow_{P_i}}} \mid Lab^* \in \mathbf{L_S}(AF^*) \wedge Lab^*{\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}} =$$
$$(\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}\}.$$

According to Definition 33, this is equal to

$$RF^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}).$$

Summing up, $Lab = L_{P1} \cup \ldots \cup L_{Pn}$ where

$$L_{Pi} \in RF^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})$$

for every i.

Turning to the reverse direction of the proof, consider a labellings $L_{P1} \cup \ldots \cup L_{Pn}$ such that, for any i,

$$L_{Pi} \in RF^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}) \tag{12}$$

According to Definition 33, AF^* is selected such that

$$AF^* \in REP^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}).$$

Taking into account that F enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$, by Proposition 17 we have that

$$F(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}) =$$
$$\{Lab^*{\downarrow_{AF\downarrow_{P_i}}} \mid Lab^* \in \mathbf{L_S}(AF^*) \wedge Lab^*{\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}} =$$
$$(\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}}\}$$

which by Definition 33 is equal to

$$RF^{\mathcal{LI}}_{\mathbf{S}}(AF{\downarrow_{P_i}}, \mathcal{LI}_{AF}(AF{\downarrow_{P_i}}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow_{\mathcal{LI}_{AF}(AF\downarrow_{P_i})\setminus AF\downarrow_{P_i}}})$$

Thus, taking into account (12), for every i it holds that

$$L_{Pi} \in F(AF{\downarrow}_{P_i}, \mathcal{LI}_{AF}(AF{\downarrow}_{P_i}), (\bigcup_{j=1\dots n, j\neq i} L_{Pj}){\downarrow}_{\mathcal{LI}_{AF}(AF{\downarrow}_{P_i})\setminus AF{\downarrow}_{P_i}})$$

which by (11) entails that $L_{P1} \cup \ldots \cup L_{Pn} \in \mathbf{L_S}(AF)$. □

In case a decomposable semantics is representable (besides weakly representable), any reduced canonical local function is the unique local function enforcing its decomposability (entailing that all reduced canonical local functions coincide).

Proposition 21. *Let $\mathcal{LI}$ be a local information function and* $\mathbf{S}$ *a representable semantics w.r.t. $\mathcal{LI}$. If* $\mathbf{S}$ *is fully decomposable under $\mathcal{LI}$, there is only one local function which enforces decomposability of* $\mathbf{S}$ *under $\mathcal{LI}$, coinciding with any reduced canonical local function $RF^{\mathcal{LI}}_{\mathbf{S}}$ of* $\mathbf{S}$ *w.r.t. $\mathcal{LI}$.*

Proof. Consider a reduced canonical local function $RF^{\mathcal{LI}}_{\mathbf{S}}$ of $\mathbf{S}$ w.r.t. $\mathcal{LI}$. If $\mathbf{S}$ is representable w.r.t. $\mathcal{LI}$, for any $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ it holds that $(AF' \setminus AF, Lab) \in P^{rep}_{\mathbf{S},\mathcal{LI}}$, thus, by Proposition 18, $(AF' \setminus AF, Lab) \in P^{real}_{\mathbf{S}}$. As a consequence, $RF^{\mathcal{LI}}_{\mathbf{S}}(AF, AF', Lab)$ is defined by the first item in Definition 33, and according to Proposition 17 its output is the same as that returned by any local function F which enforces decomposability of $\mathbf{S}$ under $\mathcal{LI}$. □

Example 9. *According to the considerations in Example 8, stable, grounded and preferred semantics are weakly representable w.r.t. inp$\mathcal{LI}$. We can then apply Proposition 20 to check whether the semantics is decomposable under inp$\mathcal{LI}$, focusing on a reduced canonical local function as per Definition 33. Since grounded and preferred semantics are also representable, Proposition 21 ensures that if the semantics turns out to be decomposable then such local function is unique.*

8 Analyzing decomposability of stable, grounded and preferred semantics with close neighboring information

For the sake of example, in this section we apply the model and the results of this paper in order to analyze the decomposability properties of stable, grounded and preferred semantics under local information concerning close neighbors only, i.e. the direct attackers and attacked arguments.

If the semantics is weakly representable w.r.t. $\mathcal{LI}$, by Proposition 20 the analysis can be based without loss of generality on a reduced canonical local function as per Definition 33, where the argumentation framework AF^* selected to represent an $(AF, AF', Lab) \in AF^{inp}_{\mathcal{LI}}$ can be constructed as explained in Example 8. If the semantics is not weakly representable, by Proposition 16 the canonical local function as per Definition 27 can be considered, which however is not guaranteed to be unitary. Whatever the case, to prove that a semantics is decomposable an explicit form of the local function is necessary to show that it enforces decomposability, i.e. that the condition in Definition 22 is satisfied for any AF and any partition of its arguments. On the other hand, to prove that the semantics is not decomposable only a counterexample AF is needed, thus only the output of the local function for the involved argumentation frameworks with input is needed.

We first consider the local information function *inp$\mathcal{LI}$*, which as mentioned in Section 1 has been the subject of a previous investigation in [5]. In this respect, the paper slightly generalizes the results of [5] for stable, grounded and preferred semantics, and also enlarges the set of semantics to which the main results can be applied[7].

Starting from stable semantics, we have shown in Example 8 that it is weakly representable under *inp$\mathcal{LI}$*. Following the construction described in the same example, it is not difficult to see that the following function satisfies the conditions in Definition 33, i.e. it is a reduced canonical local function:

$$RF^{inp\mathcal{LI}}_{\mathbf{ST}}(AF, AF', Lab) = \begin{cases} \{ \quad Lab' \in \mathfrak{L}(AF) \mid \forall \alpha \in \mathcal{A}, Lab'(\alpha) \neq \texttt{undec} \wedge \\ \qquad Lab'(\alpha) = \texttt{in} \;\; \mathit{iff} \\ \qquad\quad \forall \beta : (\beta, \alpha) \in att',\; (Lab \cup Lab')(\beta) = \texttt{out} \wedge \\ \qquad Lab(\alpha) = \texttt{out} \;\; \mathit{iff} \\ \qquad\quad \exists \beta : (\beta, \alpha) \in att' \wedge (Lab \cup Lab')(\beta) = \texttt{in}\} \\ \text{if } Lab \text{ does not include the label } \texttt{undec} \\ \\ \emptyset \text{ otherwise} \end{cases}$$

where $AF = (\mathcal{A}, att)$ and $AF' = (\mathcal{A}', att')$.

Adapting the proofs in [5], it can be shown that the above local function enforces decomposability of $\mathbf{ST}$ under *inp$\mathcal{LI}$*. In particular, referring to Definition 22, any $Lab \in \mathbf{L}_{AF}(\mathbf{ST})$ satisfies the conditions of complete labellings and does not assign

[7] In [5] the semantics have to be *complete-compatible*, i.e. satisfy a relatively articulated set of constraints. This excludes for instance admissible semantics, i.e. identified by admissible labellings, that instead can be analyzed with the results of the present paper. This specific issue is however outside the scope of this discussion.

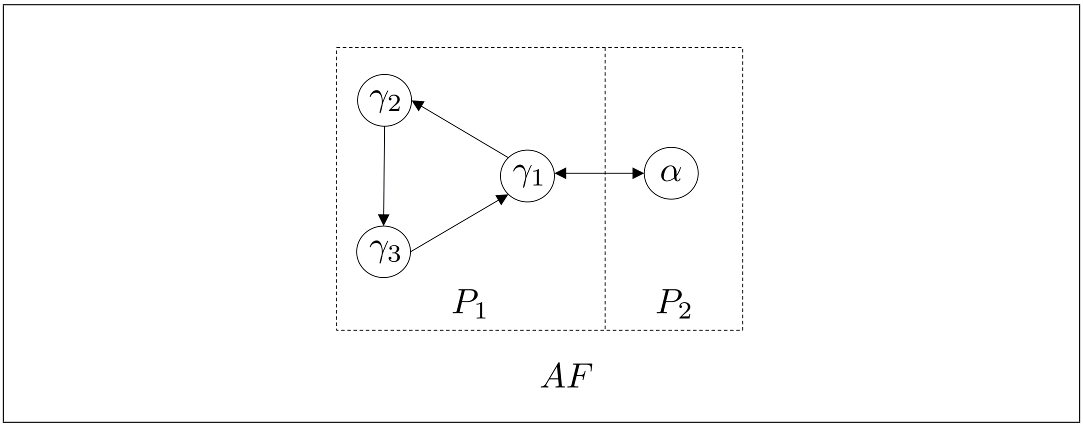


Figure 4: A counterexample for decomposability of **GR** and **PR** under $inp\mathcal{LI}$.

the label `undec`, and it can be proved that this holds iff the conditions of $RF_{\mathbf{ST}}^{inp\mathcal{LI}}$ are locally satisfied in any subframework $AF{\downarrow_{P_i}}$.

Summing up, **ST** is fully decomposable under $inp\mathcal{LI}$. Moreover, by Proposition 8 it is decomposable under any $\mathcal{LI}$ such that $inp\mathcal{LI} \preceq \mathcal{LI}$, e.g. $Binp\mathcal{LI}$, $inpout\mathcal{LI}$, $inp - k - \mathcal{LI}$ for all $k \geq 1$.

Let us now turn to grounded and preferred semantics, that have been shown to be representable in Example 8. It has also been shown in [5] that they are not decomposable under $inp\mathcal{LI}$, by means of a counterexample depicted in Figure 4. Here the arguments of AF are partitioned according to $\mathcal{P} = \{P_1, P_2\}$, where $P_1 = \{\gamma_1, \gamma_2, \gamma_3\}$ and $P_2 = \{\alpha\}$. In our setting, we determine the output of a reduced canonical local function $RF_{\mathbf{S}}^{inp\mathcal{LI}}$, with $\mathbf{S} \in \{\mathbf{GR}, \mathbf{PR}\}$ for the argumentation frameworks with input potentially involved in AF (as above, the argumentation frameworks selected to represent them are selected as in Example 8). Let $AF_1 \equiv AF{\downarrow_{P_1}}$, i.e. $AF_1 = (\{\gamma_1, \gamma_2, \gamma_3\}, \{(\gamma_1, \gamma_2), (\gamma_2, \gamma_3), (\gamma_3, \gamma_1)\})$, $AF_1' \equiv inp\mathcal{LI}_{AF}(AF_1)$, i.e. $AF_1' = (\{\gamma_1, \gamma_2, \gamma_3, \alpha\}, \{(\gamma_1, \gamma_2), (\gamma_2, \gamma_3), (\gamma_3, \gamma_1), (\alpha, \gamma_1)\})$, $AF_2 = AF{\downarrow_{P_2}}$, i.e. $AF_2 = (\{\alpha\}, \emptyset)$, and $AF_2' \equiv inp\mathcal{LI}_{AF}(AF_2)$, i.e. $AF_2' = (\{\alpha, \gamma_1\}, \{(\gamma_1, \alpha)\})$. We obtain both for the grounded and preferred semantics:

- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{in})\}) = \{\{(\gamma_1, \texttt{out}), (\gamma_2, \texttt{in}), (\gamma_3, \texttt{out})\}\}$
- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{out})\}) = \{\{(\gamma_1, \texttt{undec}), (\gamma_2, \texttt{undec}), (\gamma_3, \texttt{undec})\}\}$
- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{undec})\}) = \{\{(\gamma_1, \texttt{undec}), (\gamma_2, \texttt{undec}), (\gamma_3, \texttt{undec})\}\}$
- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \texttt{in})\}) = \{\{(\alpha, \texttt{out})\}\}$

- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \texttt{out})\}) = \{\{(\alpha, \texttt{in})\}\}$
- $RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \texttt{undec})\}) = \{\{(\alpha, \texttt{undec})\}\}$

We now consider the combinations of compatible local labellings obtained by applying $RF_{\mathbf{S}}^{inp\mathcal{LI}}$ to AF_1 and AF_2, i.e. the global labellings $L_{P1} \cup L_{P2}$ such that

$$L_{P1} \in RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_1, AF_1', L_{P2}{\downarrow_{\{\alpha\}}}) \wedge L_{P2} \in RF_{\mathbf{S}}^{inp\mathcal{LI}}(AF_2, AF_2', L_{P1}{\downarrow_{\{\gamma_1\}}})$$

It can be checked that two global labellings are obtained, i.e.

$$Lab_1 = \{(\alpha, \texttt{in}), (\gamma_1, \texttt{out}), (\gamma_2, \texttt{in}), (\gamma_3, \texttt{out})\}$$

and

$$Lab_2 = \{(\alpha, \texttt{undec}), (\gamma_1, \texttt{undec}), (\gamma_2, \texttt{undec}), (\gamma_3, \texttt{undec})\}$$

The labelling Lab_1 is the (unique) preferred labelling of AF, while Lab_2 is the grounded labelling. As a consequence, neither with $\mathbf{S} = \mathbf{GR}$ or with $\mathbf{S} = \mathbf{PR}$ the condition of Definition 22 is satisfied by $RF_{\mathbf{S}}^{inp\mathcal{LI}}$, i.e. $RF_{\mathbf{S}}^{inp\mathcal{LI}}$ does not enforce decomposability of $\mathbf{S}$ under *inp*$\mathcal{LI}$ (entailing that neither $\mathbf{GR}$ or $\mathbf{PR}$ is decomposable).

Interestingly enough, for the stable semantics $RF_{\mathbf{ST}}^{inp\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{out})\}) = RF_{\mathbf{ST}}^{inp\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{undec})\}) = RF_{\mathbf{ST}}^{inp\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \texttt{undec})\}) = \emptyset$, thus the only possible combination of compatible local labellings correctly correspond to the unique stable labelling Lab_1.

Given the counterexample against decomposability of $\mathbf{GR}$ or $\mathbf{PR}$ under *inp*$\mathcal{LI}$, it is interesting to verify whether the same example is handled correctly by increasing the local information exploited. Let us then consider the local information function *inpout*$\mathcal{LI}$, which also involves attacked arguments besides attackers, as well as the relevant attacks with both directions.

If *inpout*$\mathcal{LI}$ is adopted, it is possible that a subframework is affected from the outside and also affects external arguments. This prevents a semantics to be representable w.r.t. *inpout*$\mathcal{LI}$. For instance, referring again to Figure 4 and the above notations, it turns out that, letting $AF_2'' \equiv inpout\mathcal{LI}_{AF}(AF_2)$, $AF_2'' = (\{\alpha, \gamma_1\}, \{(\gamma_1, \alpha), (\alpha, \gamma_1)\})$. Considering then an argumentation framework with input $(AF_2, AF_2'', Lab) \in AF_{\mathcal{LI}}^{inp}$, there is no AF^* that represents it under *inpout*$\mathcal{LI}$. In fact, for any AF^* such that $AF_2'' \subseteq AF^*$, $inpout\mathcal{LI}_{AF^*}(AF^* \setminus AF_2)$ includes the argument α, thus it cannot be the case that $inpout\mathcal{LI}_{AF^*}(AF^* \setminus AF_2) = AF^* \setminus AF_2$.

Since no semantics is representable w.r.t. *inpout*$\mathcal{LI}$, we can then refer to the canonical local function as per Definition 27, as it has been done in Example 5. In particular, for any relevant (AF, AF', Lab), $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF, AF', Lab)$ returns all the possible labellings $Lab^{*}\downarrow_{AF}$, where Lab^{*} is prescribed by the semantics in an argumentation framework AF^{*} such that $(AF, AF', Lab) \in RAF^{inp}_{inpout\mathcal{LI},AF^{*},\mathbf{S}}$, and $Lab^{*}\downarrow_{AF'\setminus AF} = Lab$.

Let us start from the grounded semantics. The following result known from the literature will be useful.

Proposition 22. *Let Lab be the grounded labelling of an argumentation framework $AF = (\mathcal{A}, att)$, and let $\alpha \in \mathcal{A}$. If $Lab(\alpha) = \texttt{out}$, then there is an argument $\beta \in \mathcal{A}$ such that $Lab(\beta) = \texttt{in}$, $(\beta, \alpha) \in att$ and it is not the case that $(\alpha, \beta) \in att$.*

Proof. The reader can refer e.g. to Lemma 3, page 800 of [3], where a corresponding result for the *grounded extension* is proved. The labelling-version follows from the correspondence between the grounded labelling and the grounded extension (see e.g. [2]). □

Considering AF_1 and AF_2 as defined above, let $AF''_1 \equiv inpout\mathcal{LI}_{AF}(AF_1)$, i.e. $AF''_1 = (\{\gamma_1, \gamma_2, \gamma_3, \alpha\}, \{(\gamma_1, \gamma_2), (\gamma_2, \gamma_3), (\gamma_3, \gamma_1), (\alpha, \gamma_1), (\gamma_1, \gamma)\})$, and $AF''_2 \equiv inpout\mathcal{LI}_{AF}(AF_2)$, i.e. $AF''_2 = (\{\alpha, \gamma_1\}, \{(\gamma_1, \alpha), (\alpha, \gamma_1)\})$.

We obtain for the grounded semantics:

- $F^{inpout\mathcal{LI}}_{\mathbf{GR}}(AF_1, AF''_1, \{(\alpha, \texttt{in})\}) = \emptyset$. In fact, for any AF^{*} such that $AF''_1 = inpout\mathcal{LI}_{AF^{*}}(AF_1)$, the conditions in the definition of complete labelling require the grounded labelling to include $\{(\gamma_1, \texttt{out}), (\gamma_2, \texttt{in}), (\gamma_3, \texttt{out})\}$, but this contradicts Proposition 22 since γ_1 counterattacks the only in-labelled attacker.

- $F^{inpout\mathcal{LI}}_{\mathbf{GR}}(AF_1, AF''_1, \{(\alpha, \texttt{out})\}) = \{\{(\gamma_1, \texttt{undec}), (\gamma_2, \texttt{undec}), (\gamma_3, \texttt{undec})\}\}$. The output labelling is achieved e.g. in a modified version of AF where an unattacked argument attacks α, and it is unique because of the conditions in the definition of complete labelling.

- $F^{inpout\mathcal{LI}}_{\mathbf{GR}}(AF_1, AF''_1, \{(\alpha, \texttt{undec})\}) = \{\{(\gamma_1, \texttt{undec}), (\gamma_2, \texttt{undec}), (\gamma_3, \texttt{undec})\}\}$ The output labelling is achieved e.g. in AF, and it is unique because of the conditions in the definition of complete labelling.

- $F^{inpout\mathcal{LI}}_{\mathbf{GR}}(AF_2, AF''_2, \{(\gamma_1, \texttt{in})\}) = \emptyset$. In fact, for any AF^{*} such that $AF''_2 = inpout\mathcal{LI}_{AF^{*}}(AF_2)$, the conditions in the definition of complete labelling require the grounded labelling to include $\{(\alpha, \texttt{out})\}$, but this contradicts Proposition 22 since α counterattacks the only in-labelled attacker.

- $F^{inpout\mathcal{LI}}_{\mathbf{GR}}(AF_2, AF'_2, \{(\gamma_1, \texttt{out})\}) = \{\{(\alpha, \texttt{in})\}\}$. In fact, the argumentation framework $(\{i, \gamma_1, \alpha\}, \{(i, \gamma_1), (\gamma_1, \alpha), (\alpha, \gamma_1)\})$ yields the output labelling, and the latter is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{GR}}^{\mathit{inpout}\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \mathtt{undec})\}) = \{\{(\alpha, \mathtt{undec})\}\}$. The output labelling is achieved e.g. in $(\{\gamma_1, \alpha\}, \{(\gamma_1, \alpha), (\alpha, \gamma_1)\})$, and it is unique because of the conditions in the definition of complete labelling.

We now consider the combinations of compatible local labellings obtained by applying $F_{\mathbf{GR}}^{\mathit{inpout}\mathcal{LI}}$ to AF_1 and AF_2. It can be checked that only one global labelling is obtained, i.e.

$$Lab = \{(\alpha, \mathtt{undec}), (\gamma_1, \mathtt{undec}), (\gamma_2, \mathtt{undec}), (\gamma_3, \mathtt{undec})\}$$

coinciding with the grounded labelling of AF.

Also with preferred semantics the example is correctly handled. We obtain for the canonical local function:

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1'', \{(\alpha, \mathtt{in})\}) = \{\{(\gamma_1, \mathtt{out}), (\gamma_2, \mathtt{in}), (\gamma_3, \mathtt{out})\}\}$. The output labelling is achieved e.g. in AF, and it is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1'', \{(\alpha, \mathtt{out})\}) = \{\{(\gamma_1, \mathtt{undec}), (\gamma_2, \mathtt{undec}), (\gamma_3, \mathtt{undec})\}\}$. The output labelling is achieved e.g. in a modified version of AF where an unattacked argument attacks α, and it is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1'', \{(\alpha, \mathtt{undec})\}) = \{\{(\gamma_1, \mathtt{undec}), (\gamma_2, \mathtt{undec}), (\gamma_3, \mathtt{undec})\}\}$ The output labelling is achieved e.g. in a modified version of AF where a self-attacking argument (without other attackers) attacks α, and it is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_2, AF_2'', \{(\gamma_1, \mathtt{in})\}) = \{\{(\alpha, \mathtt{out})\}\}$. In fact, the output labelling is achieved e.g. in $(\{\gamma_1, \alpha\}, \{(\gamma_1, \alpha), (\alpha, \gamma_1)\})$, and it is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \mathtt{out})\}) = \{\{(\alpha, \mathtt{in})\}\}$. In fact, the argumentation framework $(\{i, \gamma_1, \alpha\}, \{(i, \gamma_1), (\gamma_1, \alpha), (\alpha, \gamma_1)\})$ yields the output labelling, and the latter is unique because of the conditions in the definition of complete labelling.

- $F_{\mathbf{PR}}^{\mathit{inpout}\mathcal{LI}}(AF_2, AF_2', \{(\gamma_1, \mathtt{undec})\}) = \emptyset$. In fact, for any AF^* such that $AF_1'' = \mathit{inpout}\mathcal{LI}_{AF^*}(AF_1)$, the conditions in the definition of complete labelling require any preferred labelling $Lab_{\mathbf{PR}}$ to include $\{(\gamma_1, \mathtt{undec})\}$. However, consider the labelling Lab' such that $Lab'(\gamma_1) = \mathtt{out}$, $Lab'(\alpha) = \mathtt{in}$ and for all the

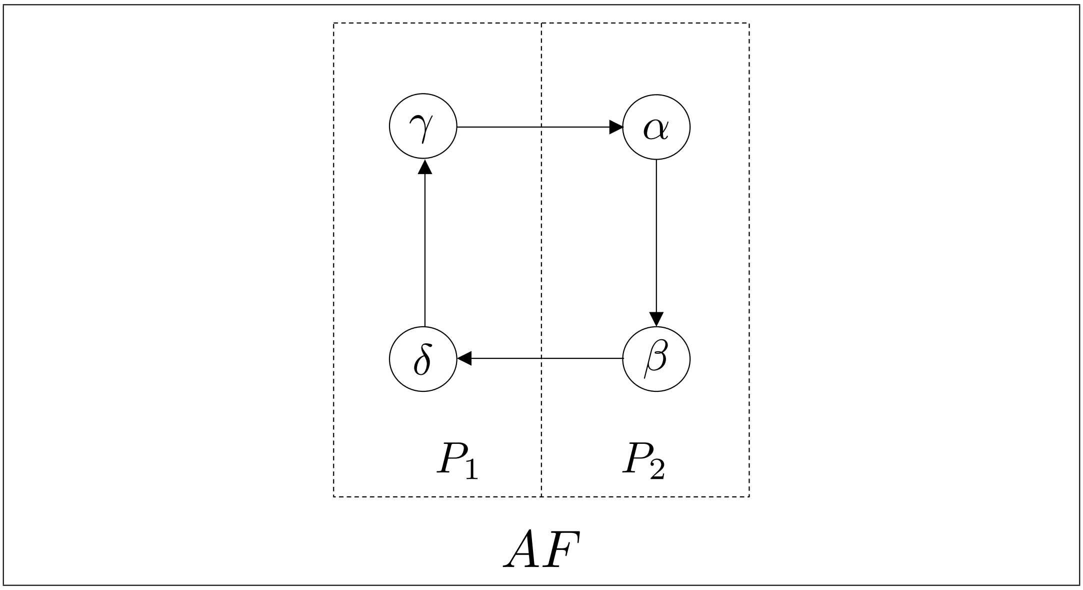


Figure 5: A counterexample for decomposability of **GR** and **PR** under *inpout$\mathcal{LI}$*.

other arguments Lab' coincides with $Lab_{\mathbf{PR}}$. Since $Lab_{\mathbf{PR}}$ is complete in AF^*, Lab' is admissible (see Definition 7): α is `in`-labelled and its unique attacker is γ_1 which is `out`-labelled, γ_1 is `out`-labelled and has the attacker α which is `in`-labelled, any other argument which is `out`-labelled has an attacker which is labelled `in` by $Lab_{\mathbf{PR}}$ and thus by Lab' too, and finally any other argument which is `in`-labelled has all of its attackers labelled `out` by $Lab_{\mathbf{PR}}$ and thus by Lab' too. But then, $Lab_{\mathbf{PR}} \sqsubseteq Lab'$ while the reverse does not hold, and Lab' is admissible. This contradicts the maximality of $Lab_{\mathbf{PR}}$ (see Proposition 2).

Combining the compatible local labellings returned by $F_{\mathbf{PR}}^{inpout\mathcal{LI}}$ for AF_1 and AF_2 yields the unique global labelling $\{(\gamma_1, \texttt{out}), (\gamma_2, \texttt{in}), (\gamma_3, \texttt{out}), (\alpha, \texttt{in})\}$ which corresponds to the unique preferred labelling of AF.

Although the above example is correctly handled, neither **GR** nor **PR** are decomposable under *inpout$\mathcal{LI}$*. To see this, consider the argumentation framework AF shown in Figure 5, and the partition $\mathcal{P} = \{P_1, P_2\}$, where $P_1 = \{\gamma, \delta\}$ and $P_2 = \{\alpha, \beta\}$. Let $AF_1 \equiv AF{\downarrow_{P_1}}$, i.e. $AF_1 = (\{\gamma, \delta\}, \{(\delta, \gamma)\})$, $AF'_1 \equiv inp\mathcal{LI}_{AF}(AF_1)$, i.e. $AF'_1 = (\{\gamma, \delta, \alpha, \beta\}, \{(\delta, \gamma), (\gamma, \alpha), (\beta, \delta\})$, $AF_2 \equiv AF{\downarrow_{P_2}} = (\{\alpha, \beta\}, \{(\alpha, \beta)\})$, and $AF'_2 \equiv inp\mathcal{LI}_{AF}(AF_2) = (\{\gamma, \delta, \alpha, \beta\}, \{(\gamma, \alpha), (\alpha, \beta), (\beta, \delta\})$.

We determine the output of the canonical local function for the argumentation frameworks with input involved in AF. First note that AF is symmetrically partitioned, thus we can focus on AF_1 and then directly derive the corresponding results for AF_2. Note that $AF_1' \setminus AF_1$ includes two arguments, namely α and β, giving rise to 9 possible labellings for them. For any AF^* such that $AF_1' = \mathit{inpout}\mathcal{LI}_{AF^*}(AF_1)$ and for any complete labelling of AF^*, the label assigned to β uniquely determines the assignment for δ and γ. In particular, according to the conditions of complete labelling, if β is `in`-labelled, then δ is `out`-labelled and γ is `in`-labelled, if β is `out`-labelled, then δ is `in`-labelled and γ is `out`-labelled, and if β is `undec`-labelled, then δ and γ are `undec`-labelled. As a consequence, for any argumentation framework with input the output of the canonical local functions of **GR** and **PR** includes one labelling at most. Moreover, also the label assigned to α is constrained by the conditions of complete labelling. In particular, taking into account that in AF^* α is attacked by γ but it can also be attacked by other arguments, if β is `in`-labelled then α can only be `out`, if β is `undec`-labelled then α cannot be `in`, while in the other case the label of α is not constrained. As a consequence, the output of the canonical local function is empty if the label appearing in the argumentation with input is $\{(\alpha, \texttt{in}), (\beta, \texttt{in})\}$, $\{(\alpha, \texttt{undec}), (\beta, \texttt{in})\}$, or $\{(\alpha, \texttt{in}), (\beta, \texttt{undec})\}$. In all of the other cases, it is easy to identify an argumentation framework AF^* with a unique complete labelling, which is thus grounded and preferred, yielding for δ and γ the induced assignment. In particular:

- for $\{(\alpha, \texttt{out}), (\beta, \texttt{in})\}$, consider $AF^* = AF_1'$
- for $\{(\alpha, \texttt{undec}), (\beta, \texttt{undec})\}$, consider $AF^* = AF_1'$ with the addition of a self-attacking (otherwise unattacked) argument which attacks β
- for $\{(\alpha, \texttt{out}), (\beta, \texttt{undec})\}$, consider $AF^* = AF_1'$ with the addition of a self-attacking (otherwise unattacked) argument attacking β and of an unattacked argument attacking α
- for $\{(\alpha, \texttt{out}), (\beta, \texttt{out})\}$, consider $AF^* = AF_1'$ where an unattacked argument attacking both α and β is added
- for $\{(\alpha, \texttt{in}), (\beta, \texttt{out})\}$, consider $AF^* = AF_1'$ with the addition of an unattacked argument attacking β
- for $\{(\alpha, \texttt{undec}), (\beta, \texttt{out})\}$, consider $AF^* = AF_1'$ with the addition of both an unattacked argument attacking β and a self-attacking (otherwise unattacked) argument attacking α.

According to the above considerations, we obtain for $\mathbf{S} \in \{\mathbf{GR}, \mathbf{PR}\}$:

- $F_{\mathbf{S}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{in}), (\beta, \texttt{in})\}) =$
  $F_{\mathbf{S}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{undec}), (\beta, \texttt{in})\}) =$
  $F_{\mathbf{S}}^{\mathit{inpout}\mathcal{LI}}(AF_1, AF_1', \{(\alpha, \texttt{in}), (\beta, \texttt{undec})\}) = \emptyset$

- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}) = \{\{(\delta, \mathtt{out}), (\gamma, \mathtt{in})\}\}$
- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{undec}), (\beta, \mathtt{undec})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{out}), (\beta, \mathtt{undec})\}) = \{\{(\delta, \mathtt{undec}), (\gamma, \mathtt{undec})\}\}$
- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{out}), (\beta, \mathtt{out})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_1, AF'_1, \{(\alpha, \mathtt{undec}), (\beta, \mathtt{out})\}) = \{\{(\delta, \mathtt{in}), (\gamma, \mathtt{out})\}\}$

And, by symmetry, it also holds:

- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{in}), (\gamma, \mathtt{in})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{undec}), (\gamma, \mathtt{in})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{in}), (\gamma, \mathtt{undec})\}) = \emptyset$
- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{out}), (\gamma, \mathtt{in})\}) = \{\{(\alpha, \mathtt{out}), (\beta, \mathtt{in})\}\}$
- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{undec}), (\gamma, \mathtt{undec})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{out}), (\gamma, \mathtt{undec})\}) = \{\{(\alpha, \mathtt{undec}), (\beta, \mathtt{undec})\}\}$
- $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{out}), (\gamma, \mathtt{out})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{in}), (\gamma, \mathtt{out})\}) =$
  $F_{\mathbf{S}}^{inpout\mathcal{LI}}(AF_2, AF'_2, \{(\delta, \mathtt{undec}), (\gamma, \mathtt{out})\}) = \{\{(\alpha, \mathtt{in}), (\beta, \mathtt{out})\}\}$

Combining the compatible local labellings returned by $F_{\mathbf{S}}^{inpout\mathcal{LI}}$ for AF_1 and AF_2 yields three global labellings for AF, namely $\{(\alpha, \mathtt{out}), (\beta, \mathtt{in}), (\delta, \mathtt{out}), (\gamma, \mathtt{in})\}$, $\{(\alpha, \mathtt{in}), (\beta, \mathtt{out}), (\delta, \mathtt{in}), (\gamma, \mathtt{out})\}$, $\{(\alpha, \mathtt{undec}), (\beta, \mathtt{undec}), (\delta, \mathtt{undec}), (\gamma, \mathtt{undec})\}$.

The first and the second are the preferred labellings of AF, the third is the grounded labelling. As a consequence, neither with $\mathbf{S} = \mathbf{GR}$ nor with $\mathbf{S} = \mathbf{PR}$ the condition of Definition 22 is satisfied by $F_{\mathbf{S}}^{inpout\mathcal{LI}}$, entailing that neither **GR** nor **PR** is decomposable.

9 Discussion and conclusion

In this paper, we have devised a model for studying the decomposability of argumentation semantics in Dung's abstract argumentation setting. The model corresponds to a generalization of the definitions introduced in a previous paper [5]: it encompasses all possible kinds of local information available for the local computations, under some mild constraints. In this general model, we have proved a monotone relationship between the degree of information available locally and the set of decomposable semantics, and we have investigated the range of capabilities of local information in allowing decomposability of semantics, by determining the sets of decomposable semantics in the two extreme situations concerning the availability of local information. Furthermore, we have investigated the construction of local functions for the computation of local labellings, by introducing a general constructive procedure independent of the specific semantics definitions. We have also applied the procedure to identify two kinds of local functions, both of them enforcing decomposability if the semantics and the local information exploited make it possible. These functions represent a reference point to prove or disprove the decomposability of a specific semantics. Finally, as an example of application of the concepts and results of the paper, we have studied the decomposability properties of stable, grounded and preferred semantics under local information concerning close neighbors.

Many future directions of this work can be envisaged, both at the level of the general model and its instantiation with specific semantics.

At the abstract level, an interesting issue concerns the possible relationship between decomposability under restricting assumptions on the possible partitions of arguments and decomposability under a local information function. For instance, the fact that a semantics is decomposable when the partition elements coincide with the strongly connected components of the argumentation framework may imply that, if the available local information includes such components (and possibly some neighboring part), the semantics turns out to be decomposable. It would be interesting to investigate this relationship in general terms.

As to the level of specific semantics, a first issue is to identify for the semantics available in the literature the canonical local function, or a reduced canonical local function, in an explicit form. This will be useful for studying decomposability under different local information functions and, possibly, determining the minimal local information sufficient to guarantee decomposability.This may in turn provide a solid basis for mixing different argumentation semantics adopted in different subframeworks. More specifically, decomposability may be a necessary requirement in the specific case where all semantics coincide. In this respect, using less information relaxes the ties among local computations and gives more flexibility in the mixing strategy.

References

[1] Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., Villata, S.: Towards artificial argumentation. AI Magazine **38**(3), 25–36 (2017)

[2] Baroni, P., Caminada, M., Giacomin, M.: An introduction to argumentation semantics. The Knowledge Engineering Review **26:4**, 365–410 (2011)

[3] Baroni, P., Dunne, P.E., Giacomin, M.: On the resolution-based family of abstract argumentation semantics and its grounded instance. Artificial Intelligence **175**(3-4), 791–813 (2011)

[4] Baroni, P., Giacomin, M.: On principle-based evaluation of extension-based argumentation semantics. Artificial Intelligence (Special issue on Argumentation in A.I.) **171**(10/15), 675–700 (2007)

[5] Baroni, P., Boella, G., Cerutti, F., Giacomin, M., van der Torre, L.W.N., Villata, S.: On the input/output behavior of argumentation frameworks. Artificial Intelligence **217**, 144–197 (2014)

[6] Baroni, P., Cerutti, F., Giacomin, M.: Constructing local functions to decompose argumentation semantics: Preliminary results. In: Proc. of the 5th Workshop on Advances In Argumentation In Artificial Intelligence (2021)

[7] Baroni, P., Giacomin, M.: Some considerations on epistemic and practical reasoning in abstract argumentation. In: Proc. of the 2nd Workshop on Advances In Argumentation In Artificial Intelligence. pp. 1–5 (2018)

[8] Baumann, R., Brewka, G.: Analyzing the equivalence zoo in abstract argumentation. In: Proc. of the 14th Int. Workshop on Computational Logic in Multi-Agent Systems (CLIMA XIV). pp. 18–33 (2013)

[9] Baumann, R., Brewka, G., Wong, R.: Splitting argumentation frameworks: An empirical evaluation. In: Theory and Applications of Formal Argumentation - First Int. Workshop (TAFA 2011). Revised Selected Papers. Lecture Notes in Computer Science, vol. 7132, pp. 17–31. Springer (2011)

[10] Baumann, R.: Splitting an argumentation framework. In: Proc. of LPNMR 2011 11th Int. Conf. on Logic Programming and Nonmonotonic Reasoning. pp. 40–53 (2011)

[11] Caminada, M.W.A.: On the issue of reinstatement in argumentation. In: Proc. of the 10th European Conference on Logics in Artificial Intelligence (JELIA 2006). pp. 111–123 (2006)

[12] Caminada, M.W.A.: A labelling approach for ideal and stage semantics. Argument & Computation **2**(1), 1–21 (2011)

[13] Caminada, M.W.A., Gabbay, D.M.: A logical account of formal argumentation. Studia Logica **93**(109) (2009)

[14] Caminada, M.W.A., Pigozzi, G.: On judgment aggregation in abstract argumentation. Journal of Autonomous Agents and Multi-Agent Systems **22**(1), 64–102 (2011)

[15] Cerutti, F., Giacomin, M., Vallati, M., Zanella, M.: A SCC recursive meta-algorithm for computing preferred labellings in abstract argumentation. In: Proc. of the 14th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR 2014). p. to appear (2014)

[16] Cerutti, F., Tachmazidis, I., Vallati, M., Batsakis, S., Giacomin, M., Antoniou, G.: Exploiting parallelism for hard problems in abstract argumentation. In: Proceedings of

the 29th AAAI Conference on Artificial Intelligence. vol. 29 (2015)

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

[18] Gabbay, D.M.: Fibring argumentation frames. Studia Logica **93**(2-3), 231–295 (2009)

[19] Giacomin, M.: Handling heterogeneous disagreements through abstract argumentation (extended abstract). In: Proc. of PRIMA 2017. pp. 3–11 (2017)

[20] Giacomin, M., Baroni, P., Cerutti, F.: Towards a general theory of decomposability in abstract argumentation. In: Proc. of 4th International Conference on Logic and Argumentation (CLAR 2021). Lecture Notes in Computer Science, vol. 13040, pp. 169–189. Springer (2021)

[21] Liao, B., Huang, H.: Partial semantics of argumentation: basic properties and empirical results. J. of Logic and Computation **23**(3), 541–562 (2013)

[22] Liao, B., Jin, L., Koons, R.C.: Dynamics of argumentation systems: A division-based method. Artificial Intelligence **175**, 1790–1814 (2011)

[23] Rienstra, T., Perotti, A., Villata, S., Gabbay, D., van Der Torre, L., Boella, G.: Multisorted argumentation frameworks. In: Theory and Applications of Formal Argumentation - First Int. Workshop (TAFA 2011). Revised Selected Papers. Lecture Notes in Computer Science, vol. 7132, pp. 231–245. Springer (2011)

[24] Villata, S., Boella, G., van Der Torre, L.: Argumentation patterns. In: Proc. of ARGMAS 2011 8th Int. Workshop on Argumentation in Multi-Agent Systems. pp. 133–150 (2011)

Received 15 June 2022

Burden of Persuasion: A Meta-argumentation Approach

Giuseppe Pisano
Alma AI – Alma Mater Research Institute for Human-Centered Artificial Intelligence, Alma Mater Studiorum—Università di Bologna, Italy
g.pisano@unibo.it

Roberta Calegari
Dipartimento di Informatica – Scienza e Ingegneria (DISI), Alma Mater Studiorum—Università di Bologna, Italy
roberta.calegari@unibo.it

Andrea Omicini
Dipartimento di Informatica – Scienza e Ingegneria (DISI), Alma Mater Studiorum—Università di Bologna, Italy
andrea.omicini@unibo.it

Giovanni Sartor
Alma AI – Alma Mater Research Institute for Human-Centered Artificial Intelligence, Alma Mater Studiorum—Università di Bologna, Italy
giovanni.sartor@unibo.it

Abstract

This work defines a burden of persuasion meta-argumentation model interpreting burden as a set of meta-arguments. Bimodal graphs are exploited to define a *meta level* (dealing with the burden) and an *object level* (dealing with standard arguments). A novel technological reification of the model supporting the burden inversion mechanism is presented and discussed.

Keywords: burdens of persuasion; argumentation; meta-argumentation

The work has been supported by the "CompuLaw" project, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 833647).

1 Introduction

In this work we discuss the model of the burden of persuasion in structured argumentation [5, 6] under a meta-argumentative approach, which leads to *(i)* a clear separation of concerns in the model, *(ii)* a simpler and more efficient implementation of the corresponding argumentation tool, *(iii)* a natural model extension for reasoning over the burden of persuasion concepts.

The work is grounded on the approaches to meta-argumentation that emphasise the inner nature of arguments and dialogues as inherently meta-logical [10, 11]. Our approach relies on those works [10, 11] that introduce only the required abstraction at the meta level. The proposed meta-argumentation framework for the burden of persuasion includes three ingredients: *(i) object-level argumentation* – to create arguments from defeasible and strict rules –, *(ii) meta-level argumentation* – to create arguments dealing with abstractions related to the burden concept using argument schemes (or meta-level rules) –, and *(iii) bimodal graphs* to define the interaction between the object level and the meta level—following the account in [10].

This work extends our previous work [13] in two main directions. First, it introduces and discusses a novel technological reification of the model supporting the burden inversion mechanism. Then, related work is discussed by positioning our contribution against the state of the art, and highlighting strengths and limitations w.r.t. other approaches—e.g., [16].

Accordingly, Section 2 introduces basic elements of the meta-argumentation framework. Section 3 formally defines the framework for the burden of persuasion introducing related argument schemes and discusses its equivalence with the model presented in [5]. Section 4 discusses a real case study in the law domain dealing with the problem of burden inversion. Finally, Section 5 presents the technological reification of the model. Related work is discussed in Section 6, whereas conclusions are drawn in Section 7.

2 Meta-argumentation framework

In this section, we introduce the meta-argumentation framework. For the sake of simplicity we choose to model our meta-argumentation framework by exploiting bimodal graphs, which are often exploited to both define meta-level concepts and understand the interactions of object-level and meta-level arguments [11, 10]. Accordingly, Subsection 2.1 presents the object-level argumentation language exploited by our model, leveraging on an ASPIC$^+$-like argumentation framework [15]. Then, Subsection 2.2 introduces bimodal argumentation graphs' main definitions. Finally, the meta-level argumentation language based on the use of argument schemes [18] is introduced in Subsection 2.3.

2.1 Structured argumentation for object-level argumentation

Let a literal be an atomic proposition or its negation.

Notation 1. *For any literal ϕ, its complement is denoted by $\bar{\phi}$. That is, if ϕ is a proposition p, then $\bar{\phi} = \neg p$, whereas if ϕ is $\neg p$, then $\bar{\phi}$ is p.*

Let us also identify burdens of persuasion, i.e., those literals whose proof requires a convincing argument. We assume that such literals are consistent (it cannot be the case that there is a burden of persuasion on both ϕ and $\bar{\phi}$).

Definition 2.1 (Burdens of persuasion)**.** *Burdens of persuasion are represented by predicates of the form bp(ϕ), stating the burden is allocated on the literal ϕ.*

Literals are put in relation with *bp* predicates through defeasible rules.

Definition 2.2 (Defeasible rule)**.** *A* ***defeasible rule*** *r has the form:*

$$\rho : \quad \phi_1, ..., \phi_n, \sim\phi'_1, ..., \sim\phi'_m \Rightarrow \psi$$

with $0 \leq n, m$, and where

- *ρ is the unique identifier for r, denoted by $\mathsf{N}(r)$;*
- *each $\phi_1, \ldots, \phi_n, \phi'_1, \ldots, \phi'_m, \psi$ is a literal or a* bp *predicate;*
- *$\phi_1, \ldots \phi_n, \sim\phi'_1, ..., \sim\phi'_m$ are denoted by Antecedent(r);*
- *ψ is denoted by Consequent(r);*
- *$\sim\phi$ denotes the weak negation (negation by failure) of ϕ—i.e., ϕ is an exception that would block the application of the rule whose antecedent includes $\sim\phi$.*

The unique identifier of a rule can be used as a literal to specify that the named rule is applicable, and its negation to specify that the rule is inapplicable, dually [9].

A superiority relation $\succ$ is defined over rules: $s \succ r$ states that rule s prevails over rule r.

Definition 2.3 (Superiority relation)**.** *A* ***superiority relation*** *$\succ$ over a set of rules Rules is a transitive, antireflexive, and antisymmetric binary relation over Rules.*

A defeasible theory consists of a set of rules and a superiority relation over the rules.

Definition 2.4 (Defeasible theory). *A* ***defeasible theory*** *is a tuple $\langle Rules, \succ\rangle$ where Rules is a set of rules, and $\succ$ is a superiority relation over Rules.*

Given a defeasible theory, we can construct arguments by chaining rules from the theory [9, 7, 17].

Definition 2.5 (Argument). *An* ***argument*** *A constructed from a defeasible theory $\langle Rules, \succ\rangle$ is a finite construct of the form: $A : A_1, \dots A_n \Rightarrow_r \phi$ with $0 \leq n$, where*

- *A is the argument's unique identifier;*
- *$A_1, \dots, A_n$ are arguments constructed from the defeasible theory $\langle Rules, \succ\rangle$;*
- *ϕ is the* conclusion *of the argument, denoted by $\mathsf{Conc}(A)$;*
- *$r : \mathsf{Conc}(A_1), \dots, \mathsf{Conc}(A_n) \Rightarrow \phi$ is the top rule of A, denoted by $\mathsf{TopRule}(A)$.*

Notation 2. *Given an argument $A : A_1, \dots A_n \Rightarrow_r \phi$ as in Definition 2.5, $\mathsf{Sub}(A)$ denotes the set of subarguments of A, i.e., $\mathsf{Sub}(A) = \mathsf{Sub}(A_1) \cup \dots \cup \mathsf{Sub}(A_n) \cup \{A\}$. $\mathsf{DirectSub}(A)$ denotes the direct subarguments of A, i.e., $\mathsf{DirectSub}(A) = \{A_1, \dots, A_n\}$.*

Preferences over arguments are defined via a last-link ordering: argument A is preferred over argument B if the top rule of A is stronger than the top rule of B.

Definition 2.6 (Preference relation). *A* ***preference relation*** *$\succ$ is a binary relation over a set of arguments $\mathcal{A}$: argument A is preferred to argument B – denoted by $A \succ B$ – iff $\mathsf{TopRule}(A) \succ \mathsf{TopRule}(B)$.*

Arguments are put in relation with each other according to the attack relation.

Definition 2.7 (Attack). *Argument A* ***attacks*** *argument B iff A undercuts or rebuts B, where*

- *A undercuts B (on B') iff $\mathsf{Conc}(A) = \neg \mathsf{N}(\rho)$ for some $B' \in \mathsf{Sub}(B)$, where ρ is $\mathsf{TopRule}(B')$*
- *A rebuts B (on B') iff either* (i) *$\mathsf{Conc}(A) = \bar{\phi}$ for some $B' \in \mathsf{Sub}(B)$ of the form $B''_1, ..., B''_M \Rightarrow \phi$ and $B' \not\succ A$, or* (ii) *$\mathsf{Conc}(A) = \phi$ for some $B' \in \mathsf{Sub}(B)$ such that $\sim\phi \in Antecedent(\mathsf{TopRule}(B'))$*

In short, arguments can be attacked either on a conclusion of a defeasible inference (*rebutting* attack) or on a defeasible inference step itself (*undercutting* attack).

Definition 2.8 (Argumentation graph)**.** *An **argumentation graph** is a tuple $\langle \mathcal{A}, \rightsquigarrow \rangle$, where $\mathcal{A}$ is the set of all arguments, and $\rightsquigarrow$ is* attack *relation over $\mathcal{A}$.*

Notation 3. *Given an argumentation graph $G = \langle \mathcal{A}, \rightsquigarrow \rangle$, we write $\mathcal{A}_G$ and $\rightsquigarrow_G$ to denote the graph's arguments and attacks, respectively.*

Now, let us introduce the notion of the $\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}$-labelling of an argumentation graph, where each argument in the graph is labelled IN, OUT, or UND, depending on whether it is accepted, rejected, or undecided, respectively.

Definition 2.9 (Labelling)**.** *Let G be an argumentation graph. An $\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}$-**labelling** L of G is a total function $\mathcal{A}_G \to \{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}$. $\mathcal{L}(\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}, G)$ denotes the set of all $\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}$-**labellings** of G.*

A labelling-based semantics prescribes a set of labellings for any argumentation graph according to some criterion embedded in its definition.

Definition 2.10 (Labelling-based semantic)**.** *Let G be an argumentation graph. A labelling-based semantics S associates with G a subset of $\mathcal{L}(\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}, G)$, denoted as $L_S(G)$.*

2.2 Object and meta level connection: bimodal graphs

In this section we recall the main definitions of bimodal graphs as the model of interaction between object and meta level. Bimodal graphs make it possible to capture scenarios where arguments are categorised in multiple levels—two in our case, the object and the meta level. Accordingly, a bimodal graph is composed of two components: an argumentation graph for the meta level and an argumentation graph for the object level, along with a relation of support that originates from the meta level and targets attacks and arguments on the object level. Every object-level argument and every object-level attack is supported by at least one meta-level argument. Meta-level arguments can only attack meta-level arguments, and object-level arguments can only attack object-level arguments.

Definition 2.11 (Bimodal argumentation graph)**.** *A **bimodal argumentation graph** is a tuple $\langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$ where*

1. *$\mathcal{A}_O$ is the set of object-level arguments*
2. *$\mathcal{A}_M$ is the set of meta-level arguments*
3. *$\mathcal{R}_O \subseteq \mathcal{A}_O \times \mathcal{A}_O$ represents the set of object-level attacks*

4. $\mathcal{R}_M \subseteq \mathcal{A}_M \times \mathcal{A}_M$ *represents the set of meta-level attacks*

5. $\mathcal{S}_A \subseteq \mathcal{A}_M \times \mathcal{A}_O$ *represents the set of supports from meta-level arguments into object-level arguments*

6. $\mathcal{S}_R \subseteq \mathcal{A}_M \times \mathcal{R}_O$ *represents the set of supports from meta-level arguments into object-level attacks*

7. $\mathcal{A}_O \cap \mathcal{A}_M = \emptyset$

8. $\forall A \in \mathcal{A}_O \; \exists \; B \in \mathcal{A}_M \; : \; (B, A) \in \mathcal{S}_A$

9. $\forall R \in \mathcal{R}_O \; \exists \; B \in A_M \; : \; (B, R) \in \mathcal{S}_R$

The object-level argument graph is represented by the couple $(\mathcal{A}_O, \mathcal{R}_O)$, while the meta-level argument graph is represented by the couple $(\mathcal{A}_M, \mathcal{R}_M)$. The two distinct components are connected by the support relations represented by $\mathcal{S}_A$ and $\mathcal{S}_R$. These supports are the only structural interaction between the meta and the object levels. Condition (8) above ensures that every object-level argument is supported by at least one meta-level argument, whereas condition (9) ensures that every object-level attack is supported by at least one meta-level argument.

Perspectives of the object-level graph can be defined as:

Definition 2.12 (Perspective)**.** *Let* $G = \langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$ *be a bimodal argumentation graph and let* L_S *be a labelling semantics. A tuple* $\langle \mathcal{A}'_O, \mathcal{R}'_O \rangle$ *is an* L_S*-perspective of* G *if* $\exists \; l \in L_S(\langle \mathcal{A}_M, \mathcal{R}_M \rangle)$ *such that*

- $\mathcal{A}'_O = \{ \; A | \exists B \in \mathcal{A}_M \;\; s.t. \; l(B) = \mathsf{IN}, (B, A) \in \mathcal{S}_A \}$
- $\mathcal{R}'_O = \{ \; R | \exists B \in \mathcal{A}_M \;\; s.t. \; l(B) = \mathsf{IN}, (B, R) \in \mathcal{S}_R \}$

Consequently, an object argument may occur in one perspective and not in another according to the results yielded by the meta-level argumentation graph. Under this setting, the role of conditions (8) and (9) becomes clear: every element in a lower level must be relevant w.r.t. the meta-level argumentation process—i.e. we can not have arguments that in no case can be part of a perspective.

2.3 Argument schemes for meta-level argumentation

A fundamental aspect to consider when dealing with a multi-level argumentation graph is how the higher-level graphs can be built starting from the object-level ones. For the purpose, in this work – following the example in [11] – we leverage on argument schemes [18]. In short, argumentation schemes are commonly-used patterns of reasoning. They can be formalised in a rule-like form [14] where every argument scheme consists of a set of conditions and a conclusion. If the conditions are met, then the conclusion holds. Each scheme comes with a set of *critical questions* (CQ), identifying possible exceptions to the admissibility of arguments derived from the schemes.

Definition 2.13 (Meta-predicate)**.** *A meta-predicate P_M is a symbol that represents a property or a relation between object-level arguments. Let be $\mathcal{M}$ the set of all P_M.*

Definition 2.14 (Object-relation meta-predicate)**.** *An object-relation meta-predicate O_M is a predicate stating the existence of a relation at the object level—e.g., attacks, preferences, and conclusions. Let be $\mathcal{O}$ the set of all O_M.*

Moving from the above definitions we can define an argument scheme as:

Definition 2.15 (Argument Scheme)**.** *An* ***argument scheme*** *s has the form:*

$$s : P_1, ..., P_n, \sim P'_1, ..., \sim P'_m \Rightarrow Q$$

with $0 \leq n, m$, and where

- *each $P_1, \ldots, P_n, P'_1, \ldots, P'_m \in \mathcal{M} \cup \mathcal{O}$, while $Q \in \mathcal{M}$*
- *$\sim P$ denotes weak negation (negation by failure) of P—i.e., P is an exception that would block the application of the rule whose antecedent includes $\sim P$*
- *we denote with CQ_s the set of critical questions associated to scheme s.*

Using argument schemes we can build meta-arguments.

Definition 2.16 (Meta-Argument)**.** *A* ***meta-argument*** *A constructed from a set of argument schemes S and an object-level argumentation graph G is a finite construct of the form: $A : A_1, \ldots A_n \Rightarrow_s P$ with $0 \leq n$, where*

- *A is the argument's unique identifier;*
- *$s \in S$ is the scheme used to build the argument;*
- *$A_1, \ldots, A_n$ are arguments constructed from S and G;*
- *P is the* conclusion *of the argument, denoted by $\mathsf{Conc}(A)$.*

$CQ(A)$ denotes the critical questions associated to scheme s. The same notation introduced for standard arguments in Notation 2 also applies to meta-arguments.

We can now define attacks over meta-arguments, or, *meta-attacks*.

Definition 2.17 (Meta-Attack). *An argument A **attacks** argument B (on B') iff either* (i) $\mathsf{Conc}(A) = \bar{P}$ *for some* $B' \in \mathsf{Sub}(B)$ *of the form* $B''_1, ..., B''_M \Rightarrow P$, *or* (ii) $\mathsf{Conc}(A) = P$ *for some* $B' \in \mathsf{Sub}(B)$ *such that* $\sim P \in Antecedent(\mathsf{TopRule}(B'))$.

The same definition of *argumentation graph* and *labellings* introduced for standard argumentation in Definitions 2.8, 2.9, 2.10 also holds for meta-arguments and for the meta level.

3 Burden of persuasion as meta-argumentation

Informally, we can say that when we talk about the notion of the burden of persuasion concerning an argument, we intuitively argue over that argument according to a meta-argumentative approach.

Let us consider, for instance, an argument A: if we allocate the burden over it, we implicitly impose the duty to prove its admissibility on A. Thus, moving the analysis up to the meta level of the argumentation process is like having two arguments, let them be F_{BP} and S_{BP}, reflecting the burden of persuasion status. According to this perspective, F_{BP} states that "the burden is not satisfied if A *fails* to prove its admissibility" – i.e. A should be rejected or undefined – and, of course, F_{BP} is not compatible with A being accepted. Alongside, S_{BP} states that "A is acceptable since it *satisfies* its burden". F_{BP} and S_{BP} have a contrasting conclusion and thus they attack each other.

Analysing the burden from this perspective makes immediately clear that the notions that the meta model should deal with are:

N1 the notion of the burden itself expressing the possibility for an argument to be allocated with a burden of persuasion (i.e., *burdened argument*)

N2 the possibility that this burden is satisfied (that is, a *burden met*) or not satisfied

N3 the possibility of making *attacks* involving burdened arguments ineffective.

The outline of that multi-part evaluation scheme for burdens of persuasion in argumentation is now visible and can be formally designed. In the following, we formally define these concepts by exploiting bimodal argument graphs as techniques for expressing the two main levels of the model – meta and object level – and the relationships between the two.

In particular, we are going to define each set of the bimodal argument graph tuple $\langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$. With respect to $\mathcal{A}_O$ and $\mathcal{R}_O$, representing respectively the set of object-level arguments and attacks, they are built accordingly to the argumentation framework discussed in Subsection 2.1. Hence, our analysis focuses on the meta-level graph $\langle \mathcal{A}_M, \mathcal{R}_M \rangle$ and on the support sets connecting the two levels ($\mathcal{S}_A$ and $\mathcal{S}_R$).

3.1 Meta-level graph

We now proceed to detail all the argumentation schemes used to build arguments in the meta-level graph. Every scheme comes along with its critical questions. As we will see in the next sections, all the critical questions have to be interpreted as kind of "presumptions": they are believed to be true during the construction and evaluation of the argumentation framework – i.e., they are not used as possible attack dimensions –, but their post hoc verification invalidates the entire solution.

Let us first introduce the basic argumentation scheme enabling the definition and representation of an argument with an allocation of the burden of persuasion (i.e., reifying **N1**). We say that an object-level argument A has the burden of persuasion on it if exists an object-level argument B such that $\mathsf{Conc}(B) = \mathit{bp}(\mathsf{Conc}(A))$. This notion is modeled through the following argument scheme:

$$\mathit{conclusion}(A, \phi), \mathit{conclusion}(B, \mathit{bp}(\phi)) \Rightarrow \mathit{burdened}(A) \tag{S0}$$

$$\textit{Is argument B provable?} \tag{$\mathsf{CQ_{S0}}$}$$

where $\mathit{bp}(\phi)$ is a predicate stating ϕ is a literal with the allocation of the burden, $\mathit{conclusion}(A, \phi)$ is a structural meta-predicate stating that $\mathsf{Conc}(A) = \phi$ holds, and $\mathit{burdened}(A)$ is a meta-predicate representing the allocation of the burden on A. Clearly, an argument produced using this scheme only holds if both the arguments A and B on which the inference is based hold—critical question $\mathsf{CQ_{S0}}$.

Analogously, we introduce the scheme S1 representing the absence of such an allocation:

$$\mathit{conclusion}(A, \phi) \Rightarrow \neg \mathit{burdened}(A) \tag{S1}$$

$$\textit{Is argument A provable? Are arguments concluding } \mathit{bp}(\phi) \textit{ not provable?} \tag{$\mathsf{CQ_{S1}}$}$$

Then, as informally introduced at the beginning of this section, we have two schemes reflecting the possibility for a burdened argument to meet or not the burden (**N2**).

$$\mathit{burdened}(A) \;\Rightarrow\; \mathit{bp\_met}(A) \tag{S2}$$

$$\mathit{burdened}(A) \;\Rightarrow\; \neg \mathit{bp\_met}(A) \tag{S3}$$

$$\textit{Is argument } A \textit{ provable?} \qquad (\mathsf{CQ}_{\mathsf{S2}})$$

$$\textit{Is argument } A \textit{ always refuted or undecidable?} \qquad (\mathsf{CQ}_{\mathsf{S3}})$$

where *bp_met* is the meta-predicate stating the burden has been met. It is important to notice that the two schemes above reach opposite conclusions from the same grounds—i.e., the presence of the burden on argument A. The discriminating elements are the critical questions they are accompanied by. In the case of S2, we have that only if a burden of persuasion on argument A exists, and A is acceptable ($\mathsf{CQ}_{\mathsf{S3}}$), then the burden is satisfied. On the other side, the validity of S3 is bound to the missing admissibility of argument A. We will see in Section 3.3 how the meta-arguments and the associated questions concur to determine the model results.

Let us now consider attacks between arguments and their relation with the burden of persuasion allocation. When a burdened argument fails to meet the burden, the only thing affecting the argument's acceptability is the burden itself—i.e., attacks from other arguments do not influence the status of the burdened argument, which only depends on its inability to satisfy the burden. The same applies to attacks issued by an argument that fails to meet the burden: the failure implies argument rejection and, as a direct consequence, the inability to effectively attack other arguments. In order to capture the nuance of discerning between effective and ineffective object-level attacks w.r.t. the concept of burden of persuasion (**N3**), we define the following scheme:

$$\mathit{attack}(B, A), \sim(\neg \mathit{bp\_met}(A)), \sim(\neg \mathit{bp\_met}(B)) \Rightarrow \mathit{effectiveAttack}(B, A) \qquad (\mathsf{S4})$$

$$\textit{Can we prove arguments } A \textit{ or } B \textit{ do not fail to meet their burden?} \qquad (\mathsf{CQ}_{\mathsf{S4}})$$

where *attack* is a structural meta-predicate stating an attack relation at the object level, whereas *effectiveAttack* is a meta-predicate expressing that an attack should be taken into consideration according to the burden of persuasion allocation. In other words, if an object-level attack involves burdened arguments, and one of these fails to satisfy the burden, then the attack is considered not effective w.r.t. the allocation of the burden.

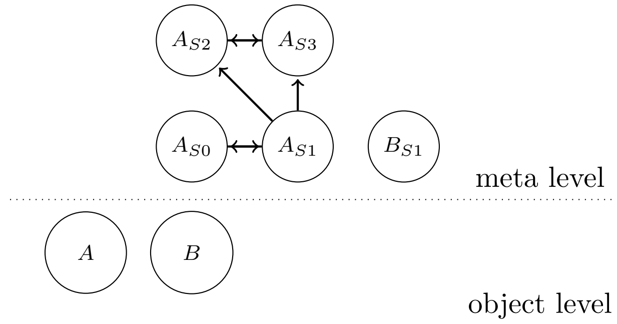
The aforementioned schemes can be used to create a meta-level graph containing all the information about constraints related to the burden of persuasion concept thus leading to a clear separation of concerns, as shown in the following example.

Example 1 (Base)**.** *Let us consider two object-level arguments A and B, concluding the literals a and $bp(a)$ respectively. Using the schemes in Subsection 3.1 we can build the following meta-level arguments:*

- *A_{S0} representing the allocation of the burden on argument A.*
- *A_{S1} and B_{S1} standing for the absence of a burden on arguments A and B respectively. The scheme used to build those arguments exploits weak negation in order to cover those scenarios where an argument concluding a* bp *literal exists at the object-level, but it is found not acceptable.*
- *A_{S2} and A_{S3} sustaining that* (i) *A was capable of meeting the burden on it,* (ii) *A was not capable of meeting its burden.*

The meta-level graph (Figure 1) points out the relations actually implicit in the notion of the burden of persuasion over an argument, where, intuitively, we argue over the consequences of A's possibly succeeding/failing to meet the burden. At the meta level, all the possible scenarios can be explored by applying different semantics over the meta-level graph.

Considering for instance Dung's preferred semantics [1], we can obtain two distinct outcomes: (1) the burden is not satisfied, i.e., argument A_{S3} is accepted, and consequently, A_{S2} is rejected, or (2) we succeed in proving A_{S2}, i.e., the burden is met and A_{S3} is rejected (A_{S0}, A_{S1} are accepted and rejected accordingly). Although the example is really simple – only basic schemes for reasoning on the burden are considered at the meta-level – it clearly demonstrates the possibility of reasoning over the burdens, since, i.e., it establishes whether or not there is a burden on a literal ϕ – argument B in the example – and enables the evaluation of the consequences of a burdened argument to meet or not its burden.



Meta-level arguments:

$A_{S0} :\Rightarrow burdened(A)$
$A_{S1} :\Rightarrow \neg burdened(A)$
$A_{S2} : A_{S0} \Rightarrow bp\_met(A)$
$A_{S3} : A_{S0} \Rightarrow \neg bp\_met(A)$
$B_{S1} :\Rightarrow \neg burdened(B)$

Object-level arguments:

$A :\Rightarrow a$
$B :\Rightarrow bp(a)$

Figure 1: Object and meta level graphs from Example 1

3.2 Object- and meta-level connection: supporting sets

Let us now define how the meta level and the object level interact. Indeed, it is not enough to reason on the consequences of the burden of persuasion allocation only concerning the burdened argument, but the results of the argument satisfying or not such a burden constraint should affect the entire object-level graph. According to the standard bimodal graph theory, defining how the object level and the meta level interact is the role of the argument support relation $\mathcal{S}_A$ and of the attack support relation $\mathcal{S}_R$, respectively. According to Definition 2.11 (Subsection 2.2), every element at level n is connected to an argument at level $n+1$ by a support edge in $\mathcal{S}_A$ or $\mathcal{S}_R$, depending on whether it is either an argument or an attack.

Let us define the support set $\mathcal{S}_A$ of meta arguments supporting object-level arguments as:

$$\begin{aligned}\mathcal{S}_A = \{(Arg_1, Arg_2) \mid\ & Arg_1 \in \mathcal{A}_M, Arg_2 \in \mathcal{A}_O,\\ & (\mathsf{Conc}(Arg_1) = bp\_met(Arg_2) \vee \mathsf{Conc}(Arg_1) = \neg burdened(Arg_2))\}\end{aligned}$$

Intuitively, an argument A at the object level is supported by arguments at the meta level claiming that either the burden on A is satisfied (S2) or there is no burden allocated on it (S1).

The set $\mathcal{S}_R$ of meta arguments supporting object-level attacks is defined as:

$$\begin{aligned}\mathcal{S}_R = \{(Arg_1, (B, A)) \mid\ & Arg_1 \in \mathcal{A}_M, (B, A) \in \mathcal{R}_O,\\ & \mathsf{Conc}(Arg_1) = \mathit{effectiveAttack}(B, A)\}\end{aligned}$$

In other words, an object-level attack is supported by arguments at the meta level claiming its effectiveness w.r.t. the burden of persuasion allocation (S4).

3.3 Equivalence with burden of persuasion semantics

The defined meta-framework can be used to achieve the same results as the original burden of persuasion labelling semantics [5].

Let us first introduce the notion of *CQ-consistency* for a bimodal argumentation graph G.

Definition 3.1 (CQ-consistency). *Let $G = \langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$ be a bimodal argumentation graph, and let $L_S(G)$ be a labelling-based semantics. P is the set of corresponding L_S-perspectives. A perspective $p \in P$ is CQ-consistent if every* IN *argument A in the corresponding meta-level labelling satisfies its critical questions ($CQ(A)$).*

Before proceeding, let us ground the Critical Questions introduced in Subsection 3.1 within the context of L_S-perspectives and labelling based semantics.

CQ_{S0} Given a L_S-perspective p and one of its labelling l, is $l(B) = \textsf{IN}$?

CQ_{S1} Given a L_S-perspective p and one of its labelling l, is $l(A) = \textsf{IN}$? If an argument B such that $\mathsf{Conc}(B) = bp(\phi)$ does exist, is $l(B) \in \{\textsf{UND}, \textsf{OUT}\}$?

CQ_{S2} Given a L_S-perspective p and one of its labelling l, is $l(A) = \textsf{IN}$?

CQ_{S2} Given all L_S-perspectives p and the set of their labellings L, does $\forall l \in L, l(A) \in \{\textsf{UND}, \textsf{OUT}\}$ hold?

CQ_{S3} Given a L_S-perspective p and one of its labelling l, are $l(A) = \textsf{IN}$ and $l(B) = \textsf{IN}$?

Using this new definition we can introduce the concept of *BP-perspective.*

Definition 3.2 (BP-perspective)**.** *Let $G = \langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$ be a bimodal argumentation graph, and P the set of its L_{stable}-perspectives [1]. We say that $p \in P$ is a BP-perspective of G iff p is CQ-consistent.*

Example 2 (Antidiscrimination law)**.** *Let us consider a case in which a woman claims to have been discriminated against in her career on the basis of her sex, as she was passed over by male colleagues when promotions came available ($ev1$), and brings evidence showing that in her company all managerial positions are held by men ($ev3$), even though the company's personnel includes many equally qualified women, having worked for a long time in the company, and with equal or better performance ($ev2$). Assume that this practice is deemed to indicate the existence of gender-based discrimination (indiciaDiscrim) and that the employer fails to provide prevailing evidence that the woman was not discriminated against ($\neg$discrim). It seems that it may be concluded that the woman was indeed discriminated against on the basis of her sex.*

Consider, for instance, the following formalisation of the European nondiscrimination law, that, in case of presumed discrimination, requires prevailing evidence that no offence was committed—i.e., $bp(\neg discrim)$:

$e1 : ev1$ $\quad$ $e2 : ev2$ $\quad$ $e3 : ev3$
$er1 : ev1 \Rightarrow indiciaDiscrim$ $\quad$ $er2 : ev2 \Rightarrow \neg discrim$ $\quad$ $er3 : ev3 \Rightarrow discrim$
$r1 : indiciaDiscrim \Rightarrow bp(\neg discrim)$

We can then build the following object-level arguments:

$A_0 :\Rightarrow ev1$	$B_0 :\Rightarrow ev2$	$C_0 :\Rightarrow ev3$
$A_1 : A_0 \Rightarrow indiciaDiscrim$	$B_1 : B_0 \Rightarrow \neg discrim$	$C_1 : C_0 \Rightarrow discrim$
$A_2 : A_1 \Rightarrow bp(\neg discrim)$		

and the following meta-level arguments:

$A_{0_{S1}} :\Rightarrow -burdened(A_0)$	$B_0 S1 :\Rightarrow -burdened(B_0)$
$A_{1_{S1}} :\Rightarrow -burdened(A_1)$	$B_{1_{S0}} :\Rightarrow burdened(B_1)$
$A_{2_{S1}} :\Rightarrow -burdened(A_2)$	$B_{1_{S1}} :\Rightarrow -burdened(B_1)$
$C_{0_{S1}} :\Rightarrow -burdened(C_0)$	$B_{1_{S2}} : B_{1_{S0}} \Rightarrow bp\_met(B_1)$
$C_{1_{S1}} :\Rightarrow -burdened(C_1)$	$B_{1_{S3}} : B_{1_{S0}} \Rightarrow \neg bp\_met(B_1)$
$C_1B_{1_{S4}} :\Rightarrow effectiveAttack(C_1, B_1)$	$B_1C_{1_{S4}} :\Rightarrow effectiveAttack(B_1, C_1)$

The resulting graph is depicted in Figure 2. In this case, at the object level, since there are indicia of discrimination (A_1), we can infer the allocation of the burden on non-discrimination (A_2). Moreover, we can build both arguments for discrimination (C_1) and non-discrimination (B_1), leading to a situation of undecidability.

At the meta level we can apply the rule S1 for every argument at the object level ($A_{0_{S1}}, A_{1_{S1}}, A_{2_{S1}}, B_0 S1, B_{1_{S0}}, C_{0_{S1}}, C_{1_{S1}}$) – where we can establish the absence of the burden for all of them –, and the rule S4 for every attack ($C_1B_{1_{S4}}, B_1C_{1_{S4}}$). By exploiting B_1 and A_2, we can also apply schema S0, and consequently rules S2 and S3. In a few words, we are concluding the meta argumentative structure given by the allocation of the burden of persuasion on argument B_1.

We can now apply the stable labelling to the meta-level graph, thus obtaining three distinct results. For clarity reasons, in the following, we ignore the arguments that are acceptable under every solution.

1. $\mathsf{IN} = \{B_{1_{S1}}, C_1B_{1_{S4}}, B_1C_{1_{S4}}\}, \mathsf{OUT} = \{B_{1_{S0}}, B_{1_{S2}}, B_{1_{S3}}\}, \mathsf{UND} = \{\}$—*i.e., B_1 is not burdened;*

2. $\mathsf{IN} = \{B_{1_{S0}}, B_{1_{S2}}, C_1B_{1_{S4}}, B_1C_{1_{S4}}\}, \mathsf{OUT} = \{B_{1_{S1}}, B_{1_{S3}}\}, \mathsf{UND} = \{\}$—*i.e., B_1 is burdened and the burden is met;*

3. $\mathsf{IN} = \{B_{1_{S0}}, B_{1_{S3}}\}, \mathsf{OUT} = \{B_{1_{S1}}, B_{1_{S2}}, C_1B_{1_{S4}}, B_1C_{1_{S4}}\}, \mathsf{UND} = \{\}$—*i.e., B_1 is burdened and the burden is not met.*

Then, the meta-level results can be reified to the object-level perspectives taking into account the CQ *we have to impose on the solutions and the results given by the perspective evaluation under the grounded semantics. Let us first consider solutions 1 and 2. They lead to the same perspective on the object-level graph—the graph remains unchanged w.r.t. the original graph. If we consider the critical questions attached to the* IN *arguments, both these solutions are not valid. Indeed, according* *to solution 1 the burden is not allocated on argument B_1, but this is in contrast with argument A_2's conclusion (A_2 is* IN *under grounded labelling)—i.e., CQ_{S1} is not satisfied. Analogously, solution 2 concludes that B_1 is allocated with the burden and its success to meet the burden, but at the same time, argument B_1 is found undecidable at the object level (B_1 is* UND *under the grounded semantics)—i.e., CQ_{S2} is not satisfied.*

The only acceptable result is the one given by solution 3. In this case, argument B_1 is not capable to meet the burden – ${B_1}_{S3}$ is IN *– and, consequently, it is rejected and deleted from the perspective. Indeed, CQ_{S3} is satisfied. As a consequence, argument C_1 is labelled* IN*. In other words, the argument for non-discrimination fails and the argument for discrimination is accepted.*

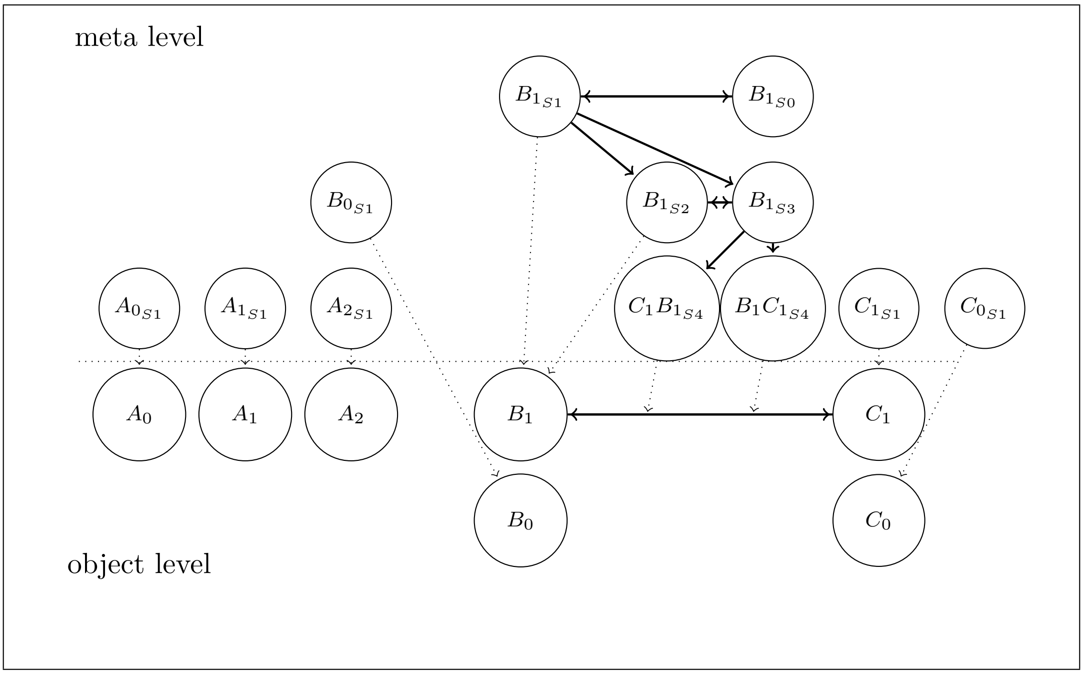


Figure 2: Argumentation graph (object- and meta- level) from Example 2

Before proceeding, let us recall the main definitions from Calegari and colleagues [5], who, in their work, present a semantics dealing with the burden of persuasion allocation on members of the argumentation language.

Definition 3.3 (BP-defeat)**.** *Given a set of burdens of persuasion* BurdPers*, A **bp-defeats** B iff there exists a subargument B' of B such that:*

1. *$\mathsf{Conc}(A) = \overline{\mathsf{Conc}(B')}$ and*
   (a) *$\mathsf{Conc}(A) \notin \mathsf{BurdPers}$, and $B' \not\succ A$, or*
   (b) *$\mathsf{Conc}(A) \in \mathsf{BurdPers}$ and $A \succ B'$.*
2. *$\mathsf{Conc}(A) = \neg \mathsf{N}(\rho)$, where ρ is $\mathsf{TopRule}(B')$.*

Definition 3.4 (Grounded BP-labelling)**.** *A grounded **BP-labelling** of an argumentation graph G, relative to a set of burdens $\mathsf{BurdPers}$, is a $\{\mathsf{IN}, \mathsf{OUT}, \mathsf{UND}\}$-labelling l s.t. the set of UND arguments is minimal and $\forall \mathsf{A} \in \mathcal{A}_G$ with $\mathsf{Conc}(A) = \phi$*

1. *$l(A) = \mathsf{IN}$ iff $\forall \mathsf{B} \in \mathcal{A}_G$ such that B bp-defeats $\mathsf{A} : l(B) = \mathsf{OUT}$*
2. *$l(A) = \mathsf{OUT}$ iff*
   (a) *$\phi \in \mathsf{BurdPers}$ and $\exists\ B \in \mathcal{A}_G$ s.t. B bp-defeats A and $l(B) \neq \mathsf{OUT}$*
   (b) *$\phi \notin \mathsf{BurdPers}$ and $\exists\ B \in \mathcal{A}_G$ such that B bp-defeats A and $l(B) = \mathsf{IN}$*
3. *$l(A) = \mathsf{UND}$ otherwise.*

Proposition 3.1. *If $\nexists A, B \in \mathcal{A}_O$ such that both A and B have a burden of persuasion on them and A is reachable from B through $\mathcal{R}_O$, the results yielded by the grounded evaluation of G's BP-perspectives are congruent with the evaluation of the object-level graph $\langle \mathcal{A}_O, \mathcal{R}_O \rangle$ under the Grounded BP-labelling as in Definition 3.4 [5].*

Proof. The burden of persuasion semantics acts like the grounded semantics, with the only difference being that the burdened arguments that would have been UND for the latter could be OUT/IN for the former. So, it is a matter of fact that burdened arguments and arguments connected to them through attack relation can change their state.

Let us consider an argumentation graph $AF\langle \mathcal{A}, \rightsquigarrow \rangle$, and let L_G be the grounded labelling resulting from the evaluation of AF under a grounded semantics. With respect to our framework, and in particular, to the bimodal argumentation graph $G = \langle \mathcal{A}_O, \mathcal{A}_M, \mathcal{R}_O, \mathcal{R}_M, \mathcal{S}_A, \mathcal{S}_R \rangle$, we have, by construction, that every node at the object level, if not burdened, has an undisputed supporting argument at the meta level (S1 or S4). As a consequence, the meta level has no influences on no burdened arguments, and – in the absence of burdened arguments – the evaluation of the object level graph under the grounded semantics would be equal to L_G. It is a matter of fact that the meta level influences only the burdened arguments' state. Accordingly, the extent of this influence and the consequences on the object-level graph will be considered in the following.

Let us consider a single argument $A \in \mathcal{A}$ allocated with the burden of persuasion, thus having the additional argument $B \in \mathcal{A}$ stating the burden on A (as depicted in Figure 1). Computing the stable semantics on the meta-level graph produces the following scenarios:

Stable.a burden on A cannot be proved;

Stable.b burden on A can be proved and the burden is met;

Stable.c burden on A can be proved and the burden is not met.

Accordingly, the stable evaluation of the meta-graph produces three different perspectives of the object level:

(i) argument A is supported—it is not burdened;

(ii) argument A is supported—it satisfies the burden;

(iii) argument A is not supported, and then it is excluded from the object-level graph—it does not meet the burden then it is refuted.

In particular, we have that Stable.a induces *(i)*, Stable.b leads to *(ii)*, while Stable.c induces *(iii)*. Let L_{BP} be this new object-level labelling (obtained by the meta-level stable semantics reification at the object level). Also, let us compare L_{BP} with the initial object-level grounded labelling L_G. Then, the following cases can occur (E is exploited for valid solutions with labelling equivalence, while C is exploited for solutions to be discarded).

- B is OUT or UND in L_G.

  E1 If *(i)* the burden is not allocated and cannot be proven, the meta level does not influence the object level supporting all unburdened arguments. CQ_{S1} is satisfied and L_{BP} is equivalent to L_G.

  C1 If *(ii)* or *(iii)*, in both cases CQ_{S0} is not satisfied—the burden is proved at the meta level and not at the object level.

- B is IN and A is OUT in L_G.

  C2 If *(i)* we have inconsistency on CQ_{S1}—the burden is proved at the object level and not at the meta level.

  C3 If *(ii)* we have inconsistency on CQ_{S2} since A is considered IN at the meta level (supported by the meta-argument) but A is OUT at the object level.

E2 If *(iii)* A is not supported, i.e., removed from the object-level graph. CQ_{S0} and CQ_{S3} are both satisfied. Then, under the grounded semantics, the removal of an OUT argument from a graph is not influent w.r.t. its evaluation, i.e., L_{BP} is equivalent to L_G.[1]

- B is IN and A is IN in L_G.

C4 If *(i)*, we have inconsistency on CQ_{S1}—the burden is proved at the object level and not at the meta level.

E3 If *(ii)*, then CQ_{S0} and CQ_{S2} are both satisfied and L_{BP} is equal to L_G.

C5 If *(iii)* we have an inconsistency because CQ_{S3} is not satisfied.

- B is IN and A is UND in L_G.

C6 If *(i)*, we have inconsistency on CQ_{S1}—the burden is proved at the object level and not at the meta level.

C7 If *(ii)*, we have an inconsistency since A is considered IN at the meta level (supported by the meta-argument) but A is UND at the object level—CQ_{S2} is not satisfied.

E4 If *(iii)* A is not supported, i.e., is removed from the object level, i.e., it can be labelled as OUT in L_{BP} (see [1]). CQ_{S0} and CQ_{S2} are satisfied.

As made evident by the proof, the reification of the meta level upon the object level generates multiple solutions: yet, only one solution for each case can be considered valid w.r.t. critical questions. Moreover, the only valid perspective coincides with the one generated from the bp-labelling in [5]—the burdened argument is labelled OUT in case of indecision (E4). Obviously, the proof can be generalised to configurations taking into account any number of burdened independent arguments—where combinations grow exponentially with the number of burdened arguments. □

4 Burden Inversion

Let us consider a situation in which one argument A is presented for a claim ϕ being burdened, and A (or one of its subarguments) is attacked by a counterargument B, of which the conclusion ψ is also burdened. Intuitively, if both arguments fail to

[1] It can trivially be proved considering that – in the grounded semantics – an OUT argument does not affect other arguments' state, i.e., it is irrelevant and can be removed; of course, also the dual proposition holds, i.e., if L_{BP} build in the meta-frameworks does not consider an argument it can be labelled as OUT in the grounded bp-labelling

satisfy the burden of persuasion, both of them are to be rejected. This is not the case if the inversion of the burden is taken into account [5]—i.e., if no convincing argument for ψ is found, then the attack fails, and the uncertainty on ψ does not affect the status of A. Accordingly, B is rejected for failing to meet its burden, thus leaving A free to be accepted also if it was not able to satisfy the burden of persuasion in the beginning.

The model we propose in this work is able to correctly deal with the inversion of the proof, as we discuss in the next example adapted from [5].

Example 3 (Inversion of the burden)**.** *Let us consider a case in which a doctor caused harm to a patient by misdiagnosing his case. Assume that there is no doubt that the doctor harmed the patient (harm), but it is uncertain whether the doctor followed the guidelines governing this case. Assume that, under the applicable law, doctors are liable for any harm suffered by their patients (liable), but they can avoid liability if they show that they exercised due care in treating the patient (dueDiligence). Let us also assume that a doctor is considered to be diligent if he/she follows the medical guidelines that govern the case (guidelines). The doctor has to provide a convincing argument that he/she was diligent* ($bp(dueDiligence)$)*, and the patient has to provide a convincing argument for the doctor's liability* ($bp(liable)$)*.*

We can formalise the case as follows:

$$\begin{array}{ll} f1 : guidelines & f2 : \neg guidelines \\ f3 : harm & r1 : \neg guidelines \Rightarrow \neg dueDiligence \\ r2 : guidelines \Rightarrow dueDiligence & r3 : harm, \sim dueDiligence \Rightarrow liable \\ bp1 : bp(dueDiligence) & bp2 : bp(liable) \end{array}$$

We can then build the following object-level arguments:

$$\begin{array}{ll} A_0 :\Rightarrow bp(dueDiligence) & A_1 :\Rightarrow bp(liable) \\ A_2 :\Rightarrow guidelines & A_3 :\Rightarrow harm \\ A_4 :\Rightarrow \neg guidelines & A_5 : A_2 \Rightarrow dueDiligence \\ A_1 : A_0 \Rightarrow indiciaDiscrim & B_1 : B_0 \Rightarrow \neg discrim \\ C_1 : C_0 \Rightarrow discrim & A_6 : A_3 \Rightarrow liable \\ A_7 : A_4 \Rightarrow \neg dueDiligence & \end{array}$$

According to the original burden semantics, the argument for the doctor's due diligence (A_5) fails to meet its burden of persuasion. Consequently, following the inversion principle, it fails to defeat the argument for the doctor's liability (A_6), which is then able to meet its burden of persuasion.

Let's now analyse the case from the meta-model perspective. Using argument schemes defined in Section 3 we can build the following meta-arguments:

$$\begin{array}{ll}
A_{0_{S1}} :\Rightarrow -\mathit{burdened}(A_0) & A_{1_{S1}} :\Rightarrow -\mathit{burdened}(A_1) \\
A_{2_{S1}} :\Rightarrow -\mathit{burdened}(A_2) & A_{3_{S1}} :\Rightarrow -\mathit{burdened}(A_3) \\
A_{4_{S1}} :\Rightarrow -\mathit{burdened}(A_4) & A_{7_{S1}} :\Rightarrow -\mathit{burdened}(A_7) \\
A_2A_{7_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_2, A_7) & A_2A_{4_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_2, A_4) \\
A_4A_{2_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_4, A_2) & \\
A_7A_{5_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_7, A_5) & A_5A_{7_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_5, A_7) \\
A_4A_{5_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_4, A_5) & A_5A_{6_{S4}} :\Rightarrow \mathit{effectiveAttack}(A_5, A_6) \\
A_{5_{S0}} :\Rightarrow \mathit{burdened}(A_5) & A_{5_{S1}} :\Rightarrow -\mathit{burdened}(A_5) \\
A_{5_{S2}} : A_{5_{S0}} \Rightarrow \mathit{bp\_met}(A_5) & A_{5_{S3}} : A_{5_{S0}} \Rightarrow \neg \mathit{bp\_met}(A_5) \\
A_{6_{S0}} :\Rightarrow \mathit{burdened}(A_6) & A_{6_{S1}} :\Rightarrow -\mathit{burdened}(A_6) \\
A_{6_{S2}} : A_{6_{S0}} \Rightarrow \mathit{bp\_met}(A_6) & A_{6_{S3}} : A_{6_{S0}} \Rightarrow \neg \mathit{bp\_met}(A_6)
\end{array}$$

Connecting the object- and meta-level arguments we obtain the graph in Figure 3. Let us now consider the extensions obtained applying stable semantics to the meta-level graph:

1. $\{A_{6_{S0}}, A_{6_{S2}}, A_{5_{S0}}, A_{5_{S3}}\}$
2. $\{A_{6_{S0}}, A_{6_{S3}}, A_{5_{S0}}, A_{5_{S3}}\}$
3. $\{A_{6_{S0}}, A_{6_{S2}}, A_{5_{S0}}, A_{5_{S2}}, A_5A_{6_{S4}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
4. $\{A_{6_{S0}}, A_{6_{S3}}, A_{5_{S0}}, A_{5_{S2}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
5. $\{A_{6_{S0}}, A_{6_{S2}}, A_{5_{S1}}, A_5A_{6_{S4}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
6. $\{A_{6_{S0}}, A_{6_{S3}}, A_{5_{S1}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
7. $\{A_{6_{S1}}, A_{5_{S0}}, A_{5_{S2}}, A_5A_{6_{S4}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
8. $\{A_{6_{S1}}, A_{5_{S1}}, A_5A_{6_{S4}}, A_5A_{7_{S4}}, A_7A_{5_{S4}}, A_4A_{5_{S4}}\}$
9. $\{A_{6_{S1}}, A_{5_{S0}}, A_{5_{S3}}\}$

The only extensions that produce a CQ-consistent perspective are the first and the second, given that all the others violate at least one of the constraints imposed by the critical questions—e.g. CQ_{S1} for $5, 6, 7, 8, 9$ and CQ_{S2} for $3, 4$. The first perspective acts exactly like the original semantics from [5]—i.e., the argument for the doctor's due diligence (A_5) fails to meet the burden ($A_{5_{S3}}$), and consequently, the argument for doctor's liability (A_6) is able to satisfy its own burden ($A_{6_{S2}}$). However, the model delivers a second result according to which both A_5 and A_6 fail to meet their burden of persuasion ($A_{6_{S3}}$ and $A_{5_{S3}}$). It is the result that we would have expected in absence of the inversion principle.

The example highlights the meta-argumentation model is able to provide both a solution that follows the inversion principle and one not considering it. When the inversion principle is taken into account the number of burdened arguments is maximised in the final extension. Accordingly, we can provide a generalisation of Property 3.1:

Proposition 4.1. *Given the results yielded by the grounded evaluation of G's BP-perspectives, the results that maximise the number of burdened arguments in the* IN *set are congruent with the evaluation of the object-level graph $\langle \mathcal{A}_O, \mathcal{R}_O \rangle$ under the grounded-bp semantics as in Definition 3.4 [5].*

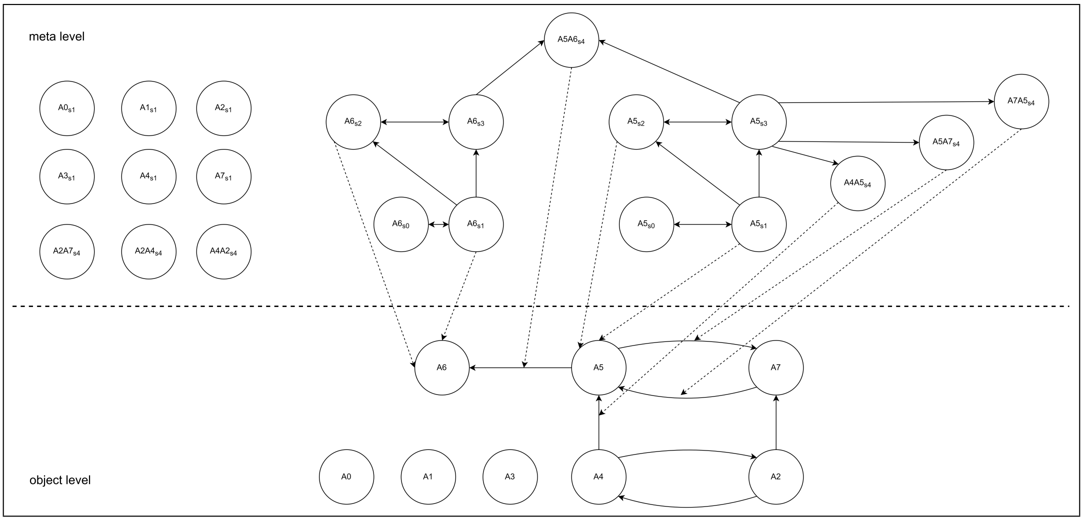


Figure 3: Argumentation graph (object- and meta- level) from Example 3

5 Technological Reification

Despite the benefits of the meta-approach discussed in Section 3 – such as clear separation of concerns, encapsulations of argumentation abstractions and naturalness in terms of human thinking – the method is quite inefficient from a computational perspective. Indeed, the meta-level evaluation leads to a stable semantics computation, with a non-polynomial complexity [8]. This is why, from a technological perspective, the model presented in Section 3 has been reified into a more efficient resolution method.

In a nutshell, the proposed approach exploits the stable semantics to explore the search space at the meta level. Then, in order to identify the final solution, the grounded assessment of the object level is taken into account—selecting the acceptable scenario according to the critical questions. The idea behind the technological refinement is exactly to leverage the information of those arguments to guide the search—i.e., to exploit the grounded assessment of the object level as an *a priori* constraint. Following this idea, the computation algorithm becomes really simple. The two argumentation levels (object and meta) are collapsed in a single graph, following [3]. Then, the graph is modified dynamically, leveraging the information on the burdened arguments. In a sense, we have a multi-stage evaluation that leads to the modification of the graph itself at every stage.

Let us consider the framework of Example 2. There, two arguments exist, namely B_1 and C_1, attacking each other. Then, another argument, A_2, concludes the presence of the burden on B_1. The grounded evaluation of this framework would lead to a single extension containing argument A_2—i.e., the burden on B_1 has been proved, and we should proceed to verify B_1's compliance with the constraint. According to the model presented in Section 3, the graph should be used to build the meta-level framework expressing all the possible outcomes the burden could lead to. Then, the one leading to an object-level perspective that satisfies all the attached *Critical Questions* would be the correct one. This kind of assessment has one major drawback: we already know from the initial grounded evaluation that B_1 does not satisfy its burden; however, through stable semantics, we explore also the scenarios in which B_1's burden is satisfied, just to discard them later using the *Critical Questions*. The main idea of the technological reification presented in this Section is exactly to use the information generated by the initial grounded assessment to produce a new graph including all the new meta-knowledge.

Let us test this approach with the theory in Example 2. We know that B_1 has a burden on it, but it has not been able to satisfy it. As in the original model, we can use this info to build the argument $B_{1_{S3}}$ using the scheme S3. Intuitively, this new argument claims that "B_1 should be rejected for not being able to defend itself" and, consequently, it throws a new attack against it. If we add these new elements to the original framework, we obtain a new framework containing both object- and meta-arguments on the same level. Its evaluation under grounded semantics leads to the expected result: B_1 is rejected, while C_1 and $B_{1_{S3}}$ are both accepted.

More generally, what we are doing is verifying the *Critical Questions* associated with a meta scheme using the grounded evaluation of the original framework. In this way, we do not need stable semantics to explore all the possible scenarios, but, instead, we can directly select the correct one. For instance, in the case of Example 2, $B_{1_{S3}}$ satisfies its critical questions, while $B_{1_{S2}}$ does not. In the case B_1 were able to satisfy its burden, then just $B_{1_{S2}}$ would have been instantiated, and consequently, no new attacks would have been introduced in the framework.

Summing up, given a constraint $bp(x)$, then for every argument A having x as its conclusion a new argument B can be introduced in the graph. This argument represents the possibility of A failing/succeeding to meet the burden—expressed by S3 and S2 in the meta-model. A and B's interaction is decided according to the A's ability to satisfy the burden under the grounded semantics:

i) iff A is OUT or UND, then B is an instance of scheme S3, and consequently an attack from B to A is introduced;

ii) iff A is IN, then B is an instance of scheme S2, then no attack is introduced.

Basically, through the first evaluation of the graph, the knowledge required to choose between schemes S3 and S2 is obtained—i.e. the stable semantics evaluation becomes superfluous.

Let us now apply the new approach to Example 3 to see whether the inversion principle is supported or not. If we consider the grounded evaluation of the object-level framework, we obtain two burdened arguments, A_5 and A_6, both failing to satisfy the persuasion constraint. According to our algorithm, we can introduce two meta-arguments based on scheme S3 in the framework, one attacking A_5, and the other A_6. The evaluation of this framework under grounded semantics would of course lead to an undesirable result—i.e. both arguments A_5 and A_6 are rejected.

The enforcement of the inversion mechanism requires a procedural evaluation of the burdened arguments—i.e., we should first evaluate those arguments whose acceptability does not depend on burdened arguments not yet evaluated, and then we apply the algorithm again until all the burdens have been evaluated. For instance, in Example 3 we should first introduce argument $A_{5_{S3}}$ in the graph, and then use the results of this new framework to evaluate the consequences on A_6. Accordingly, the dependencies among burdened arguments are respected—i.e., we enforce the inversion principle.

More formally, given an argumentation framework $AF = \langle \mathcal{A}, \rightsquigarrow \rangle$ along with its grounded extension E_G, we can define the set of burdens to evaluate B_e as

$$\{A_0 \in E_G | \mathsf{Conc}(A_0) = bp(a) \ and \ \nexists b \in E_G \ s.t. \\ \mathsf{Conc}(b) = bp\_met(A_1) \ or \ \neg bp\_met(A_1) \ with \ Conc(A_1) = a\}$$

Then we can define the reduction R_{B_e} of B_e as:

$$\{bp(a) \in B_e \mid \nexists bp(b) \in B_e \ s.t. \ a \ is \ reachable \ from \ b \ through \rightsquigarrow\}$$

In simpler terms, the reduction set contains all the burdens on the arguments whose status does not depend on other burdened arguments. Then, given an AF and its

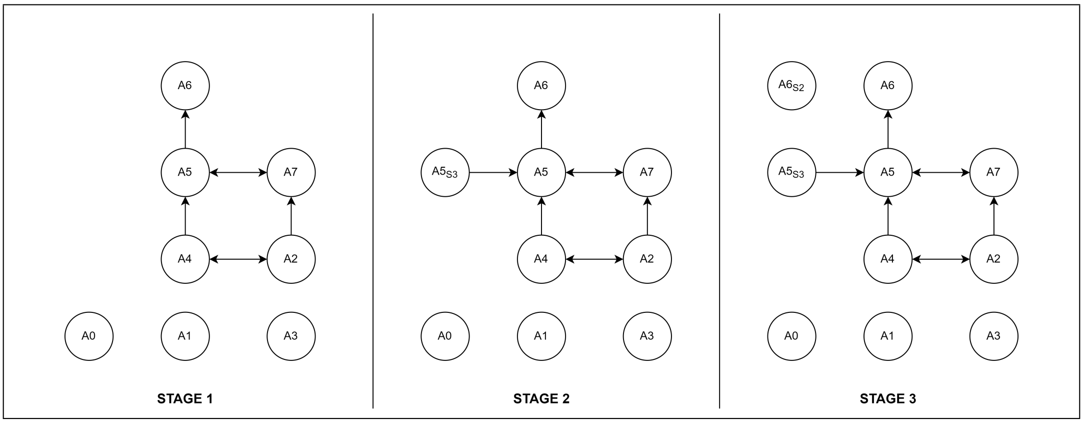


Figure 4: Staged evaluation of Example 3

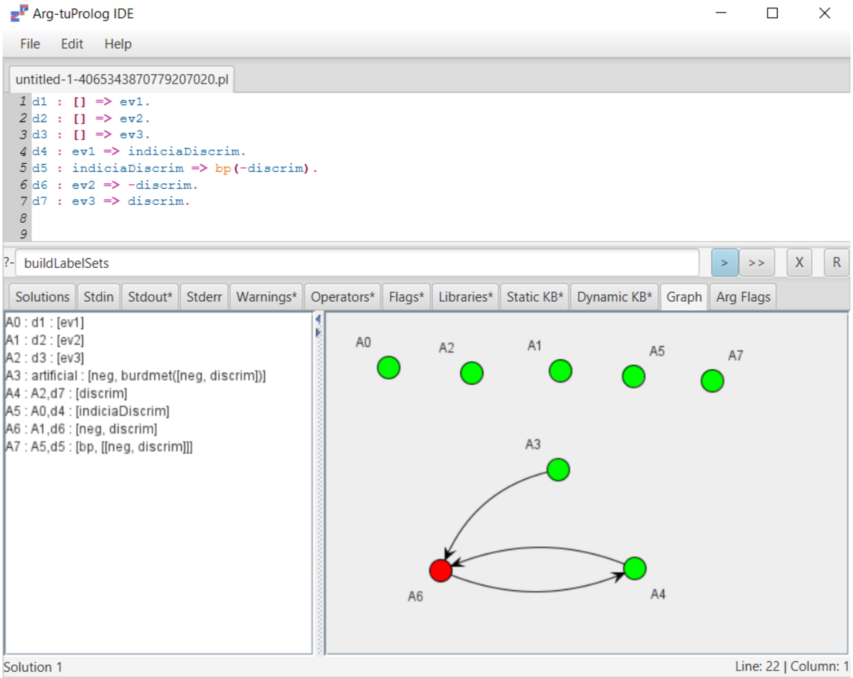
grounded extension we can use the reduction set to produce a new framework AF_1 containing the meta-arguments for the burdens in the set. We can then recursively apply the same procedure on AF_1 until no elements remain to be evaluated in the reduction set. Understandably, the procedure requires the absence of cycles in the burdened arguments in order to derive a partial ordering over the burdens to evaluate. When all the elements in B_e are independent, the reduction set R_{B_e} is the same as B_e—i.e., the procedure is a generalisation of the naive algorithm introduced at the beginning of this section and used in the evaluation of Example 2.

Figure 4 shows Example 3's evaluation steps. The graph on the left is obtained from the initial theory. We can compute the set of burdens ($\{A_0, A_1\}$) and its reduction ($\{A_0\}$). The new knowledge is used to build the framework in the middle by adding an instance of scheme S3 relative to argument A_5 and its attack. Again, we compute the set of burdens ($\{A_0\}$) and its reduction ($\{A_0\}$), and use it to instantiate scheme S2 in the graph on the right. Now the set of burdens to evaluate is empty and we have our final result: argument A_5 fails to satisfy its burden and it is rejected, thus making it possible for A_6 to satisfy its burden.

5.1 Implementation in Arg2P

The algorithm has been tested and implemented in the Arg2P framework[2] [4, 12]. Please note that the equivalence of the optimised procedure with the formal model presented in the paper has for now only been conjectured, thus remaining still unproven. Figure 5 shows the tool evaluation of the example discussed in Example 2.

[2] http://arg2p.apice.unibo.it/



```
d1 : [] => ev1.
d2 : [] => ev2.
d3 : [] => ev3.
d4 : ev1 => indiciaDiscrim.
d5 : indiciaDiscrim => bp(-discrim).
d6 : ev2 => -discrim.
d7 : ev3 => discrim.
```

```
?- buildLabelSets
```

```
A0 : d1 : [ev1]
A1 : d2 : [ev2]
A2 : d3 : [ev3]
A3 : artificial : [neg, burdmet([neg, discrim])]
A4 : A2,d7 : [discrim]
A5 : A0,d4 : [indiciaDiscrim]
A6 : A1,d6 : [neg, discrim]
A7 : A5,d5 : [bp, [[neg, discrim]]]
```

Figure 5: Arg2P evaluation of Example 2

So, the entire process is based on grounded semantics and reachability checking—both polynomial complexity [8]. The algorithm requires $m+1$ evaluation stages to end – where m is the number of connected burdened arguments –, then the final complexity is polynomial.

6 Related

Our approach relies on the work from [10, 11] introducing the required abstraction at the meta level. In particular, the first formalisation of meta-argumentation synthesising bimodal graphs, structured argumentation, and argument schemes in a unique framework is presented in [10]. There, a formal definition of the meta-ASPIC framework is provided as a model for representing object arguments. Along the same line, bimodal graphs are exploited in [11] for dealing with arguments sources' trust. In [11] ASPIC+ is used instead of meta-ASPIC at the object level and on a set of meta-predicates related to the object level arguments and the schemes in the meta level, as in our approach. Both [10] and [11] use critical questions for managing attacks at the meta level.

Our framework and its model mix the two approaches by exploiting bimodal graphs in ASPIC+ and defining all the burdens abstractions at the meta-level. The reification of the meta level at the object level allows the concept of the burden of persuasion to be properly dealt with—i.e., arguments burdened with persuasion have to be rejected when there is uncertainty about them. As a consequence, those arguments become irrelevant to the argumentation framework including them: not only do they fail to be included in the set of accepted arguments, but they also are unable to affect the status of the arguments they attack.

An interesting connection with our work could be drowned with the multi-sorted argumentation networks proposed in [16], and their reification in the modal fibring approach from [2]. The main idea of their work is to allow different parts of a framework – called cells – to be evaluated under different semantics. In a nutshell, a set of arguments is a multi-sorted extension only if it is the union of the extensions computed on the qualified arguments – i.e., arguments not defeated and defended from attacks coming from other cells – of the single cells composing the framework. The modal fibring approach from [2] allows every cell to be represented as a separate argumentation framework, with the possibility of modality used to express inter-cell attacks within these frameworks. Their work could appear similar to the bimodal approach in the way different graphs are used to derive the final results, but there is an important difference to consider: the nature of the relation used to connect the different graphs. Bimodal graphs exploit a support relation to model the dependency of an N-level argument on an N+1-level argument, while multi-sorted networks are based on inter-cell attacks. A naive transposition of our work in a multi-sorted setting would require three steps:

1. the use of the supports to build the attack set connecting meta and object level in order to compose a single graph made of two cells (object and meta);

2. enumeration of the multi-sorted extensions using grounded semantics for the object-cell and stable semantics for the meta-cell;

3. evaluation of the extensions using the *Critical Questions* connected to the meta-argument in them.

However, the transposition would bring no real benefits, while at the same time losing the encapsulation and clarity given by the multi-level structuring of the problem.

7 Conclusions

In this paper we present a meta-argumentation approach for the burden of persuasion in argumentation, discussing interconnections with the state of the art. We show how this model easily deals with all the nuances of burdens such as reasoning over the concept of the burden itself, thus leading to a full-fledged, interoperable framework open to further extensions. Moreover, the model correctly deals with the inversion of the burden.

Future research will be devoted to studying the properties of our meta framework and the connection of our framework with meta-ASPIC for argumentation. We also plan to inquire about the way in which our model fits into legal procedures and enables their rational reconstruction.

References

[1] Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *The Knowledge Engineering Review*, 26(4):365–410, 2011.

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

[3] Guido Boella, Dov M. Gabbay, Leendert van der Torre, and Serena Villata. Meta-argumentation modelling I: Methodology and techniques. *Studia Logica*, 93(2–3):297, 2009.

[4] Roberta Calegari, Giuseppe Pisano, Andrea Omicini, and Giovanni Sartor. Arg2P: An argumentation framework for explainable intelligent systems. *Journal of Logic and Computation*, 32(2):369–401, March 2022. Special Issue from the 35th Italian Conference on Computational Logic (CILC 2020).

[5] Roberta Calegari, Regis Riveret, and Giovanni Sartor. The burden of persuasion in structured argumentation. In Juliano Maranhão and Adam Zachary Wyner, editors, *ICAIL'21: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, ICAIL'21, pages 180–184. ACM, June 2021.

[6] Roberta Calegari and Giovanni Sartor. A model for the burden of persuasion in argumentation. In Serena Villata, Jakub Harašta, and Petr Křemen, editors, *Legal Knowledge and Information Systems. JURIX 2020: The Thirty-third Annual Conference*, volume 334 of *Frontiers in Artificial Intelligence and Applications*, pages 13–22, Brno, Czech Republic, 9-11 2020. IOS Press.

[7] Martin Caminada and Leila Amgoud. On the evaluation of argumentation formalisms. *Artificial Intelligence*, 171(5—6):286–310, 2007.

[8] Markus Kröll, Reinhard Pichler, and Stefan Woltran. On the complexity of enumerating the extensions of abstract argumentation frameworks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 1145–1152, Melbourne, Australia, 2017.

[9] Sanjay Modgil and Henry Prakken. The *ASPIC*$^{+}$ framework for structured argumentation: a tutorial. *Argument & Computation*, 5(1):31–62, 2014.

[10] Jann Müller, Anthony Hunter, and Philip Taylor. Meta-level argumentation with argument schemes. In *International Conference on Scalable Uncertainty Management*, volume 8078 of *Lecture Notes in Computer Science*, pages 92–105, Washington, DC, USA, 2013. Springer.

[11] Gideon Ogunniye, Alice Toniolo, and Nir Oren. Meta-argumentation frameworks for multi-party dialogues. In *International Conference on Principles and Practice of Multi-Agent Systems*, volume 11224 of *Lecture Notes in Computer Science*, pages 585–593, Tokyo, Japan, 2018. Springer.

[12] Giuseppe Pisano, Roberta Calegari, Andrea Omicini, and Giovanni Sartor. A mechanism for reasoning over defeasible preferences in Arg2P. In Stefania Monica and Federico Bergenti, editors, *CILC 2021 – Italian Conference on Computational Logic. Proceedings of the 36th Italian Conference on Computational Logic*, volume 3002 of *CEUR Workshop Proceedings*, pages 16–30, Parma, Italy, 7-9 September 2021.

[13] Giuseppe Pisano, Roberta Calegari, Andrea Omicini, and Giovanni Sartor. Burden of persuasion in argumentation: A meta-argumentation approach. In Marcello D'Agostino, Fabio Aurelio D'Asaro, and Costanza Larese, editors, *Advances in Argumentation in Artificial Intelligence 2021*, volume 3086 of *CEUR Workshop Proceedings*, pages 5:1–5:19, February 2022. Proceedings of the Workshop on Advances in Argumentation in Artificial Intelligence (AI^{3} 2021), co-located with the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, November 29, 2021.

[14] Henry Prakken. AI & Law, logic and argument schemes. *Argumentation*, 19(3):303–320, 2005.

[15] Henry Prakken. An abstract framework for argumentation with structured arguments. *Argument and Computation*, 1(2):93–124, 2010.

[16] Tjitze Rienstra, Alan Perotti, Serena Villata, Dov M. Gabbay, and Leendert W. N. van der Torre. Multi-sorted argumentation. In Sanjay Modgil, Nir Oren, and Francesca Toni, editors, *Theorie and Applications of Formal Argumentation – First International Workshop, TAFA 2011*, volume 7132 of *Lecture Notes in Computer Science*, pages 215–231. Springer, 2011.

[17] Gerard Vreeswijk. Abstract argumentation systems. *Artificial Intelligence*, 90(1–2):225–279, 1997.

[18] Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes.* Cambridge University Press, United Kingdom, 2008.

Received 31 May 2022

Explaining Classifiers' Outputs with Causal Models and Argumentation

Antonio Rago, Fabrizio Russo, Emanuele Albini and Francesca Toni
Department of Computing, Imperial College London, UK
{antonio, fabrizio, emanuele, ft}@imperial.ac.uk

Pietro Baroni
Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
pietro.baroni@unibs.it

Abstract

We introduce a conceptualisation for generating argumentation frameworks (AFs) from causal models for the purpose of forging explanations for models' outputs. The conceptualisation is based on reinterpreting properties of semantics of AFs as *explanation moulds*, which are means for characterising argumentative relations. We demonstrate our methodology by reinterpreting the property of *bi-variate reinforcement* in *bipolar AFs*, showing how the extracted bipolar AFs may be used as relation-based explanations for the outputs of causal models. We then evaluate our method empirically when the causal models represent (Bayesian and neural network) machine learning models for classification. The results show advantages over a popular approach from the literature, both in highlighting specific relationships between feature and classification variables and in generating counterfactual explanations with respect to a commonly used metric.

1 Introduction

The field of explainable AI (XAI) has in recent years become a major focal point of the efforts of researchers, with a wide variety of models for explanation being proposed (see [1] for an overview). More recently, incorporating a causal perspective into explanations has been explored by some, e.g. [2, 3, 4]. The link between causes and explanations has long been studied [5]; indeed, the two have even been equated under a broad sense of the concept of "cause" [6] and causal models have been advocated as “explanations or understanding of how data are generated” [7]. Furthermore, some see causal reasoning as underpinning how humans explain to one another [8]. Also, research from the social sciences [9] has indicated the value of causal links, particularly in the form of counterfactual reasoning, within explanations, and that the importance of such information surpasses that of probabilities or statistical relationships for users. Given that “looking at how humans explain to each other can serve as a useful starting point for explanation in AI” [9], it does makes sense to draw explanations for AI models from causal models. However, it is also broadly understood that different users may need different forms of explanations [10], taking into account their cognitive abilities, their background and their specific goals when seeking explanations of AI systems, and work within the social sciences clearly points to humans favouring seemingly non-causal forms of explanations in some contexts, in particular: “the majority of what might look like causal attributions turn out to look like *argumentative* claim-backings”[11], and “people use *reasons* to explain or justify decisions already taken and beliefs already held” [12].

Meanwhile, *computational argumentation* (see [13, 14] for recent overviews) has received increasing interest in recent years as a means for providing explanations of the outputs of a number of AI models, e.g. recommender systems [15], classifiers [16], Bayesian networks [17] and *PageRank* [18]. Furthermore, several works focus on the power of argumentation to provide a bridge between explained models and users, validated by user studies [19, 20]. While *argumentative explanations* are wide-ranging in their format and application (see [21, 22] for recent surveys), the links between causality and argumentative explanations have remained largely unexplored to date. In this paper, we aim to fill this gap and bring causality and argumentation together to support the XAI vision, focusing on the explanation of outputs of machine learning classifiers.

Specifically, we introduce a conceptualisation for generating *argumentation frameworks* (AFs) with any number of dialectical relations as envisaged in [23, 24], from causal models for the purpose of forging explanations for the models’ outputs. Like [25], we focus not on explaining by features, but instead by relations, hence the use of argumentation as the underpinning explanatory mechanism. After covering the most relevant work in the literature (Section 2) and giving the necessary background (Section 3), we show how properties of argumentation semantics from the literature can be reinterpreted to serve as *explanation moulds*, i.e. means for characterising argumentative relations (Section 4). Then (in Section 5) we propose a way to define explanation moulds based on inverting properties of argumentation semantics. Briefly, the idea is to detect, inside a causal model, the satisfaction of the conditions specified by some semantics property: if these conditions are satisfied by some influence in the causal model, then the influence can be assigned an explanatory role by casting it as a dialectical relation, whose type is in correspondence with the detected property. The identified dialectical relations compose, altogether, an argumentation framework. We demonstrate our methodology by reinterpreting the property of *bi-variate reinforcement* [26] from *bipolar AFs* [27] and then showing in (Section 6) how the extracted bipolar AFs may be used as counterfactual explanations for the outputs of causal models representing different classification methods. We then provide an empirical assessment of these explanations (Section 7), demonstrating how they can provide some important insights on the differences between different models' functionalities, while outperforming a popular approach from the literature along a counterfactual metric. Finally, we conclude, indicating potentially fruitful future work (Section 8).

Overall, we make the following main contributions:

- We propose a novel concept for defining relation-based explanations for causal models by inverting properties of argumentation semantics.
- We use this concept to define a novel form of *reinforcement explanation* (RX) for causal models.
- We show deployability of RXs with two machine-learning models, from which causal models are drawn.
- We evaluate our proposal empirically: although preliminary, this evaluation shows promise and indicates directions for future work.

This work extends [28, 29] significantly, with Section 7 being completely new and the other sections being extended and improved.

2 Related Work

A dominant approach for model-agnostic explainability of AI models is the use of *feature attribution* methods, which assign a *signed* value to each feature (in input) to represent their importance towards the output of a classification model, for each of the inputs. LIME [30] and SHAP [31] are popular attribution methods, using different techniques to assess each feature's importance by measuring the outcome of changes to inputs. In a nutshell, LIME is based on sampling *perturbations* of the reference input, while SHAP is based on the notion of Shapley values from game theory, assessing the effect of the presence of a feature when added to all possible sets of other features (in practice a sampling over the possible *permutations* of features is used for an approximate evaluation since the exact calculation would be too costly for large sets of features). Alternatively, another model-agnostic approach is the use of *counterfactuals*, e.g. as in [32, 33, 34], in which a modified input which would result in the change in the classification is given. In the literature, feature attribution methods have been used to generate counterfactual explanations [35], and vice versa [36]. Various studies [37, 38, 39] have have highlighted how feature attribution methods (including SHAP) are often mis-interpreted and overly trusted. In line with [9], we regard counterfactual explanations as some of the most useful for understanding model behaviour. Hence, in this work, we analyse feature attribution explanations in a counterfactual manner as a baseline against which we assess our approach, demonstrating the advantages of incorporating causal information to explanations.

The role of causality within explanations for AI models has received increasing attention of late. [2] define a framework for determining the causal effects between features and predictions using a variational autoencoder. The detection of causal relations and explanations between arguments within text has also proven effective within NLP [40]. [3] give causal explanations for neural networks (NNs) in that they train a separate NN by masking features to determine causal relations (in the original NN) from the features to the classifications. Generative causal explanations of black box classifiers [41] are built by learning the latent factors involved in a classification, which are then included in a causal model. [42] take a different approach, proposing a general framework for constructing structural causal models with deep learning components, allowing tractable counterfactual inference. Other approaches towards explaining NNs, e.g., [43, 44], take into account causal relations when calculating features' attribution values for explanation. Meanwhile, [4] introduce causal explanations for reinforcement learning models based on [5]. We take a different approach, drawing argumentative explanations from causal models.

Computational argumentation has been widely used in the literature as a mechanism for explaining AI models, from data-driven explanations of classifiers' outputs [45], powered by AA-CBR [46], to the explanation of the *PageRank* algorithm [47] via bipolar AFs [18]. The outputs of Bayesian networks have been explained by SAFs [17], while decision-making [48] and scheduling [49] have also been targeted. Property-driven explanations based on bipolar [20] and tripolar [50] AFs have been extracted for recommendations, where the properties driving the extraction are defined in the orthodox manner (with respect to the resulting frameworks), rather than inversions thereof, as we propose. Other forms of argumentation have also proven effective in providing explanations for recommender systems [15], decision making [51] and planning [52]. Our proposal in this paper adds to this line of work providing novel forms of argumentative explanations, but drawn from causal models.

Various works have explored the links between causality and argumentation. [53] shows that a propositional argumentation system in a full classical language is equivalent to a causal reasoning system, while [54] develops a formal theory combining "causal stories" and evidential arguments. Somewhat similarly to us, [55] present a method for extracting argumentative explanations for the outputs of causal models. However, their method requires more information than the causal model alone, namely, ontological links, and the argumentation supplements the rule-based explanations, rather than being the main constituent, as is the case in our approach.

Despite the clear potential of causality towards XAI, many of the approaches for generating explanations for AI models have neglected causality as a potential drive for explainability. Some of the most popular methods, as discussed earlier, are heuristic and model-agnostic [30, 31], and, although they are useful, particularly with regards to their wide-ranging applicability, they neglect *how* models are determining their outputs and therefore the underlying causes therein. This has arguably left a chasm between how explanations are provided by models at the forefront of XAI technology and what users actually require from explanations [56]. On the other hand, while causal models provide the raw material for explanation, the latter is not limited to the selection of a set of appropriate causes [57]. We aim to address these problems by delivering explanations to users which are directly driven by, but not limited to, causal models themselves.

3 Background

Our method relies upon causal models and some notions from computational argumentation. We provide core background for both.

Causal models. A *causal model* [58] is a triple $\langle U, V, E \rangle$, where:

- U is a (finite) set of *exogenous variables*, i.e. variables whose values are determined by external factors (outside the causal model);
- V is a (finite) set of *endogenous variables*, i.e. variables whose values are determined by internal factors, namely by (the values of some of the) variables in $U \cup V$;
- each variable may take any values in its associated *domain*; we refer to the domain of $W_i \in U \cup V$ as $\mathcal{D}(W_i)$;
- E is a (finite) set of *structural equations* that, for each endogenous variable $V_i \in V$, define V_i's values as a function f_{V_i} of the values of V_i' *parents* $PA(V_i) \subseteq U \cup V \setminus \{V_i\}$.

Example 1. *Let us consider a simple causal model $\langle U, V, E\rangle$ comprising $U = \{U_1, U_2\}$, $V = \{V_1, V_2\}$ and for all $W_i \in U \cup V$, $\mathcal{D}(W_i) = \{\top, \bot\}$. Figure 1i (we ignore Figure 1ii for the moment: this will be discussed later in Section 5) visualises the variables' parents, and Table 1 gives the combinations of values for the variables resulting from the structural equations E (amounting to $V_1 = U_1 \wedge \neg U_2$ and $V_2 = V_1$). This may represent a group's decision on whether or not to enter a restaurant, with variables U_1:* "margherita" is spelt correctly on the menu, not like the drink*; U_2:* there is pineapple on the pizzas*; V_1:* the pizzeria seems to be legitimately Italian*; and V_2:* the group chooses to enter the pizzeria*.*

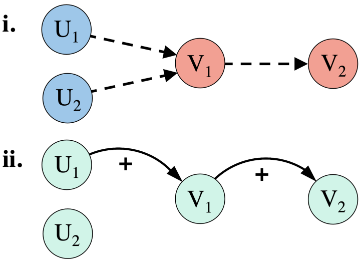


Figure 1: (i) Variables and parents for Example 1, with parents indicated by dashed arrows (for example $\{U_1, U_2\} = PA(V_1)$, i.e. U_1 and U_2 are the parents of V_1). (ii) SAF explanation (see Section 4) for the assignment to exogenous variables $\mathbf{u} \in \mathcal{U}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \top$.

U_1	U_2	V_1	V_2
$\top$ margherita	$\top$ pineapple	$\bot$ ∼Italian	$\bot$ ∼enter
$\top$ margherita	$\bot$ ∼pineapple	$\top$ Italian	$\top$ enter
$\bot$ margarita	$\top$ pineapple	$\bot$ ∼Italian	$\bot$ ∼enter
$\bot$ margarita	$\bot$ ∼pineapple	$\bot$ ∼Italian	$\bot$ ∼enter

Table 1: Combinations of values ($\top$ or $\bot$) resulting from the structural equations for Example 1. Here we also indicate the intuitive reading of the assignment of values to variables according to the illustration in Example 1 (for example, the assignment of $\top$ to U_1 may be read as *"margherita" is spelt correctly on the menu* – simply given as 'margherita' in the table, and the assignment of U_2 to $\bot$ may be read as *there is no pineapple on the pizzas* – simply given as '∼pineapple' in the table).

Given a causal model $\langle U, V, E\rangle$ where $U = \{U_1, \ldots, U_i\}$, we denote with $\mathcal{U} = \mathcal{D}(U_1) \times \ldots \times \mathcal{D}(U_i)$ the a set of all possible combinations of values of the exogenous variables (realisations). With an abuse of notation, we refer to the value of any variable $W_i \in U \cup V$ given $\mathbf{u} \in \mathcal{U}$ as $f_{W_i}[\mathbf{u}]$: if W_i is an exogenous variable, $f_{W_i}[\mathbf{u}]$ will be its assigned value in $\mathbf{u}$; if W_i is an endogenous variable, it will be the value dictated by the structural equations in the causal model.

We use the *do* operator [59] to indicate *interventions*, i.e., for any variable $V_i \in V$ and value thereof $v_i \in \mathcal{D}(V_i)$, $do(V = v_i)$ implies that the function f_{V_i} is replaced by the constant function v_i, and for any variable $U_i \in U$ and value thereof $u_i \in \mathcal{D}(U_i)$, $do(U_i = u_i)$ implies that U_i is assigned u_i.

Argumentation. In general, an *argumentation framework* (AF) is any tuple $\langle \mathcal{A}, \mathcal{R}_1, \ldots, \mathcal{R}_l \rangle$, with $\mathcal{A}$ a set (of *arguments*), $l > 0$ and $\mathcal{R}_i \subseteq \mathcal{A} \times \mathcal{A}$, for $i \in \{1, \ldots, l\}$, (binary and directed) *dialectical relations* between arguments [23, 24]. In the abstract argumentation [60] tradition, arguments in these AFs are unspecified *abstract* entities that can be instantiated differently to suit different settings of deployment. Several specific choices of dialectical relations can be made, giving rise to specific AFs instantiating the above general definition, including *abstract AFs* (AAFs) [60], with $l = 1$ (and $\mathcal{R}_1$ a dialectical relation of *attack*, referred to later as $\mathcal{R}_-$), *support AFs* (SAFs) [61], with $l = 1$ (and $\mathcal{R}_1$ a dialectical relation of *support*, referred to later as $\mathcal{R}_+$), and *bipolar AFs* (BAFs) [27], with $l = 2$ (and $\mathcal{R}_1$ and $\mathcal{R}_2$ dialectical relations of *attack* and *support*, respectively, referred to later as $\mathcal{R}_-$ and $\mathcal{R}_+$).

The meaning of AFs (including the intended dialectical role of the relations) may be given in terms of *gradual semantics* (e.g. see [24, 62] for BAFs), defined, for AFs with arguments $\mathcal{A}$, by means of mappings $\sigma : \mathcal{A} \to \mathbb{V}$, with $\mathbb{V}$ a given set of *values* of interest for evaluating arguments.

The choice of gradual semantics for AFs may be guided by *properties* that the mappings σ should satisfy (e.g. as in [26, 62]). We will utilise, in Section 5, a variant of the property of *bi-variate reinforcement* for BAFs from [26].

4 From Causal Models to Explanation Moulds and Argumentative Explanations

In this section we see the task of obtaining *explanations* for causal models' assignments of values to variables as a two-step process: first we define *moulds* characterising the core ingredients of explanations; then we use these moulds to obtain, automatically, (instances of) AFs as argumentative explanations. Moulds and explanations are defined in terms of *influences* between variables in the causal model, focusing on those from parents to children given by the causal structure underpinning the model, as follows.

Definition 1. *Let $M = \langle U, V, E\rangle$ be a causal model. The* influence graph *corresponding to M is the pair $\langle \mathcal{V}, \mathcal{I}\rangle$ with:*

- *$\mathcal{V} = U \cup V$ is the set of all (exogenous and endogenous) variables;*
- *$\mathcal{I} \subseteq \mathcal{V} \times \mathcal{V}$ is defined as $\mathcal{I} = \{(W_1, W_2)|W_1 \in PA(W_2)\}$ (referred to as the set of* influences*).*

Note that, while straightforward, the concept of influence graph (closely related to the notion of causal diagram [63]) is useful as it underpins much of what follows.

Next, the idea underlying explanation moulds is that, typically, inside the causal model, some variables affect others in a way that may not be directly understandable or even cognitively manageable by a user. The influence graph synthetically expresses which variables affect which others but does not give an account of how the influences actually occur in the context (namely, the values given to the exogenous variables) that a user may be interested in. Thus, the perspective we take is that each influence can be assigned an explanatory role, indicating how that influence is actually working in that context. The explanatory roles ascribable to influences can be regarded as a form of explanatory knowledge which is user specific: different users may be willing (and/or able) to accept explanations built using different sets of explanatory roles as they correspond to their understanding of how variables may affect each other. We assume that each explanatory role is specified by a *relation characterisation*, i.e. a Boolean logical requirement, which can be used to mould the explanations to be presented to the users by indicating which relations play a role in the explanations.

Definition 2. *Given a causal model $\langle U, V, E\rangle$ and its corresponding influence graph $\langle \mathcal{V}, \mathcal{I}\rangle$, an* explanation mould *is a non-empty set:*

$$\{c_1, \ldots, c_m\}$$

where for all $i \in \{1, \ldots, m\}$, $c_i : \mathcal{U} \times \mathcal{I} \rightarrow \{\top, \bot\}$ is a relation characterisation*, in the form of a Boolean condition expressed in some formal language. Given some $\mathbf{u} \in \mathcal{U}$ and $(W_1, W_2) \in \mathcal{I}$, if $c_i(\mathbf{u}, (W_1, W_2)) = \top$ we say that the influence (W_1, W_2)* satisfies *c_i for $\mathbf{u}$.*

Note that we are not prescribing any formal language for specifying relation characterisations, as several such languages may be suitable.

Given an assignment $\mathbf{u}$ to the exogenous variables, based on an explanation mould, we can obtain an AF including, as (different) dialectical relations, the influences satisfying the (different) relation characterisations for the given $\mathbf{u}$. Thus, the choice of relation characterisations is to a large extent dictated by the specific form of AF the intended users expect. Before defining argumentative explanations formally, we give an illustration.

Example 1 (Cont.). *Let us imagine a situation where one would like to explain the behaviour of the causal model from Figure 1i and Table 1 with a SAF (see Section 3). We thus require one single form of relation (i.e. support) to be extracted from the corresponding influence graph* $\langle \{U_1, U_2, V_1, V_2\}, \{(U_1, V_1), (U_2, V_1), (V_1, V_2)\}\rangle$*. In order to define the explanation mould for such a situation, we note that the behaviour defining this relation could be characterised as changing the state of* rejected *arguments that it supports to* accepted *when the supporting argument's state is* accepted*. In our simple causal model,* accepted *arguments may amount to variables assigned to value* $\top$ *and* rejected *arguments may amount to variables assigned to value* $\bot$*. Thus, the intended behaviour can be captured by a relation characterisation* c_s *such that, given* $\mathbf{u} \in \mathcal{U}$ *and* $(W_1, W_2) \in \mathcal{I}$*:*

$$\begin{aligned} &c_s(\mathbf{u}, (W_1, W_2)) = \top \textit{ iff} \\ &(f_{W_1}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}] = \top \wedge f_{W_2}[\mathbf{u}, do(W_1 = \bot)] = \bot) \vee \\ &(f_{W_1}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}] = \bot \wedge f_{W_2}[\mathbf{u}, do(W_1 = \top)] = \top). \end{aligned}$$

Then, for the assignment to exogenous variables $\mathbf{u} \in \mathcal{U}$ *such that* $f_{U_1}[\mathbf{u}] = \top$ *and* $f_{U_2}[\mathbf{u}] = \bot$*, we may obtain the SAF in Figure 1ii (visualised as a graph with nodes as arguments and edges indicating elements of the support relation). For illustration, consider* $(U_1, V_1) \in \mathcal{I}$ *for this* $\mathbf{u}$*. We can see from Table 1 that* $f_{V_1}[\mathbf{u}] = \top$ *and also that* $f_{V_1}[\mathbf{u}, do(U_1 = \bot)] = \bot$ *and thus from the above it is clear that* $c_s(\mathbf{u}, (U_1, V_1)) = \top$ *and thus the influence is of the type of support that* c_s *characterises. Meanwhile, consider* $(U_2, V_1) \in \mathcal{I}$ *for the same* $\mathbf{u}$*: the fact that* $f_{U_2}[\mathbf{u}] = \bot$ *and* $f_{V_1}[\mathbf{u}] = \top$ *means that* $c_s(\mathbf{u}, (U_2, V_1)) = \bot$ *and thus the influence is not cast as a support. Indeed, if we consider the first and second rows of Table 1, we can see that* U_2 *being true actually causes* V_1 *to be false, thus it is no surprise that the influence is not cast as a support and plays no role in the resulting SAF. If we wanted for this influence to play a role, we could, for example, choose to incorporate an additional relation of attack into the explanation mould, to generate instead BAFs (see Section 3) as argumentative explanations. This example thus shows how explanation moulds must be designed to fit causal models depending on external explanatory requirements dictated by users. It should be noted also that some explanation moulds may be unsuitable to some causal models, e.g. the explanation mould with the earlier* c_s *would not be directly applicable to causal models with variables with non-binary or continuous domains.*

In general, AFs serving as argumentative explanations can be generated as follows.

Definition 3. *Given a causal model $\langle U, V, E\rangle$, its corresponding influence graph $\langle \mathcal{V}, \mathcal{I}\rangle$, some $\mathbf{u} \in \mathcal{U}$ and an explanation mould $\{c_1, \ldots, c_m\}$, an* argumentative explanation *is an AF $\langle \mathcal{A}, \mathcal{R}_1, \ldots \mathcal{R}_m\rangle$, where*

- *$\mathcal{A} \subseteq \mathcal{V}$, and*
- *$\mathcal{R}_1, \ldots, \mathcal{R}_m \subseteq \mathcal{I} \cap (\mathcal{A} \times \mathcal{A})$ such that, for any $i = 1 \ldots m$, $\mathcal{R}_i = \{(W_1, W_2) \in \mathcal{I} \cap (\mathcal{A} \times \mathcal{A}) | c_i(\mathbf{u}, (W_1, W_2)) = \top\}$.*

Note that we have left open the choice of $\mathcal{A}$ (as a generic, possibly non-strict subset of $\mathcal{V}$). In practice, $\mathcal{A}$ may be the full $\mathcal{V}$, but we envisage that users may prefer to restrict attention to some variables of interest (for example, excluding variables not "involved" in any influence satisfying the relation characterisations).

Example 1 (Cont.). *The behaviour of the causal model from Figure 1i and Table 1 for $\mathbf{u}$ such that $f_{U_1}[\mathbf{u}] = \top$ and $f_{U_2}[\mathbf{u}] = \bot$, using the explanation mould $\{c_s\}$ given earlier, can be captured by either of the two SAFs (argumentative explanations) below, depending on the choice of $\mathcal{A}$:*

- *the SAF in Figure 1ii, where every variable is an argument;*
- *the SAF with the same support relation but U_2 excluded from $\mathcal{A}$, as it is not "involved" and thus does not contribute to the explanation.*

Both SAFs explain that $f_{V_1}[\mathbf{u}] = \top$ is supported by $f_{U_1}[\mathbf{u}] = \top$, in turn supporting $f_{V_2}[\mathbf{u}] = \top$. Namely, the causal model recommends that the group should enter the pizzeria because the pizzeria seems legitimately Italian, given that "margherita" is spelt correctly on the menu. Note that the pineapple not being *on the pizza could also be seen as a support towards the pizzeria being legitimately Italian, the inclusion of which could be achieved with a slightly more complex explanation mould.*

5 Inverting Properties of Argumentation Semantics: Reinforcement Explanations

The choice (number and form) of relation characterisations in explanation moulds is crucial for the generation of explanations concerning the value assignments to endogenous variables in the causal models. Even after having decided which argumentative relations to include in the AF/argumentative explanation, the definition of the relation characterisations is non-trivial, in general. In this section we demonstrate a novel concept for utilising properties of gradual semantics for AFs for the definition of relation characterisations and the consequent extraction of argumentative explanations.

The common usage of these properties in computational argumentation can be roughly equated to: *if a semantics, given an AF, satisfies some desirable properties, then the semantics is itself desirable (for the intended context, where those properties matter).* We propose a form of inversion of this notion for use in our XAI setting, namely: *if some desirable properties are identified for the gradual semantics of (still unspecified) AFs, then these properties can guide the definition of the dialectical relations underpinning the AFs.* For this inversion to work, we need to identify first and foremost a suitable notion of gradual semantics for the AFs we extract from causal models. Given that, with our AFs, we are trying to explain the results obtained from underlying causal models, we cannot impose just any gradual semantics from the literature, but need to make sure that we capture, with the chosen semantics, the behaviour of the causal model itself. This is similar, in spirit, to recent work to extract (weighted) BAFs from multi-layer perceptrons (MLPs) [64], using the underlying computation of the MLPs as a gradual semantics, and to the proposals to explain recommender systems (RSs) via tripolar AFs [50] or BAFs [20], using the underlying predicted ratings by the RSs as a gradual semantics.

A natural semantic choice for causal models, since we are trying to explain why endogenous variables are assigned specific values in their domains given assignments to the exogenous variables, is to use the assignments themselves as a gradual semantics. Then, the idea of inverting properties of semantics to obtain dialectical relations in AFs can be recast to obtain relation characterisations in explanation moulds as follows: *given an influence graph and a selected value assignment to exogenous variables, if an influence satisfies a given, desirable property, then the influence can be cast as part of a dialectical relation in the resulting AF.*

Naturally, for this inversion to be useful, we need to identify useful properties from an explanatory viewpoint. We will illustrate this concept with the property of *bi-variate reinforcement* for BAFs [26], which we posit is generally intuitive in the realm of explanations. Bi-variate reinforcement is defined when the set of values $\mathbb{V}$ for evaluating arguments is equipped with a *pre-order* $<$. Intuitively, bi-variate reinforcement states that[1] strengthening an attacker (a supporter) cannot strengthen (cannot weaken, respectively) an argument it attacks (supports, respectively), where strengthening an argument amounts to increasing its value from $v_1 \in \mathbb{V}$ to $v_2 \in \mathbb{V}$ such that $v_2 > v_1$ (whereas weakening an argument amounts to decreasing its value from such v_2 to v_1). In our formulation of this property, we require that increasing the value of variables represented as attackers (supporters) can only decrease

[1] Here, we ignore the intrinsic *basic strength* of arguments used in the formal definition in [26].

(increase, respectively) the values of variables they attack (support, respectively).

Property 1. *Given a causal model $\langle U, V, E\rangle$ such that, for each $W_i \in U \cup V$, the domain $\mathcal{D}(W_i)$ is equipped with a pre-order $<$,*[2] *and given its corresponding influence graph $\langle \mathcal{V}, \mathcal{I}\rangle$, an argumentative explanation $\langle \mathcal{A}, \mathcal{R}_-, \mathcal{R}_+\rangle$ for $\mathbf{u} \in \mathcal{U}$ satisfies* causal reinforcement *iff for any $(W_1, W_2) \in \mathcal{I}$ where $w_1 = f_{W_1}[\mathbf{u}]$, for any $w_- \in \mathcal{D}(W_1)$ such that $w_- < w_1$, and for any $w_+ \in \mathcal{D}(W_1)$ such that $w_+ > w_1$:*

- *if $(W_1, W_2) \in \mathcal{R}_-$, then $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \leq f_{W_2}[\mathbf{u}]$ and $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \geq f_{W_2}[\mathbf{u}]$;*
- *if $(W_1, W_2) \in \mathcal{R}_+$, then $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \geq f_{W_2}[\mathbf{u}]$ and $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \leq f_{W_2}[\mathbf{u}]$.*

We can then invert this property to obtain an explanation mould. In doing so, we introduce slightly stricter conditions to ensure that influencing variables that have no effect on influenced variables do not constitute both an attack and a support, a phenomenon which we believe would be counter-intuitive from an explanation viewpoint.

Definition 4. *Given a causal model $\langle U, V, E\rangle$ such that, for each $W_i \in U \cup V$, the domain $\mathcal{D}(W_i)$ is equipped with a pre-order $<$, and given its corresponding influence graph $\langle \mathcal{V}, \mathcal{I}\rangle$, a* reinforcement explanation mould *is an explanation mould $\{c_-, c_+\}$ such that, given some $\mathbf{u} \in \mathcal{U}$ and $(W_1, W_2) \in \mathcal{I}$, letting $w_1 = f_{W_1}[\mathbf{u}]$:*

- *$c_-(\mathbf{u}, (W_1, W_2)) = \top$ iff:*
  1. *$\forall w_+ \in \mathcal{D}(W_1)$ such that $w_+ > w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \leq f_{W_2}[\mathbf{u}]$;*
  2. *$\forall w_- \in \mathcal{D}(W_1)$ such that $w_- < w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \geq f_{W_2}[\mathbf{u}]$;*
  3. *$\exists_{\geq 1} w_+ \in \mathcal{D}(W_1)$ or $\exists_{\geq 1} w_- \in \mathcal{D}(W_1)$ satisfying strictly the inequality conditions in points 1 and 2 above.*
- *$c_+(\mathbf{u}, (W_1, W_2)) = \top$ iff:*
  1. *$\forall w_+ \in \mathcal{D}(W_1)$ such that $w_+ > w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_+)] \geq f_{W_2}[\mathbf{u}]$;*
  2. *$\forall w_- \in \mathcal{D}(W_1)$ such that $w_- < w_1$, it holds that $f_{W_2}[\mathbf{u}, do(W_1 = w_-)] \leq f_{W_2}[\mathbf{u}]$;*

[2] With an abuse of notation we use the same symbol for all pre-orders.

3. *$\exists_{\geq 1} w_+ \in \mathcal{D}(W_1)$ or $\exists_{\geq 1} w_- \in \mathcal{D}(W_1)$ satisfying strictly the inequality conditions in points 1 and 2 above.*

We call any argumentative explanation resulting from the explanation mould $\{c_-, c_+\}$ a reinforcement explanation *(RX).*

Note that, as for generic argumentative explanations, we do not commit in general to any choice of $\mathcal{A}$ in RXs.

Proposition 1. *Any RX satisfies causal reinforcement.*

Proof. Follows directly from the definition of Property 1 and Definition 4. □

The satisfaction of the property of causal reinforcement indicates how RXs could be used counterfactually, given that the results of changes to the variables' values on influenced variables are guaranteed. For example, if a user is looking to increase an influenced variable's value, supporters (attackers) indicate variables whose values should be increased (decreased, respectively). In the following sections, we will explore the potential of this capability when causal models provide abstractions of classifiers whose output needs explaining.

6 Reinforcement Explanations for Classification

In this section, we first instantiate causal models for two families of classifiers commonly used in the literature. We then demonstrate how RXs can be used to explain these classifiers in a counterfactual manner, supplementing their structure with weights on the relations, which allows RXs to be compared with *feature attribution* methods (see Section 2).

The two families of classifiers that we use to instantiate causal models are Bayesian network classifiers (BCs) and classifiers built from feed-forward NNs. Given some assignments to *input variables* $\mathbf{I}$ (from the variables' domains), these classifiers can be seen as determining the most likely value for *classification variables*, which, in this paper, we assume to be binary, in a given set $\mathbf{C}$. Thus, the classification task may be seen as a mapping $\mathcal{M}(\mathbf{x})$ returning, for assignment $\mathbf{x}$ to input variables, either 1 or 0 (for the classification variables in $\mathbf{C}$) depending on whether the probability exceeds a given threshold θ. We summarise the classification process in Figure 2. Note that, in the case of NNs, the probabilities may result from using, e.g., a softmax activation for the output layer. Furthermore, note that for the purposes of this paper, the underpinning details of these classifiers and how they can be obtained are irrelevant and will be ignored. In other words, we treat the classifier

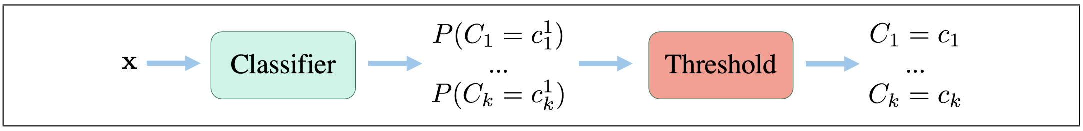


Figure 2: A schematic view of classification by BCs and NNs. We assume $\mathbf{C} = \{C_1, \ldots, C_k\}$, for $k \geq 1$, with each C_i a binary classification variable, with values c_i^1 and c_i^0, such that $P(C_i = c_i^0) = 1 - P(C_i = c_i^1)$; c_i is the value for C_i whose probability P exceeds the threshold (θ). Assuming that the threshold is suitably chosen so that c_i is uniquely defined for each C_i, the classifier can be equated to the function $\mathcal{M}$ such that $\mathcal{M}(\mathbf{x}) = (c_1 \ldots, c_k)$.

as a black-box, as standard in much of the XAI literature, and explain its outputs in terms of its inputs.

We represent the classification task by a (naive) BC or by a NN with the following causal model:

Definition 5. *A* causal model for a naive BC or classifier built from a NN *is a causal model $\langle U_C, V_C, E_C \rangle$, where:*

- *U_C consists of the input variables $\boldsymbol{I}$ of the classifier, with their respective domains;*
- *$V_C = \boldsymbol{C}$ such that, for each $C_i \in \boldsymbol{C}$, $\mathcal{D}(C_i) = \{c_i^1, c_i^0\}$;*
- *E_C corresponds to the computation of the probability values $P(C_i = c_i^1)$) by the classifier (see Figure 2).*

$\mathcal{I}_C = U_C \times V_C$ represents the influences in the causal model for the classifier; these are such that the exogenous variables U_C are densely connected to the endogenous variables V_C. In line with our assumptions for RXs, we assume that the variables' domains are equipped with a pre-order.

As discussed in Section 2, the purpose of feature attribution methods is to assign a signed importance value to each feature for a given input. Our motivation for this work is to explore an alternative direction, namely to interpret changes in outcomes with a *causal* lens and produce explanations that follow human intuition when presented to the users, while still maintaining feature attribution methods' goal of characterising the impact of each feature on a classification.

We aim to characterise (and rank) features based on their potential to change the outcome of the model. Our ingredients are: (i) The model outcome for the example to be explained in the form of a probability; (ii) A function to select the direction of change in the domain of the variable intervened; (iii) Interventions over the domain of the variables and the change in model probability resulting from them.

To arrive at the formulation for the counterfactual feature importance we propose, we introduce three functions that will help us scan features for their "counterfactual capabilities". They all refer to generic input variable U_j and classification variable C_i, for a given realisation $\mathbf{u}$ of the input variables. Note that for every variable $U_j \in U_C$, we assume that $\mathcal{D}(U_j)$ is finite and totally ordered and for each $u \in \mathcal{D}(U_j)$ we denote as $pos(u) \geq 1$ the natural number corresponding to its position in the ordering.

Potential Change in Outcome quantifies the change in probability of C_i given an intervention assigning the value u' to U_j:

$$\Delta f_{\mathbf{u},u'}^{(U_j,C_i)} = |f_{C_i}[\mathbf{u}|do(U_j = u')] - f_{C_i}[\mathbf{u}]|.$$

Relation Sign Function returns a positive or negative sign depending on the type of relation between U_j and C_i:

$$\delta(U_j, C_i, \mathbf{u}) = \begin{cases} 1 & \text{if } c_+(\mathbf{u}, (U_j, C_i)) = \top \\ -1 & \text{if } c_-(\mathbf{u}, (U_j, C_i)) = \top \\ 0 & \text{otherwise} \end{cases}$$

Domain Subset Function selects the subset of the domain of U_j to be considered to achieve a change in the classification outcome of C_i with respect to the one given by $\mathbf{u}$. The selection takes into account the threshold θ and the relation sign function δ:

$$\gamma(U_j,C_i,\mathbf{u}) = \begin{cases} \{u' \in \mathcal{D}(U_j) | u' > f_{U_j}[\mathbf{u}]\} & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) < 0 \\ \{u' \in \mathcal{D}(U_j) | u' < f_{U_j}[\mathbf{u}]\} & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) > 0 \\ \emptyset & \text{if } (f_{C_i}[\mathbf{u}] - \theta) * \delta(U_j,C_i,\mathbf{u}) = 0 \end{cases}$$

The idea is that the function γ selects the possible values of U_j which are greater than the current one in $\mathbf{u}$ in two cases: $f_{C_i}[\mathbf{u}]$ is above the threshold and U_j is an attacker; $f_{C_i}[\mathbf{u}]$ is below the threshold and U_j is a supporter. Analogously, γ selects the possible values of U_j which are lower than the current one in $\mathbf{u}$ in two cases: $f_{C_i}[\mathbf{u}]$ is above the threshold and U_j is a supporter; $f_{C_i}[\mathbf{u}]$ is below the threshold and U_j is an attacker.

On this basis, we formulate in the following our notion of counterfactual feature importance.

Counterfactual Importance ranks the input features based on the amount of change in probability that a value close to the current one can bring, provided that it produces a change in classification:

$$\omega(U_j,C_i,\mathbf{u}) = \sum_{u' \in \gamma(U_j,C_i,\mathbf{u})} \frac{\Delta f_{\mathbf{u},u'}^{(U_j,C_i)} * \mathbb{1}((\theta - f_{C_i}[\mathbf{u}|do(U_j = u')]) \cdot (\theta - f_{C_i}[\mathbf{u}]) < 0)}{|pos(u') - pos(f_{U_j}[\mathbf{u}])|} \tag{1}$$

where $\mathbb{1}()$ is the indicator function taking value 1 if the expression in brackets is true and 0 otherwise. Note also that we assume by convention that $\omega(U_j,C_i,\mathbf{u}) = 0$ when $\gamma(U_j,C_i,\mathbf{u}) = \emptyset$.

The rationale behind the formulation is as follows: The sum includes a term for every possible value u' that can be used for an intervention on U_j coherently with the expected direction of change (these values are returned by $\gamma(U_j,C_i,\mathbf{u})$). Each of these values contributes to the sum proportionally to the potential change in probability of C_i (namely $\Delta f_{\mathbf{u},u'}^{(U_j,C_i)}$) but only if it causes a change in the final classification, i.e. if the threshold is crossed in the desired direction (i.e. the difference between θ and f_{C_i} changes sign). Therefore, the indicator function filters the *"wanted"* changes and the interventions not producing a change are disregarded. Moreover, each of these terms is weighted according to the distance of u' from the current values of U_j: the greater the distance, the greater the denominator, the lower the contribution to the importance. This will improve the ranking of the variables that produce *actionable* changes, which are closest to the current input $\mathbf{u}$.

In representing classifiers as causal models and generating importance values for the relations of the resulting RXs, we are now able to directly compare RXs experimentally with feature attribution methods.

7 Experimental Evaluation

In this section we provide an empirical evaluation of our approach, focusing our evaluation on the property of causal reinforcement for RXs. The main research questions we aim to address are:

1. Can the attacks and supports in RXs be put in correspondence with positive and negative, respectively, *polarity* in feature attribution techniques?

2. Can relation importance in RXs be put in correspondence with the *magnitude* of the values associated to features in feature attribution techniques?

To answer both questions we compare RXs with a prominent feature attribution technique (i.e. SHAP [31], where, for the experiments, we use version 0.35.0 of the publicly available SHAP library). Concretely, we use SHAP in two ways: in Section 7.1 to extract *reasons for and against* classifications by classifiers (in comparison with supports and attacks in RXs); and in Section 7.2 as a way to determine reasons' importance as (the absolute values of) feature attribution values computed by SHAP (in comparison with our notion of relation importance). The sign of these feature attribution values is used to determine the sign of the reasons themselves.

In our experiments we use two publicly available datasets (FICO [65] and COMPAS [66]) and two different models, a naive BC and a NN, in line with Section 5. We implement the naive BCs using the scikit-learn implementation and the NNs using CASTLE [67]. For both datasets, there is a single, binary classification variable. We discretised continuous features using equally-sized bins limiting them to a maximum of 10 for FICO. For COMPAS, we used the existing variable domains, given that the variables are discrete (with a minimum of two values and a maximum of 17 values). Also, since Definition 4 and the definition of importance work under the assumption that variables' domains are ordered, a random ordering was generated for all variables with no inherent order. Some comments on the effect of this arbitrary ordering will be provided later. Additional details on the datasets are given in Table 2. Here, we can see how in the FICO dataset all features are continuous (and thus their domain is equipped with a natural total order) while in COMPAS 50% of the features lack an inherent order. We will show the consequences of this difference between the datasets in the results.

For each of the datasets, we trained a Naive BC and a NNs with 1 hidden layer and 32 hidden neurons. We trained the NN for a maximum of 200 epochs and with learning rate of 0.0005 and patience on the validation loss of 50 epochs. The naive BCs were fitted using Laplace estimation from the training set with $\alpha = 0.1$. Classification metrics for the two types of models on the two datasets (when trained on 75% of the samples and tested on the remaining 25%) are reported in Table 3. Note how the different models have similar performances on the same dataset. Note also that model performance optimisation was not the focus of this work and that we kept models as standard as possible.

7.1 Causal Reinforcement Analysis

In order to understand whether our RXs are able to handle different models while also unveiling differences in the way RXs operate when compared with SHAP, we measured: the prevalence of relations (i.e. the percentage of occurrences for each method) and agreement (between the two methods).

		FICO	**COMPAS**
	Number of samples	10,458	6,950
	Number of features	23	12
Size of Domain	Minimum	4	2
	Average	7	4.6
	Maximum	10	17
	Number of ordinal features	23	6
	% of ordinal features	100%	50%

Table 2: Dataset details. The number of samples for the dataset is the total. The number of features does not include the target classification variable. The size of the domains for the two datasets consist of deciles (where enough data were available) for continuous features and the original categories for categorical features. The number and % of ordinal features represents the features with a natural ordering, e.g. continuous, or with naturally ordered categories.

	FICO		**COMPAS**	
(*)	NN	NBC	NN	NBC
ROC-AUC	0.783	0.771	0.78	0.79
Accuracy	71.7%	71.9%	70.5%	71.6%
F1 Score	71.6%	71.9%	70.2%	71.5%
Precision	71.7%	71.9%	70.7%	72.6%
Recall	71.6%	71.8%	69.6%	70.5%

Table 3: Performances of the models. (*) NN (Neural Network) or NBC (Naive Bayesian network Classifier).

Prevalence of relations. We extracted RXs and SHAP explanations for all samples in the testing part of the two datasets and measured: for RXs, the percentage of influences in the causal models for the two models contributing attacks and supports, and, for SHAP, the percentage of negative and positive reasons.

The results are shown in Table 4. We note that there are large discrepancies across models and types of explanations for each of the two datasets, in contrast with similar performances by the classifiers (see Table 3). This is somewhat not surprising, as it could be a consequence of very different workings by the (very different) models to obtain classifications, and provides part of the motivation for the experiments in Section 7.2 to verify faithfulness of the explanations to the models, counterfactually. We also note that the total percentages of negative and positive attribution values established by SHAP are greater for FICO than for COMPAS, while the total percentages of influences that become part of the attack and support relations in RXs are considerably higher for NNs than NBCs independently of the dataset. This reflects the inner workings of the two models: NNs leverage the orderings over variables' domains since they assign weights that get multiplied with the value of the input variable, whose ordering (its value) has, therefore, a big influence on the final output. BCs on the other hand, mostly disregard these orderings since they calculate the probability of the classification variables for specific values of the input variables if categorical, or for a group/bucket of values, if numeric. BCs will therefore disregard ordering within the bucket, while across buckets the only link to the original ordering could come through the conditional probabilities, with a much less direct effect given that the value of the variable would be modified according to the class frequency in that band. In the case of the FICO dataset, whose continuous variables are all equipped with a natural ordering, RXs result in larger attack and support relations than for BCs, whereas in COMPAS, where some variables have been artificially and arbitrarily ordered to obtain RXs, the difference in relation size across the models is not so dramatic, somewhat confirming the expected dependence of RXs on the existence of natural orderings. Table 4 gives insight into the interactions between data, model, and RXs. For the FICO data, where all variables are numeric and hence have natural ordering, the difference between the amount of relations identified in NN and BC is much more significant than in the case of COMPAS (for FICO the difference is between 87.3% and 28.3%, while for COMPAS it is between 77.3% and 64.2%). This is to say that NN does a better job at leveraging numeric variables and shows an increased power to extract RXs that reflect the model behaviour for a given dataset, noting also that RXs need natural ordering to work at their best. NN does not support the extraction of many more relations than BCs for the COMPAS data instead, since there are not many natural orderings to leverage in the first place. Note that we do not assume that the larger the number of relations extracted the better. Instead, what we deem important is that the relationships that the model actually finds in the data are extracted for explanatory purposes. Investigation of this from different angles is provided in the following sections, highlighting how RXs are very effective at representing models that have extracted relationships from ordered variables in the data. Concerning the split between positive and negative reasons for SHAP, there seems to be a clear dominance of the former across datasets and models, but no clear pattern emerges for supports and attacks in RXs. We note though there are discrepancies in the +/- splits across the two different explanation methods, showing that they work differently and begging for further exploration of faithfulness in Section 7.2.

Agreement between RXs and SHAP. We also conducted a finer-grained analysis of the differences between the two forms of explanation, focusing on how many influences/reasons with opposite sign the two methods extract and on extracted influences/reasons versus ignored ones. Table 5 shows the results.

We note that SHAP and RXs agree less than 40% of the time for FICO and less than 20% for COMPAS. To understand this, we firstly looked at the cases where the models were establishing relations/reasons of opposite sign (Strong Disagree, i.e. + vs -) and noticed that this happened around 50% of the time for FICO NN and only 10% of the time for FICO NBC. Of course, this is a consequence of the number of extracted influences/reasons overall for this dataset, as seen in Table 4. Still, the amount of strong disagreement is quite high, but it does make intuitive sense when we think about the inner workings of the two explanation methods: for SHAP, a positive reason means that the current value of the corresponding variable is in favour of the current model output; according to Definition 4, instead, supporting variables are those whose values above the current one increase the probability of the value of the target classification variable (to be explained). In other words, our causal reinforcement definition focuses on the projection of possible changes to a variable that are guaranteed to have the expected behaviour on the target. At a general level this shows that apparently simple and superficially similar explanations elements may actually allow quite different interpretations. In our case, the generic idea of positive and negative influence can correspond to instances with significantly different meanings. Conveying the correct meaning to the users is obviously a crucial and nontrivial issue in this respect. Since we are assuming a context where users ascribe a counterfactual meaning to explanations, this observation brought us to the set of experiments in the next section, where we analyse the usefulness of Definition 4 for counterfactual purposes.

7.2 Causal Reinforcement for Counterfactuals

The second set of experiments assesses how we can apply Definition 4 to extract intuitive and actionable counterfactual behaviour from our models. One method for providing such an assessment is to compare with attribution methods functioning as counterfactual explanation methods, e.g. as in [68], a set-up which we use, along with the importance measure defined in Section 6. In doing so we evaluate the counterfactual nature of our explanations (see the relevant discussion in Section 2).

Again, we consider the same models and datasets, in comparison to SHAP, but this time we focus on applying interventions to input variables and observing the change in the models' outputs (or classification). To do this we couple Definition 4 with the counterfactual feature importance ω from (1) of how important the relations

		FICO		COMPAS	
		NN	NBC	NN	NBC
	−	30.3%	22.2%	21.9%	22.3%
SHAP	+	46.1%	66.1%	40.6%	40.3%
	Total	**76.4%**	**88.2%**	**62.5%**	**62.5%**
	−	43.8%	10.9%	47.5%	40.6%
RXs	+	43.5%	17.4%	29.7%	23.6%
	Total	**87.3%**	**28.3%**	**77.3%**	**64.2%**

Table 4: Prevalence of relations. Here + and − indicate, respectively, support and attack relations in RXs and positive and negative attribution values in SHAP. Totals do not sum up to 100% given that there can be influences/features that the methods do not extract.

	FICO		COMPAS	
RXs vs SHAP	NN	NBC	NN	NBC
Strong Disagree	52.9%	10.1%	30.2%	21.1%
Weak Disagree	47.1%	89.9%	69.8%	78.9%
Disagree	**60.2%**	**69.8%**	**84%**	**92.8%**
Agree	39.8%	30.2%	16%	7.2%

Table 5: Relation Agreement Summary. The 'Strong Disagree' row looks at influences/reasons that both RXs and SHAP extract, but with opposite signs (+ vs -, as per caption of Table 4). The 'Weak Disagree' row looks at the influences/reasons that one method extracts while the other does not (+ or - vs influences/reasons not extracted). 'Strong' and 'Weak Disagree' sum up to 100% and split the total of disagreements shown in the 'Disagree' row, while the 'Agree' row gives the remainders, i.e. the extracted influences/reasons with the same sign across explanation methods.

established by the models are. Concretely, we used the absolute value of $\omega(U_j,C_i,\mathbf{u})$ to select the input variables U_j to change in order to achieve a change in classification C_i (counterfactual output).

Definition 4 is useful in selecting the direction of change, given the current classification. Given that all input variables have categorical domains in this setting (after discretisation), we had to choose how many steps to move away from the current value u of U_j. We focused first on the most actionable change recommendations that the receiver of a model decision and explanation could want. Hence we analysed the change in classification for setting u' one step away from its current value $f_{U_j}[\mathbf{u}]$ (i.e. $|pos(u') - pos(f_{U_j}[\mathbf{u}])| = 1$). We did the same for SHAP. For both methods, we changed the sets of the most important features, increasing their size from 1 to 5 (Top $U_j = 1,\ldots,5$, respectively) according to either SHAP or RXs.

The results are presented in Figure 3. In the FICO-NN setting, where the monotonic relationships are strong and well captured by the model, RXs perform well and outperform SHAP in all scenarios. In particular, a higher number of classification changes is achieved when allowing a greater number of variables to be changed while this does not happen in the case of SHAP. For FICO-NBC the situation is less clear-cut, though it can be observed that RXs do better than SHAP in the case where only one step away from the current value is allowed. It can be argued that this case is the most actionable and therefore relevant counterfactually. For COMPAS, RXs perform worse than SHAP in most cases. This again is expected given the mix of purely categorical and ordinal features in the data as well as the lower average number of categories. For the not naturally ordered variables we had to enforce a random ordering for the purposes of this tests, and this has evidently had an impact.

8 Conclusions & Future Work

We have introduced a novel approach for extracting AFs from causal models in order to explain the latter's outputs. We have shown how explanation moulds can be defined for particular explanatory requirements in order to generate argumentative explanations. We focused, in particular, on inverting the existing property of argumentation semantics of bi-variate reinforcement to create an explanation mould, before demonstrating how the resulting *reinforcement explanations* (RXs) can be used to explain causal models representing different machine-learning-based classifiers. We then performed an empirical evaluation of RXs, analysing the differences between the relations in RXs and the reasons for and against classification produced by the popular SHAP method [31]. We also introduced a preliminary measure of importance over the relations in RXs and used it to assess the counterfactuality of RXs. A deeper investigation on the notion of importance at a general level and the study of further, possibly more appropriate, definitions of this measure represent an important direction of future work.

Our preliminary empirical evaluation suggests that our approach outperforms SHAP in the cases where the conditions for its applicability are satisfied, and provides the basis for discussing the suitability of different approaches in different contexts. Our results also highlight the need for different explanation mechanisms depending on the users' needs. For instance, actionable explanations, concerning

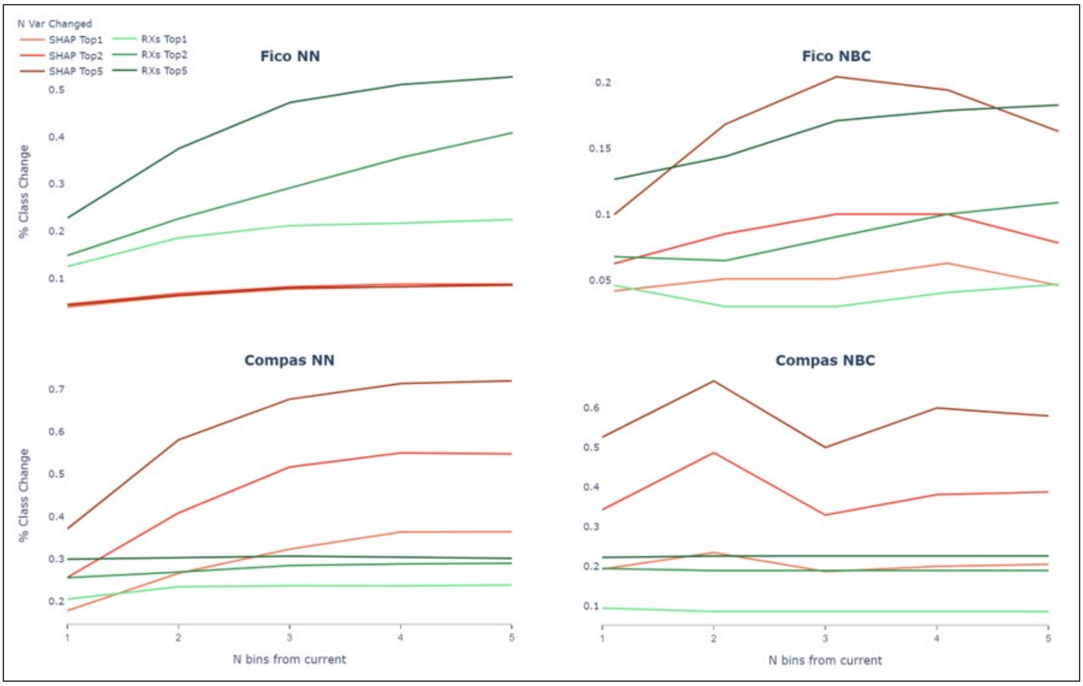


Figure 3: Proportion of successful counterfactual classification changes achieved by number of input variables changed (Top 1 to 5). The x axis represents the number of bins away from the current value i.e. distance from current position $(|pos(u') - pos(f_{U_j}[\mathbf{u}])|)$ for each changed input variable U_j. The different shades of green are for changing the one, two and five most important variables for RXs, while the reds are for SHAP.

how to change the input of a model to get a different output, may not fit feature attribution techniques, and, in general, a one-size-fits-all approach to explanations cannot achieve this.

One of the most promising aspects of our work is the vast array of directions for future work it suggests. Clearly, the wide-ranging applicability of causal models broadens the scope of explanation moulds and argumentative explanations well beyond machine learning models, and we plan to undertake an investigation into other contexts in which they may be useful, for example for decision support in healthcare.

We also plan to study inversions of different properties of argumentation semantics and different forms of AFs to understand their potential, e.g. *counting* for AAFs [69]. Within the context of explaining machine learning models, we plan to assess RXs' suitability for different data structures and different classifiers, considering in particular deeper explanations, e.g. including influences amongst input variables and/or intermediate, in addition to input and output, variables, in the spirit of [70, 25]. This may be aided by the deployment of methods for the extraction of more sophisticated causal models from classifiers, e.g., [67] for NNs.

Finally, while we posit that, when properly defined, the meaning and explanatory role of the dialectical relations can be rather intuitive at a general level, providing effective explanations to users through AFs will require the investigation of proper presentation and visualization methods, possibly tailored to users' competences and goals and to different application domains.

Acknowledgments

Toni was partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934). Further, Russo was supported by UK Research and Innovation (grant number EP/S023356/1), in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). Baroni was partially supported by the research project GNCS-INdAM CUP_E55F22000270001. Finally, Rago and Toni were partially funded by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors listed, and may differ, in particular, from the views and opinions expressed by J.P. Morgan or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction.

References

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Computing Surveys 51 (5) (2019) 93:1–93:42.

[2] D. Alvarez-Melis, T. S. Jaakkola, A causal framework for explaining the predictions of black-box sequence-to-sequence models, in: Proc. EMNLP, 2017, pp. 412–421.

[3] P. Schwab, W. Karlen, CXPlain: Causal explanations for model interpretation under uncertainty, in: Proc. NeurIPS, 2019, pp. 10220–10230.

[4] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: Proc. AAAI, 2020, pp. 2493–2500.

[5] J. Y. Halpern, J. Pearl, Causes and explanations: A structural-model approach: Part 1: Causes, in: UAI, 2001, pp. 194–202.

[6] J. Woodward, Explanation, invariance, and intervention, Philosophy of Science 64 (1997) S26–S41.

[7] J. Pearl, Causality, Cambridge university press, 2009.

[8] M. M. A. de Graaf, B. F. Malle, How people explain action (and autonomous intelligent systems should too), in: Proc. AAAI Fall Symposia, 2017, pp. 19–26.

[9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[10] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, Y. Zhang, AI explainability 360: Impact and design, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 12651–12657.
URL https://ojs.aaai.org/index.php/AAAI/article/view/21540

[11] C. Antaki, I. Leudar, Explaining in conversation: Towards an argument model, Europ. J. of Social Psychology 22 (1992) 181–194.

[12] H. Mercier, D. Sperber, The Enigma of Reason – A New Theory of Human Understanding, Penguin Books, 2018.

[13] K. Atkinson, P. Baroni, M. Giacomin, A. Hunter, H. Prakken, C. Reed, G. R. Simari, M. Thimm, S. Villata, Towards artificial argumentation, AI Magazine 38 (3) (2017) 25–36.

[14] P. Baroni, D. Gabbay, M. Giacomin, L. van der Torre (Eds.), Handbook of Formal Argumentation, College Publications, 2018.

[15] J. C. Teze, L. Godo, G. R. Simari, An argumentative recommendation approach based on contextual aspects, in: Proc. SUM, 2018, pp. 405–412.

[16] A. Dejl, P. He, P. Mangal, H. Mohsin, B. Surdu, E. Voinea, E. Albini, P. Lertvittayakumjorn, A. Rago, F. Toni, Argflow: A toolkit for deep argumentative explanations for neural networks, in: Proc. AAMAS, 2021, pp. 1761–1763.

[17] S. T. Timmer, J. C. Meyer, H. Prakken, S. Renooij, B. Verheij, Explaining Bayesian networks using argumentation, in: Proc. ECSQARU, 2015, pp. 83–92.

[18] E. Albini, P. Baroni, A. Rago, F. Toni, PageRank as an argumentation semantics, in: Proc. COMMA, 2020, pp. 55–66.

[19] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: Proc. AAMAS, 2019, pp. 1033–1041.

[20] A. Rago, O. Cocarascu, C. Bechlivanidis, F. Toni, Argumentation as a framework for interactive explanations for recommendations, in: Proc. KR, 2020, pp. 805–815.

[21] K. Cyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative XAI: A survey, in:

Proc. IJCAI, 2021, pp. 4392–4399.

[22] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and Explainable Artificial Intelligence: A Survey, Knowledge Eng. Rev. 36 (2) (2021).

[23] D. M. Gabbay, Logical foundations for bipolar and tripolar argumentation networks: preliminary results, J. Log. Comput. 26 (1) (2016) 247–292.

[24] P. Baroni, G. Comini, A. Rago, F. Toni, Abstract games of argumentation strategy and game-theoretical argument strength, in: Proc. PRIMA, 2017, pp. 403–419.

[25] E. Albini, A. Rago, P. Baroni, F. Toni, Relation-based counterfactual explanations for bayesian network classifiers, in: Proc. IJCAI, 2020, pp. 451–457.

[26] L. Amgoud, J. Ben-Naim, Weighted bipolar argumentation graphs: Axioms and semantics, in: Proc. IJCAI, 2018, pp. 5194–5198.

[27] C. Cayrol, M.-C. Lagasquie-Schiex, On the acceptability of arguments in bipolar argumentation frameworks, in: Proc. ECSQARU, 2005, pp. 378–389.

[28] A. Rago, F. Russo, E. Albini, P. Baroni, F. Toni, Forging argumentative explanations from causal models, in: Proceedings of the 5th Workshop on Advances in Argumentation in Artificial Intelligence 2021 co-located with the 20th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2021), Milan, Italy, November 29th, 2021, 2021.
URL http://ceur-ws.org/Vol-3086/paper3.pdf

[29] A. Rago, P. Baroni, F. Toni, Explaining causal models with argumentation: the case of bi-variate reinforcement, in: Proceedings of the 19th International Conference on Principles of Knowledge Representation and Reasoning, KR 2022, Haifa, Israel. July 31 - August 5, 2022, 2022.
URL https://proceedings.kr.org/2022/52/

[30] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proc. ACM SIGKDD, 2016, pp. 1135–1144.

[31] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: Proc. NeurIPS, 2017, pp. 4765–4774.

[32] S. Wachter, B. D. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, CoRR abs/1711.00399 (2017). `arXiv:1711.00399`.
URL http://arxiv.org/abs/1711.00399

[33] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, 2020, pp. 607–617. `doi:10.1145/3351095.3372850`.
URL https://doi.org/10.1145/3351095.3372850

[34] K. Kanamori, T. Takagi, K. Kobayashi, Y. Ike, K. Uemura, H. Arimura, Ordered counterfactual explanation by mixed-integer linear optimization, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on

Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, 2021, pp. 11564–11574.
URL https://ojs.aaai.org/index.php/AAAI/article/view/17376

[35] Y. Ramon, D. Martens, F. J. Provost, T. Evgeniou, A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, LIME-C and SHAP-C, Adv. Data Anal. Classif. 14 (4) (2020) 801–819. doi:10.1007/s11634-020-00418-3.
URL https://doi.org/10.1007/s11634-020-00418-3

[36] E. Albini, J. Long, D. Dervovic, D. Magazzeni, Counterfactual Shapley additive explanations, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, 2022, pp. 1054–1070. doi:10.1145/3531146.3533168.
URL https://doi.org/10.1145/3531146.3533168

[37] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, S. A. Friedler, Problems with Shapley-value-based explanations as feature importance measures, in: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 5491–5500.
URL http://proceedings.mlr.press/v119/kumar20e.html

[38] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. M. Wallach, J. W. Vaughan, Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning, in: R. Bernhaupt, F. F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjøn, S. Zhao, B. P. Samson, R. Kocielnik (Eds.), CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020, ACM, 2020, pp. 1–14. doi:10.1145/3313831.3376219.
URL https://doi.org/10.1145/3313831.3376219

[39] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment, in: M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, G. Zanfir-Fortuna (Eds.), FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020, ACM, 2020, pp. 648–657. doi:10.1145/3351095.3375624.
URL https://doi.org/10.1145/3351095.3375624

[40] Y. Son, N. Bayas, H. A. Schwartz, Causal explanation analysis on social media, in: Proc. EMNLP, 2018, pp. 3350–3359.

[41] M. R. O'Shaughnessy, G. Canal, M. Connor, C. Rozell, M. A. Davenport, Generative causal explanations of black-box classifiers, in: Proc. NeurIPS, 2020.

[42] N. Pawlowski, D. C. de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, in: Proc. NeurIPS, 2020.

[43] A. Chattopadhyay, P. Manupriya, A. Sarkar, V. N. Balasubramanian, Neural network attributions: A causal perspective, in: Proc. ICML, 2019, pp. 981–990.

[44] T. Heskes, E. Sijben, I. G. Bucur, T. Claassen, Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models, in: Proc. NeurIPS,

2020.

[45] O. Cocarascu, A. Stylianou, K. Cyras, F. Toni, Data-empowered argumentation for dialectically explainable predictions, in: Proc. ECAI, 2020, pp. 2449–2456.

[46] K. Cyras, K. Satoh, F. Toni, Abstract argumentation for case-based reasoning, in: Proc. KR, 2016, pp. 549–552.

[47] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, WWW: Internet and Web Inf. Syst. 54 (1999-66) (1998) 1–17.

[48] L. Amgoud, H. Prade, Using arguments for making and explaining decisions, Artificial Intelligence 173 (3-4) (2009) 413–436.

[49] K. Cyras, D. Letsios, R. Misener, F. Toni, Argumentation for explainable scheduling, in: Proc. AAAI, 2019, pp. 2752–2759.

[50] A. Rago, O. Cocarascu, F. Toni, Argumentation-based recommendations: Fantastic explanations and how to find them, in: Proc. IJCAI, 2018, pp. 1949–1955.

[51] Q. Zhong, X. Fan, X. Luo, F. Toni, An explainable multi-attribute decision model based on argumentation, Exp. Syst. Appl. 117 (2019) 42–61.

[52] N. Oren, K. van Deemter, W. W. Vasconcelos, Argument-based plan explanation, in: Knowledge Engineering Tools and Techniques for AI Planning, Springer, 2020, pp. 173–188.

[53] A. Bochman, Propositional argumentation and causal reasoning, in: Proc. IJCAI, 2005, pp. 388–393.

[54] F. Bex, An integrated theory of causal stories and evidential arguments, in: Proc. ICAIL, 2015, pp. 13–22.

[55] P. Besnard, M. Cordier, Y. Moinard, Arguments using ontological and causal knowledge, in: Proc. FoIKS, 2014, pp. 79–96.

[56] A. Ignatiev, Towards trustable explainable AI, in: Proc. IJCAI, 2020, pp. 5154–5158.

[57] M. J. Nathan, Causation vs. causal explanation: Which is more fundamental?, Foundations of Science (2020). `doi:https://doi.org/10.1007/s10699-020-09672-2`.

[58] J. Pearl, Reasoning with cause and effect, in: Proc. IJCAI, 1999, pp. 1437–1449.

[59] J. Pearl, The do-calculus revisited, in: Proc. UAI, 2012, pp. 3–11.

[60] P. M. Dung, On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and n-Person Games, Artificial Intelligence 77 (2) (1995) 321–358.

[61] L. Amgoud, J. Ben-Naim, Evaluation of arguments from support relations: Axioms and semantics, in: Proc. IJCAI, 2016, pp. 900–906.

[62] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: Proc. AAAI, 2018, pp. 1736–1743.

[63] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (4) (1995) 669–710.

[64] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: Proc. AAAI, 2021, pp. 6463–6470.

[65] FICO, Fico xml challenge found at community.fico.com/s/xml (2017).

URL https://community.fico.com/s/explainable-machine-learning-challenge

[66] D. S. ProPublica, Compas recidivism risk score data and analysis (2016).
URL https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analys

[67] T. Kyono, Y. Zhang, M. van der Schaar, CASTLE: regularization via auxiliary causal graph discovery, in: Proc. NeurIPS, 2020.

[68] A. White, A. d'Avila Garcez, Measurable counterfactual local explanations for any classifier, in: Proc. ECAI, 2020, pp. 2529–2535.

[69] L. Amgoud, J. Ben-Naim, Axiomatic foundations of acceptability semantics, in: Proc. KR, 2016, pp. 2–11.

[70] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, A. Mordvintsev, The building blocks of interpretability, Distill 3 (3) (2018) e10.

Received 26 May 2022

The Logic of the Arguer. Representing Natural Argumentative Discourse in Adpositional Argumentation

Marco Benini
University of Insubria[*]
marco.benini@uninsubria.it

Federico Gobbo
University of Amsterdam[†]
f.gobbo@uva.nl

Jean H.M. Wagemans
University of Amsterdam[‡]
j.h.m.wagemans@uva.nl

Abstract

In this paper, we show how to represent natural argumentative discourse through Adpositional Argumentation, a uniform framework for expressing linguistic and pragmatic aspects of such discourse on various levels of abstraction. Starting from representing the utterer and the utterance, we expand to claims and minimal arguments, finally focusing on complex argumentation in three different structures: convergent (many premises), divergent (many conclusions), and serial (an argument whose premise is the conclusion of another argument). An innovative feature of the framework is that it enables the analyst to provide a granular description of natural argumentative discourse, thus letting the logic of the arguer dynamically unfold while the discourse is presented without enforcing any particular interpretation.

We wish to thank two anonymous reviewers for their insightful remarks and helpful suggestions to improve the paper. Additionally, Federico Gobbo wants to thank the students of the course in Constructive Adpositional Grammars at the Faculty of Cognitive Science and Psychology at the AMU Poznań who, in the months Oct-Nov 2022, were challenged to annotate some of the examples presented here using linguistic and argumentative adtrees.

[*] DISAT, via Valleggio 11, IT-22100 Como, Italy

[†] ACLC, Spuistraat 134, NL-1012 VB Amsterdam, The Netherlands

[‡] ACLC, Spuistraat 134, NL-1012 VB Amsterdam, The Netherlands

Vol. 10 No. 3 2023
Journal of Applied Logics — IfCoLog Journal of Logics and their Applications

1 Introduction

Natural argumentative discourse can be defined as a piece of natural language resulting from someone's effort to convince an interlocutor or audience of the acceptability of a particular point of view. As the the name indicates, the main feature of such discourse is the presence of argumentation as a means to establish or increase that acceptability within the context of a disagreement—for a short overview of definitions of argument(ation) see Wagemans (2019) [30].

Disagreements may arise within a great many different contexts, and the characteristics of a concrete piece of natural argumentative discourse usually vary with the specific settings or the sub-genre within which it is produced, e.g., a court case, a scientific paper, or a conversation in the pub. Since the 1950s, scholars in the field of Argumentation Theory (AT) have studied a great many such sub-genres of argumentative discourse, describing the rules and conventions that govern the exchange of arguments within each specific setting. Making use of concepts, theories, and models from the longstanding traditions of logic, dialectic, and rhetoric, they have developed a rich set of insights on the constituents of various types of arguments, the micro and macro structure of different sub-genres of argumentative discourse, as well as the stylistic features thereof. In combination with normative standards regarding the validity, reasonableness, and effectiveness of argumentation, these insights are used for providing theoretically informed analyses and evaluations of argumentative texts and discussions—for a comprehensive survey of historical backgrounds, approaches, and applications see van Eemeren et al. (2014) [27]; for a concise overview of the philosophy of argument see Wagemans (2021b) [31].

Approaching the subject from a different angle, Computational Argumentation (CA) developed since the 1990s from a branch of Artificial Intelligence into an independent field of research. So far, researchers in CA have developed various computational models of argument that are used, for example, in developing tools for argument mapping, argument mining, and computer-aided human decision-making—for an overview of the development of the discipline, see Bench-Capon and Dunne (2017) [3]; for a representative collection of recent work, see Modgil et al. (2019) [20] and Prakken et al. (2020) [23].

Until now, there is hardly any interaction between the fields of AT and CA. Their quiet coexistence is reflected in the fact that researchers in CA have only used a small part of the plethora of insights developed by researchers in AT, while the latter shy away from abstract models and formal tools as such. A possible reason for this lack of interaction is the methodological distance between the humanities and the sciences: since the insights developed within AT, although profound and detailed, are expressed in a rather informal way, they are not easily transferred in models suitable for computation. And since these models are abstract and formal, they are difficult to apply by researchers in the humanities. As a result, many insights potentially useful for researching natural argumentative discourse are ignored or misunderstood.

A second observation we make is that the development of tools and models of natural argumentative discourse requires a formalization of linguistic material, which implies that *part of the information is lost.* While such a loss of information is not necessarily or always a problem, the requirement of formalization as such does create the challenge of finding the right balance between, on the one hand, the level of linguistic detail to be incorporated in the tool or model and, on the other hand, its robustness from a formal point of view.

While we acknowledge that it is not always possible for discourse expressed in natural language to be completely unambiguous, we firmly believe that increasing the level of detail in the formalization can drastically reduce the possible sources of disagreement about the interpretation. To identify the interpretative issues in the text as precisely as possible, it needs to be formalized as rigorously as possible without resulting in a loss of relevant details or a decrease in the richness of the information that can be represented. After all, in natural argumentative discourse, it is not uncommon to refer to arguments previously stated, or parts of them, to enhance the cohesion of the whole argumentation. It is, therefore, essential for the analyst at any stage to be able to represent detailed information in case it is ever needed in subsequent stages of the analysis.

Against this background, Adpositional Argumentation has been developed as a comprehensive framework for representing interpretations of natural argumentative discourse. Adpositional Argumentation is rigorous in its formalism and directly based on the linguistic material expressed in the discourse. Each level of abstraction is clearly stated, so that part of the information may be hidden without running the risk of being lost. Table 1 offers an overview of the levels of abstraction represented in Adpositional Argumentation, which will be illustrated in the following sections of this paper.

Current approaches in AT only seem suitable for formalization at the the surface level of the argumentation. Walton et al. (2008) [33], for instance, conceive an 'argumentation scheme' as consisting of a set of statements (a conclusion and one or more premises), occasionally formalizing elements within these statements (such as 'A' for authority or 'C' for consequence). The widely used model by Toulmin (1958) [25], to mention another example, contains a claim, datum, warrant, backing, rebuttal, and qualifier, which are connected in a specific way. However, both Walton's and Toulmin's conceptualizations of argumentation do not give any cue on how to analyze the internal structure of each element functioning in the argumentation. Except for

symbol	*level of abstraction*	*domain of reference*	*main references*
λ, Ω, ω	argumentation structure	pragmatics	this paper
$\alpha, \beta, \gamma, \delta$	argument form and type	pragmatics	Wagemans (2019) [30] Gobbo et al. (2019) [12]
σ, π	statements in arguments	pragmatics	Wagemans (2016) [28] Gobbo et al. (2019) [12]
ρ, ξ, φ	voice and utterance	pragmatics	Gobbo et al. (2022) [16]
ϵ	valency structure	syntax (& semantics)	Gobbo & Benini (2013,2011) [11, 10]
μ	word structure	morphology (& semantics)	Gobbo & Benini (2011) [10]

Table 1: Overview of the levels of abstraction in Adpositional Argumentation

Toulmin's qualifier, there is no explicit representation of the linguistic constituents of an argumentation, neither on the morphosyntactic nor the semantic level. To represent the relevant information contained in natural argumentative discourse, we need a deeper level of formalization of the linguistic material and the argumentative fabric.

In CA, rigorous formalizations such as those based on Dung's (2005) [6] notion of argumentation frameworks abstract away from the information contained inside an argument, such as the the distinction between conclusions and premises, as well as from the linguistic material used to express it—for the state-of-the-art on that field, see at least Baroni, Toni, and Verheij, 2020) [2]. In other formal approaches, minimal arguments are often treated as atoms, i.e., they cannot be broken down to analyze specific linguistic details or modes of expression. Inference Anchoring Theory (IAT), for instance, works on the level of illocutions and provides information on the speaker, speech act, and propositional content. However, it does not enable the analyst to label more fine-grained discourse elements, such as subjects and predicates of propositions or the voice entity and the voice predication—see, e.g., Budzynska et al. (2016) [4].

The above-mentioned problems of insufficient formalization of relevant insights, on the side of AT, and loss of information, on the side of CA, are especially salient because natural argumentative discourse, like any other communication expressed in natural language, may be interpreted in many ways. This does not only apply to the interpretations provided by different audience or readers, but also for those provided by different analysts of the same piece of discourse. Now, sometimes, disagreements about the interpretation of the discourse are easily solved, for instance, because there happens to be a misunderstanding on the part of one of the interpreters that, once pointed out, is immediately labeled as such. There are, however, also disagreements that need to be solved by discussing specific aspects of the interpretation or even its methodology. In this case, it helps if the analysts can justify their reconstruction of the discourse in a detailed and unambiguous way.

The specific aim of this paper is to illustrate how Adpositional Argumentation can provide a representation of complex argumentation and to discuss how such a representation provides insights into the logic of the arguer, which dynamically unfolds while the discourse is presented. To this end, Section 2 is an introduction to the fundamental notion of 'adpositional tree' ('adtree'). We outline its background in the Philosophy of Information and explain its general structure, before delving into the levels of abstraction introduced in Table 1. We start from morphology and syntax, used to represent linguistic information in natural language. Then, in Section 3, we turn to the pragmatic levels of abstraction, from the utterance to argumentation. We explain the basic notions of voice and utterance and differentiate between the representation of explanation and argumentation. In Section 4, we zoom in on the representation of individual arguments, using the argument categorization framework of the Periodic Table of Arguments (PTA) to represent their essential characteristics. We then move from the level of abstraction of individual arguments to the level of complex argumentation structures. Section 5 shows how to formalize the notions of convergent, divergent, and serial argumentation and represent them in adpositional trees. Finally, in Section 6, we reflect on how the analyst can provide an interpretation of the logic of the arguer based on the representation of the linguistic and pragmatic information contained in natural argumentative discourse.

2 Abstract and linguistic adpositional trees

2.1 Abstract adpositional trees

Within the Philosophy of Information, Floridi (2011) [7] defines *observables* as data with a structure imposed on them. In this way, data can be treated as variables, on different levels of abstraction. Data do not speak per se: a structure is needed to pass from the level of data to that of information. Once information is established, the analysts can give their respective interpretations. If these interpretations are directly accepted by the counterpart, we are in the realm of explanation; otherwise, if one part has to convince the counterpart of the acceptability and validity of the interpretation, we are in the realm of argumentation. In the latter case, acceptance—if it happens—comes only after the counterpart has been convinced. Richer information, that is, a more granular and refined structure imposed on the data, minimizes the risk of misunderstanding between the parts involved, as their interpretations share the same foundations as explicitly as possible.

Within Adpositional Argumentation, the main tool to represent observable linguistic material—and the argumentative information they carry on—is the adpositional tree (adtree). In general, trees are the obvious way to represent hierarchical information about natural language, especially in the field of syntax: they are less liberal than graphs and more human-readable than linear formulas in capturing the deep structure underlying word order, called by Chomsky (1965) 'surface structure' [5]. However, there is no general agreement on the optimal form of trees to represent information that is grounded in natural language material, depending on the grammatical theory adopted—for a recent overview, see Müller (2020) [21].

Adtrees keep the general standpoint that recursion is possible; however, putting all information explicitly can be inconvenient for the analyst. For example, if the focus is on the pragmatic levels of abstraction, such as utterances and argumentation, as illustrated in Figure 1, triangles (△) in leaves may hide morphological and syntactic information.

In its minimal form, the adtree represents two elements and their relationship, with one element 'ruling' the other. As pictured in Figure 1, conventionally, the ruler is called 'governor' (*gov*) and it is put on the leaf on the right side; conversely, the leaf on the left side hosts the ruled element, which is called 'dependent' (*dep*). Their connecting relation is represented as an adposition (*adp*), i.e., something that stays in-between: it can be a linguistic preposition, a conjunction, or an argument type, depending on the level of analysis.

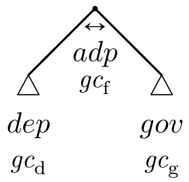


Figure 1: The standard abstract adtree

It is important to distinguish the observable linguistic material elements by their function, as natural language is ambiguous and the the same element may have different functions depending on the context. For this reason, adpositions, governors, and dependents are equally labeled by 'grammar characters' (gc). The word 'grammar' here goes well beyond the linguistic denotation as it means a set of rules for transforming the functions respectively of the tree or sub-tree indicated by the character. The grammar character of the adposition is the final result (f) of the interaction between the grammar characters of the governor (g) and the dependent (d). Finally, the small arrow of the adposition indicates information prominence, i.e., whether the governor is more prominent than the dependent, or vice versa. Prominence is a level of abstraction which is different from the asymmetrical relation between the governor and the dependent. In an adtree, information prominence goes from the most prominent to the least prominent, regardless of the relation between governor and dependent—for a comprehensive explanation, see Gobbo and Benini (2011) [10]. As will become apparent in the following sections, the actual values of characters and prominence depend on the concrete type of observables represented in the adtree.

Adtrees were introduced initially to give an account of linguistic information of written natural language material, mainly morphological and syntactic. However, they were also used to express information on different levels of abstraction in pragmatics, such as Searle's speech act taxonomy—see Gobbo and Benini (2011) [10] for details. The latter includes argumentation, which is the focus of the representation framework of Adpositional Argumentation and this paper in particular.

2.2 From abstract to linguistic information

In the present context, the data are the linguistic material contained in the piece of natural argumentative discourse to analyze, while their information is represented in the form of adpositional trees. For instance, a grammar character in a linguistic adtree may indicate the part-of-speech, such as a verb (I) or a noun (O), while one in an argumentative adtree may indicate the type of statement expressed in a conclusion or a premise of an argument, for example, a statement of fact (F) or a statement of value (V).

Adtrees distinguish between the governor-dependent relation and the direction of information prominence. Figure 2 shows a linguistic example by contrast: a hypothetical person A. is evaluated in her or his possibility to pay the bill; if people consider A. rich, the fact that A. rich is more prominent (left); vice versa, it will be A.'s possibility to pay to be more prominent (right). This distinction is evident from the choice of the linguistic adpositions 'and' and 'but' respectively.

On the linguistic level, the distinction between the governor-dependent relation and information prominence may be under-specified, such as in the genitive case in Latin and Greek. In particular, genitives may sometimes be interpreted both subjectively and objectively. A standard example is the Latin nominal syntagm

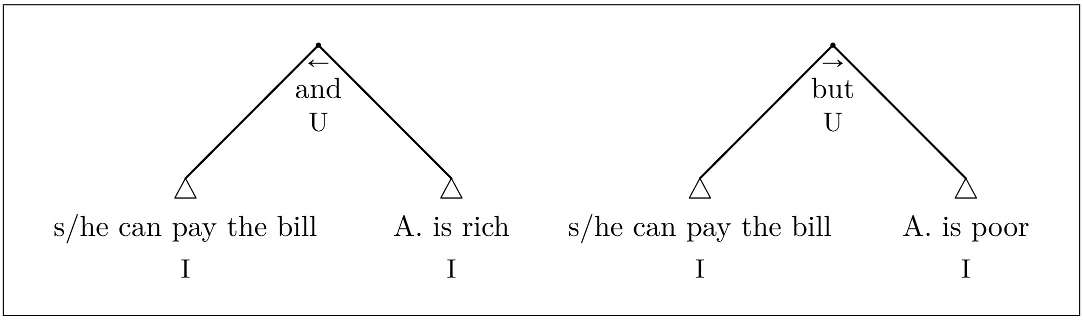


Figure 2: Example of information prominence in contrast

amor matris (mother's love), in Figure 3.

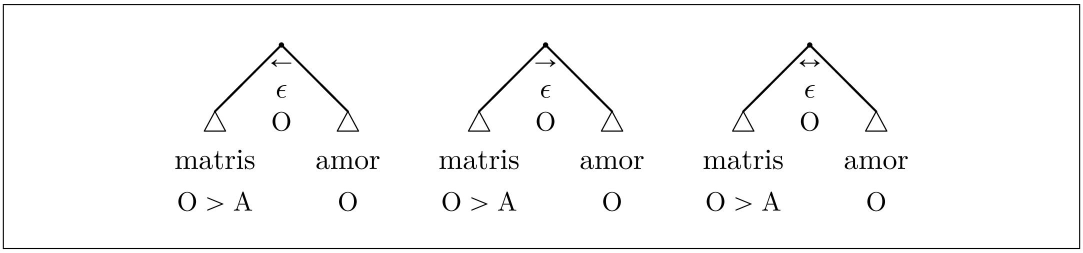


Figure 3: Example of information prominence in contrast

If the genitive is subjective, the meaning is *mater amat*, i.e. 'the mother loves (her children)' (adtree on left); by contrast, if the genitive is objective, the meaning *filii matrem amant*, i.e. 'children love their mother' (adtree in the center). Disambiguation is possible only if the context is known: if the context is not at disposal, information prominence will be represented by a left-right arrow (↔), to indicate under-specification (adtree on the right).

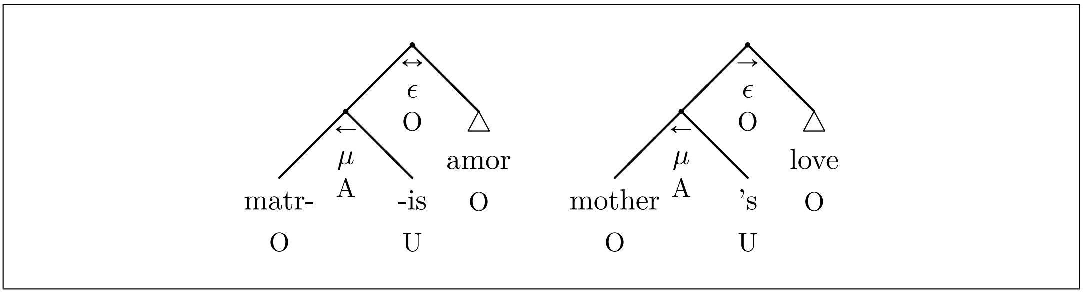


Figure 4: Example of linguistic morphosyntactic adtree

Figure 4 unhides the morphological information (μ) of the word *matris* (mother's) expressed in Figure 3 in the compact form: O>A. The morphosyntactic adtree of the English correspondent is provided on the right, for the reader's sake. However, please note that while the information prominence in Latin is underspecified (←) as explained before, by default in the English syntagm 'mother's love' the genitive is subjective, henceforth prominence is on the mother, which stays as the nominal (O) part of the dependent. For more details on linguistic adtrees and their transformations, see Gobbo and Benini [10].

3 Pragmatic adpositional trees

3.1 The concept of voice: Who is saying what?

The pragmatic level of analysis focuses on how language is *used* for various purposes, such as explaining what someone does not yet know or convincing them of something they do not yet accept. This level presupposes not only the presence of linguistic material—the observable data, 'something' that is uttered—but also the presence of an utterer, i.e., 'someone' performing the act of uttering the linguistic material, such as 'says' or 'writes'. It is important to underline the fact that utterers are observables too, i.e., they are not only imagined in the mind of the analyst but they are a real part of the information, and therefore they need to be represented explicitly. In other words, the utterance in its most general form ('something that someone says') includes the indication of who is saying what, and this can completely change the interpretation; in fact, the utterer rules the actual content of what is said: therefore, in the adpositional tree, the actual content depends on the utterer and the way he or she expresses the content itself. In fact, in analyzing natural language examples in real or fictional worlds, we cannot dismiss the role of the utterer; otherwise, we lose information, with the risk of inserting unnecessary biases in the analysis. For instance, the common offering 'have some wine' would intend something completely different if it is said by a friend during dinner or by the March Hare to Alice during the Mad Tea Party in Wonderland, as, in the latter case, on the table "there was nothing but tea" (Chapter VII).

In Adpositional Argumentation, the layer of the act of uttering is an adtree indicated with the adposition φ_x. The act of uttering is conventionally called 'voice', following a convention in narrative studies [16]. As illustrated in Table 1, the adposition φ_x is more abstract than ϵ and μ, respectively representing syntax and morphology—and encapsulate, in their leaves, most of the semantics. For this reason, φ_x, indicating the uttering, shall appear as a governor of the uttered content, representing the fact that the uttered content depends on the existence of the utterance. The act of uttering is constituted by two elements: an utterer and a sign of predication—Figure 5.

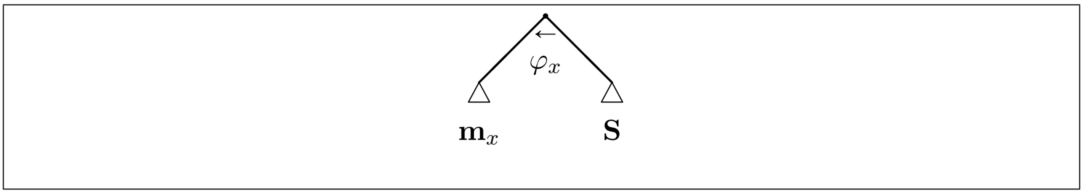


Figure 5: Abstract adtree for voice

The sign of predication is generically indicated as a verb of saying (S) while the utterer is indicated with a lower-case letter of the Latin alphabet showing the order of appearance in the discourse or text (generically: m), and marked with an index x, a natural number indicating the distance from the author, whose voice is indicated as a_0. Conventionally, the leaves of levels of abstraction above morphosyntax, i.e., pragmatic and argumentative, are represented in bold. Finally, if needed, its information prominence can be inverted, for instance, when the focus is on the utterer instead of the predication. The concept of voice has been introduced and widely discussed in Gobbo, Benini, and Wagemans (2022) [16].

3.2 From explanatory to argumentative information

As we remarked above, there are various discourse genres, such as *explanation* and *argumentation*. Within pragmatics, as the study of the use of language for various purposes, the attribution of these genre labels is based on the utterer's anticipation of the epistemic and doxastic status of the addressee. In short, when the utterer anticipates a lack of knowledge on the part of the addressee, the discourse produced is explanatory, and when the utterer anticipates a lack of acceptance, it is argumentative. Since the linguistic marker 'because' functions in both genres, it may only become clear from the context which of the two is instantiated. The utterance 'The cake tastes like carton because it does not contain sugar', for instance, counts as an explanation if it is clear from the the context that the addressee agrees that the cake tastes like carton but does not yet know why this is the case.

Within Adpositional Argumentation, these two main types of information are represented on the left branches of an adtree with the voice as a right branch, as the content—be it explanatory or argumentative—depends on the voice. When the utterer is explaining something, the relation between the act of uttering and the actual content is indicated by the Greek letter ρ_x. When the utterer is directly conveying argumentation, the relation between the act of uttering and the actual content is indicated by the Greek letter ξ_x; conventionally, we name it as an act of expressing a viewpoint. Figure 6 illustrates the respective abstract pragmatic adtrees, where the content is signaled by the dots ($\dots$).

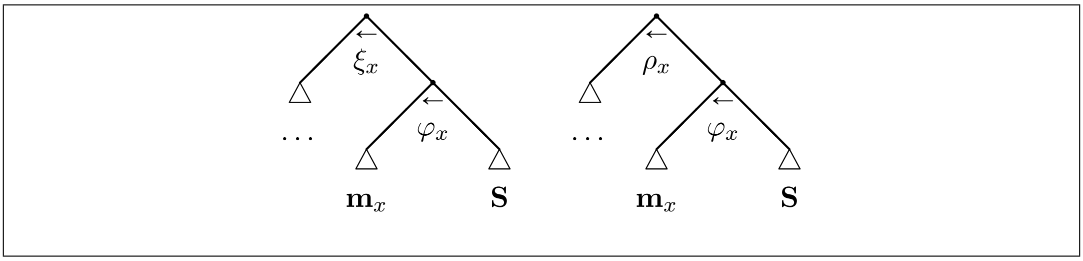


Figure 6: Abstract pragmatic adtrees for viewpoint (left), and reported speech (right)

In natural argumentative discourse, it may also occur that someone reports (ρ_x) the viewpoint of someone else (ξ_y), as in Figure 7. The reported content may be an explanation or an argumentation.

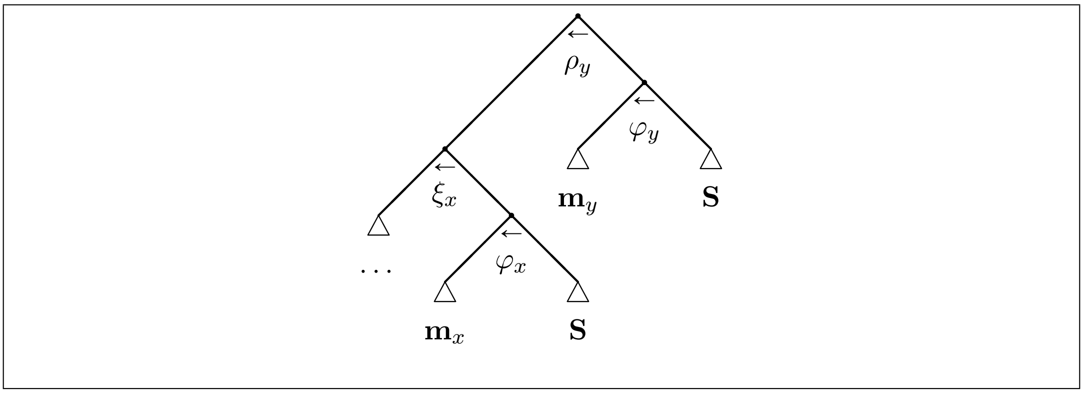


Figure 7: Abstract adtrees for reporting someone else's viewpoint

Figure 8 offers an example of a reported explanation. The sentence 'George said the cake tastes like carton because it does not contain sugar' is an example of a reported explanation. When annotating natural argumentative discourse, it is important to acknowledge the parts that are not argumentative but merely explanatory. Those parts may be annotated by linguistic adtrees, without referring to any argumentation framework such as the PTA, whose representation in the form of adtrees is illustrated in the next sections. In particular, the 'because' in the sentence should not be treated as argumentative, but just as a linguistic indicator, in this case, a unifier (U) of the two phrases 'the cake tastes like carton' and 'It does not contain sugar'. For an extensive explanation of valency in linguistic adtrees, represented by superscripts and subscripts, such as in the grammar characters I_2^2, E_2, O_1 in Figure 8, please see Gobbo and Benini (2013,2011) [11, 10].

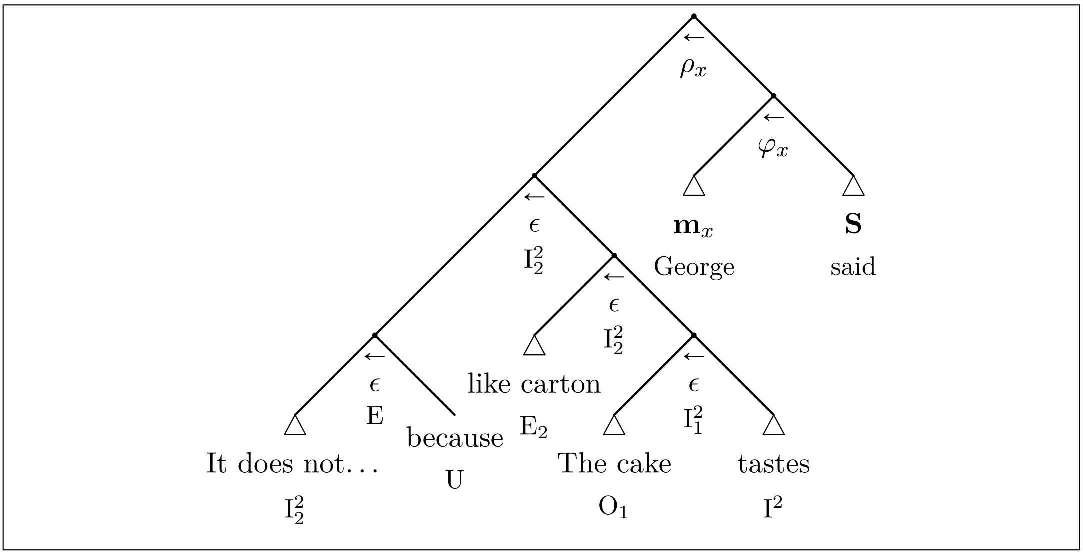


Figure 8: Exemplar adtree of reported speech of an explanation

It may also occur that someone reports (ρ_x) the viewpoint of someone else (ξ_y), including one or more arguments. A concrete example of such reported argumentation may be found in the opening lines of an exercise from a textbook on argumentation, already analyzed in Gobbo, Benini, and Wagemans (2022) [16]:

> In his article "Plagiarism: A rich tradition in science", editor John Lowell argues, referring to an article by dr. P. Smith, that Copernicus was also guilty of plagiarism:

In this case, the corresponding adtree has a sub-tree with the reported argumentation, as pictured in Figure 9.

Figure 9 shows the structure of reported speech without delving into the analysis of the subsequent argumentation: the utterer (a_0), being the author's voice (φ_0) reports (ρ_0) that the voice entity 'editor John Lowell' (b_1) argues about the accusation of Copernicus being guilty of plagiarism. While linguistic details of the voice entity b_1 are left hidden (△), the adtree also shows part of the linguistic structure of the voice entity's predication, distinguishing the the verb 'argues', which governs the circumstantial 'In his article "Plagiarism: A rich tradition in science",'.

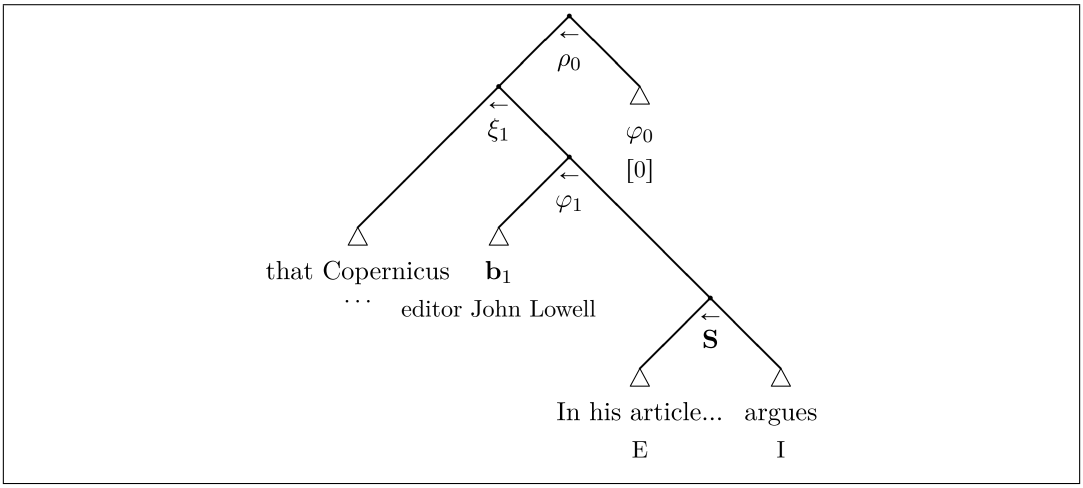


Figure 9: Reported speech of a voice

It is worth remarking that viewpoints and reported speech are represented as adtrees when the information of who is saying what is explicitly stated in the text; otherwise, it is always possible to represent the viewpoint ξ_0 of the author a_0 as the governor of the linguistic material included in the argumentative adtree in the dependent 'as it is'. Finally, adtrees can represent the extreme case of the author referring to themselves in the third person, as Caesar did in *De bello Gallico*, with a ρ_0 for the reporting and an m_0 for the voice subject, whose distance from the author is, in this case, zero.

For the sake of simplicity, in the following, we will consider viewpoint as the default indication of the utterance, that is, all the adtrees presented in the next sections will be ruled by the utterer putting forward an argumentation, unless indicated otherwise. For more details on how to represent more complex structures of reported speech, see Gobbo, Benini, and Wagemans (2022) [16].

4 Representing claims and minimal arguments

In this paper, with the term 'argumentation' we indicate the fabric of arguments put forward in a discourse or text expressed in natural language, while 'argument' is reserved for a single element of that fabric. An argumentation generally consists of a collection of premises, a collection of conclusions, and a way to relate them: all these pieces are observables, as they can be recognized in the piece of natural argumentative discourse at hand. As described and discussed in Section 3, it is also essential to analyze the the context in which the argumentation is uttered.

Before analyzing how complex argumentation could be represented and interpreted, it is worth considering the simple case where the collections of premises and conclusions are minimal. Indeed, a collection is a structure that groups together and coordinates the involved elements. Hence, leaving out the grouping structure for the moment allows us to simplify the study of argumentation greatly. Also, elements may be either atomic assertions (called *statements* in the following) or arguments themselves: again, it is far simpler to assume that the elements of a collection are atomic. These two hypotheses provide a fair point to start describing how argumentative adtrees are constructed. So, in this section, we assume the above collections to contain at most one statement, addressing the general case in Section 5.

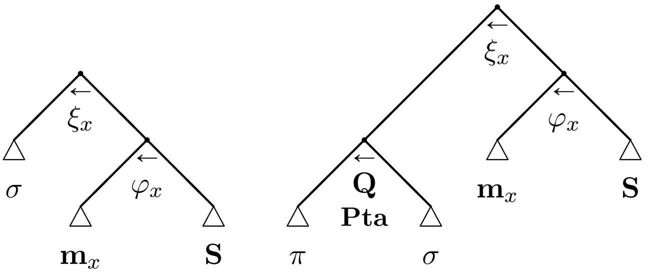


Figure 10: Abstract adtrees of a claim (left) and a minimal argument (right)

In the first place, we observe that there is no object whose acceptability the arguer aims to establish or increase when the conclusion is absent. Thus, by the very meaning of the notion, without a conclusion there is no argument. Therefore, there are only two cases for a simple argument: first, one conclusion with no premise; second, one conclusion and one premise. The former case is called a *claim*, i.e., an unsupported statement, while the latter is a *minimal argument.* Figure 10 illustrates the respective abstract adtrees in which the Q indicates a generic quadrant $(\alpha, \beta, \gamma, \delta)$.

A claim is then a statement that is atomic with respect to the argumentative level of abstraction. It is represented as a leaf in the argumentative part of the adtree, and functionally it may act as a premise or conclusion for another argument. As a side note, observe how a claim may be the root of an adtree which further analyses its internal structure with respect to another level of abstraction, for example, its linguistic representation. Hence, the natural interpretation of a claim A in isolation is the sequent $\vdash A$ in the logic of the arguer, while it becomes an assumption when used as a premise, so an element in the left-hand side of a sequent. These two uses are special cases of convergent and divergent arguments, as explained in the next section.

Within Adpositional Argumentation, the premise is indicated by the Greek letter π and the conclusion by σ. The prominence is identified by the shape of the argument: retrogressive (←) when it is 'σ because π'; progressive (→) when its shape is 'π then σ'. Since a minimal argument contains a conclusion σ and a premise π, but the conclusion is necessary while the premise is optional, it is natural to think that σ rules over π. This fact is reflected in the adtree representation where the governor, the right leaf, is σ, and the dependent, the left leaf, is π. Thus, the representation privileges the retrogressive normal form of an argument: σ because π. Consequently, its intended interpretation in the logic of the arguer is the sequent $\pi \vdash \sigma$.

An apparent problem with the intended interpretation $\pi \vdash \sigma$ is that the arguer states this sequent because it holds by some inference rule r: this rule r is **not** an observable, and in most cases in the real world, it is unknown not only by the analyst or the counterpart in the discourse, but even by the arguer. Therefore, according to the principle that an adtree must represent the argument 'as it is', as close as possible to what can be observed, the adposition in the root of the adtree for a minimal argument has to identify the 'inference rule' as objectively as possible, according to what is observable.

4.1 Representing minimal arguments

Because the inference rule cannot always be precisely identified from the observables, we need a more fine-grained analysis of the content of the statements functioning as the conclusion and the premise of the argument. For this reason, we represent minimal arguments in terms of the argument categorization framework of the Periodic Table of Arguments (PTA)—see Wagemans (2016,2019,2020,2023) [28, 30, 29, 32]). This framework conceptualizes an 'argument type' as a collection of instantiated values of three different parameters (form, substance, and lever). The determination of the first parameter, the argument form, requires breaking down the statements functioning as the conclusion and the premise of the argument into a subject, indicated with small roman letters (a, b, ...), and a predicate, indicated with capital roman letters (X, Y, ...). The determination of the third parameter, the argument lever, provides the inference rule and thus indicates an aspect of the logic of the arguer. For the present purposes, we refer to Table 2 for an overview of the configurations of the subjects and predicates in the four basic argument forms (named α, β, γ, δ) distinguished within the theoretical framework of the PTA and their corresponding levers.

From a structural point of view, the abstract argumentative adtrees correspond-

name	*conclusion* σ	*premise* π	*retrogressive normal form* *of minimal arguments*	*lever*
α	**a** is **X**	**a** is **Y**	**a** is **X**, because **a** is **Y**	**X** R **Y**
β	**a** is **X**	**b** is **X**	**a** is **X**, because **b** is **X**	**a** R **b**
γ	**q**(**a** is **X**)	**r**(**b** is **Y**)	**q**(**a** is **X**), because **r**(**b** is **Y**)	**q** R **r**
δ	**q**(**a** is **X**) [is **T**]	**q**(**a** is **X**) is **Z**	**q** [is **T**], because **q** is **Z**	[**T**] R **Z**

Table 2: Overview of abstract minimal argument retrogressive forms

ing to the three forms α, β, and γ are very similar, as Figure 11 illustrates.

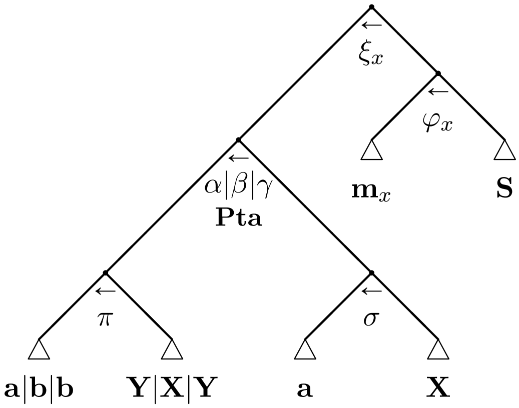


Figure 11: Abstract argumentative adtrees: α, β, and γ quadrants

The symmetry of α, β, and γ arguments is not found in δ arguments. This is because, in the latter, the arguer supports the acceptability of the conclusion by attributing an external property to it: the conclusion is deemed acceptable, for instance, because some authority endorses it or because not accepting it leads to bad things. If we indicate the acceptability of the conclusion as 'T' and the external property attributed to it as 'Z', we can represent the form of δ arguments as 'q [is T], because q is Z'. It is important to note that the predicate T correlated to the statement q which represents the subject and the predicate as a whole, under the form: [is T] has **no** relation to the operator of truth in logic $\top$, but should be read as 'is trustworthy'.

The above difference between, on the one hand, α, β, and γ arguments and, on the other hand, δ arguments is reflected in the lever, which represents an aspect of the logic of the arguer, namely the inference rule. In the first three forms, the lever is found in the components of the premise and the conclusion, namely a, X and b, Y. The relation between these pieces of the statements is what allows to relate the premise with the conclusion, and the relation is generally found in the semantics.

On the contrary, the lever of a δ argument does not depend on the components of the conclusion. The arguer aims to establish or increase the acceptability of the conclusion by predicating something of it as a whole. The lever is thus a relationship between the external property (Z) and the trustworthiness (T) of the conclusion as such, which is usually not expressed in the linguistic material—see also Table 1. As mnemotechnics, we say that α, β, and γ arguments provide a first-order relation, while δ envisages a second- or higher-order relation. The reader is adverted that such terminology has **no** logical value.

4.2 Two examples of minimal arguments

The first example illustrates first-order relations, while the second one will clarify how δ arguments work. The statement 'I think *Interstellar* is great', which contains the claim '*Insterstellar* is great', referring to the feature film directed by Christopher Nolan in 2014—see Figure 12.

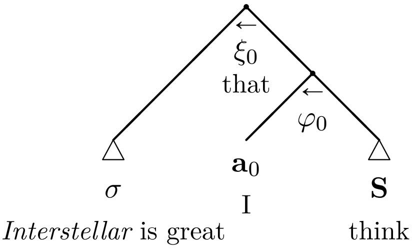


Figure 12: 'I think that *Interstellar* is great'

Admittedly, it is generally more effective to argue through something more substantial than a simple claim. In general, the arguer, who is also the utterer, in this case, supports the previous statement with a subsequent one, adding a statement such as 'It is directed by Nolan'. What we obtain is a minimal argument: '*Interstellar* is great because it is directed by Nolan', which is represented in Figure 13.

In the example, when observing the two statements prima facie, we note that the argument form is α, with the following distribution of subjects and predicates: 'Interstellar (b) is great (X), because it (b) is directed by Nolan (Y)'. The lever is thus a relationship between the predicates X and Y. What concrete relationship that is, is something for the analyst to decide, as this aspect of the logic of the arguer is not included in the linguistic material. The values of the parameters form (α) and

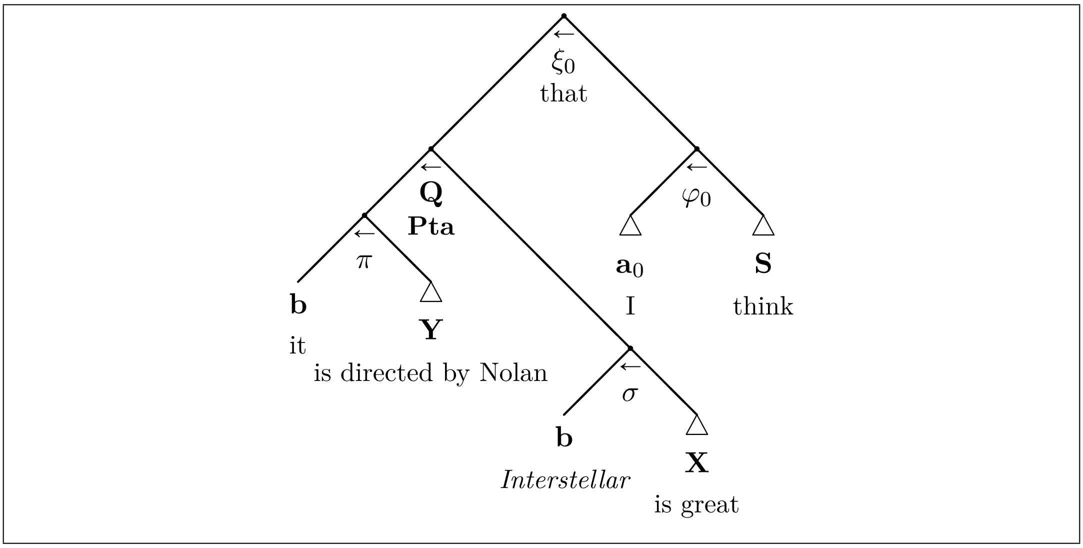


Figure 13: ‘I think that *Interstellar* is great because it is directed by Nolan’

substance VF, expressed via the pair of argumentative characters of value (V) of the conclusion σ and fact (F) of the premise π reduce the number of possibilities here, as the framework of the PTA suggests it to be an ’argument from criterion’ (Cr). This means that the relationship between the predicates is such that the predicate of the premise functions as a criterion for the predicate of the conclusion: ’being directed by Nolan is a criterion for being great’.

The second example illustrates how the δ arguments may imply the introduction of a new voice. Consider the minimal argument ‘Infinity is not unique because Cantor said so’: it is clear that the ‘so’ particle is an anaphora, in other words, it is a way to avoid to repeat linguistic material already expressed previously, in this case ‘Infinity is not unique’. In adtrees, anaphoras are indicated by the arrow that turns back to the right ↻, immediately followed by the target addressed by the place marker—analogously, cataphoras, i.e. anticipations of linguistic material explained later in the text, shall be indicated in adtrees as ↺. The δ authority (Au) argument represented in Figure 14 has the conclusion ‘Infinity is not unique’ in the governor and the premise ‘Cantor said so’ in the dependent. The premise contains as a subject the conclusion by means of the anaphoric (↻) ‘so’ and as a predicate the voice ‘Cantor said’.

Observe that the statement ‘infinity is not unique’ is established as a whole and not in force of its components: in this very aspect lies the unique feature of the δ arguments—Figure 14.

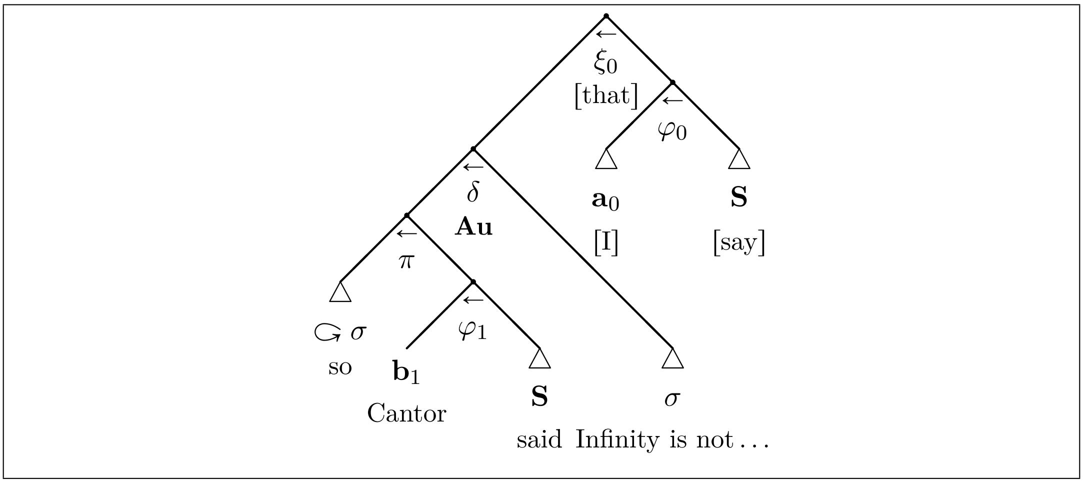


Figure 14: The adtree of 'Infinity is not unique because Cantor said so'

Figure 15 shows the argumentative levels of abstraction of the two examples, in contrast, hiding all linguistic details.

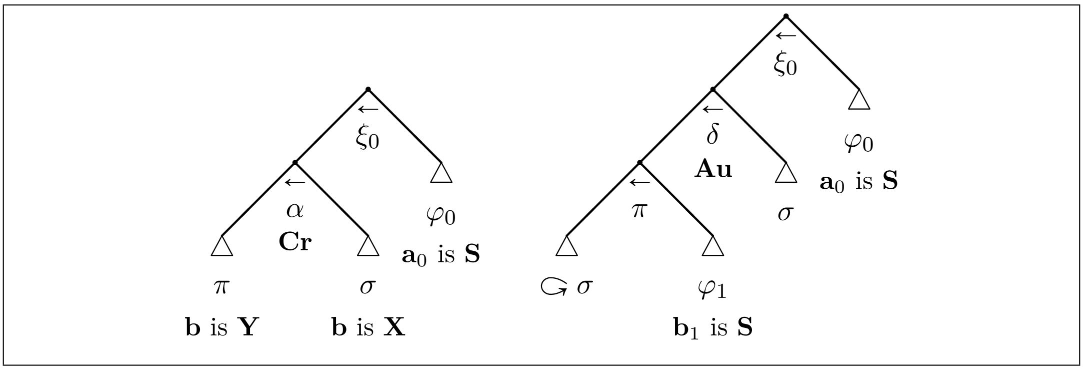


Figure 15: Argumentative adtrees of the two examples in contrast

Argument types of all four forms α, β, γ, and δ, are based on a lever of some sort—see Table 2. However, while form and substance are pieces of information based on observables, the the lever is not always—or rather: usually not—explicitly present in the linguistic material. In this case, the PTA is used as a heuristic for formulating the lever, which is made possible by its conventional validity as a classification framework based on taxonomies of argument types (argument schemes, fallacies, and other means of persuasion) from the informal traditions of dialectic and rhetoric. This information is enough to identify the potential attacks by the the counterpart in the argumentative discourse on the solidity of the argument lever—see Hinton and Wagemans, (2021) [17]. In other words, while the representation of minimal arguments in Adpositional Argumentation is not enough to identify the logic of the arguer in use, it allows attacking the argument since it makes explicit the observable nature of the inference. We can therefore conclude that the representation injects the tradition of AT in a solid formalism through the PTA, mitigating the gap between AT and formal approaches such as CA exposed in the Introduction.

5 Representing complex argumentation

Real-world argumentation is often complex: it may contain multiple premises, sometimes multiple conclusions, and one argument may use a conclusion of another argument as a premise, yielding a chain of arguments—see, e.g. Freeman (2011) [8]. A proper representation of natural argumentative discourse has to cope with all these cases. We call *convergent* complex argumentation with one conclusion and many premises, while conversely more conclusions driven by one premise will be called *divergent*. Finally, complex argumentation using as a premise the conclusion of another argument is called *serial*. The representation suggests a way to interpret the argumentation, providing clues on the logic of the arguer, which could and should be identified to see how an argumentation conveys acceptance or refusal.

A concrete argumentation, i.e., one that is expressed in natural language, may contain conclusions and premises that can be convergent, divergent, and serial at the same time. The guiding principle ruling composition is that the premises are in the dependent part of an adtree, while the conclusions are in the governor part; finally, the adposition specifies how the complex argumentation is constructed, and thus how it should be represented. In the following, the three cases of complex argumentation are discussed in detail and separately. However, the formalism allows for smoothly composing the representations of the three cases, if needed.

5.1 Convergent argumentation

Convergent argumentation is characterized by having more than one premise. The key idea to represent them is to combine all the premises into a single one.

To obtain a sound representation of all the premises without introducing new information beyond the observables, one has to remark that the premises are ordered in the textual exposition of the argument, thus there is an observable list of premises $\pi_1, \dots, \pi_n$ with $n > 1$. The representation, depicted in Figure 16, divides the list into two parts π_x and π_y, in a process detailed below, and assigns the prominence accordingly. The adposition is completed by a λ symbol to indicate the combination

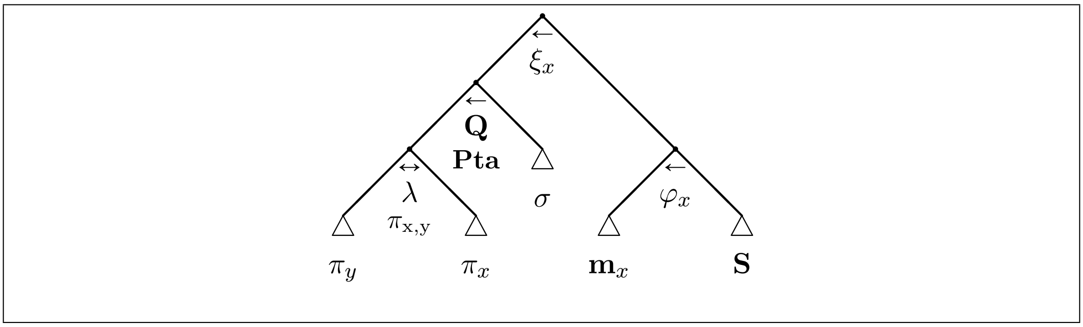


Figure 16: The abstract adtree of convergent argumentation

operation, and the label $\pi_{x,y}$ to remind both the elements and the order in which they have been combined (as usual, these pieces of information are omitted when they can be reconstructed from the context).

The interpretation of the λ operation within an argument as in Figure 16 is a sequent $\pi_{x_1}, \ldots, \pi_{x_n} \vdash \sigma$ in the logic of the arguer, where $\pi_{x_1,\ldots,x_n}$ is (the label of) the adtree grouping all the premises. The order of the premises is a consequence of the dependencies among them. Indeed, the logic of the arguer is generally unknown and not observable, thus the analyst has to determine whether two premises π_x and π_y are independent, so $\pi_x \leftrightarrow \pi_y$, or if π_x depends on π_y, thus $\pi_x \leftarrow \pi_y$, or vice versa. This piece of information is sometimes present in the text, so it may be observable, but it could also be added by the analysts, based on their experience and understanding, in which case the adtree is not objective, but represents the point of view of the analyst. In the following, we assume fair representations, which incorporate observable dependencies among premises only.

To better understand what dependency is in this context, consider the argument "the number n is odd (σ) because $n - 1$ is even (π_1), n is a natural number (π_2), and n is strictly positive (π_3)". It is clear that $\pi_1 \leftarrow \pi_3$ and $\pi_3 \leftarrow \pi_2$ since the subtraction on naturals would be undefined unless $n > 0$, and in turn $n > 0$ makes no sense in a number system without an order and 0. Hence, the right way to order the premises by their dependency would be $\pi_{2,3,1}$ which is the right order a mathematical analyst should impose on the combination. Observe how putting π_1 in evidence as the first uttered premise emphasizes its importance in conveying the validity of the conclusion, which is not a proper argumentative aspect but rather pertains to the pragmatic level of abstraction.

Moreover, dependencies among the premises provide insight into the the logic of the arguer. In fact, the structure of the left-hand side of a sequent distinguishes logics in which assumptions are collected in sets, e.g., classical logic with the LK presentation, see Schwichtenberg et al. (1996) [24], from those in which assumptions are represented as a partial order, e.g., dependent types, see Martin-Löf [19] and Homotopy Type Theory [26]. Of course, other structures (multisets, linear orders, etc.) are possible, and they hint at specific families of logics.

Independence of premises provides a hint on how attacking the combination may elicit information about the arguer's logic. If the counterpart attacks a variant adtree, in which the independent premises are permuted, and the arguer defends its original argument refuting the permuted representation, the variant is observably not admissible in the logic of the arguer, thus revealing a hidden dependency. This fact suggests that studying the transformations of argumentative adtrees, like the permutation of independent premises or conclusions, is a powerful instrument to better understand them, and to provide formal clues to orient the dialogue and clarify the arguments. But this lies beyond the scope of the present work. The specifications of the argument types in the PTA in a convergent argument, see Figure 16, are obtained following the analogy with Chemistry: a minimal argument is analogous to an atom of matter, while a complex argumentation structure is analogous to a molecule. Hence, the quadrant is usually γ since the lever is rarely found. However, there are exceptions: for example 'The Blues Brothers is a cult movie (σ) because it has superb music on the score (π_1) and it stars John Belushi at his best (π_2)' can be identified as an α argument, and the 'molecule' is composed of two atoms which are both St (Standard), so the adposition becomes (α, St^2)—see Figure 17. The St^2 part denotes the 'raw formula' for the argument type, analogously to $\mathsf{H_2O}$ which denotes water in Chemistry: it describes the general form of the argument, while its inner structure is represented in the relationship between the dependent and the governor.

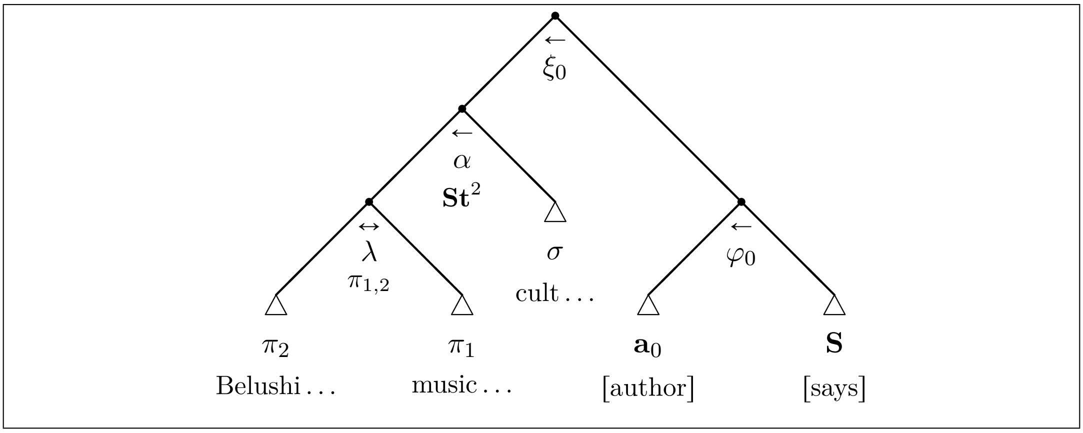


Figure 17: The adtree of 'The Blues Brothers is a cult movie because...and...'

In the general case, because a complex argumentation may mix statements of value, policy, and fact, and furthermore can combine them into involved structures, the raw formula for precisely identifying the 'molecule' is still an open problem to be addressed in the future: a brief discussion can be found in the Conclusion.

In the end, it is worth observing that a combined element may be an argument itself rather than a statement. For example 'Lily wears a raincoat (σ) because it's very cloudy so it may rain (π_1) and, if it rains and she is not well covered then Lily could get a cold (π_2)'. Both premises π_1 and π_2 are arguments: π_1 is 'it may rain (σ_{1a}) because it's very cloudy (π_{1a})', and π_2 is 'Lily could get a cold (σ_{2a}) because it rains (π_{2a}) and (λ) Lily is not well covered (π_{2b})'.

In general, using arguments in place of statements models *hypothetical reasoning*: the premise which is an argument $\pi_a \vdash \sigma_a$ tells that $\pi_a \vdash \sigma_a$ is assumed to be valid in order to deduce the conclusion, even if the arguer does not establish the premise π_a. In the example, π_2 has been asserted, and its premises may be attacked: for example, the counterpart may reply 'Since Lily already has six layers of clothes on, she is well covered'.

5.2 Divergent argumentation

Divergent argumentation is characterized by having multiple conclusions grouped together so that the premise aims at establishing all of them. Analogously to premises in convergent argumentation, see Section 5.1, the conclusions are ordered as a list $\sigma_1, \ldots, \sigma_n$ by the text material. Thus, the way to represent them is the same as for the combination of premises in convergent argumentation, as shown in Figure 18, including the analysis of dependencies among conclusions.

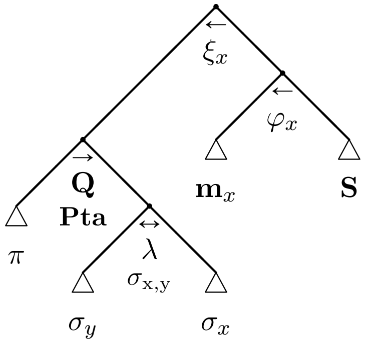

$\xleftarrow{}$ ξ_x
$\xleftarrow{}$ φ_x
$\mathbf{m}_x$
S
$\xrightarrow{}$ **Q** **Pta**
π
$\leftrightarrow$ λ $\sigma_{\text{x,y}}$
σ_y
σ_x

Figure 18: The abstract adtree of divergent argumentation

Reminding that an argument is interpreted as a sequent $\pi \vdash \sigma$ from the interpretation of the premise to the interpretation of the conclusion in the logic of the arguer, when the conclusion is $\sigma_1, \ldots, \sigma_n$, it should be interpreted in the product of σ_x and σ_y, according to the notation in Figure 18. When σ_x and σ_y are independent, we could reasonably expect that the product is Cartesian, that is, conjunction; when σ_y depends on σ_x, we should expect the product to be an amalgamation, like the Σ operator in Martin-Löf's *An Intuitionistic Theory of Types* [19].

In general, the optimal guess for the product is the categorical product of σ_x and σ_y in the category of statements whose arrows are sequents. However, this is an educated guess at best, since the underlying category may not have all the finite limits. Hence, divergent arguments provide deeper but uncertain clues on the logic of the arguer. How to devise an attack strategy to elicit stronger information about the nature of the product in the logic of the arguer is still a work in progress.

Consider the argument 'The house is cold and we cannot cook hot food because the gas supply is broken'. Clearly, it is a divergent argument from the premise 'the gas supply is broken' (π) to the conclusions 'The house is cold' (σ_1) and 'we cannot cook hot food' (σ_2). Also, the conclusions are factually independent. Hence, the argument is represented as in Figure 19.

Observe how prominence between the conclusions σ_1 and σ_2 has been left underspecified, since they are independent. However, if this text is a fragment of a phone conversation of a house owner complaining to a gas company, we could suppose that the heating problem would be more relevant in the rest of the call.

A critical feature one needs in order to interpret arguments as sequents, and, at the same time, to have a notion of product, is that the apparently trivial δ argument "S because S" must hold, which tells, when S is a collection of premises/conclusions, that the product is the reification of structure on the collection of the premises. This link between premises and conclusions is required in Adpositional Argumentation: the requirement is imposed by using the same constructor λ both in convergent and divergent arguments.

An important observation is about incoherent collections of premises/conclusions *in the logic of the arguer*: they will never be formed by the arguer; however, the counterpart may form a counter-argument having arbitrary premises/conclusions to attack the arguer's argument, even when these are incoherent for the arguer. This kind of attack is effective to understand what the arguer considers non-admissible, creating observables, in the form of replies from the arguer, about inner aspects of the logic of the arguer.

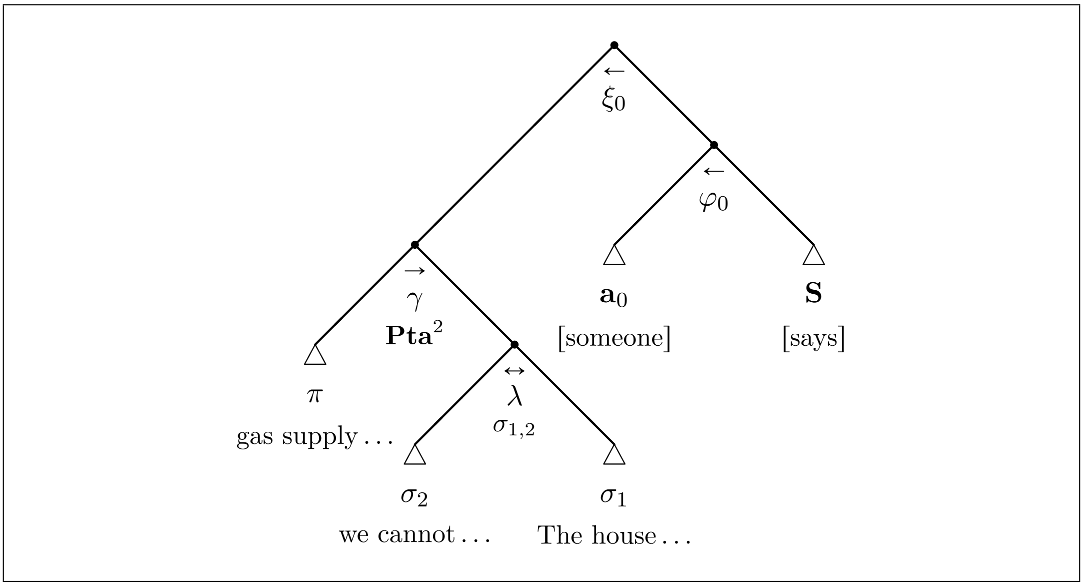


Figure 19: The adtree for 'The house is cold and . . . '

5.3 Serial argumentation

A serial argument composes two arguments by using a conclusion of the first as a premise for the second one. So, if the first argument is 'Ξ, Δ because π_1' and the second argument is 'σ_2 because Ξ, Δ', then the serial argument is usually summarised as showing 'σ_2 because π_1', hiding the extra premises Δ', the extra conclusions Δ, and their link Ξ.

The serial argument is represented in Figure 20: the right subtree is the viewpoint, while the left subtree, marked by a Ω to indicate serial composition, has the 'σ_2 because Ξ, Δ'' argument as its governor, and the 'Ξ, Δ because π_1' argument as its dependent. The Ξ statement acts as an independent conclusion in the left branch, and as the governor premise in the right branch. The $\omega(\pi_1, \sigma_2)$ indicates the prominent premises and conclusions of the serial argument.

The intended interpretation of serial composition is a logical cut, as in Negri et al. (2001) [22]: indeed, if $\pi_1 \vdash \Xi$ and $\Xi \vdash \sigma_2$, then $\pi_1 \vdash \sigma_2$, in its simplest form. When the first argument is divergent, i.e., $\pi_1 \vdash \Xi \wedge \Delta$, or the second argument is convergent, i.e., $\Xi, \Delta' \vdash \sigma_2$, the serial composition hides, but does not discard, the extra premises/conclusions, i.e., the Δ's. This is the usual way in which serial arguments are written down, possibly using the Δ's in a subsequent argument, which is eventually treated duplicating the representations of the arguments $\pi_1 \vdash \Xi \wedge \Delta$

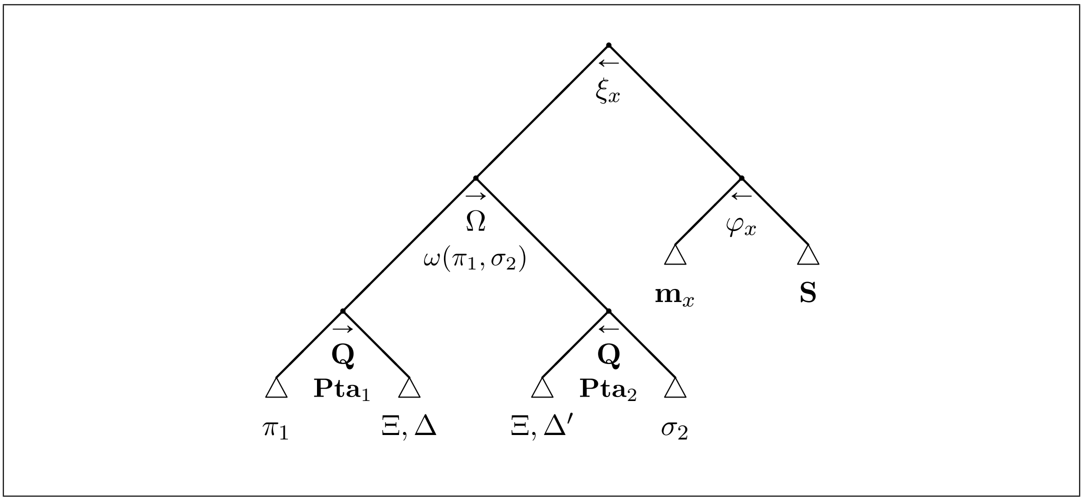

$\overleftarrow{\xi_x}$
$\overrightarrow{\Omega}$
$\omega(\pi_1, \sigma_2)$
$\overleftarrow{\varphi_x}$
$\mathbf{m}_x$
S
$\overrightarrow{\mathbf{Q}}$
$\mathbf{Pta}_1$
$\overleftarrow{\mathbf{Q}}$
$\mathbf{Pta}_2$
π_1
Ξ, Δ
Ξ, Δ'
σ_2

Figure 20: The abstract adtree of serial argumentation

or $\Xi, \Delta' \vdash \sigma_2$ and reordering the combined elements under the λ's. In this respect, the adopted representation has the purpose to strictly follow the linguistic material: indeed, a serial argument with convergent/divergent components usually emphasizes the connecting component Ξ, and hides the Δ and Δ' premises and conclusions in the composed argument. For example, consider the argument a driver made to the insurance company: 'The car crashed into the tree because the car was skidding, and it was skidding because the road was wet; then, the car crashed into the tree because the road was wet'. There are two arguments, 'the car was skidding because the road was wet' (A_1) and 'the car crashed into the tree because the car was skidding' (A_2); they are serially composed to obtain 'the car crashed into the tree because the the road was wet' using the pivot 'the car was skidding', conventionally marked by Ξ. The corresponding representation is shown in Figure 21.

A more complex example is 'The car crashed into the tree because I touched the brakes and the car was skidding, and it was skidding because the road was wet and I lost control, then the car crashed into the tree because the road was wet'. Here, differently from the previous example, there is a convergent and a divergent argument involved in the pivot Ξ. The corresponding adtree is shown in Figure 22.

Interpreting the serial composition of arguments as a cut is natural, but it does not tell that in the logic of the arguer the cut rule is admissible, but rather that the specific instance represented in the the serial argument can be observed and thus it can be carried on in the logic of the arguer.

Moreover, the sequent $\pi_1 \vdash \sigma_2$ is not necessarily the exact result of the serial

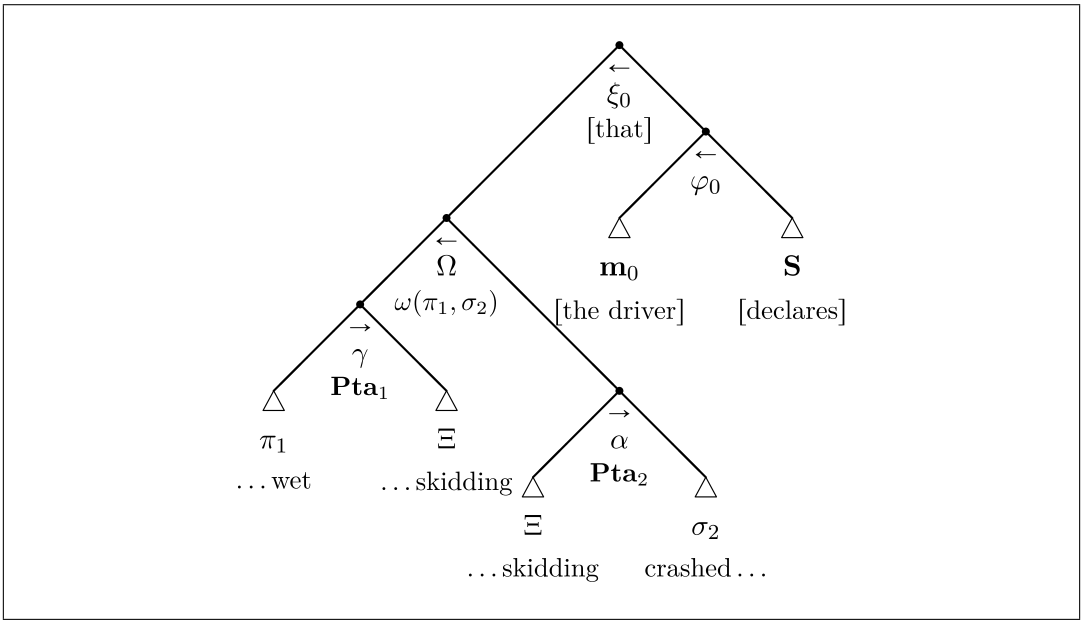


Figure 21: Adtree for 'the car crashed...', simple version

composition: depending on the logic of the arguer, the sequent could be different, eventually involving (parts of) the Δ's. Therefore, it has been indicated by $\omega(\pi_1, \sigma_2)$ in the representing adtree, where the ω operation yields the resulting sequent from the 'cut' of the two represented arguments, which are really the left and right subtrees. For example, if the arguer reasons using linear logic (see Girard (1987) [9]), or dependent type theory (see Martin-Löf (1975) [19, 26]) or a paraconsistent logic (see Avron et al. (2018) [1]) the resulting sequence may differ from $\pi_1 \vdash \sigma_2$, involving, e.g, further premises on which π_1 depends on.

To further clarify, when the second argument is $\Xi \vdash \sigma_2$, which is the usual way in which serial composition is written down, Ξ appears to be an independent premise. Nevertheless, this is not always the case: for example, in homotopy type theory [26], Ξ may depend on (a part of) π_1, thus the second argument should be really understood as $\pi_1, \Xi \vdash \sigma_2$. However, hiding this fact is an essential ingredient to make neat, compact, and vividly understandable proofs in that theory: the 'logic' of that theory requires hiding dependencies to support intuition and clarify reasoning.

A crucial point in understanding serial composition is that using a serial argument as the premise or conclusion of another argument one has to *extract* the composed arguments. The representation constructs an argument S which contains the arguments A_1 and A_2 to compose using the pivot Ξ. The result of the Ω op-

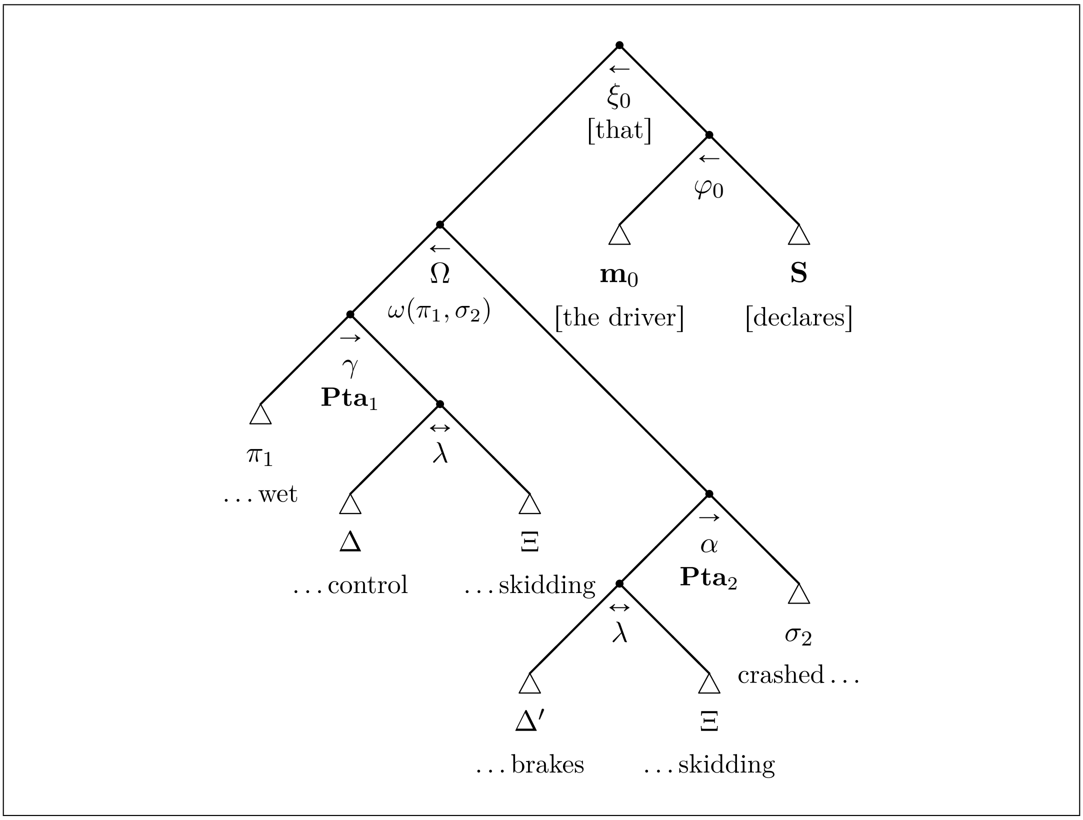


Figure 22: Adtree for 'the car crashed. . . ', complex version

erator is a complex adtree containing all the pieces of information to reconstruct the composition, but the result, which, as discussed above is really $\omega(\pi_1, \sigma_2)$. To make this argument explicit in the representation, one needs a further inference that takes a Ω adtree as a premise and concludes with an adtree representing the result. Reprising the previous example in Figure 21, the complete representation of the serial argument is shown in Figure 23, where the δ inference is responsible for providing the conclusion that 'the car crashed into the tree because the road was wet'. In summary, the whole argument reads '[the driver] [declares] [that] the car crashed into the tree (σ) because the road was wet (π), since (δ) the car was skidding (Ξ) because the road was wet (π_1), and the car crashed into the tree (σ_2) because the car was skidding (Ξ)'.

Therefore, the chosen representation closely adheres to the observable textual material, although its interpretation may significantly deviate because of the logic of the arguer: the Ξ may not be independent, the dependency being hidden from

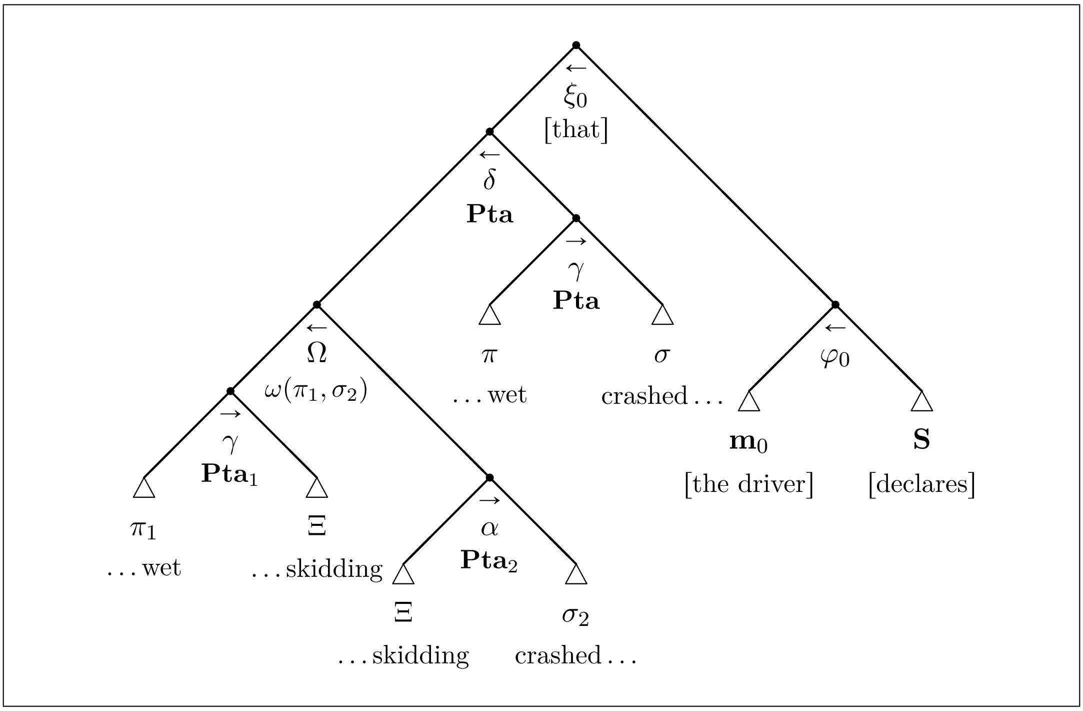


Figure 23: Adtree for ‘the car crashed. . . ’, final

the observer, which resolves into having hidden premises in the sequent resulting from the serial composition; also, the cut applied to obtain the serial composition is not necessarily the classical one. In these cases, the ω operation, which is marked but unspecified in the representation, has to be filled in to understand the arguer's argument. Of course, this is an evident point of attack, which may lead to clarify or to make evident a fallacy in the logic of the arguer. It is worth remarking that the δ extracting the final argument of a serial composition is responsible for making explicit the $\omega(\pi_1, \sigma_2)$ in the representation, i.e., for providing the result of the cut as it appears in the textual representation.

6 Conclusion

In this paper, we have illustrated a way to represent natural argumentative discourse in a formalism, the one of adpositional trees (see Section 2), which is uniform among many levels of abstraction, from the morphosyntactic (linguistic) to the argumentative (pragmatic) one. After showing, in Section 3, how the textual exposition of argumentation, possibly complemented by explanations and voices, can be included as part of its representation, we moved to considering claims and minimal arguments in Section 4. While these subjects have also been treated in previous works by the authors, the intended interpretation of minimal arguments as sequents is made explicit here for the first time. Another novelty is that, building upon this intended interpretation, the notions of convergent, divergent, and serial argumentation have been introduced, represented, and interpreted. Their representation in adpositional trees has been modeled after that of the minimal argument to allow for arbitrary compositions of these argumentative structures, which are the fundamental ones.

Natural argumentative discourse is expressed in linguistic material, which eventually is the place where argumentation can be observed in real-world use. The logic of the arguer is used to convince the addressee of the validity and acceptability of the argumentation. This is a dynamic process: the logic of the arguer does not only show in the observables but mainly in what is inferred from them.

In the first place, forming an attack on a given argument has a double purpose: contesting its validity (direct attack), but also a better understanding of how its logic works (indirect attack). A systematic way to improve understanding is to consider variant arguments, i.e., natural language rewording of the argument in order to clarify them, and to propose them to the arguer: their acceptability provides clues on the logic of the arguer, specifically about which structural properties of the logic could be observed, which ultimately leads to an identification of the logic itself. Devising such inquiring strategies has been hinted at in the present article, but not developed.

Systematically deriving these strategies requires to identify linguistic variants of the same argument that may validate or confute hypotheses about the structural properties of logic: variants are obtained by transforming the original argument to test whether it is equivalent or acceptable for the arguer. What the reasonable transformations are, and how to orient them towards testing specific properties is still an open problem, and the subject of further research.

In a similar vein, there is no one-to-one mapping from the levers of the minimal arguments listed in the argument categorization framework of the Periodic Table of Arguments (PTA) to the levers in complex argumentation. As in Chemistry, in which a molecule is composed of many different atoms, complex argumentation derives its convincing force from many different minimal arguments, i.e., ways to transport the acceptability from the premises to the conclusions. In this respect, the adpositional representation shows the fine structure by which this transport of validity is performed, but a synthetic way to denote it, as the raw formula for a molecule in Chemistry, is still under development.

On a similar note, so far, Adpositional Argumentation has focused on representing monological discourse. Some aspects of the representation of complex argumentation are therefore still open. In particular, the dynamics of attacking and defending an argument in a dialogue (or a polylogue, in the sense of Lewiński and Aakhus (2014) [18]), require more investigation. Also, the relationship between the *representation* of natural argumentative discourse and its *evaluation* is only touched upon briefly in this paper, namely in identifying the points of attack, and is left for future work.

Acknowledgements

This article builds on previous works of the authors, in particular [12, 13, 14, 15, 16] whose outcomes are summarised in Sections 2,3 and 4, while the content of Section 5 is new or significantly extends previous achievements.

References

[1] Arnon Avron and Ofer Arieli and Anna Zamansky. *Theory of effective propositional paraconsistent logics*, volume 75 of *Studies in Logic.* College Publications, 2018. Mathematical Logic and Foundations.

[2] Pietro Baroni and Francesca Toniand and Bart Verheij. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games: 25 years later. *Argument & Computation*, 11(1-2):1–14, 2020.

[3] Trevor J.M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10):619–641, 2007.

[4] Katarzyna Budzynska and Mathilde Janier and Chris Reed and Patrick Saint-Dizier. Theoretical foundations for illocutionary structure parsing. *Argument and Computation*, 7(1):91–108, 2016.

[5] Noam Chomsky. *Aspects of the Theory of Syntax.* The MIT Press, Cambridge, 1965.

[6] Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[7] Luciano Floridi. *The Philosophy of Information.* Oxford University Press, 2011.

[8] James B. Freeman. *Argument Structure: Representation and Theory.* Springer, 2011.

[9] Jean-Yves Girard. Linear logic. *Theoretical Computer Science*, 50:1–101, 1987.

[10] Federico Gobbo and Marco Benini. *Constructive Adpositional Grammars. Foundations of Constructive Linguistics.* Cambridge Scholars Publishing, 2011.

[11] Federico Gobbo and Marco Benini. Dependency and valency. from structural syntax to constructive adpositional grammars. In *Computational Dependency Theory*, pages 113–135, 2013.

[12] Federico Gobbo and Marco Benini and Jean H.M. Wagemans. Annotation with adpositional argumentation: Guidelines for building a gold standard corpus of argumentative discourse. *Intelligenza Artificiale*, 13(2):155–172, 2019.

[13] Federico Gobbo and Marco Benini and Jean H.M. Wagemans. Complex arguments in adpositional argumentation. volume 3086 of *Advances in Argumentation in Artificial Intelligence*. CEUR Workshop Proceedings, 2021.

[14] Federico Gobbo and Jean H.M. Wagemans. Adpositional Argumentation (AdArg): A new method for representing linguistic and pragmatic information about argumentative discourse. In *Actes JIAF 2019*, pages 101–107, 2019.

[15] Federico Gobbo and Jean H.M. Wagemans. A method for reconstructing first-order arguments in natural language. In *Proceedings of AI*IA 2018*, number 2296, pages 27–41. CEUR Workshop Proceedings, 2019.

[16] Federico Gobbo and Marco Benini and Jean H.M. Wagemans. More than relata refero: Representing the various roles of reported speech in argumentative discourse. *Languages*, 7(1), 2022.

[17] Martin Hinton and Jean H.M. Wagemans. Evaluating reasoning in natural arguments: A procedural approach. *Argumentation*, 36:61–84, 2022.

[18] Marcin Lewinski and Mark Aakhus. Argumentative polylogues in a dialectical framework: A methodological inquiry. *Argumentation*, 28(2):161–185, May 2014.

[19] Per Martin-Löf. An intuitionistic theory of types: Predicative part. In H.E. Rose and J.C. Shepherdson, editors, *Logic Colloquium '73*, volume 80 of *Studies in Logic and the Foundations of Mathematics*, pages 73–118. Elsevier Science Publisher B.V., 1975.

[20] Sanjay Modgil and Katarzyna Budzynska and John Lawrence, editors. *Computational Models of Argument*, 2018.

[21] Stefan Müller. *Grammatical theory.* Number 1 in Textbooks in Language Sciences. Language Science Press, Berlin, 2020.

[22] Sara Negri and Jan von Plato and Aarne Ranta. *Structural Proof Theory.* Cambridge University Press, 2001.

[23] Henry Prakken and Stefano Bistarelli and Francesco Santini and Carlo Taticchi, editors. *Computational Models of Argument*, 2020.

[24] Helmut Schwichtenberg and Anne Sjerp Troelstra. *Basic Proof Theory*, volume 43 of *Cambridge Tracts in Theoretical Computer Science.* Cambridge University Press, 1996.

[25] Stephen Toulmin. *The uses of argument.* Cambridge University Press, Cambridge, 1958.

[26] The Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics.* `https://homotopytypetheory.org/book`, Institute for Advanced Study, 2013.

[27] Frans H. van Eemeren and Bart J. Garssen and Erik C.W. Krabbe and A. Francisca Snoeck Henkemans and H. Bart Verheij and Jean H.M. Wagemans. *Handbook of Argumentation Theory.* Springer, Dordrecht, 2014.

[28] Jean H.M. Wagemans. Constructing a Periodic Table of Arguments. In *Argumentation,*

Objectivity, and Bias: Proceedings of the 11th International Conference of the OSSA, 2016.

[29] Jean H.M. Wagemans. Why missing premises can be missed. In *Evidence, Persuasion and Diversity: Proceedings of the 12th International Conference of the OSSA*, 2020.

[30] Jean H.M. Wagemans. Four basic argument forms. *Research in Language*, 17:57–69, 2019.

[31] Jean H.M. Wagemans. *The Cambridge Handbook of the Philosophy of Language*, chapter The Philosophy of Argument. Cambridge University Press, 2021.

[32] Jean H.M. Wagemans. How to identify an argument type? on the hermeneutics of persuasive discourse. *Journal of Pragmatics*, 203:117–129, 2023.

[33] Douglas N. Walton and Christopher Reed and Fabrizio Macagno. *Argumentation Schemes.* Cambridge University Press, Cambridge, 2008.

Received 17 June 2022

A PWK-style Argumentation Framework and Expansion

Massimiliano Carrara
FISPPA Department, University of Padua, Italy
massimiliano.carrara@unipd.it

Filippo Mancini
FISPPA Department, University of Padua, Italy
filippo.mancini@unipd.it

Wei Zhu
Department of Philosophy, University of Regensburg, Germany
wei.zhu@psk.uni-regensburg.de

Abstract

In this article we consider argumentation as an epistemic process performed by an agent to extend and revise her beliefs and gain knowledge, according to the information provided by the environment. Such a process can also generate the suspension of the claim under evaluation. How can we account for such a suspension *phenomenon* in argumentation process? We propose: (1) to distinguish two kinds of suspensions – critical suspension and non-critical suspension – in epistemic change processes; (2) to introduce a Paraconsistent Weak Kleene logic (PWK) based belief revision theory which makes use of the notion of topic to distinguish the two kinds of suspensions previously mentioned, and (3) to develop a PWK-style argumentation framework and its expansion. By doing that, we can distinguish two kinds of suspensions in an epistemic process by virtue of the notion of topic.

We are highly grateful to Marcello D'Agostino, Fabio Aurelio D'Asaro, and Costanza Larese for their remarkable efforts in editing this issue. The friendly interaction and helpful advice provided by Fabio Aurelio D'Asaro made the publishing process pleasant. We would like to express our gratitude to the two anonymous referees of this issue for their insightful remarks and suggestions, which significantly improved the quality of the article. Massimiliano Carrara was partially funded for this research by the CARIPARO Excellence Project (CARR_ ECCE20_ 01): *Polarization of irrational collective beliefs in post-truth societies. How anti-scientific opinions resist expert advice, with an analysis of the antivaccination campaign* (PolPost). Wei Zhu was funded by the Postdoctoral Stipend from the Fritz Thyssen Foundation, Germany, for the project (Ref. 40.22.0.015PH)*Suspension and Computational Errors in Paraconsistent Weak Kleene Belief Update.*

Keywords: Suspension; Abstract Argumentation; Paraconsistent Weak Kleene logic (PWK); PWK Belief Revision; *Off-topic* Interpretation in PWK

1 Introduction

There is a close connection between belief revision and argumentation.[1] Here we consider one specific aspect of argumentation where this connection is viewed as an epistemic process performed by an agent to improve her beliefs and gain knowledge by acquiring some new information from the external environment: the suspension of the claim under evaluation.[2]

Such an aspect, suspension, is characterized by the absence of belief and disbelief concerning a proposition ϕ.[3] We consider it as a state of absence of judgements on propositions or arguments in the reasoning process. Moreover, following [16] we propose to distinguish two kinds of suspensions: *non-critical* suspension and *critical* suspension. When an agent neither believes nor disbelieves (or reject) certain information, such a suspension is non-critical. It is non-critical because the agent can still form a judgment or continue to process an argument as long as she gains more information from her environment. As we are going to see, such a kind of suspension can be modeled through the standard AGM model for belief revision.[4] Instead, a critical suspension occurs when an agent gains some *irrelevant*, *meaningless*, *off-topic* and even *malicious* information from the environment. This suspension cannot be held in the subsequent epistemic process, and should be filtered and set apart from the argumentation process.[5]

To better understand the two cases of *suspension* consider the following analogy with non-critical and critical errors in computation. In a computational program, a non-critical error stops the computation program partially, and this error can be fixed in the subsequent computation process. Instead, a critical error stops the program completely and this error cannot be fixed.[6] One can see the two types of suspension in terms of the two types of computational errors: *non-critical suspension* corresponds to the non-critical error, whereas *critical suspension* corresponds to the critical error.

[1] See e.g. [24], [36], [6], [8], and [3]. For a survey of argumentation theory (specifically in Artificial Intelligence), see e.g. [11], [33], and [20].

[2] See [3] where belief revision and argumentation are related and compared as two formal approaches to model reasoning processes.

[3] Some recent views consider it as a question-directed (or inquisitive) attitude [25, 26, 27].

[4] See e.g. [1].

[5] For a further discussion see §4.1 below.

[6] On this kind of computational errors see e.g. [32].

Now, in general a belief revision process, as modeled in AGM, uses classical propositional logic as its background logic. However, classical propositional logic assumes that each propositional variable has a classical truth-value – i.e. true or false. Hence, it excludes the possibility that an agent's belief state permanently stops because some propositions fail to obtain a truth-value. Moreover, it assumes that each proposition is on-topic. But in the case of a critical suspension an agent stops reasoning because it obtains some meaningless, off-topic information. This problem suggests us to change the background logic of the current belief revision theory and to make it able to filter the information so as to prevent potential critical errors from occurring during the belief revision process. Thus, in this article we develop an expansion of AGM theory based on a Paraconsistent Weak Kleene logic (PWK), where the third value of PWK is read as *off-topic*, and we conceive a PWK-style argumentation framework that is capable of distinguishing the two kinds of above mentioned suspensions in argumentation.

The present paper is organized as follows. In §2 we introduce the PWK logic and the *off-topic* interpretation of its non-classical truth-value, **u**. Also, we discuss how such an interpretation can be used to account for two kinds of suspensions occurring in argumentation: critical suspension and non-critical suspension. In §3 we present PWK belief revision (PWK-BR). In §4, we put forth a PWK abstract argumentation framework and its expansion, which is capable of distinguishing critical suspension and non-critical suspension. Finally, in §5 we make some concluding remarks.

2 PWK and the Off-topic Interpretation

In this section we will introduce two of the main elements we will need to develop our proposal: the Paraconsistent Weak Kleene logic and the off-topic interpretation of its non-classical truth value, i.e. **u**.

Traditionally, Kleene's three-valued logics divide into two families: strong and weak.[7] Weak Kleene logics, WK3, originate from weak tables (see table 1, below). Arguably, the two most important WK3 are **(author?)** [13][8] and [29]'s logics (B and H, respectively), which differ in the designated values they take on.[9] B assumes that classical truth is the only value to be preserved by valid inferences. H includes

[7] See [31].

[8] Translated in [14].

[9] There is an increasing interest in WK3. To give some examples, [19] develop sequent calculi for WK3, [34] introduce a cut-free calculus (a hybrid system between a natural deduction calculus and a sequent calculus) for PWK, [17] explores some connections between H and Graham Priest's *Logic of Paradox*, LP, and [18] focus on logical consequence in PWK.

also the non-classical value among the designated ones. Thus, it turns out that H, or better, PWK, is the paraconsistent counterpart of B. Precisely, PWK corresponds to the so-called Halldén's *internal logic*, that is a logic that includes the standard propositional connectives, but cannot express the meaningfulness of its own statements. Instead, Halldén's *external logic* extends PWK with a unary connective that allows to build statements such as "ϕ is meaningful".[10] In what follows, we will use PWK. Thus, let us briefly introduce it.

2.1 PWK

The language of PWK is the standard propositional language, L. Given a nonempty countable set $\mathsf{Var} = \{p, q, r, \dots\}$ of atomic propositions, the language is defined by the following Backus-Naur Form:

$$\Phi_L ::= p \mid \neg\phi \mid \phi \vee \psi \mid \phi \wedge \psi \mid \phi \supset \psi$$

We use $\phi, \psi, \gamma, \delta \dots$ to denote arbitrary formulas, $p, q, r, \dots$ for atomic formulas, and $\Gamma, \Phi, \Psi, \Sigma, \dots$ for sets of formulas. Propositional variables are interpreted by a valuation function $V_a : \mathsf{Var} \longmapsto \{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$ that assigns one out of three values to each $p \in \mathsf{Var}$. The valuation extends to arbitrary formulas according to the following definition:

Definition 2.1 (Valuation)**.** A valuation $V : \Phi_L \longmapsto \{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$ is the unique extension of a mapping $V_a : \mathsf{Var} \longmapsto \{\mathbf{t}, \mathbf{u}, \mathbf{f}\}$ that is induced by the tables from Table 1.

Table 1 provides the full *weak tables* from **(author?)** [31, §64], that obtain "by supplying [the third value] throughout the row and column headed by [the third value]".[11] Note that in PWK, like in the others WK3, negation works like in Strong Kleene logics, whereas conjunction and disjunction work differently. The way **u** transmits is usually called *contamination* (or *infection*), since the value propagates from any $\phi \in \Phi_L$ to any construction $k(\phi, \psi)$, independently from the value of ψ (here, k is any complex formula made out of some occurrences of both ϕ and ψ and

[10] Notice that [29] calls C_0 what we call PWK.

[11] It is clear by table 1 that $\wedge$ and $\supset$ can be defined in terms of $\neg$ and $\vee$ as usual, namely $\phi \wedge \psi = \neg(\neg\phi \vee \neg\psi)$ and $\phi \supset \psi = \neg\phi \vee \psi$. Nevertheless, we prefer to introduce them all as primitives for the sake of clarity.

ϕ	$\neg\phi$
$\mathbf{t}$	$\mathbf{f}$
$\mathbf{u}$	$\mathbf{u}$
$\mathbf{f}$	$\mathbf{t}$

$\phi \vee \psi$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$
$\mathbf{t}$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{t}$
$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$
$\mathbf{f}$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$

$\phi \wedge \psi$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$
$\mathbf{t}$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$
$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$
$\mathbf{f}$	$\mathbf{f}$	$\mathbf{u}$	$\mathbf{f}$

$\phi \supset \psi$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$
$\mathbf{t}$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{f}$
$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$	$\mathbf{u}$
$\mathbf{f}$	$\mathbf{t}$	$\mathbf{u}$	$\mathbf{t}$

Table 1: Weak tables for logical connectives in Φ_L

whatever combination of $\vee$, $\wedge$, $\supset$, and $\neg$). To better capture the way $\mathbf{u}$ works in combination with the other truth-values, let us introduce the following definition:

Definition 2.2. For any $\phi \in \Phi_L$, *var* is a mapping from Φ_L to the power set of Var, which can be defined inductively as follows:

- $var(p) = \{p\}$,
- $var(\neg\phi) = var(\phi)$,
- $var(\phi \vee \psi) = var(\phi) \bigcup var(\psi)$,
- $var(\phi \wedge \psi) = var(\phi) \bigcup var(\psi)$,
- $var(\phi \supset \psi) = var(\phi) \bigcup var(\psi)$.

Then, the following fact expresses *contamination* very clearly:

Fact 2.1 (Contamination)**.** For all formulas ϕ in Φ_L and any valuation V:

$$V(\phi) = \mathbf{u} \quad \textit{iff} \quad V_a(p) = \mathbf{u} \text{ for some } p \in var(\phi).$$

The logical consequence relation of PWK is defined as preservation of non-false values – i.e. the designated values are both $\mathbf{u}$ and $\mathbf{t}$. In other words:

Definition 2.3. $\Gamma \vDash_{\mathsf{pwk}} \Delta$ iff there's no interpretation $\mathbb{I}$ such that:

$$\mathbb{I}(\phi) \neq \mathbf{f} \text{ for all } \phi \in \Gamma \text{ and } \mathbb{I}(\psi) = \mathbf{f} \text{ for some } \psi \in \Delta.$$

PWK is reflexive and transitive. It is also monotonic (i.e. if $\Gamma \vDash_{\mathsf{pwk}} \Delta$ then $\Gamma \cup \{\alpha\} \vDash_{\mathsf{pwk}} \Delta$), but given the behaviour of conjunction in the premise side, PWK has a 'non-monotonic flavour', in the sense that, for example, $p \vDash_{\mathsf{pwk}} p$ but $p \wedge q \nvDash_{\mathsf{pwk}} p$. Further, note that the inclusion of all the atoms of a premise set Γ in a conclusion set Δ guarantees that if $\Gamma \vDash_{\mathsf{cl}} \Delta$ then $\Gamma \vDash_{\mathsf{pwk}} \Delta$, where $\vDash_{\mathsf{cl}}$ is the classical consequence relation.

2.2 Off-topic Interpretation for u

Recently, the third value **u** of WK3 – initially understood as *nonsense*, *meaninglessness* or *undefined* – has been studied in more depth. A recent proposal by [10] suggests to read **u** of WK3 as *off-topic*. More specifically: Beall proposes to "[...] read the value 1 not simply as *true* but rather as *true and on-topic*, and similarly 0 as *false and on-topic*. Finally, read the third value 0.5 as *off-topic*" [10, p. 140]. [12] Thus, *What is a topic?* is arguably a crucial question for his proposal. Unfortunately, **(author?)** [10] is silent about that. But we can make some assumptions and develop his proposal in order to make it suitable for our purposes.

We assume that topics can be represented by sets.[13] We use bold letters for topics, such as $\mathbf{s}$, $\mathbf{t}$, etc. $\subseteq$ is the inclusion relation between topics, so that $\mathbf{s} \subseteq \mathbf{t}$ expresses that $\mathbf{s}$ is included into (or is a subtopic of) $\mathbf{t}$.[14] Given that, we define a *degenerate* topic as one that is included in every topic. Also, we define the overlap relation between topics as follows: $\mathbf{s} \cap \mathbf{t}$ iff there exists a non-degenerate topic $\mathbf{u}$ such that $\mathbf{u} \subseteq \mathbf{s}$ and $\mathbf{u} \subseteq \mathbf{t}$. Further, it is assumed that every meaningful sentence α comes with a *least* subject matter, represented by $\tau(\alpha)$. $\tau(\alpha)$ is the unique topic which α is about, such that for every topic α is about, $\tau(\alpha)$ is included into it.

[12] Interestingly, a similar proposal comes from [21] and [22] where it is provided an informational semantics for three values, in which **u** is interpreted as *informationally indeterminate*.

[13] This is a natural assumption. As discussed in **(author?)** [30], topics are represented by sets in all the main approaches you can find in the literature. In this paper we take no position with respect to what exactly a topic is, that is whether a set of sets of proposition (a partition of the logical space), a set of objects, etc. We just set some constraints about how topics behave and how they relate to sentences.

[14] The inclusion relation, $\subseteq$, is usually taken to be reflexive, so that every topic includes itself.

Thus, we say that α is *exactly* about $\tau(\alpha)$.[15] But α can also be *partly* or *entirely* about other topics: α is entirely about $\mathbf{t}$ iff $\tau(\alpha) \subseteq \mathbf{t}$, whereas α is partly about $\mathbf{t}$ iff $\tau(\alpha) \cap \mathbf{t}$.

Next, we assume the following conditions concerning how topics behave with respect to the logical connectives:

1. $\tau(\phi \wedge \psi) = \tau(\phi) \cup \tau(\psi)$.
2. $\tau(\phi \vee \psi) = \tau(\phi) \cup \tau(\psi)$.
3. $\tau(\neg\phi) = \tau(\phi)$.

As shown in **(author?)** [15, §2], from these assumptions we can also prove that the topic of a complex sentence boils down to the union of the topics of its atomic components.

Further, not only do sentences have a topic, but also sets of sentences do. More in detail, we have the following:

Definition 2.4. Given a set S of sentences of Φ_L, i.e. $S \subseteq \Phi_L$, the topic of S, that is $\tau(S)$, is such that $\tau(S) = \bigcup\{\tau(\phi) \mid \phi \in S\}$.

Then, since both theories and arguments can be represented by sets of sentences, we can legitimately speak about their topics: the topic of an argument (or theory) is the union of all and only the (least) topics of each of its sentences. Thus, as for sentences, given any argument A we say that: A is exactly about $\tau(A)$; A is entirely about $\mathbf{t}$ just in case $\tau(A) \subseteq \mathbf{t}$; and A is partly about $\mathbf{t}$ just in case $\tau(A) \cap \mathbf{t}$. Moreover, as shown by **(author?)** [15, Corollary 2.2], what a set of sentences S is about boils down to the union of what the atomic components of each claims in S are about: that is, $\tau(S) = \bigcup\{\tau(p) \mid p \in var(S)\}$, where $var(S)$ is the set of all and only the atomic variables occurring in the sentences that belong to S.

Finally, let us set a reference (or discourse) topic, τ_R, that is the topic that one or more agents discuss/argue about. Then, a sentence ϕ, or an argument A, or a theory T are off-topic with respect to τ_R iff $\tau(\phi), \tau(A), \tau(T) \nsubseteq \tau_R$ – i.e. iff ϕ, A and T are not entirely about τ_R. Given such a regimentation of the notion of topic and Beall's off-topic interpretation of $\mathbf{u}$, our aim now is to use them to get an argumentation framework based on PWK.

[15] Throughout this paper, when we talk about the topic of a sentence we mean its least topic. In case we want to refer to one of its topics that is not the least one, we will make it clear.

2.3 Off-topic and Critical/Non-critical Suspensions

In §3 we integrate the off-topic interpretation of **u** into a PWK belief revision theory, based on the standard AGM model. But before we do that, it is important to point out the reason behind the development of our framework. Such an integration allows us to distinguish two kinds of suspensions that may occur in an epistemic process of change of beliefs. Since an argumentation can be represented by a set of sentences, in line with Definition 2.4 we assume that an agent's argumentation process has a topic – i.e. the reference topic.

Let's take an example. Suppose that an argumentation process is about the topic represented by the question "How many stars are there?". An argument like "There is an infinite number of stars in the universe because it is infinite in space" is an on-topic one in the argumentation process, which should participate in the argumentation process. However, an argument like "Alice is in wonderland because I read about it in a book" is an off-topic one in the argumentation process, which should be filtered and set apart from the argumentation process. Let us make an example to show how an off-topic argument can be harmful to the reasoning process. Suppose there are three arguments in the argumentation process whose topic is "How many stars are there":

(1) "100 stars are in the sky"

(2) "Alice is in wonderland or there are no stars in the sky"

(3) "Alice is not in wonderland"

If we do not set apart off-topic arguments from on-topic ones, from (2) and (3) we can derive "there are no stars in the sky", which is in conflict with (1). If we set apart (2) and (3) from the argumentation process as off-topic arguments, we can derive that "100" is the conclusion.[16]

Given a reference topic, an epistemic agent's argument can be either on-topic or off-topic with respect to it. If the argument is off-topic, we get a critical-suspension of the conclusion of the argumentation process. In other words, the epistemic agent assigns **u** to the claim that is meant to be the conclusion of the argument at stake. If it is on-topic, the conclusion might be believed, disbelieved or non-critically suspended, depending on how the argument works and on there being other good arguments attacking such conclusion – i.e. depending on the argumentation framework in which the epistemic agent performs her argumentation process. In particular, a conclusion is non-critically suspended just in case it generates a contradiction, that

[16] We express our gratitude to a referee who proposed this example.

is if we can draw both that conclusion and its negation from our set of beliefs. In that case, the suspension is non-critical in the sense that the claim under evaluation is neither believed nor disbelieved, and it remains available to be processed in a further argumentation process where new (on-topic) information is acquired.

3 A PWK Based Belief Revision Theory

The next step is to enter belief-revision. This is the process through which an ideal rational agent revises her own beliefs to get an ever-improving understanding of the world, i.e. a better representation of it. How does this process work? There is a well-known formal account that gives a model of it: the AGM theory. Here, we aim at developing a different version of belief revision: a PWK based belief revision theory (PWK-BR). Now, since our PWK-BR is based on (and can be seen as an expansion of) the AGM theory, let us start by quickly introducing AGM.

3.1 AGM Theory

Among all the belief revision theories, AGM theory is widely recognized as a milestone. It was initialized by [1] and soon developed by [28]. The main question of AGM belief revision theory is: in order to accommodate new information which is contradictory to an agent's own beliefs, how to get rid of the inconsistency as well as minimizing the information loss? To solve this problem, a worked out formal epistemology of belief revision theory is required. Basically, such a theory needs consider the following essential components, which are: a formal representation of epistemic states; a classification of the epistemic attitudes; an account of the epistemic inputs and a classification of epistemic changes; and a criterion of rationality. Thus, the main framework of AGM theory can be listed as follows:

1. An agent's belief state is formalized as a belief set Θ, which is closed under the consequence operation Cn. Since AGM theory adopts classical propositional logic, Cn is $\vDash_{\mathsf{cl}}$ in this regard. Specifically, the definitions of the consequence operation and belief set are as follows.

   Definition 3.1 (Consequence Operation Cn)**.** A consequence operation on a language $\mathcal{L}$ is a function Cn that takes each subset of $\mathcal{L}$ to another subset of $\mathcal{L}$, such that:

(a) $A \subseteq Cn(A)$.

(b) $Cn(A) = Cn(Cn(A))$.

(c) If $A \subseteq B$, *then* $Cn(A) \subseteq Cn(B)$.

Definition 3.2 (Belief Set Θ)**.** Θ is a set of sentences. It is a belief set if it is closed under Cn. That is, $\Theta = Cn(\Theta)$.

2. An agent has three kinds of epistemic attitudes, which are: belief, disbelief and suspension. Suspension in fact is not an attitude but a lack of attitude, called non-attitude. For writing convenience, we call it is an attitude. These attitudes are exclusive and exhaustive. Hence, a sentence is believed, disbelieved or kept in suspension.

3. In AGM, an epistemic input is regarded to be external, in terms of a new sentence from the environment.

4. Three basic kinds of epistemic change operators are expansion, contraction, and revision. Since the aim is to model the process of belief-revision, some operations on Θ representing the belief changes can be defined. In the AGM account, there are three: expansion (+), contraction (−), and revision ($*$).

    - Expansion models the addition of a belief, say α, when nothing is removed: Θ is replaced by $\Theta + \alpha$, that is the smallest logically closed set containing both Θ and α. Thus, $\Theta + \alpha = \{\beta : \Theta \cup \{\alpha\} \vDash \beta\}$, where $\vDash$ denotes the selected consequence relation.
    - Contraction models the removal of a belief. This is not just to delete α from Θ. Since the result must be logically closed, we may have to delete other things as well. From Θ we get $\Theta - \alpha$, that is a set such that $\Theta - \alpha \subseteq \Theta$ and that $\alpha \notin \Theta - \alpha$, but this change can be accomplished in different ways – i.e. there are many sets $\Theta - \alpha$ satisfying these conditions. The AGM account does not give an explicit definition of contraction but gives a set of axioms that $\Theta - \alpha$ must satisfy, the so-called basic AGM postulates.
    - Finally, revision models the addition of a belief to Θ when other sentences have to be removed to ensure that the resulting set of beliefs, $\Theta * \alpha$, is consistent. As for contraction, also revision has been axiomatically characterized.

5. The rational criterion of AGM belief revision theory is the principle of information economy, which requires an agent to accommodate new information and at the same time to minimize the loss of the original beliefs. This criterion is resulted from the fact that data are valuable. It is better to preserve as much data as possible. To ensure this criterion, AGM postulates are developed to regulate the performance of the belief change operators.

Let us now turn to our different belief revision theory: PWK-BR.

3.2 Belief States in PWK-BR

Differently from the AGM belief set, in PWK-BR an agent's belief state concerns a topic. An agent's epistemic attitude toward a given proposition α depends on whether α is on-topic or off-topic with respect to a given reference topic – i.e. the topic of the argumentation process she is performing. If α is on-topic, the agent can believe it, disbelieve it, or keep it in non-critical suspension. If α is off-topic, the agent would keep it in critical suspension. Non-critical suspension and critical suspension are two exclusive attitudes:

1) If α is in a non-critical suspension, α is still available to be believed or disbelieved by the agent in a subsequent process of belief revision triggered by new information.

2) If α is off-topic – i.e. it is a piece of irrelevant information –, then it should be isolated from the current belief change process and kept in critical suspension, with no chance to change its belief-status, unless the reference topic is changed.

Let us then define a belief state in PWK-BR:

Definition 3.3 (Belief State in PWK-BR)**.** An agent's belief state is a triple $\langle \Theta, \Delta, \Sigma \rangle$. Θ, Δ and Σ are all sets of propositions of $\Phi_{\mathcal{L}}$ (i.e. Θ, Δ, $\Sigma \subseteq \Phi_{\mathcal{L}}$), such that:

a) a belief set is defined as $\Theta = \{\alpha : \Theta \vDash_{\mathsf{pwk}} \alpha, \alpha \in \Phi_{\mathcal{L}}\} \smallsetminus \{\alpha : \alpha \text{ is off-topic}, \alpha \in \Phi_{\mathcal{L}}\}$, that is Θ is PWK-logically closed and does not have any off-topic proposition as member;

b) a non-critical suspension set is defined as $\Delta \subseteq \{\beta : \beta \text{ is on-topic}\}$ and $\Delta = \Delta \bigcup \{\neg\beta \mid \beta \in \Delta\}$, for any $\beta \in \Delta$;

c) a critical suspension set is defined as $\Sigma \subseteq \{\gamma : \gamma \text{ is off-topic}\}$;

d) are exclusive, but not necessarily exhaustive: $\Theta \cap \Delta = \Theta \cap \Sigma = \Delta \cap \Sigma = \varnothing$ and $\Theta \cup \Delta \cup \Sigma \subseteq \Phi_L$.

3.3 Expansion, Contraction and Revision in PWK-BR

In PWK-BR, expansion, contraction and revision are three operations that take both a belief state and a proposition as input, and output a new belief state. Specifically, we define such operators as follows:

Definition 3.4 (Expansion $\oint^{+}$)**.** The expansion of a belief state $\langle \Theta, \Delta, \Sigma \rangle$ with respect to a new proposition ϕ is represented by an operator defined from $\langle \langle \mathscr{P}(\Phi_L), \mathscr{P}(\Phi_L), \mathscr{P}(\Phi_L) \rangle, \Phi_L \rangle$ to $\langle \mathscr{P}(\Phi_L), \mathscr{P}(\Phi_L), \mathscr{P}(\Phi_{\mathcal{L}}) \rangle$, such that:

$$\oint^{+}(\langle \Theta, \Delta, \Sigma \rangle, \phi) = \begin{cases} \langle \Theta + \phi, \Delta, \Sigma \rangle & \textit{if } \phi \textit{ is on-topic,} \\ \langle \Theta, \Delta, \Sigma \cup \{\phi\} \rangle & \textit{if } \phi \textit{ is off-topic.} \end{cases}$$

where + is the AGM-expansion.[17]

Definition 3.5 (Contraction $\oint^{-}$)**.** The contraction of a belief state $\langle \Theta, \Delta, \Sigma \rangle$ with respect to a new proposition ϕ is represented by an operator defined from $\langle \langle \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}) \rangle, \Phi_{\mathcal{L}} \rangle$ to $\langle \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}) \rangle$, such that:

$$\oint^{-}(\langle \Theta, \Delta, \Sigma \rangle, \phi) = \begin{cases} \langle \langle \Theta, \Delta \rangle - \phi, \Sigma \rangle & \textit{if } \phi \textit{ is on-topic,} \\ \langle \Theta, \Delta, \Sigma \rangle & \textit{if } \phi \textit{ is off-topic.} \end{cases}$$

where − is the AGM-contraction.

Definition 3.6 (Revision $\oint^{*}$)**.** The revision of a belief state $\langle \Theta, \Delta, \Sigma \rangle$ with respect to a new proposition ϕ is represented by an operator defined from $\langle \langle \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}) \rangle, \Phi_{\mathcal{L}} \rangle$ to $\langle \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}), \mathscr{P}(\Phi_{\mathcal{L}}) \rangle$, such that:

[17] $\mathscr{P}$ denotes a power set, which applies to all its occurrences in this article.

$$\oint^{*}(\langle\Theta,\Delta,\Sigma\rangle,\phi)=\begin{cases}\langle\langle\Theta,\Delta\rangle*\phi,\Sigma\rangle & \textit{if}\ \phi\ \textit{is on-topic,}\\ \langle\Theta,\Delta,\Sigma\cup\{\phi\}\rangle & \textit{if}\ \phi\ \textit{is off-topic.}\end{cases}$$

where $*$ is the AGM-revision.

All the AGM postulates can be preserved in PWK-BR. Therefore, PWK-BR counts as an extension of AGM theory. This is ensured by the following theorem, the proof of which can be found in **(author?)** [16]:

Theorem 3.1. AGM postulates agree with the PWK-BR.

Proof. According to the definitions, AGM operators are adopted to deal with the on-topic part of PWK belief change. $+$, $-$, and $*$ are embedded into $\{\oint^{+},\oint^{-},\oint^{*}\}$. As long as AGM operators follow AGM postulates, $\{\oint^{+},\oint^{-},\oint^{*}\}$ do as well. Therefore, AGM postulates, which regulate $\{\oint^{+},\oint^{-},\oint^{*}\}$, also support this PWK belief change framework based on $\{\oint^{+},\oint^{-},\oint^{*}\}$. □

4 PWK Abstract Argumentation Framework and Expansion

4.1 Motivating Ideas

In this section we put forward our proposal to account for suspension in a PWK-based argumentation process. Our main considerations are as follows.

First, suspensions should be analyzed and identified in an argumentation theory. Given that (1) both belief revision and argumentation theory are important approaches in knowledge representation to formalize epistemic processes, and that (2) suspensions are identified and distinguished in a PWK belief revision theory, suspensions can be considered in argumentation theory just as they are considered in belief revision theory (recall the discussion in §2.3). One distinction between a belief revision process and an argumentation process lies in their starting points. A belief revision process assumes a consistent set of propositions, while an abstract argumentation process starts with a set of arguments related by binary attack relations. We put forth two suggestions regarding the two different types of suspensions. 1) A non-critical suspension in an abstract argumentation framework occurs when all arguments in the framework are self-attacking. For instance, an argument is self-attacking if its conclusion contradicts one of its premise. (See [9, 4, 12] for recent discussions about self-attacking arguments.) This is problematic because self-attacking arguments cannot be used to justify any other argument. To address this issue, the attack relations connected to these self-attacking arguments should be removed, except for their own self-attack loop. When an argumentation process is suspended in this way, no conflict-free subset of the framework exists. As a result, there are no admissible, grounded, ideal, preferred, or stable extensions in the framework. 2) A critical suspension in an abstract argumentation framework occurs when certain arguments are irrelevant to the topic being discussed in the argumentation process. The reason why a critical suspension in an abstract argumentation framework is important is that it can use up the computational resources and lead the argumentation process to arrive at an incorrect conclusion. In this case, the attack relations of these off-topic arguments should be set apart from the argumentation process.

Second, to analyze suspensions in an argumentation process we can take [16]'s proposal as a plausible approach. It analyzes suspensions in an epistemic change process on the basis of PWK logic with Beall's off-topic interpretation and AGM theory. Similarly, we can consider two kinds of suspensions in an argumentation process by relying on the notion of topic. As discussed in §2.3, our assumption is that an argumentation process has a topic – the reference topic – corresponding to a set of answers to a specific question. Any off-topic epistemic inputs would result in a major interruption of the argumentation process because it is important that it stays on topic without introducing any unrelated information.

Third, it is a feasible task to account for suspensions in argumentation theory by bringing together Dung's abstract argumentation theory and PWK-BR. Dung's argumentation theory and AGM theory have been integrated in a whole comprehensive framework corresponding to the AGM-style abstract argumentation theory (see e.g. [5, 6, 7, 8]). Thus, we claim that a PWK-style abstract argumentation framework can be developed in a similar way from PWK-BR and Baumann, Brewka and Linker's works.

Last, a PWK-style abstract argumentation framework is worth investigating. It is not just an aimless technical integration of all the previously mentioned works, but an integrated view that enables us to account for different kinds of suspensions in argumentation. Since suspension is an important *phenomenon* actually occurring in argumentation processes, the development of a PWK-style argumentation framework is a worthwhile enterprise.

4.2 PWK Abstract Argumentation Framework

Given the discussions above, we propose a PWK abstract argumentation framework that has a topic t on the basis of Dung's abstract argumentation framework and some recent proposals concerning integrating Dung's abstract argumentation theory with AGM theory.[18] Let's start by outlining some fundamental definitions of argumentation frameworks before defining a PWK abstract argumentation framework.

Definition 4.1 (Argumentation framework AF [2])**.** An argumentation framework AF is a pair $\langle \mathsf{Ar}, \mathsf{att} \rangle$ in which Ar is a finite set of arguments and $\mathsf{att} \subseteq \mathsf{Ar} \times \mathsf{Ar}$.

Given an argumentation framework $\mathsf{AF} = \langle \mathsf{Ar}, \mathsf{att} \rangle$ and $\mathsf{Args} \subseteq \mathsf{Ar}$, for any arguments $a, b \in \mathsf{Ar}$, $(a, b) \in \mathsf{att}$ is to be read as "a attacks b"; a attacks Args iff there is $b \in \mathsf{Args}$ such that $(a, b) \in \mathsf{att}$; Args attacks a iff there is $b \in \mathsf{Args}$ such that $(b, a) \in \mathsf{att}$; Args attacks $\mathsf{Args}' \subseteq \mathsf{Ar}$ iff there are $a \in \mathsf{Args}, b \in \mathsf{Args}'$, such that $(a, b) \in \mathsf{att}$.

Definition 4.2 (PWK abstract argumentation framework)**.** A PWK abstract argumentation framework is a triple $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t \rangle$ where Ar is a finite set of abstract arguments, $\mathsf{att} \subseteq \mathsf{Ar} \times \mathsf{Ar}$ is the attack relation, and t is a set of topics, such that:[19]

1. For any arguments $a, b \in \mathsf{Ar}$, a attacks b if $(a, b) \in \mathsf{att}$.
2. $a \in \mathsf{Ar}$ is an on-topic argument if a's topic belongs to t – i.e., $\tau(a) \in t$.
3. $b \in \mathsf{Ar}$ is an off-topic argument if b's topic does not belong to t – i.e., $\tau(b) \notin t$.

Let us explain some assumptions concerning the definition above. First of all, we take every argument $a \in \mathsf{Ar}$ to be an abstract atomic argument, which means we do not assume any specific structure on such arguments. This is in line with [23]. As [3] remarks:

> While the word *argument* may recall several intuitive meanings, abstract argumentation frameworks are not (even implicitly or indirectly) bound to any of them: an abstract argument is not assumed to have any specific structure but, roughly speaking, an argument is anything that may

[18] See [1, 23, 6, 8]

[19] We will use the symbol $\mathscr{PAF}$ to denote the set of all PWK argumentation frameworks.

> attack or be attacked by another argument, where, again, no specific meaning is ascribed to the notion of attack.
>
> [3, p. 12]

In an abstract argumentation framework, the arguments are indeed abstract, which means that they can be adapted to suit different theories about arguments. We take the notion of argument in this abstract way in our framework.

Next, we have made an important assumption for the PWK abstract argumentation framework: that is, any argument $a \in \mathsf{Ar}$ is about only one unique topic. The reason for making such an assumption is intuitive: we want to keep the idea simple enough to be understood. This is helpful for us to clarify our ideas. Indeed, such an assumption limits the possibility that a can be about several different topics. We will confine our discussion to this limited scope in this article.

In order to express being "off-topic", we assume a set of topics, t, in a PWK abstract argumentation framework, which is a collection of abstract single topics. For a similar reason, we do not assume that t has any specified structures to express the connections between topics. As any argument corresponds to one topic, it either belongs to t, or does not belong to t. For brevity, we use $a \in_t t$ to denote that an argument a's topic is included in t. In other words, $a \in_t t$ if and only if $\tau(a) \in t$.[20] Hence we can express what an on-topic argument is. That is, a is on-topic of t iff $a \in_t t$; otherwise a is off-topic.

Last, we preserve [23]'s abstract relation att in the Definition 4.2: we do not assume any specific meaning to the notion *attack*. It just has a form of a binary relation between arguments and does not embody any form of evaluation ([35]). In a PWK abstract argumentation framework, att can be between any arguments, regardless of their being on-topic or off-topic.

Let us make an example concerning Definition 4.2.

Example 4.1. Let a PWK argumentation framework be $\langle \mathsf{Ar}, \mathsf{att}, t \rangle$, where $\mathsf{Ar} = \{a, b, c, d\}$, $\mathsf{att} = \{(a, b), (b, d), (d, a)\}$, and $a, b, c \in_t t$, $d \notin_t t$.

This example shows a PWK argumentation framework that has four arguments a, b, c, d and a set of topics t, where a, b, c are on-topic arguments and d is an off-topic argument. a attacks b; b attacks d; d attacks a.

[20] $\in_t$ can be specified in different ways, according to a specific theory of topic. For example, it can be defined by a set of judgment rules that recognize whether an argument belongs to the set of topics t or not.

Next, given the discussion in the previous section, we can establish a classification for two types of suspension within a PWK framework $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$.

Definition 4.3 (Suspension)**.** Let $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK argumentation framework.

1. A suspension is classified as non-critical if all the on-topic arguments in Ar attack themselves, meaning that for any $a \in \mathsf{Ar}$ and $a \in_t t$, $(a, a) \in \mathsf{att}$.
2. A suspension is classified as critical if certain arguments in Ar deviate from the argumentation topic, that is, there exists $b \in A$, such that $b \notin_t t$.

What are the outcomes resulting from these two classifications of suspension? To understand this better, we can define the extensions of a PWK framework.

Definition 4.4 (Extension)**.** Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ and $\mathsf{Args} \subseteq \mathsf{Ar}$.

1. Args is a conflict-free on-topic extension if and only if Args does not attack itself and for any $a \in \mathsf{Args}$, a is on-topic. That is, $(a, b) \notin \mathsf{att}$ and $a \in_t t$ for all $a, b \in \mathsf{Args}$.
2. Args is an admissible on-topic extension if and only if Args is a conflict-free on-topic extension and Args defends all its elements.
3. Args is a complete on-topic extension if and only if Args is a conflict-free on-topic extension and the set of on-topic arguments defended by Args is equal to Args.
4. Args is a preferred on-topic extension if abd only if Args is an admissible on-topic extension and for no admissible on-topic extension Args', $\mathsf{Args} \subsetneq \mathsf{Args}'$.
5. Args is a grounded on-topic extension if and only if Args is the minimal complete on-topic extension. That is, Args is an complete on-topic extension and there is no complete $\mathsf{Args}' \subseteq \mathsf{Ar}$, such that $\mathsf{Args}' \subsetneq \mathsf{Args}$.
6. Args is a stable on-topic extension if and only if Args is a complete on-topic extension that attacks any on-topic argument in $\mathsf{Ar} \setminus \mathsf{Args}$.

Lemma 4.1. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK argumentation framework. It is considered to be in a non-critical suspension if and only if there does not exist any conflict-free on-topic extension for $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$.

Proof. From the left to the right: when $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$ is kept in a non-critical suspension, then for any $a \in \mathsf{Ar}$ and $a \in_t t$ such that $(a, a) \in \mathsf{att}$. As a result, it is impossible to include any such a in a conflict-free on-topic $\mathsf{Args} \subseteq \mathsf{Ar}$, because of the presence of $(a, a) \in \mathsf{att}$. This implies that there are no conflict-free on-topic extensions possible for Ar. From the right to the left: when $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$ does not contain any conflict-free on-topic extensions, then for any $a \in A$ and $a \in_t t$ such that the singleton set $\{a\}$ is not conflict-free. Therefore, it follows that a attacks itself, meaning that $(a, a) \in \mathsf{att}$. □

This outcome is comprehensible because if every on-topic argument within an argumentation framework attacks itself, then it becomes impossible to draw any conclusions from them. Non-critical suspension are considered problematic, because it makes the argumentation framework uninformative by undermining all of the arguments and limiting its ability to provide rational conclusions. Therefore, it is important to prevent non-critical suspension to ensure that the argumentation framework remains informative.

Lemma 4.2. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK argumentation framework. Let any argument $a \in \mathsf{Ar}$ such that $a \notin_t t$, then there does not exist any conflict-free on-topic extension.

Proof. The proof is evident. In case there are no on-topic arguments within the framework, it is impossible to have any conflict-free on-topic extensions. □

Compared to non-critical suspensions, critical suspensions can be less apparent if we do not evaluate whether an argument is on-topic or off-topic. Lemma 4.2 shows that if all arguments in the framework are under critical suspension, then it becomes impossible to have any conflict-free on-topic extensions, thereby undermining the framework.

Lemma 4.3. Any abstract argumentation framework $\langle \mathsf{Ar}, \mathsf{att}\rangle$ can be extended to a PWK abstract argumentation framework $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$ by specifying a set of topics t and a membership relation $\in_t$ between an argument and t.

Proof. Let $\langle \mathsf{Ar}, \mathsf{att}\rangle$ be an abstract argumentation framework and t be a set of topics. $\langle \mathsf{Ar}', \mathsf{att}', t\rangle$ is derived from $\langle \mathsf{Ar}, \mathsf{att}\rangle$ if $\langle \mathsf{Ar}', \mathsf{att}', t\rangle$ satisfies the following conditions:

1. $\mathsf{Ar}' = \mathsf{Ar}$ and $\mathsf{att}' = \mathsf{att}$;

2. for any argument $a \in \mathsf{Ar}'$, a is on-topic if $\tau(a) \in t$; otherwise, a is off-topic.

Since $\langle \mathsf{Ar}', \mathsf{att}', t\rangle$ satisfies Definition 4.2, it is a PWK argumentation framework, which extends $\langle \mathsf{Ar}, \mathsf{att}\rangle$ by specifying a set of topics t. □

To see this lemma clear, let us make the following example that shows a PWK argumentation framework generated by specifying a set of topics t. As we discuss before, we try to keep t as simple as possible, and thus we do not specify a method that generates a t. In the following example, t is generated by selecting some topics from the arguments' topics. This is not necessarily the only way to generate a t.

Example 4.2. Let an abstract argumentation framework be $\langle \mathsf{Ar}, \mathsf{att}\rangle$, where $\mathsf{Ar} = \{a, b, c, d\}$, $\mathsf{att} = \{(a, b), (b, c), (d, d)\}$. Let $t = \{\tau(a)\} \cup \{\tau(b)\}$. Then a PWK abstract argumentation framework is $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$, where $a, b \in_t t$ are on-topic. c, d are on-topic if $\tau(c), \tau(d) \in_t t$.

Note that we do not delete any argument and any attack relations to derive a PWK abstract argumentation framework from any abstract argumentation framework. We just add a set of topics t that distinguishes on-topic arguments from off-topic ones.

4.3 PWK Argumentation Expansion

To expand a PWK argumentation framework, we use the method described in [6] and define a kind of σ-kernel that makes constrains on the attack relation by removing from att certain attack relations that are related with off-topic arguments. To do that, let us introduce the extension-based semantics, and the definition of σ-kernel, which is a sub-framework of $\langle \mathsf{Ar}, \mathsf{att}\rangle$ that meets specific requirements regarding att.

According to [6, 8], given any $\mathsf{AF} = \langle \mathsf{Ar}, \mathsf{att}\rangle$, σ is a function that assigns to AF a set of sets of *arguments* denoted by $\sigma(\mathsf{AF}) \subseteq 2^{\mathsf{Ar}}$. Generally, there are six basic kinds of σ extensions: conflict-free, admissible, complete, preferred, grounded, and stable extensions. For example, a conflict-free extension of AF is $cf(\mathsf{AF})$: $\mathsf{Args} \in cf(\mathsf{AF})$ iff for all $a, b \in \mathsf{Args}$, $(a, b) \notin \mathsf{att}$. Following this idea, we can define an on-topic extension for Paf as follows.

Definition 4.5 (On-topic extension)**.** Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK abstract argumentation framework and $\mathsf{Args} \subseteq \mathsf{Ar}$. Then, an on-topic extension for Paf, $on(\mathsf{Paf})$, is such that $\mathsf{Args} \in on(\mathsf{Paf})$ iff for any $a, b \in \mathsf{Args}$, $a, b \in_t t$.

As a result, $on(\mathsf{Paf})$ is a set of any subset of Ar that contains only the on-topic arguments. Next, we can define a σ-kernel, $k(\sigma)$, from $\mathscr{PAF}$ to $\mathscr{PAF}$ by removing certain attack relations from a Paf, such that $\mathsf{Paf}^k(\sigma) = \langle \mathsf{Ar}, \mathsf{att}^{k(\sigma)}, t\rangle$. In particular, let us define a t-kernel, namely $k(t)$, which removes the attack relations from Paf that are related with off-topic arguments. Before that, let us define the following relations between any PWK argumentation frameworks.

Definition 4.6. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ and $\mathsf{Paf}^* = \langle \mathsf{Ar}^*, \mathsf{att}^*, t\rangle$ be two PWK argumentation frameworks that have the same set of topics t.

1. $\mathsf{Paf} \subseteq_t \mathsf{Paf}^*$ if and only if $\mathsf{Ar} \subseteq \mathsf{Ar}^*$, $\mathsf{att} \subseteq \mathsf{att}^*$.

2. $\mathsf{Paf} =_t \mathsf{Paf}^*$ if and only if $\mathsf{Paf} \subseteq_t \mathsf{Paf}^*$ and $\mathsf{Paf}^* \subseteq_t \mathsf{Paf}$.

Definition 4.7 ($k(t)$)**.** Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK abstract argumentation framework, and $k(t)$ is an t kernel function, such that $\mathsf{Paf}^{(k(t))} = \langle \mathsf{Ar}, \mathsf{att}^{k(t)}, t\rangle$ and $\mathsf{att}^{k(t)} = \mathsf{att} \setminus \{(a,b) \mid a \notin_t t \text{ or } b \notin_t t\}$.

Next, we shall define a set of on-topic models for a PWK argumentation framework Paf, called $k(t)$-models ($Mod^{k(t)}$). A model of a Paf is an argumentation framework related to Paf that satisfies certain conditions.

Definition 4.8. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK argumentation framework. The set of $k(t)$-models of Paf is defined as $Mod^{k(t)}(\mathsf{Paf}) = \{\mathsf{Paf}^* \mid \mathsf{Paf}^{k(t)} \subseteq_t \mathsf{Paf}^{*k(t)}\}$.

To understand Definition 4.8 better, let us take an example.

Example 4.3. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ be a PWK argumentation framework, where $\mathsf{Ar} = \{a, b, c\}$, $\mathsf{att} = \{(b,a), (b,c)\}$, and $a, b \in_t t$. Then $\mathsf{Paf}^* = \langle \mathsf{Ar}^*, \mathsf{att}^*, t\rangle \in Mod^{k(t)}(\mathsf{Paf})$, where $\mathsf{Ar}^* = \{a, b, c\}$, $\mathsf{att}^* = \{(b,a), (c,b)\}$, and $a, b \in_t t$.

Next, we consider how a PWK abstract argumentation framework $\langle \mathsf{Ar}, \mathsf{att}, t\rangle$ expands itself with respect to another framework $\langle \mathsf{Ar}^*, \mathsf{att}^*, t^*\rangle$ argumentation framework under a same topic. We denote such an operator as $+^{k(t)}$.

Definition 4.9. Let $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ and $\mathsf{Paf}^* = \langle \mathsf{Ar}^*, \mathsf{att}^*, t^*\rangle$ be two PWK abstract argumentation frameworks. A function $\mathsf{Paf} +^{k(t)} \mathsf{Paf}^*$ is a $k(t)$-expansion if and only if $Mod^{k(t)}(\mathsf{Paf} +^{k(t)} \mathsf{Paf}^*) = Mod^{k(t)}(\mathsf{Paf}) \cap Mod^{k(t)}(\mathsf{Paf}^*)$ and $t = t^*$.

Definition 4.9 constrains the expansion of PWK argumentation to a specific set of topics, represented by $t = t^*$. As a result, the set of topics remains the same after the expansion. Moreover, the attack relations that are related to only on-topic arguments are preserved. After the expansion, the off-topic arguments are still considered off-topic, under the same set of topics $t = t^*$.

Theorem 4.1. For any PWK argumentation framework $\mathsf{Paf} = \langle \mathsf{Ar}, \mathsf{att}, t\rangle$ and $\mathsf{Paf}^* = \langle \mathsf{Ar}^*, \mathsf{att}^*, t^*\rangle$, there exists an $\mathsf{Paf}' = \langle \mathsf{Ar}', \mathsf{att}', t'\rangle$, such that
$Mod^{k(t)}(\mathsf{Paf}') = Mod^{k(t)}(\mathsf{Paf} +^{k(t)} \mathsf{Paf}^*)$ if $t = t^* = t'$. *Moreover, if* $Mod^{k(t)}(\mathsf{Paf} +^{k(t)} (\mathsf{Paf}^*) \neq \varnothing$, *then* $\mathsf{Paf}'^{k(t)} = \mathsf{Paf}^{k(t)} \cup \mathsf{Paf}^{*k(t)}$.

Proof. According to Definition 4.9, for any $k(t)$ expansion, $Mod^{k(t)}(\mathsf{Paf} +^{k(t)} \mathsf{Paf}^*) = Mod^{k(t)}(\mathsf{Paf}) \cap Mod^{k(t)}(\mathsf{Paf}^*)$ and $t = t^*$. Therefore, if the intersection $Mod^{k(t)}(\mathsf{Paf}) \cap Mod^{k(t)}(\mathsf{Paf}^*) \neq \langle \varnothing, \varnothing, t\rangle$. Then for any $\mathsf{Paf}^\circ \in Mod^{k(t)}(\mathsf{Paf}) \cap Mod^{k(t)}(\mathsf{Paf}^*)$, $Mod^{k(t)}(\mathsf{Paf}^\circ)$ equals to $Mod^{k(t)}(\mathsf{Paf} +^{k(t)} \mathsf{Paf}^*)$. □

Two PWK argumentation frameworks can be incorporated through PWK argumentation expansion under the same set of topics. Such expansion is different from a PWK belief set expansion operation, because in PWK belief set expansion, the off-topic sentences are collected in Σ. However, for a PWK argumentation framework expansion, all the off-topic arguments are kept in the argument set Ar in an isolated way: that is, the attack relations between any off-topic argument and any on-topic argument are deleted to avoid the off-topic arguments from a argumentation process which is around a topic.

5 Concluding Remarks

In this article we have presented the basic elements of a PWK-style argumentation framework that extends the abstract argumentation framework and makes a distinction between two kinds of suspension. the AGM belief revision model with two kinds of suspension. What is next? In future works we would like to have a *precise* model to distinguish whether an argument is on-topic or off-topic, by using a game-theoretic semantics, as we have done in other works [15].

References

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

Pietro Baroni, Martin Caminada, and Massimiliano Giacomin. An introduction to argumentation semantics. *The knowledge engineering review*, 26(4):365–410, 2011.

Pietro Baroni, Eduardo Fermé, Massimiliano Giacomin, and Guillermo Ricardo Simari. Belief Revision and Computational Argumentation: A Critical Comparison. *Journal of Logic, Language and Information*, 31(4):555–589, December 2022.

Inconsistent Datalog Knowledge Bases. Sets of attacking arguments for inconsistent datalog knowledge bases. *Computational Models of Argument: Proceedings of COMMA 2020*, 326:419, 2020.

Ringo Baumann. Normal and strong expansion equivalence for argumentation frameworks. *Artificial Intelligence*, 193:18–44, 2012.

Ringo Baumann and Gerhard Brewka. AGM meets abstract argumentation: expansion and revision for Dung frameworks. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Ringo Baumann and Gerhard Brewka. The equivalence zoo for Dung-style semantics. *Journal of Logic and Computation*, 28(3):477–498, 2018.

Ringo Baumann and Felix Linker. AGM meets abstract argumentation: Contraction for Dung frameworks. In *European Conference on Logics in Artificial Intelligence*, pages 41–57. Springer, 2019.

Ringo Baumann and Stefan Woltran. The role of self-attacking arguments in characterizations of equivalence notions. *Journal of Logic and Computation*, 26(4):1293–1313, 2016.

Jc Beall. Off-topic: A new interpretation of weak-Kleene logic. *The Australasian Journal of Logic*, 13(6):136–142, 2016.

Trevor JM Bench-Capon and Paul E Dunne. Argumentation in artificial intelligence. *Artificial intelligence*, 171(10-15):619–641, 2007.

Vivien Beuselinck, Jérôme Delobelle, and Srdjan Vesic. A principle-based account of self-attacking arguments in gradual semantics. *Journal of Logic and Computation*, 33(2):230–256, 2023.

D. Bochvar. On a three-valued calculus and its application in the analysis of the paradoxes of the extended functional calculus. *Matamaticheskii Sbornik*, 4:287–308, 1938.

Dimitri Anatolevich Bochvar and Merrie Bergmann. On a three-valued logical calculus and its application to the analysis of the paradoxes of the classical extended functional calculus. *History and Philosophy of Logic*, 2(1-2):87–112, 1981.

Massimiliano Carrara, Filippo Mancini, and Wei Zhu. A topic game theoretical semantics (TGTS) for PWK. *Manuscript Submitted*, 2022.

Massimiliano Carrara and Wei Zhu. Computational errors and suspension in a PWK epistemic agent. *Journal of Logic and Computation*, 31(7):1740–1757, 2021.

R. Ciuni. Conjunction in paraconsistent weak Kleene logic. In P. Arazim and M. Dancák, editors, *Logica Yearbook 2014*, pages 61–76, London, 2015. College Publications.

R. Ciuni and M. Carrara. Characterizing logical consequence in paraconsistent weak Kleene. In L. Felline, A. Ledda, F. Paoli, and E. Rossanese, editors, *New Developments in Logic and the Philosophy of Science*, pages 165–176, London, 2016. College Publications.

M. E. Coniglio and M.I. Corbalan. Sequent calculi for the classical fragment of Bochvar and Halldén's nonsense logic. In D. Kesner and Petrucio, V., editors, *Proceedings of the 7th LSFA Workshop*, Electronic Proceedings in Computer Science, pages 125–136, 2012.

M. D'Agostino and S. Modgil. Classical logic, argument and dialectic. *Artificial Intelligence*, 262:15–51, 2018.

Marcello D'agostino. Informational semantics, non-deterministic matrices and feasible deduction. *Electronic Notes in Theoretical Computer Science*, 305:35–52, 2014.

Marcello D'Agostino. An informational view of classical logic. *Theoretical Computer Science*, 606:79–97, 2015.

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

Marcelo A Falappa, Gabriele Kern-Isberner, and Guillermo R Simari. Explanations, belief revision and defeasible reasoning. *Artificial Intelligence*, 141(1-2):1–28, 2002.

Jane Friedman. Question-directed attitudes. *Philosophical Perspectives*, 27(1):145–174, 2013.

Jane Friedman. Suspended judgment. *Philosophical studies*, 162(2):165–181, 2013.

Jane Friedman. Why suspend judging? *Noûs*, 51(2):302–326, 2017.

Peter Gärdenfors. Knowledge in flux. modeling the dynamics of epistemic states. *Studia Logica*, 49(3):421–424, 1990.

Sören Halldén. *The Logic of Nonsense.* Uppsala Universitets Arsskrift, Uppsala, 1949.

Peter Hawke. Theories of aboutness. *Australasian Journal of Philosophy*, 96(4):697–723, 2018.

Stephen Cole Kleene, NG de Bruijn, J de Groot, and Adriaan Cornelis Zaanen. *Introduction to Metamathematics*, volume 483. van Nostrand, New York, 1952.

John McCarthy. A basis for a mathematical theory of computation. 1962.

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

Francesco Paoli and Michele Pra Baldi. Proof theory of paraconsistent weak kleene logic. *Studia Logica*, 108(4):779–802, 2020.

Henry Prakken and Gerard Vreeswijk. Logics for defeasible argumentation. *Handbook of philosophical logic*, pages 219–318, 2002.

Nicolás D Rotstein, Martín O Moguillansky, Marcelo A Falappa, Alejandro Javier García, and Guillermo Ricardo Simari. Argument theory change: revision upon warrant. *COMMA*, 172:336–347, 2008.

Received 11 July 2022