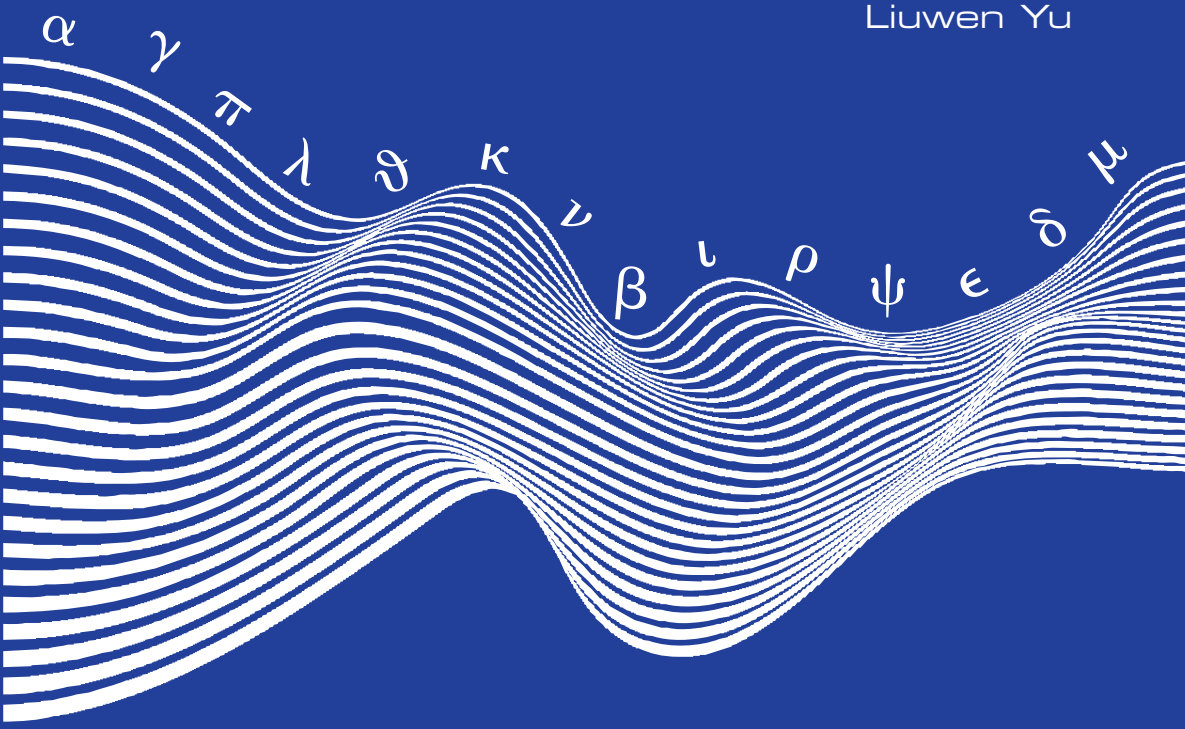
LNGAI 5

Logics for New-Generation AI 2025

Fifth International Workshop. December 1-5 2025, Luxembourg

Editors
Beishui Liao
Antonino Rotolo
Leendert van der Torre
Liuwen Yu



Logics for New-Generation AI

Fifth International Workshop

1-5 December 2025, Luxembourg

Volume 1
Proceedings of the First International Workshop, Hangzhou, 2021
Beishui Liao, Jieting Luo and Leendert van der Torre, eds

Volume 2
Proceedings of the Second International Workshop, Zhuhai, 2022
Beishui Liao, Réka Markovich and Yì N. Wáng, eds

Volume 3
Logics for AI and Law. Joint Proceedings of the Third International Workshop on Logics for New-Generation Artificial Intelligence and the International Workshop on Logic, AI and Law, September 8-9 and 11-12, 2023, Hangzhou, Zhuhai, 2023
Bruno Bentzen, Beishui Liao, Davide Liga, Réka Markovich, Bin Wei, Minghui Xiong and Tianwen Xu, eds

Volume 4
Proceedings of the Fourth International Workshop, Hangzhou, 2024
Beishui Liao, Jun Pang and Tjitze Rienstra, eds

Volume 5
Proceedings of the Fifth International Workshop, Luxembourg, 2025
Beishui Liao, Antonino Rotolo, Leendert van der Torre and Liuwen Yu, eds

Logics for New-Generation AI

Fifth International Workshop

1-5 December 2025, Luxembourg

Edited by

Beishui Liao

Antonino Rotolo

Leendert van der Torre

Liuwen Yu

ISBN 978-1-84890-495-8

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

http://www.collegepublications.co.uk

Original cover design by Laraine Welch

Preface

With the development of several new directions of AI, including agentic AI, explainable AI, ethical AI and knowledge-based AI, the corresponding directions of logical research are gaining momentum: causal reasoning, reasoning with norms and values, formal and computational argumentation, and knowledge graph reasoning, etc. This volume continues the series Logics for New Generation Artificial Intelligence (LNGAI), which accompanies the national key project "Research on Logics for New Generation Artificial Intelligence" (No. 20&ZD047, 2021–2025). In an open and dynamic environment, the main challenges for modelling such kinds of reasoning are to deal with information that is typically incomplete, uncertain, dynamic and conflicting, and to effectively explain the results and procedures of reasoning to ordinary human beings. The fifteen papers in this volume report recent advances in research on related topics. Together, these contributions can be broadly grouped into two complementary directions: foundational developments in logical systems, and application-driven research connecting logic with learning, causality, law, and ethics.

The first group of papers advances the theoretical boundaries of non-monotonic reasoning, algebraic semantics, and proof theory. Several contributions focus on the structural properties of logical systems. Andrew Lewis-Smith and Zhiguang Zhao introduce a new Kripke-style relational semantics for Monoidal T-norm Logic (MTL) by extracting it from the ordinal-sum algebraic representation of MTL, and prove that this semantics is sound and complete for the logic. Zhiguang Zhao establishes a complete axiomatization of hybrid tense logic over the real line by extending the canonical model of the rational line with irrational points using a concept of locality to capture Dedekind completeness. Lifei Wang, Zhe Yu, and Zhe Lin investigate the Intuitionistic Epistemic Logic (IEL) by establishing its algebraic completeness and finite model property (FMP), and by designing a sound, complete, and terminating labeled sequent system for IEL and a natural extension of it with an adjoint diamond operator. In the realm of normative reasoning, Andrea De Domenico et al. develop an algebraic semantics for Input/Output (I/O) logic by introducing and studying slanted (co-)Heyting algebras, which are equivalent to distributive lattices with subordination relations, to provide a deontic interpretation

for conditional obligations. Shuwen Wu compares three systems—Kreuger's definitional reflection, Schroeder-Heister's LI, and French's LKR—that restrict the identity rule or related reflexive principles. This paper emphasizes that all three share a common methodological insight: reflexivity is not harmless, and weakening or abandoning the identity rule can restore desirable proof-theoretic properties such as determinism, cut admissibility, and especially consistency in theories of truth. Yuxin Sun and Yuping Shen show that two-valued logic programs can be equivalently translated into propositional logic by extending Clark's completion with loop formulas, proving that the "Completion + Loop Formulas" pattern also applies to this nonmonotonic formalism. Stipe Pand ˇzi´c proposes first-order default justification logic to model exceptions and undercutting directly in the object language, avoiding unintended default extensions and enabling a principled integration of learning outputs with symbolic reasoning. Finally, within argumentation theory, Caren Al Anaissy et al. provide a principle-based robustness analysis of labeling-based bipolar argumentation semantics, evaluating variants of support under dynamic changes.

The second group of papers illustrates how logical frameworks interact with machine learning, causal modeling, and real-world reasoning tasks. A major focus is the neuro-symbolic gap. Aditya Kar, Emiliano Lorini, and Timothée Masquelier map Binary Spiking Neural Networks onto binary causal models to enable SAT/SMT-based abductive explanations. Krzysztof Pancerz et al. propose "readable twins" that transform deep learning models into rough-set flow graphs to provide human-interpretable, global explanations of model behavior. In the domain of ensemble learning, Sheng Wei and Beishui Liao propose A^2C, an adaptive ensemble classifier that models base classifiers' predictions as arguments and uses a dynamic argumentation framework to resolve conflicts and improve accuracy over soft-voting ensembles. Muyun Shao, Siyi Liu, and Beishui Liao develop a causal approach to contrastive explanation in abstract argumentation, adapting Halpern-Pearl causality to define actual cause explanations.

On the application side, Liuwen Yu et al. introduce DiSCo-RAD, a reasoning alignment model that explains how constitutional court majorities can form despite deep internal disagreement by aligning two complementary routes—a normative coalition-formation route and an argumentation route—to show when both perspectives yield the same coalition, thereby revealing how judges with incompatible reasoning can still support the same outcome. Addressing ethical AI, Julian Alfredo Mendez and Timotheus Kampik present the AR fairness metamodel, a formal framework designed to abstractly represent, analyze, and compare different fairness scenarios by formally defining fairness notions, which is then instantiated using the Tiles framework to support the operationalization and evaluation of these fairness definitions in various application contexts. Yini Huang and Beishui Liao extend the Jiminy Advisor ethical argumentation framework with a relevant-argument removal semantics that identifies how each argument contributes to a final obligation.

All papers in this volume have undergone careful peer review by members of the programme committee. We would like to thank the authors for their contributions, and the reviewers for their diligent and constructive reviews.

Beishui Liao, Zhejiang University
Antonino Rotolo, University of Bologna
Leon van der Torre, University of Luxembourg, Zhejiang University
Liuwen Yu, Luxembourg Institute of Science and Technology
November 17, 2025

Contents

A Causal Approach to Contrastive Explanation in Abstract Argumentation

Muyun Shao, Siyi Liu, Beishui Liao

Zhejiang University, Hangzhou, China

Abstract

A contrastive explanation is an account that clarifies why a particular event happens instead of an alternative. This paper proposes a method to construct explanations in abstract argumentation by selecting actual causes that explain why the target argument has a particular label instead of other labels. Our methodology is inspired by Causal Calculus and develops a novel formalism for counterfactual reasoning in abstract argumentation. We translate an argumentation framework into a rule-based system that characterizes the acceptance conditions of arguments. In this model, intervention is formalized as the replacement of specific rules with new axioms. This formalism enables us to express the key counterfactual condition – specifically, AC2(a^m) – from the theory of actual causality, in the setting of abstract argumentation. Together with two other conditions, Actuality and Minimality, this counterfactual requirement guides the selection of actual causes used to generate contrastive explanations in abstract argumentation.

Keywords: Contrastive Explanation, Abstract Argumentation, Actual Causality, Counterfacutal Reasoning, Explainable AI.

1 Introduction

Within eXplainable AI (XAI), Abstract Argumentation (AA) is a well-established tool for explanation generation [17,19,36,38], while the argumentative process itself has also been studied in its own right as a subject of explanation [16,24,34,37]. Explanations in AA seek to identify the causes behind the acceptability status of an argument. Most existing approaches define these causes as sets of arguments [5,11,12,20,23,24,37], and adopt key notions from causal explanation literature – such as *sufficiency*, which requires that the acceptance of the cause guarantees the acceptance of the argument in the question [7,12] – as rational criteria for cause selection.

In contrast to prior studies which are mainly inspired by causal reasoning, this work presents a formal adoption of the HP-definition, developed by Joseph Halpern and Judea Pearl [22], in the context of AA. The HP-definition of actual causality is built on structural equation model (SEM) [21,22]. A SEM is a directed acyclic graph where nodes denote variables and arrows denote the causal dependence relation between variables. An event is understood as

a variable X which takes a value x, denoted as $X = x$. An actual cause of an event $Y = y$ is defined as a set of events E that satisfy three conditions. *Actuality* requires that E should be the case in the actual world. *Counterfactual* requires that there should be at least one counterfactual situation where E is not the case and the effect $Y = y$ does not hold. And *Minimality* requires that the set E should not include any superfluous events.

The main challenge in applying the HP-definition to formulate the explanations in AA lies in characterizing reasoning about counterfactual. To overcome this difficulty, we develop an interventionist approach to counterfactual reasoning in AA, where counterfactual situations are specified by intervening with the labels of some arguments. Our formalism is grounded in the logical causality developed by Alexander Bochman [8]. We translate an AF into a set of rules, which take two possible forms. The first form ($A \Rightarrow \mathbf{v}(\alpha)$) states that condition A is sufficient for argument α to be labeled $\mathbf{v}$; the second form ($\top \Rightarrow \mathbf{v}(\alpha)$) asserts that α is labeled $\mathbf{v}$ unconditionally. An Intervention is a process which revises the set of rules. An intervention forces the label $\mathbf{v}'$ onto the argument α by removing the rules related to α, and adding a new rule ($\top \Rightarrow \mathbf{v}'(\alpha)$) says that the argument α should be labeled $\mathbf{v}'$ without any condition. Counterfactual situations are then defined with respect to an actual labelling. If the actual label of argument α is $\mathbf{v}$, then intervening to change it to another label $\mathbf{v}'$ defines a class of counterfactual situations.

The above formalization sets foundations for building contrastive explanations in AA. The contrastive question form we consider includes four elements: an argumentation framework $\mathcal{F}$, a labelling $\mathcal{L}$ of the framework, the label $\mathbf{v}$ of an argument α to be explained, and the expected label of that argument which is different from its actual label. Two instances of the contrastive question form are considered, they differ in whether the explainee expect that the argument α has another particular label $\mathbf{v}'$, or the explainee is just wondering why the argument α is not assigned with other labels. We clarify that these two types of questions can be unified in one single question form without losing generality. To answer these questions is to select actual causes that satisfy three conditions of the HP-definition. We rewrite the conditions of HP-definition in AA, and the counterfactual condition to which we refer is the AC2(a^{m}) version.

In summary, this paper's contributions are threefold:

(i) A system for intervention and counterfactual reasoning in AA;

(ii) A method for generating contrastive explanation using actual causality;

(iii) A demonstration of the new framework's effectiveness in cause selection via examples.

The structure of this paper is as follows. In Section 2, we establish the concepts from AA and causal reasoning that are required for our analysis. Section 3 then presents a counterfactual reasoning framework in AA. Leveraging this formalism, Section 4 formalizes the notion of an actual cause in AA to generate contrastive explanations. Our work is contextualized within the broader literature in Section 5, and concluded with future directions in Section 6.

2 Preliminaries

2.1 Abstract Argumentation

An argumentation framework (AF) is a directed graph $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, where $\mathcal{A}$ is a finite set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ represents the attack relation [18]. The universe of AF is represented by the set $\mathcal{UF}$, while the set of arguments that appear within $\mathcal{UF}$ is denoted as $\mathcal{UA}$.

Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF. For $\alpha, \beta \in \mathcal{A}$, we say that α *attacks* β if $(\alpha, \beta) \in \mathcal{R}$. For $\alpha \in \mathcal{A}$ and $E \subseteq \mathcal{A}$, we say that α *attacks* E if there is an argument in E that is attacked by α, and we say that E *attacks* α if there is an argument in E that attacks α. The following notations are used to achieve conciseness:

- $\alpha^+ = \{\beta \in \mathcal{A} \mid \alpha \text{ attacks } \beta\}$ and $\alpha^- = \{\beta \in \mathcal{A} \mid \beta \text{ attacks } \alpha\}$;
- $E^+ = \{\alpha \in \mathcal{A} \mid E \text{ attacks } \alpha\}$ and $E^- = \{\alpha \in \mathcal{A} \mid \alpha \text{ attacks } E\}$.

Definition 2.1 Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF. A set $E \subseteq \mathcal{A}$ is conflict-free in $\mathcal{F}$, denoted by $E \in cf(\mathcal{F})$, iff for all $\alpha, \beta \in E$, it holds that $(\alpha, \beta) \notin \mathcal{R}$. A set E defends an argument α iff for all $\beta \in \alpha^-$, there exists $\gamma \in E$, s.t. $(\gamma, \beta) \in \mathcal{R}$.

Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF. The characteristic function of the AF $\mathcal{F}$ is a function Γ s.t. for every $E \subseteq \mathcal{A}$, we have $\Gamma_{\mathcal{F}}(E) = \{a \in E \mid E \text{ defends } a\}$. The argumentation semantics is a function σ, associating with AF $\mathcal{F}$ a subset of $2^{\mathcal{A}}$, denoted as $\mathcal{E}_\sigma(\mathcal{F})$.

The acceptability of an argument is prescribed by specific argumentation semantics (cf. [3] for an overview). Given an AF, each semantics returns a set of sets of acceptable arguments called extensions. Classical semantics includes admissible, complete, grounded, preferred, and stable semantics (abbr. ad, co, gr, pr, st). And we refer to an extension under the semantics $\sigma \in \{ad, co, pr, gr, st\}$ as σ-extension.

Definition 2.2 Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $E \in cf(\mathcal{F})$.

(i) $E \in ad(\mathcal{F})$ iff E defends all its elements;

(ii) $E \in co(\mathcal{F})$ iff $E \in ad(F)$ and any $\alpha \in \mathcal{A}$ defended by E is in E;

(iii) $E \in gr(\mathcal{F})$ iff E is $\subseteq$-minimal in $co(\mathcal{F})$;

(iv) $E \in pr(\mathcal{F})$ iff E is $\subseteq$-maximal in $co(\mathcal{F})$;

(v) $E \in st(\mathcal{F})$ iff E attacks each $\alpha \in \mathcal{A} \setminus E$.

Argumentation semantics can be defined through two parallel methods: the *extension-based method*, which uses semantics to define the acceptability criteria of arguments; and the *labelling-based method*, which formulates semantics in terms of σ-labelling (cf. [2] for an overview).

Definition 2.3 Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, a labelling for $\mathcal{F}$ is a total function $\mathcal{L} : \mathcal{A} \mapsto \{\text{in}, \text{out}, \text{und}\}$ that assigns a label to each argument, denoted as $\mathcal{L}_\sigma(\mathcal{F})$ for $\sigma \in \{ad, co, pr, gr, st\}$.

Let $\text{in}(\mathcal{L})$, $\text{out}(\mathcal{L})$, and $\text{und}(\mathcal{L})$ be the sets of arguments labelled with in,

out, and und respectively, we use the triple $\langle \text{in}(\mathcal{L}), \text{out}(\mathcal{L}), \text{und}(\mathcal{L})\rangle$ to represent the labelling $\mathcal{L}$.

Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, a labelling $\mathcal{L}$ of $\mathcal{F}$ is said to be *admissible* ($\mathcal{L} \in \mathcal{L}_{ad}(\mathcal{F})$) iff $\forall \alpha \in \text{in}(\mathcal{L}) \cup \text{out}(\mathcal{L})$ it holds that: (i) $\mathcal{L}(\alpha) = \text{out}$ iff $\exists(\beta, \alpha) \in \mathcal{R}$ such that $\mathcal{L}(\beta) = \text{in}$; and (ii) $\mathcal{L}(\alpha) = \text{in}$ iff $\forall(\beta, \alpha) \in \mathcal{R}$, $\mathcal{L}(\beta) = \text{out}$ holds. Moreover, $\mathcal{L}$ is a *complete labelling* ($\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})$) iff conditions (i) and (ii) hold for all arguments in $\mathcal{A}$. Given an AF $\mathcal{F}$, and the set of complete labellings $\mathcal{L}_{co}(\mathcal{F})$, the following conditions determine whether a labelling is *grounded*, *preferred*, or *stable*:

(i) $\mathcal{L} \in \mathcal{L}_{gr}(\mathcal{F})$ iff $\text{in}(\mathcal{L})$ is $\subseteq$-minimal in $\bigcup_{\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})} \text{in}(\mathcal{L})$;

(ii) $\mathcal{L} \in \mathcal{L}_{pr}(\mathcal{F})$ iff $\text{in}(\mathcal{L})$ is $\subseteq$-maximal in $\bigcup_{\mathcal{L} \in \mathcal{L}_{co}(\mathcal{F})} \text{in}(\mathcal{L})$;

(iii) $\mathcal{L} \in \mathcal{L}_{st}(\mathcal{F})$ iff $\mathcal{L} \in \mathcal{L}_{ad}(\mathcal{F})$ and $\text{und}(\mathcal{L}) = \emptyset$.

Between complete semantics and complete labelling, there is a bijective mapping [2, Proposition 3] (the proof appears in [15, Theorem 9 and 10]).

Theorem 2.4 *Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $E \subseteq \mathcal{A}$, for each co-extension E there is a unique co-labelling $\mathcal{L} = \langle \{\alpha \mid \alpha \in E\}, \{\beta \mid \beta \in E^+\}, \{\gamma \mid \gamma \in \mathcal{A} \setminus (E \cup E^+)\}\rangle$; and for each co-labelling $\mathcal{L}$, there is a unique co-extension, that is* $\text{in}(\mathcal{L})$.

2.2 Causal Reasoning

In this section, we introduce a causal reasoning formalism that integrates *causal calculus* [8] with Pearl's *structural equation model* (SEM) [29] and the Halpern-Pearl definition (HP definition) of *actual causality* [22]. To ensure consistency and improve readability, we adjust the notation in Definition 2.5 to match that of Section 3 and then reinterpret Definition 3.8.

2.2.1 Causal Calculus

The atomic reasoning pattern of causal calculus, called a *proposition*, has the form of $X = x$, where X is a variable and x is the value assigned to X. *Causal rules* are then introduced as inference rules to express causal dependencies. A causal rule specifies how a set of antecedent propositions (the causes or premises) leads to a consequent proposition (the effect or conclusion).

Definition 2.5 [Causal Rule] A causal rule has the form $A \Rightarrow (X = x)$, where A is a set of propositions and $X = x$ a proposition.

The interpretation of the causal rule $A \Rightarrow (X = x)$ is " a set of propositions A causes proposition $X = x$". A *causal theory* is a collection of causal rules.

Definition 2.6 [Causal Theory] A causal theory Δ is a set of causal rules.

Given that a set of propositions B is true, the consequence of a causal theory Δ is defined as $\Delta(B)$, where Δ is understood as a monadic operator. In the following, we use Δ-operator to denote its operational aspect. $\Delta(B)$ denotes the set of propositions that are directly caused by B, which is defined as:

$$\Delta(B) = \{X = x \mid A \Rightarrow (X = x),\ A \subseteq B\}$$

The boolean semantics of a causal theory is defined based on the notion of *valuation*. A valuation is a function which assigns either 1 ('truth') or 0 ('falsity') to every proposition of the language. A valuation v is called a *model* of the causal theory Δ, if it is compatible with Δ such that the Δ-operator does not derive contradictory conclusions from v as premises.

2.2.2 Structural Equation Model

Pearl's structural equation model (SEM) [29] can be viewed as an instantiation of the causal calculus [9]. A SEM characterizes causal dependencies among variables, which are categorized as *exogenous* (whose values are externally determined) or *endogenous* (whose values are determined by other variables). The values of endogenous variables are fully determined by a set of functions F, which explicitly define the structural dependencies.

Definition 2.7 [Structural Equation Model] A SEM is a triple $M = (U, V, F)$, where

- $U = \{U_1, U_2, \ldots, U_n\}$ is a finite set of exogenous variables;
- $V = \{V_1, V_2, \ldots, V_n\}$ is a finite set of endogenous variables whose values are determined by the values of other variables in $U \cup V$;
- F is a set of functions $\{f_1, f_2, \ldots, f_n\}$ such that each f_i is a mapping from $U \cup (V \backslash \{V_i\})$ to V_i.

Given a SEM $M = (U, V, F)$, a setting $\vec{u}$ of M is the values of exogenous variables, which can be represented as a conjunction of primitive propositions $U_1 = u_1 \wedge U_2 = u_2 \ldots \wedge U_n = u_n$, meaning that every exogenous variable U_i is assigned a value u_i. A SEM together with a setting $\vec{u}$ constitutes a *causal model*. Let $X = x$ be a proposition, $(M, \vec{u}) \models X = x$ iff the value of X is x once the value of exogenous variables is set to $\vec{u}$.

Definition 2.8 [Causal Model] A causal model is a pair $(M, \vec{u})$ where M is a SEM and $\vec{u}$ is a setting of M.

The interventionist semantics for counterfactual reasoning is defined using SEMs. A formula of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k](X = x)$ states that variable X takes the value x after an intervention that sets the set of variables $\vec{Y}$ to the values $\vec{y}$. Here, the notation $[\vec{Y} \leftarrow \vec{y}]$ abbreviates $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]$ for simplicity [1].

2.2.3 Actual Causality

The HP-definition of causality is introduced based on SEMs [22]. The types of events that are allowed as actual causes are those of the form $[X_1 = x_1 \wedge \ldots X_k = x_k]$ (also abbreviated as $\vec{X} = \vec{x}$), and the events that are caused (called the effect) are events $Y = y$. The definition of actual cause comprises three key clauses:

[1] Vectors (e.g., $\vec{Y}$, $\vec{y}$) represent sets of variables and their value assignments, while plain symbols (e.g., Y, y) represent single variables and values, respectively.

- **Actuality:** Specifies that both the causes and the effect must hold true in the actual causal model.
- **Counterfactual:** There must exist an alternative assignment $\overrightarrow{W} \leftarrow \overrightarrow{w}$, called the *witness*, such that under the witness, changing the values of $\overrightarrow{X}$ from $\overrightarrow{x}$ to some $\overrightarrow{x'}$ will cause the non-occurrence of effect.
- **Minimality:** Requires that the set of causes should be minimal, meaning that it should not include any superfluous elements.

Definition 2.9 [Actual Cause] $\overrightarrow{X} = \overrightarrow{x}$ is an actual cause of $Y = y$ in the causal setting $(M, \overrightarrow{u})$ if the following three conditions hold:

AC1. $(M, \overrightarrow{u}) \models (\overrightarrow{X} = \overrightarrow{x})$ and $(M, \overrightarrow{u}) \models (Y = y)$;
AC2. See Definition 2.10;
AC3. $\overrightarrow{X}$ is minimal; there is no strict subset $\overrightarrow{X'}$ of $\overrightarrow{X}$ such that $\overrightarrow{X'} = \overrightarrow{x'}$ satisfies conditions AC1 and AC2, where $\overrightarrow{x'}$ is the restriction of $\overrightarrow{x}$ to the variables in $\overrightarrow{X}$.

Three variants of the counterfactual condition AC2 have been proposed in [22]. Here, we focus specifically on the simplest modified version AC2($\mathrm{a^m}$) [2]. This condition states that, when the variables in $\overrightarrow{W}$ are held fixed at their actual values $\overrightarrow{w}$, there must exist at least one counterfactual situation in which $\overrightarrow{X}$ is assigned some alternative value $\overrightarrow{x'}$ – different from its actual value – such that the effect $Y = y$ no longer holds.

Definition 2.10 [AC2($\mathrm{a^m}$)] There is a set of variables $\overrightarrow{W}$ in V such that $\overrightarrow{Z}$ and $\overrightarrow{W}$ is a partition of V with $\overrightarrow{X} \subseteq \overrightarrow{Z}$, and a setting $\overrightarrow{x'}$ of the variables in $\overrightarrow{X}$ such that if $(M, \overrightarrow{u}) \models \overrightarrow{W} = \overrightarrow{w}$, then we have

$$(M, \overrightarrow{u}) \models [\overrightarrow{X} \leftarrow \overrightarrow{x'}, \overrightarrow{W} \leftarrow \overrightarrow{w}]\neg(Y = y).$$

3 Counterfactual Reasoning in Abstract Argumentation

This section formalizes counterfactual reasoning in AA via a four-step procedure. We first define the *syntax* of a language for reasoning with argument-label pairs, where atomic propositions consist of arguments and their labels. Next, we translate the acceptance condition of the arguments into *acceptability rule*, thereby characterizing the AFs via sets of rules called *acceptability theories*. Interventions are then defined as modifications to the acceptability theory that enforce specific labels on certain arguments – a process referred to as *revision*. Finally, we introduce two *counterfactual operators* ($>$ and $\succ$) and define their truth conditions based on the revision mechanism.

[2] The other two variants are not in the scope of this paper. We will leave it for future discussion.

3.1 Interpreting Argumentation Framework into Acceptability Theory

We start by defining the language for argument labelling. The atomic propositions of the language are pairs of arguments and their labels expressed by the form $\mathbf{v}(\alpha)$ [3], where α is an argument variable and $\mathbf{v} \in \{\text{in}, \text{out}, \text{und}\}$ is the label of α. In the following, we use $\mathbb{V}$ to denote the set $\{\text{in}, \text{out}, \text{und}\}$. The set of logical connectives is the adequate set of connectives $\{\neg, \wedge\}$. The interpretation of $\neg\mathbf{v}(\alpha)$ is "the label of α is not $\mathbf{v}$". Other connectives such as $\vee$ are defined based on these two. [4] Moreover, an additional symbol $\top$ is involved in the language to denote unconditional true propositions.

Definition 3.1 [Syntax] Given an universal set of arguments $\mathcal{UA}$, with α ranging over $\mathcal{UA}$ and $\mathbf{v}$ ranging over $\mathbb{V}$, we define the syntax of language $L_{\mathcal{UA}}$ by the following rule:

$$\varphi := \mathbf{v}(\alpha) \mid \top \mid \neg\varphi \mid \varphi \wedge \varphi$$

The truth condition of formulas in $L_{\mathcal{UA}}$ is defined with respect to an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and a labelling $\mathcal{L}$ of $\mathcal{F}$. Here we refer to the basic notion of labelling, which is simply a total mapping from $\mathcal{A}$ to $\mathbb{V}$. An atomic proposition $\mathbf{v}(\alpha)$ is true in $\mathcal{L}$ of $\mathcal{F}$ if the label that $\mathcal{L}$ assigns to α is exactly $\mathbf{v}$. The truth condition of compound formulas involving connectives of $\neg, \wedge$ is defined the same way as they were in the propositional language.

Definition 3.2 [Truth Condition] Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF and $\mathcal{L}$ be a labelling of $\mathcal{F}$, the truth conditions of formulas are defined with respect to these two indices as follows:

- $\top$ is unconditionally true.
- $\mathbf{v}(\alpha)$ is true in $\mathcal{L}$ of $\mathcal{F}$ iff $\mathcal{L}(\alpha) = \mathbf{v}$.
- $\neg\mathbf{v}(\alpha)$ is true in $\mathcal{L}$ of $\mathcal{F}$ iff $\mathcal{L}(\alpha) \neq \mathbf{v}$.
- $\neg\varphi$ is true in $\mathcal{L}$ of $\mathcal{F}$ iff φ is not true in $\mathcal{L}$ of $\mathcal{F}$.
- $\varphi \wedge \psi$ is true in $\mathcal{L}$ of $\mathcal{F}$ iff both φ and ψ are true in $\mathcal{L}$ of $\mathcal{F}$.

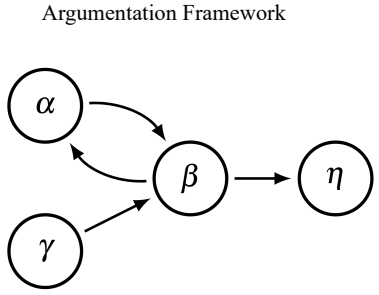
A set of propositions A is true in $\mathcal{L}$ of $\mathcal{F}$ iff every proposition in A is true in $\mathcal{L}$ of $\mathcal{F}$.

Based on the formulas of $L_{\mathcal{UA}}$, AFs could be translated into sets of rules which represents the acceptance conditions of arguments in the framework. These rules, called *acceptability rules*, take the form of $r : (A \Rightarrow \mathbf{v}(\alpha))$, where r is the index of the rule, A is a non-empty set of propositions called the *premises* and $\mathbf{v}(\alpha)$ is a proposition called the *effect*. *Axioms* are acceptability rules with no premise, which has the form of $r : (\top \Rightarrow \mathbf{v}(\alpha))$.

---

[3] Bold $\mathbf{v}$: assignment of $\{\text{in}, \text{out}, \text{und}\}$ in AF; Regular v: assignment (typically binary) in causal theories.

[4] $\varphi \vee \psi$ is defined as $\neg(\neg\varphi \wedge \neg\psi)$.



Acceptability Theory

α	r_1: (in(β) ⇒ out(α)) r_2: (out(β) ⇒ in(α))
β	r_3: ((in(α) ∨ in(γ)) ⇒ out(β)) r_4: ((out(α) ∧ out(γ)) ⇒ in(β))
γ	r_5: (⊤ ⇒ in(γ))
η	r_6: (in(β) ⇒ out(η)) r_7: (out(β) ⇒ in(η))

Fig. 1. The Running Example ($\mathcal{F}_1$ and $\Delta_{\mathcal{F}_1}$)

Definition 3.3 [Acceptability Rule] An acceptability rule has two possible forms, $r : (A \Rightarrow \mathbf{v}(\alpha))$, and $r : (\top \Rightarrow \mathbf{v}(\alpha))$. r is the index of the rule, A is a non-empty set of propositions and $\mathbf{v}(\alpha)$ is an atomic proposition. [5]

A collection of acceptability rules forms a *acceptability theory*. Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, the corresponding acceptability theory $\Delta_{\mathcal{F}}$ could be obtained by translating the acceptance conditions of arguments expressed by the relation $\mathcal{R}$. Briefly, an argument is labelled in if all its attackers are labelled out, and is labelled out if at least one attacker is labelled in. It is labelled und if it can not be labelled in or out.

The translation process divides the arguments into two groups. Initial arguments (with no ancestors) are always labelled in, so their acceptability rules are defined as axioms ($\top \Rightarrow \text{in}(\alpha)$). The acceptance conditions of non-initial arguments are expressed as two acceptability rules, one rule expresses the condition of labelling in, the other expresses the condition of labelling out. There will be no rule expressing the condition of labelling und since it is only an alternative label that will be assigned to arguments in case that the label in and the label out can not be assigned to arguments. We will leave this for Definition 3.8.

Definition 3.4 [Acceptability Theory] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, let $parent(\alpha) = \{\beta \mid (\beta, \alpha) \in \mathcal{R}\}$, an acceptability theory $\Delta_{\mathcal{F}}$ is a set of acpetability rules such that:

- For any initial argument α,

(i) $(r : \top \Rightarrow \text{in}(\alpha)) \in \Delta_{\mathcal{F}}$.

- For any non-initial argument β,

(ii) $(r : (\bigwedge_{\gamma \in parent(\beta)} \text{out}(\gamma)) \Rightarrow \text{in}(\beta)) \in \Delta_{\mathcal{F}}$,

(iii) $(r : (\bigvee_{\gamma \in parent(\beta)} \text{in}(\gamma)) \Rightarrow \text{out}(\beta)) \in \Delta_{\mathcal{F}}$.

- Only the rules above are contained in $\Delta_{\mathcal{F}}$.

Example 3.5 Consider the AF $\mathcal{F}_1$ in Fig. 1 [6]. The corresponding acceptabil-

[5] If $A = \{\mathbf{v}'(\beta)\}$ is a singleton, we write $\mathbf{v}'(\beta) \Rightarrow \mathbf{v}(\alpha)$ for simplicity.

[6] See also the discussion of (ir)relevant arguments in [30], within the same framework.

ity theory $\Delta_{\mathcal{F}_1}$ is obtained by translating the acceptability conditions of each argument listed in the right-hand table of Fig. 1. The acceptability rule for the initial argument γ is the axiom r_5. α, β, η are non-initial arguments, their acceptability rules include rules (r_1, r_3, r_6) expressing the conditions of labelling out, and rules (r_2, r_4, r_7) expressing the conditions of labelling in.

Given an acceptability theory $\Delta_{\mathcal{F}}$ and a labelling $\mathcal{L}$ of $\mathcal{F}$, we define the *derivation* of $(\Delta_{\mathcal{F}}, \mathcal{L})$ by the symbol $\models$. A proposition $\mathbf{v}(\alpha)$ is derivable from $(\Delta_{\mathcal{F}}, \mathcal{L})$, denoted as $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \mathbf{v}(\alpha)$, if there is an axiom $r : (\top \Rightarrow \mathbf{v}(\alpha))$, or there is a rule $r' : (A \Rightarrow \mathbf{v}(\alpha))$ such that $r' \in \Delta_{\mathcal{F}}$, and the premises of r' coincide with $\mathcal{L}$, so that $\mathbf{v}(\alpha)$ could be obtained by applying the rule r'.

Definition 3.6 [Derivation] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$, its corresponding acceptability theory $\Delta_{\mathcal{F}}$ and a labelling $\mathcal{L}$ of $\mathcal{F}$. Let $\mathbf{v}(\alpha)$ be a proposition. The derivation $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \mathbf{v}(\alpha)$ holds iff one of the following conditions holds: [7]

- $\exists r \in \Delta_{\mathcal{F}}$, $r : (\top \Rightarrow \mathbf{v}(\alpha))$;
- $\exists r \in \Delta_{\mathcal{F}}$, $r : (A \Rightarrow \mathbf{v}(\alpha))$, s.t. A is true in $\mathcal{L}$ of $\mathcal{F}$.

Example 3.7 [Continued Example 3.5] Given $\Delta_{\mathcal{F}_1}$ in Example 3.5, let $\mathcal{L}_1 = \langle \{\beta\}, \{\alpha\}, \{\gamma, \eta\} \rangle$.

- The derivation $(\Delta_{\mathcal{F}_1}, \mathcal{L}_1) \models \text{in}(\gamma)$ holds because there is an axiom $r_5 : (\top \Rightarrow \text{in}(\gamma))$ states that (γ) could be labeled in.
- The derivation $(\Delta_{\mathcal{F}_1}, \mathcal{L}_1) \models \text{out}(\alpha)$ holds because there is an acceptability rule $r_1 : (\text{in}(\beta) \Rightarrow \text{out}(\alpha))$, such that $\mathcal{L}_1(\beta) = \text{in}$, so r_1 could be applied to derive $\text{out}(\alpha)$.
- The derivation $(\Delta_{\mathcal{F}_1}, \mathcal{L}_1) \models \text{in}(\beta)$ **does not hold** because even if there is an acceptability rule $r_4 : (\text{Out}(\alpha) \wedge \text{Out}(\gamma) \Rightarrow \text{in}(\beta))$, the premises of r_4 is not true in $\mathcal{L}_1$ of $\mathcal{F}_1$ since $\mathcal{L}_1(\gamma) = \text{und}$. So r_4 chould not be applied to derive $\text{in}(\beta)$, and there exists no other rule which could be applied to derive $(\text{in})\beta$.

The solution concept of acceptability theories, which corresponds to admissible labelling in AF, is the notion of *acceptability model*. Acceptability models are labellings satisfying the condition that every label in (resp.out) assgined to arguments should be derived from the acceptability theory $\Delta_{\mathcal{F}}$ and the labelling itself. The label und could be (*illegally*) assigned to any argument without any condition.

Definition 3.8 [Acceptability Model] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its corresponding acceptability theory $\Delta_{\mathcal{F}}$. A labelling $\mathcal{L}$ is an acceptability model of the acceptability theory $\Delta_{\mathcal{F}}$ iff

- $\forall \alpha \in \mathcal{A}$, if $\mathcal{L}(\alpha) = \text{in}$, then $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \text{in}(\alpha)$;
- $\forall \alpha \in \mathcal{A}$, if $\mathcal{L}(\alpha) = \text{out}$, then $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \text{out}(\alpha)$;
- $\mathcal{L}(\alpha) = \text{und}$ iff $\mathcal{L}(\alpha) \neq \text{in}$ and $\mathcal{L}(\alpha) \neq \text{out}$.

[7] The derivable relation between $(\Delta_{\mathcal{F}}, \mathcal{L})$ and compound formulas, such as $\mathbf{v}(\alpha) \vee \mathbf{v}'(\beta)$, could be defined. But it is out of the scope of this paper.

As we are more interested in the complete labelling, Definition 3.9 provides a more strict condition. An Argument α should be labelled in (resp. out) once there is an acceptability rule $r : (A \Rightarrow \text{in}(\alpha))$ (resp. $r : (A \Rightarrow \text{out}(\alpha))$) such that A is true in $\mathcal{L}$ of $\mathcal{F}$. The label und could be assigned to an argument α only if $\text{in}(\alpha)$ and $\text{out}(\alpha)$ are not derivable from $(\Delta_{\mathcal{F}}, \mathcal{L})$.

Definition 3.9 [Complete Acceptability Model (CAM)] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its corresponding acceptability theory $\Delta_{\mathcal{F}}$. A labelling $\mathcal{L}$ is an complete acceptability model of the acceptability theory $\Delta_{\mathcal{F}}$ iff $\mathcal{L}$ satisfying the following conditions:

- $\forall \alpha \in \mathcal{A}$, if $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \text{in}(\alpha)$, then $\mathcal{L}(\alpha) = \text{in}$;
- $\forall \alpha \in \mathcal{A}$, if $(\Delta_{\mathcal{F}}, \mathcal{L}) \models \text{out}(\alpha)$, then $\mathcal{L}(\alpha) = \text{out}$;
- $\mathcal{L}(\alpha) = \text{und}$ iff $\mathcal{L}(\alpha) \neq \text{in}$ and $\mathcal{L}(\alpha) \neq \text{out}$. [8]

Proposition 3.10 *Each acceptability model is an admissible labelling. Each complete acceptability model is a complete labelling.*

Example 3.11 [Continued Example 3.5] Given $\Delta_{\mathcal{F}_1}$ in Example 3.5. $\mathcal{L}_2 = \langle \{\gamma\}, \{\beta\}, \{\alpha, \eta\} \rangle$ is an acceptability model of $\Delta_{\mathcal{F}_1}$ because for $\mathcal{L}_2(\gamma) = \text{in}$ and $\mathcal{L}_2(\beta) = \text{out}$, $(\Delta_{\mathcal{F}_1}, \mathcal{L}_2) \models \text{in}(\gamma)$ and $(\Delta_{\mathcal{F}_1}, \mathcal{L}_2) \models \text{out}(\beta)$ hold. But $\mathcal{L}_2$ is not a CAM because $(\Delta_{\mathcal{F}_1}, \mathcal{L}_2) \models \text{in}(\eta)$ and $(\Delta_{\mathcal{F}_1}, \mathcal{L}_2) \models \text{in}(\alpha)$ hold, while we have $\mathcal{L}_2(\eta) \neq \text{in}$ and $\mathcal{L}_2(\alpha) \neq \text{in}$.

3.2 Intervention and Counterfactual

We will then focus on defining counterfactual situation and the truth condition for counterfactual statement. The notion of *intervention* is introduced to force particular labels onto arguments. An intervention I is a set of atomic propositions that are forced to be true. When some labels fixed by I are not included in models of acceptability theory, I is considered to construct a class of counterfactual situations.

Definition 3.12 [Intervention] An intervention I is a set of atomic propositions. The domain of I is defined as $dom(I) = \{\alpha \mid \exists \mathbf{v},\ \mathbf{v}(\alpha) \in I\}$.

A basic requirement for interventions is that it should be *consistent*, which avoids multiple labels assigned to a single argument.

Definition 3.13 [Consistency] An intervention I is consistent iff for all argument α, if $\mathbf{v}(\alpha) \in I$, then there exists no $\mathbf{v}' \neq \mathbf{v}$, such that $\mathbf{v}'(\alpha) \in I$.

Our general definition of intervention does not exclude the situation where an intervention I is already a subset of acceptability models. In that case, I itself is part of the actual situation, and counterfactual situations are acceptability models that contradict I. The Definition 3.14 formalizes this idea by defining the negations of an intervention I (denoted as $\neg I$). Negations of an intervention I are also interventions whose domain coincides with I, but there

[8] Item 3 is technically unnecessary, but we preserve it for better understanding.

is at least one argument whose label contradicts the label I assigns to that argument.

Definition 3.14 [Negation of Intervention] Let I, I' be two consistent interventions, I' is a negation of I iff:

- $dom(I) = dom(I')$;
- $\exists \alpha \in dom(I)$, $\exists \mathbf{v}, \mathbf{v}'$ s.t. $\mathbf{v} \neq \mathbf{v}'$ and $\mathbf{v}(\alpha) \in I, \mathbf{v}'(\alpha) \in I'$.

The process of modifying an acceptability theory to ensure that an intervention I holds in all of its CAMs is referred to as *revision*. A revision is a functional procedure that for every argument in the domain of I, acceptability rules whose effect is a proposition related to that argument are excluded out from the theory. And new axioms which say that propositions in I should be unconditionally true are added to the theory. The revised acceptability theory $\Delta_{\mathcal{F}} * I$ ensures that for every CAM $\mathcal{L}$ of the theory, I is true in $\mathcal{L}$ of $\mathcal{F}$.

Definition 3.15 [Revised Acceptability Theory] Let $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ be an AF, $\Delta_{\mathcal{F}}$ be its corresponding acceptability theory and I be an intervention. A revision of $\Delta_{\mathcal{F}}$ by I is a revised acceptability theory $\Delta_{\mathcal{F}} * I$ obtained by the following steps:

- $\Delta'_{\mathcal{F}} = \Delta_{\mathcal{F}} - \{(A \Rightarrow \mathbf{v}(\alpha)), (\top \Rightarrow \mathbf{v}(\alpha)) \mid \alpha \in dom(I)\}$;
- $\Delta_{\mathcal{F}} * I = \Delta'_{\mathcal{F}} \cup \{(\top \Rightarrow \mathbf{v}(\alpha)) \mid \mathbf{v}(\alpha) \in I\}$.

Example 3.16 [Continued Example 3.5] Given $\Delta_{\mathcal{F}_1}$ in Example 3.5. Let $I = \{\text{und}(\alpha)\}$ be an intervention. The revised acceptability theory $\Delta_{\mathcal{F}_1} * I$ is obtained by removing the rules r_1 and r_2, then adding an axiom $r_8 : (\top \Rightarrow \text{und}(\alpha))$.

Counterfactual statements are evaluated in revised acceptability theories. In contrast to SEM, a causal model $(M, \vec{u})$ has only one value for each variable, but abstract AFs could have multiple complete labellings, where one argument could have different labels. This difference gives rise to two levels of counterfactuality in AA. *Strict counterfactual* requires that the effect holds in all CAMs of the revised acceptability theory, while *weak counterfactual* only requires that the effect holds in one CAM of that theory.

The counterfactual defined below aligns with Bochman's definition [9, Definition 2.14]: a generalized concept based on acceptability theory. These two definitions do not involve an actual labelling, but instead examines the consequences of an intervention I within the context of a given acceptability theory. For the the counterfactual condition that does address an actual labeling, please refer to Definition 4.3.

Definition 3.17 [Strict Counterfactual] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its corresponding acceptability theory $\Delta_{\mathcal{F}}$. Let I be an intervention and $\mathbf{v}(\alpha)$ be an atomic proposition. A strict counterfactual has the form $I >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$. $I >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ holds if $\mathbf{v}(\alpha)$ holds in all CAMs of the revised acceptablity theory

$\Delta_{\mathcal{F}} * I$.[9]

Definition 3.18 [Weak Counterfactual] Given an AF $\mathcal{F} = (\mathcal{A}, \mathcal{R})$ and its corresponding acceptability theory $\Delta_{\mathcal{F}}$. Let I be an intervention and $\mathbf{v}(\alpha)$ be an atomic proposition. A weak counterfactual has the form $I \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$. $I \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ holds if $\mathbf{v}(\alpha)$ holds in at least one CAM of the revised acceptability theory $\Delta_{\mathcal{F}} * I$.

Example 3.19 [Continued Example 3.5] Given $\Delta_{\mathcal{F}_1}$ in Example 3.5.

- The strict counterfactual $\text{out}(\gamma) >_{\Delta_{\mathcal{F}_1}} \text{out}(\eta)$ **does not hold** because the revised acceptability theory $\Delta_{\mathcal{F}_1} * \{\text{out}(\gamma)\}$ has three CAMs: $\mathcal{L}_3 = \langle\{\beta\}, \{\alpha, \gamma, \eta\}, \emptyset\rangle$, $\mathcal{L}_4 = \langle\{\alpha, \eta\}, \{\beta, \gamma\}, \emptyset\rangle$ and $\mathcal{L}_5 = \langle\emptyset, \{\gamma\}, \{\alpha, \beta, \eta\}\rangle$. The effect $\text{out}(\eta)$ holds only in $\mathcal{L}_3$ but not in $\mathcal{L}_4$ and $\mathcal{L}_5$, which suggests that the weak counterfactual $\text{out}(\gamma) \succ_{\Delta_{\mathcal{F}_1}} \text{out}(\eta)$ holds.
- The strict counterfactual $\text{out}(\alpha) >_{\Delta_{\mathcal{F}_1}} \text{in}(\eta)$ holds because the revised acceptability theory $\Delta_{\mathcal{F}_1} * \{\text{out}(\alpha)\}$ has only one CAM $\mathcal{L}_6 = \langle\{\gamma, \eta\}, \{\alpha, \beta\}, \emptyset\rangle$, and the effect $\text{in}(\eta)$ holds in $\mathcal{L}_6$. This follows that the weak counterfactual $\text{out}(\alpha) \succ_{\Delta_{\mathcal{F}_1}} \text{in}(\eta)$ holds and another weak counterfactual $\text{out}(\alpha) \succ_{\Delta_{\mathcal{F}_1}} \text{out}(\eta)$ **does not hold**.

The above two binary operations do not satisfy *Monotonicity*.

Proposition 3.20 (Failed Monotonicity) *For any acceptability theory $\Delta_{\mathcal{F}}$, two interventions I, I' such that $I \subset I'$, it holds that:*

- $I >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ *does not imply* $I' >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$.
- $I \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ *does not imply* $I' \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$.

A key reason is that I' could assign values that contradict with the R-relation to auxiliary arguments. A solution is to restrict the superset I' be compatible with a CAM of $\Delta_{\mathcal{F}} * I$.

Definition 3.21 [Coherent Expansion] For any acceptability theory $\Delta_{\mathcal{F}}$ and any intervention I, I' is a coherent expansion of I iff

- $I \subset I'$;
- There is a CAM $\mathcal{L}$ of $\Delta_{\mathcal{F}} * I$ s.t. $\forall \mathbf{v}(\alpha) \in I'$, $\mathcal{L}(\alpha) = \mathbf{v}$.

Proposition 3.22 (Revised Monotonicity) *For any acceptability theory $\Delta_{\mathcal{F}}$, intervention I, I' such that I' is a coherent expansion of I, it holds that:*

- $I >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ *imples* $I' >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$.
- $I \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ *imples* $I' \succ_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$.

4 Contrastive Explanation

This section focuses on defining actual causes in Abstract Argumentation (AA) using the counterfactual reasoning framework introduced in Section 3, and employing it to provide explanation for contrastive questions. Before formalizing

[9] If $I = \{\mathbf{v}'(\alpha)\}$ is a singleton, we also write $\mathbf{v}'(\beta) >_{\Delta_{\mathcal{F}}} \mathbf{v}(\alpha)$ for simplicity.

the structure of contrastive questions and actual causes, we first present a brief overview of contrastive questions and motivate their selection as the subject of this study.

A contrastive explanation aims to clarify why a particular event occurred rather than a specific alternative. As Peter Lipton suggests, questions are often motivated by unexpected outcomes [25]. Consequently, questions arising in such contexts tend to be contrastive in nature. While ordinary questions focus solely on explaining an event that actually happened, a contrastive question introduces an additional element: a foil event that did not occur, yet was expected to happen. This additional element plays a role in the selection of the causes, as demonstrated by the following example:

> In Lewis's example, we can explain why he went to Monash rather than to Oxford in 1979 by pointing out that only Monash invited him because the invitation to Monash was a cause of his trip and that invitation would not have been a cause of a trip to Oxford if he had taken one. On the other hand, Lewis's desire to go to places where he has good friends would not explain why he went to Monash rather than Oxford, since he has friends in both places, and so the desire would have been part of either causal history. [25, p.255]

We believe that this additional element remains useful for identifying relevant causes when explaining the acceptability of the argument. Therefore, this paper focuses primarily on contrastive questions as the subjects of explanation. The following two types of contrastive questions are considered in this Section. They differ in whether the explainee expect that the argument α has another particular label $\mathbf{v}'$, or the explainee is just wondering why the argument α has the $\mathbf{v}$ instead of any other label $\neg\mathbf{v}$ [10] .

(i) For an explainee who has a mind of his own, he may ask: Why does the argument α have the label $\mathbf{v}$ instead of $\mathbf{v}'$ on a labelling $\mathcal{L}$ of the AF $\mathcal{F}$?

(ii) For an innocent explainee, he may ask: Why does the argument α have the label $\mathbf{v}$ instead of any other label $\neg\mathbf{v}$ on a labelling $\mathcal{L}$ of AF $\mathcal{F}$?

These two types of contrastive question are unified in the following contrastive form.

Definition 4.1 [Contrastive Question form] A contrastive question form is a 4-tuple $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$ with either $\mathbf{e} = \neg\mathbf{v}$, or $\mathbf{e} \neq \mathbf{v}$ ranging over $\mathbb{V}$, where $\mathcal{F}$ is an AF, $\mathcal{L}$ is a labelling presented by an computational process, $\mathbf{v}(\alpha) \in \mathcal{L}$ is the actual label of α and $\mathbf{e}(\alpha)$ is the label of a that is expected by the explainee.

To construct an explanation of a contrastive question form $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$ is to select an actual cause which explains why the actual label $\mathbf{v}(\alpha)$ holds in $\mathcal{L}$, and counterfactually the foil $\mathbf{e}(\alpha)$ holds. An actual cause is defined as an intervention satisfying the three conditions of HP-definition.

[10] The interpretation of $\neg\mathbf{v}$ is illustrated in Definition 3.2.

Definition 4.2 [Actual Cause in AA] An actual cause of the contrastive question form $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$ is a consistent intervention I satisfying the following conditions:

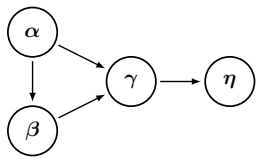
- (Actuality) I is true in $\mathcal{L}$ of $\mathcal{F}$;
- (Counterfactual) See Definition 4.3;
- (Minimality) I is minimal; there is no strict subset of I satisfies the above two conditions.

The conditions of actuality and minimality are defined the same way as in the HP-definition. Condition AC2($\mathrm{a^m}$) in AA is reformulated using the counterfactual operator $\succ$, as defined in Section 3 [11] .

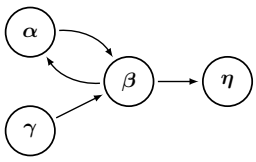
Definition 4.3 [Counterfactual Condition AC2($\mathrm{a^m}$) in AA] Given a contrastive question form $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$. An intervention I satisfy AC2($\mathrm{a^m}$) in AA iff there is another intervention I', called the witness, such that I' satisfies *Actuality* and $dom(I') \cap dom(I) = \emptyset$, and there exists an intervention $\neg I$, which is a negation of I such that

$$\neg I \cup I' \succ \mathbf{e}(\alpha).$$

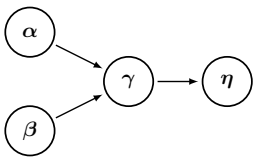
To illustrate our ideas, we use examples to demonstrate how these three conditions affect the selection of actual causes. Example 4.4 is inspired by *preemption* and Example 4.6 is inspired by *overdetermination*. These two are well-known examples in causal reasoning [22,30]. Example 4.5 is the continued analysis of the running example presented in Section 3.



(a) Preemption in AA ($\mathcal{F}_2$)



(b) The Running Example ($\mathcal{F}_1$)



(c) Overdetermination in AA ($\mathcal{F}_3$)

Example 4.4 Consider the AF $\mathcal{F}_2$ in Figure 2a. Given the contrastive question form $Q_1 = \{\mathcal{F}_2, \mathcal{L}_7, \mathrm{in}(\eta), \neg\mathrm{in}(\eta)\}$ and the actual labelling $\mathcal{L}_7 = \langle\{\alpha, \eta\}, \{\beta, \gamma\}, \emptyset\rangle$. The intervention $I_1 = \{\mathrm{in}(\alpha)\}$ satisfies the counterfactual condition AC2($\mathrm{a^m}$) in AA with respect to a negation $\neg I_1 = \{\mathrm{out}(\alpha)\}$ and a witness $I_1' = \{\mathrm{out}(\beta)\}$. The intervention $I_2 = \{\mathrm{in}(\beta)\}$ does not satisfy *Actuality* because $\mathrm{in}(\beta)$ is not part of the actual labelling, although there is a $\neg I_2 = \{out(\beta)\}$ and a non-actual witness $I_2' = \{\mathrm{out}(\alpha)\}$, the weak counterfactual $\neg I_2 \cup I_2' \succ \neg\mathrm{in}(\eta)$ holds. So $I_1 = \{\mathrm{in}(\alpha)\}$ is an actual cause of Q_1, but I_2 is not an actual cause of Q_1.

[11] Here we use $\succ$ since we follows the idea of the original HP-definition that there need to be only one counterfactual situation where the contrary of the effect holds.

Example 4.5 Consider the AF $\mathcal{F}_1$ in Figure 2b. Given the contrastive question form $Q_2 = \{\mathcal{F}_1, \mathcal{L}_8, \text{in}(\eta), \text{out}(\eta)\}$ and the actual labelling $\mathcal{L}_8 = \langle\{\gamma, \alpha, \eta\}, \{\beta\}, \emptyset\rangle$, two interventions $I_3 = \{\text{in}(\gamma)\}$ and $I_4 = \{\text{in}(\alpha)\}$ both satisfy *Actuality*. I_3 satisfies the counterfactual condition AC2(a^m) in AA with a witness $I_3' = \emptyset$ and a negation $\neg I_3 = \{\text{out}(\gamma)\}$.

$$\neg I_3 \cup I_3' \succ_{\Delta_{\mathcal{F}_1}} \text{out}(\eta)$$

But I_4 does not satisfy the counterfactual condition AC2(a^m) in AA for any witness I_4' and any negation $\neg I_4$.

$$\neg I_4 \cup I_4' \not\succ_{\Delta_{\mathcal{F}_1}} \text{out}(\eta)$$

So I_3 is an actual cause of Q_2, but I_4 is not an actual cause of Q_2. [12]

Example 4.6 Consider the AF $\mathcal{F}_3$ in Figure 2c. Given the contrastive question form $Q_3 = \{\mathcal{F}_3, \mathcal{L}_9, \text{in}(\eta), \text{out}(\eta)\}$ and the actual labelling $\mathcal{L}_9 = \langle\{\alpha, \beta, \eta\}, \{\gamma\}, \emptyset\rangle$, interventions $I_5 = \{\text{in}(\alpha)\}$ and $I_6 = \{\text{in}(\beta)\}$ do not satisfy the counterfactual condition AC2(a^m) in AA, because either $\text{in}(\alpha)$ or $\text{in}(\beta)$ ensures $\text{in}(\eta)$. $I_7 = \{\text{in}(\alpha), \text{in}(\beta)\}$ satisfies the counterfactual condition since there is a negation $\neg I_7 = \{\text{out}(\alpha), \text{out}(\beta)\}$, such that $\neg I_7 \cup \emptyset \succ_{\Delta_{\mathcal{F}_3}} \text{out}(\eta)$. So I_7 is an actual cause of Q_3, while I_5, I_6 are not.

As suggested in [22], sufficiency follows from AC2(a^m), as well as other properties that have been proposed in the literature. We emphasize two propositions to show that within the context of AA, Sufficiency and Relevance could be derived from the definition of actual cause in AA. We believe that these two properties are essential for a good characterization of explanations based on actual causality.

Proposition 4.7 (Sufficiency) *If I is an actual cause of the contrastive question form $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$, then for any intervention I' which satisfies* Actuality*, we have*

$$I \cup I' > \mathbf{v}(\alpha).$$

Proposition 4.8 (Relevance) *If I is an actual cause of the contrastive question form $Q = (\mathcal{F}, \mathcal{L}, \mathbf{v}(\alpha), \mathbf{e}(\alpha))$, then $\forall \mathbf{v}'(\beta) \in I$, β is relevant to α (that is, there exists a path from β to α).*

5 Related Work

Research on explanation in Abstract Argumentation (AA) has evolved along two main trajectories: using AA to explain other AI domains, and explaining the outcomes of AA itself [17,38]. Positioning our work within this second stream, this section reviews prior research on: explanations in AA, the formal

[12] While Definition 26 in [14] deems either α or γ *sufficient* for the acceptance of η (since both defend η and themselves), our analysis demonstrates that only $\text{in}(\gamma)$ qualifies as the actual cause explaining η's acceptance. This case serves to illustrate a fundamental distinction between our methodological approach and the *sufficient* notion presented in [14].

treatment of causality therein, and specifically, works that leverage causal reasoning for explanatory purposes. We thus focus exclusively on literature at this intersection, excluding broader discussions of XAI or general causal theory.

Explanation methods constitute a natural development of the justification inherent in extension-based (or labelling-based) method. As these semantics (or labellings) have already provided a basic account of the acceptability of arguments, research has shifted towards formulating explanations that capture properties like Monotonicity [35,37], σ-basic [5,37], and Sufficiency [7,14]. These explanation methods—broadly categorized as strong [37], presumptive [10,34], iterative [5], tree-like [20], and root explanation methods [23,24]—are each grounded in distinct intuitions and defined in respective literature [27]. To ensure comparability, it is general to preserve the standard conception of explanations as sets of arguments, facilitating a unified analysis across different approaches.

Causal reasoning is increasingly used to explain autonomous decision-making systems [4,26,28]. The use of causal reasoning to explain the acceptability status of arguments in AA has led to two distinct conceptualizations of counterfactual: one constructs a counterfactual situation by modifying the original AF (e.g., by adding or removing arguments or attacks) [10,31,33]; while the other considers the counterfactual situation as an alternative evaluation outcome of the same AF without altering the AF's structure [1,6,9]. Our work aligns with the second perspective by modeling counterfactual as alternative labellings of an AF, while the key distinction lies in the treatment of causal rule. The prior work often encodes causal relationships as internal inference rules within the structured arguments [6,9]. We operate directly on the AA by formulating causal rule via acceptability rule – a novel concept we introduce that defines the causal dependencies of the acceptance status of arguments, which differs from using the acceptance conditions in ADFs (abstract dialectical frameworks) for explanation [32].

Contrastive explanations have been studied in the context of both abstract and structured argumentation. The most prominent approach is from Borg and Bex, who introduced a method for generating such explanations by pinpointing the differences between a fact and a foil [13]. Similarly, Besnard et al. discuss how to explain "why α is accepted and not β" [7]. However, a commonality of these works is that they provide explanations based on defined properties, without leveraging the formal notion of counterfactuals from causal theory, and our paper fills this void by integrating this key causal concept.

6 Conclusion and Future Work

This paper introduces a causality-based approach to explain the acceptability of arguments in AA. Inspired by Causal Calculus, we develop a framework for counterfactual reasoning via interventions. Utilizing this framework, we formalize the HP-definition of actual causality to construct contrastive explanations in AA. In contrast to previous work, which captures the principles or properties of causal explanations, our method presents a formalized system for modeling

counterfactual reasoning and formulating contrastive explanations through a causal approach within the context of AA.

Future work will proceed in four key directions. Firstly, while the HP-definition introduces three variants of AC2 conditions – AC2(a) + AC2(b^o), AC2(a) + AC2(b^u), and AC2(a^m) – this paper has only addressed AC2(a^m). Future research should systematically compare how these variants characterize counterfactual conditions in AA and influence explanation generation. Secondly, this paper focuses exclusively on the acceptability model and the complete acceptability model (CAM), which correspond to the admissible and complete labellings in AFs. More labelllings (e.g., grounded, preferred, and (semi-)stable labellings) are not considered here but are identified as a direction for future work, where the current model could be extended to capture their requirements. Thirdly, philosophical literature distinguishes contrastive questions based on whether facts and foils are compatible or incompatible; this study focused solely on incompatible cases, leaving compatible ones open for exploration in AA and their corresponding explanatory frameworks. Finally, the relationship – both correlative and differential – between the proposed approach and existing methods warrants deeper investigation.

Acknowledgements. The authors are thankful to the anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (No. 62576309).

References

[1] Alfano, G., S. Greco, F. Parisi and I. Trubitsyna, *Counterfactual and semifactual explanations in abstract argumentation: Formal foundations, complexity and computation*, in: *Proceedings of the Twenty-First International Conference on Principles of Knowledge Representation and Reasoning*, 2024, pp. 14–26.

[2] Baroni, P., M. Caminada and M. Giacomin, *An introduction to argumentation semantics*, The Knowledge Engineering Review **26** (2011), pp. 365–410.

[3] Baroni, P., M. Caminada and M. Giacomin, *Abstract argumentation frameworks and their semantics*, in: *Handbook of Formal Argumentation*, College Publications, London, England, 2018, 1 edition pp. 159–236.

[4] Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai*, Information Fusion **58** (2020), pp. 82–115.

[5] Baumann, R. and M. Ulbricht, *Choices and their consequences - explaining acceptable sets in abstract argumentation frameworks*, in: *Proceedings of the Eighteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2021, pp. 110–119.

[6] Bengel, L., L. Blümel, T. Rienstra and M. Thimm, *Argumentation-based causal and counterfactual reasoning*, in: *CEUR Workshop Proceedings - 1st International Workshop on Argumentation for eXplainable AI (ArgXAI)*, 2022.

[7] Besnard, P., S. Doutre, T. Duchatelle and M.-C. Lagasquie-Schiex, *Explaining semantics and extension membership in abstract argumentation*, Intelligent Systems with Applications **16** (2022), p. 200118.

[8] Bochman, A., "A logical theory of causality," Mit Press, 2021.

[9] Bochman, A., F. Cerutti and T. Rienstra, *Causation and argumentation*, Journal of Applied Logics **12** (2025), pp. 713–786.

[10] Booth, R., D. Gabbay, S. Kaci, T. Rienstra and L. van der Torre, *Abduction and dialogical proof in argumentation and logic programming*, in: *ECAI*, 2014, pp. 117–122.
[11] Borg, A. and F. Bex, *A basic framework for explanations in argumentation*, IEEE Intelligent Systems **36** (2021), pp. 25–35.
[12] Borg, A. and F. Bex, *Necessary and sufficient explanations for argumentation-based conclusions*, in: *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, 2021, pp. 45–58.
[13] Borg, A. and F. Bex, *Contrastive explanations for argumentation-based conclusions*, in: *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022, p. 1551–1553.
[14] Borg, A. and F. Bex, *Minimality, necessity and sufficiency for argumentation and explanation*, International Journal of Approximate Reasoning **168** (2024), p. 109143.
[15] Caminada, M. W. A. and D. M. Gabbay, *A logical account of formal argumentation*, Studia Logica **93** (2009), pp. 109–145.
[16] Čyras, K., D. Birch, Y. Guo, F. Toni, R. Dulay, S. Turvey, D. Greenberg and T. Hapuarachchi, *Explanations by arbitrated argumentative dispute*, Expert Systems with Applications **127** (2019), pp. 141–156.
[17] Čyras, K., A. Rago, E. Albini, P. Baroni and F. Toni, *Argumentative XAI: A survey*, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021, pp. 4392–4399.
[18] Dung, P. M., *On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games*, Artificial Intelligence **77** (1995), pp. 321–357.
[19] Engelmann, D., J. Damasio, A. R. Panisson, V. Mascardi and R. H. Bordini, *Argumentation as a method for explainable ai : A systematic literature review*, in: *2022 17th Iberian Conference on Information Systems and Technologies*, 2022, pp. 1–6.
[20] Fan, X. and F. Toni, *On computing explanations in argumentation*, in: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 15 **29**, 2015, pp. 1496–1502.
[21] Halpern, J. Y., *A modification of the halpern-pearl definition of causality*, in: *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 3022–3033.
[22] Halpern, J. Y., "Actual Causality," MIT Press, 2016.
[23] Liao, B. and L. Van Der Torre, *Explanation semantics for abstract argumentation*, in: *Computational Models of Argument*, 2020, pp. 271–282.
[24] Liao, B. and L. Van Der Torre, *Attack-defense semantics of argumentation*, in: *Computational Models of Argument*, 2024, pp. 133–144.
[25] Lipton, P., *Contrastive explanation*, Royal Institute of Philosophy Supplement **27** (1990), pp. 247–266.
[26] Lipton, P., *What good is an explanation?*, in: *Explanation: Theoretical Approaches and Applications*, Springer, Dordrecht, 2001 pp. 43–59.
[27] Liu, S., Z. Gao, B. Liao and C. Chen, *On pluralistic methods for explaining argument acceptance in abstract argumentation*, in: *Logic and Argumentation*, 2025, pp. 235–253.
[28] Miller, T., *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence **267** (2019), pp. 1–38.
[29] Pearl, J., "Causality: Models, Reasoning, and Inference," Cambridge university press, 2009.
[30] Pisano, G., H. Prakken, G. Sartor and R. Liepina, *Modelling cause-in-fact in legal cases through defeasible argumentation*, in: *Proceedings of the 20th International Conference on Artificial Intelligence and Law*, 2025, pp. 268–277.
[31] Rienstra, T., "Argumentation In Flux (Modelling Change in the Theory of Argumentation)," Ph.D. thesis, Université Montpellier II-Sciences et Techniques du Languedoc; Université du Luxembourg (2014).
[32] Rienstra, T., J. Heyninck, G. Kern-Isberner, K. Skiba and M. Thimm, *Explaining argument acceptance in ADFs*, in: *The First International Workshop on Argumentation for eXplainable AI (ArgXAI)*, 2022.
[33] Sakama, C., *Counterfactual reasoning in argumentation frameworks*, Frontiers in Artificial Intelligence and Applications **266** (2014), pp. 385–396.

[34] Sakama, C., *Abduction in argumentation frameworks*, Journal of Applied Non-Classical Logics **28** (2018), pp. 218–239.
[35] Saribatur, Z. G., J. P. Wallner and S. Woltran, *Explaining non-acceptability in abstract argumentation*, in: *European Conference on Artificial Intellignece*, 2020, pp. 881–888.
[36] Sklar, E. I. and M. Q. Azhar, *Explanation through argumentation*, in: *Proceedings of the 6th International Conference on Human-Agent Interaction*, 2018, pp. 277–285.
[37] Ulbricht, M. and J. P. Wallner, *Strong explanations in abstract argumentation*, in: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021, pp. 6496–6504.
[38] Vassiliades, A., N. Bassiliades and T. Patkos, *Argumentation and explainable artificial intelligence: A survey*, The Knowledge Engineering Review **36** (2021), p. e5.

A^2C: An Adaptive Argumentation-based Classifier for Robust Ensemble Learning

Sheng Wei

ZLAIRE
Zhejiang University

Beishui Liao

ZLAIRE
Zhejiang University

Abstract

Ensemble classifiers are effective yet often opaque, lacking principled mechanisms to resolve conflicting predictions. This paper introduces the **Adaptive Argumentation-based Classifier (A^2C)**, a novel framework that enhances soft-voting ensembles with a dynamic argumentation layer. In A^2C, classifier outputs are modeled as arguments that engage in a formal debate to reconcile disagreements. We propose a new argument strength metric that jointly considers each model's predictive probability and historical reliability. The resulting debate produces an adaptive corrective adjustment to the ensemble's output, with greater influence when model uncertainty is high. Experiments on three public UCI datasets demonstrate that A^2C consistently outperforms a strong weighted soft-voting baseline, achieving up to 1.43% higher accuracy and superior F1, precision, and recall scores. These results suggest that A^2C offers a principled pathway toward more accurate, robust, and interpretable ensemble learning systems.

Keywords: Computational Argumentation, Ensemble Learning, Soft Voting, Conflict Resolution.

1 Introduction

Ensemble learning has emerged as a cornerstone of modern machine learning, consistently achieving state-of-the-art(SOTA) performance across a broad spectrum of classification tasks [12]. By aggregating predictions from multiple diverse models, techniques such as Random Forests [10] and weighted soft voting [18] effectively reduce variance and mitigate the biases of individual classifiers. Soft voting [17], in particular, leverages confidence scores (i.e., predicted probabilities) from each model, often yielding superior results compared to simple majority voting. However, its standard formulation, a straightforward weighted average of probabilities, can be suboptimal, as it lacks a principled

mechanism to resolve sharp conflicts among base classifiers. This limitation becomes particularly problematic when a highly confident but incorrect model dominates a weakly confident yet correct majority.

Simultaneously, the increasing complexity of these models has given rise to the 'black box' problem. This lack of transparency is a critical barrier to adoption in high-risk domains such as medical diagnostics and credit scoring, where understanding the rationale behind a decision is as important as the decision itself. While the field of Explainable AI (XAI) has introduced powerful *post-hoc* explanation techniques such as Variable Importance(VI)[14], Local interpretable model-agnostic explanations (LIME) [15], SHapley Additive exPlanations(SHAP)[19], etc. These methods analyse the model from the outside. They do not fundamentally alter the model's internal decision-making process to be more transparent. There remains a significant need for models that are not only accurate but also inherently interpretable.

To address these two challenges of suboptimal conflict resolution and inherent opacity, we propose the **A**daptive **A**rgumentation-based **C**lassifier (A^{2}C in brief). Unlike existing methods, our A^{2}C model introduces an uncertainty-based dynamic arbitration mechanism. We reframe the task of aggregating predictions as a formal debate, governed by the principles of computational argumentation [11]. In our hybrid structure, the predictions from each base classifier are instantiated as arguments, each with an initial strength derived from a novel blend of the model's real-time confidence and its historical performance. These arguments then attack and support one another within a formal framework. The final justification degree of each argument is used to compute a corrective bonus, which adaptively refines the initial soft-voting scores. Crucially, this argumentation-driven correction has the most influence precisely when the baseline ensemble is the most uncertain, thus targeting the model's weakest points without interfering in cases of high confidence.

This paper makes the following key contributions:

- **A Novel Hybrid Architecture:** We propose A^{2}C, a model that synergistically integrates a weighted soft-voting ensemble with a formal argumentation framework, creating a new meta-reasoning layer for classification.
- **A Dynamic Adjudication Mechanism:** We introduce an uncertainty-aware approach. The ensemble's uncertainty scales the argumentation module. A hybrid metric defines argument strength using predictive probability and historical F1 scores.
- **Comprehensive Empirical Validation:** We demonstrate through experiments on three public UCI datasets that A^{2}C consistently outperforms a strong, weighted soft-voting baseline, achieving an accuracy improvement of up to 1.43%.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 details the proposed A^{2}C framework. Section 4 describes the experimental setup, followed by a discussion of the results in Section 5. Finally, Section 6 concludes the paper and suggests directions for future work.

2 Background and Related Work

Our research is situated at the intersection of two major fields: ensemble learning and computational argumentation. In this section, we first review the foundational concepts from each domain and then discuss prior work that has sought to combine them, positioning our A^2C model within the current SOTA.

2.1 Ensemble Learning and Soft Voting

Ensemble methods are based on the principle that combining multiple models, often called base or weak learners, can lead to a single, more robust model with better generalization performance [12]. The 'wisdom of crowds' effect is achieved through two primary mechanisms: bagging, which reduces variance by training models on different subsets of the data (e.g., Random Forests [10]), and boosting, which reduces bias by training models sequentially, with each new model focusing on the mistakes of its predecessors.

Voting classifiers represent a third, conceptually simpler category of ensemble techniques. In a *hard voting* classifier, the final prediction is the class label that receives the majority of votes from the base models. A more nuanced approach is *soft voting*, which is the direct baseline for our work. In a soft voting ensemble, the final prediction is derived from the average of the predicted probabilities from each classifier. For a set of classifiers $J = \{1, ..., m\}$ and a set of classes $C = \{1, ..., k\}$, the predicted class $\hat{y}$ for an instance $\mathbf{x}$ is given by:

$$\hat{y} = \arg\max_{i \in C} \sum_{j=1}^{m} w_j p_{ij}$$

where w_j is the weight assigned to classifier j, and p_{ij} is the probability that classifier j predicts for class i. These weights are typically determined based on each classifier's performance on a validation set. Despite its effectiveness, the weighted average is a linear and relatively simple aggregation rule, which may not adequately capture the complex relationships and conflicts between model predictions.

2.2 Computational Argumentation

Computational argumentation is a subfield of artificial intelligence that provides formal models for reasoning with conflicting information [9]. The foundational work in this area is Dung's Abstract Argumentation Framework (AAF) [11], defined as a pair $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a set of abstract arguments and $\mathcal{R}$ is an 'attack' relation between them. Classical *extension-based semantics* (often referred to as set-based semantics) are used to determine which subsets of arguments (extensions) can be collectively accepted.

In contrast to this binary accept/reject classification, a family of approaches known as *graded semantics* has emerged. Seminal work in this area includes approaches based on bipolar frameworks (which include both attack and support relations) [5] and social network-inspired models [13]. These semantics assign a numerical score—an acceptability value or justification degree—to each argument. This score offers a more fine-grained evaluation of an argument's

standing, which is particularly well-suited for applications in machine learning where inputs, such as classifier probabilities, are inherently numerical. Our A^{2}C framework adopts this graded semantics approach to quantify the final justification of each argument.

2.3 Argumentation in Machine Learning

The synergy between machine learning and argumentation is a growing area of research. One major trend focuses on leveraging argumentation for XAI. In this paradigm, argumentation is used to generate human-readable explanations for the decisions of black-box models. The work by Amgoud and colleagues is particularly notable, with formalisms proposed for explaining classifiers with arguments [2,3] and for generating robust, sample-based explanations [7,6].

A second trend, more aligned with our work, involves creating hybrid models where argumentation is integral to the classification task itself. Early works explored using argumentation for model selection or to build classifiers based on argumentative inference [16].

Our work is most closely related to the recent argumentation-based ensemble model proposed by Abchiche-Mimouni et al. [1]. Their approach involves extracting explicit classification rules from each base model and using a structured argumentation framework to resolve conflicts between these rules, ultimately using only the 'winning' rules for prediction. While both approaches leverage argumentation to adjudicate classifier conflicts, our A^{2}C model introduces two key novelties. First, we operate at the level of abstract predictions rather than extracted rules, making our framework applicable to any classifier that outputs probabilities without requiring a rule extraction step. Second, and most importantly, we propose a **dynamic adjudication mechanism** where the argumentation module's influence is weighted by the baseline ensemble's uncertainty. This allows for targeted intervention only when necessary, a feature not present in prior models. Due to this fundamental difference in the adjudication mechanism, a direct experimental comparison is a compelling direction for future work.

3 The Proposed Adaptive Argumentation-based Classifier (A^{2}C)

To address the limitations of standard ensemble methods, we introduce the A^{2}C. Our framework enhances a conventional weighted soft-voting ensemble by adding a reasoning layer based on computational argumentation. This layer adjudicates disagreements among base classifiers in a dynamic, context-aware manner. The entire process can be broken down into five sequential stages, as detailed below.

3.1 Overall Architecture

The A^{2}C architecture operates as a two-pathway system that converges at a final adjudication stage. As shown in Figure 1, when an input instance $\mathbf{x}$ is received, it is processed in parallel by:

(i) A baseline weighted soft-voting ensemble, which computes an initial probability distribution over the classes.

(ii) An argumentation module, which translates classifier predictions into a formal argumentation framework, resolves the ensuing debate, and outputs a corrective bonus vector.

Finally, the dynamic adjudicator combines these two outputs, using the uncertainty of the baseline to determine the influence of the argumentation module on the final prediction.

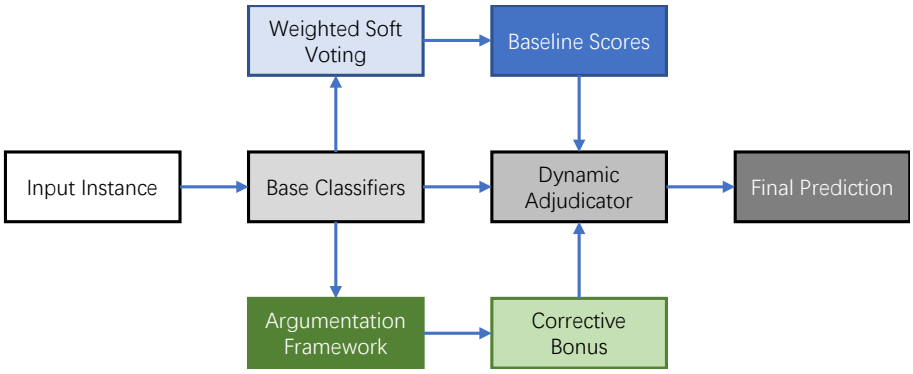


Fig. 1. The overall architecture of the A^{2}C framework, showing the parallel baseline and argumentation pathways converging at the dynamic adjudicator.

3.2 Stage 1: Argument Generation

Let $M = \{m_1, ..., m_n\}$ be the set of n trained and calibrated base classifiers. For a given input instance $\mathbf{x}$, each classifier $m_j \in M$ produces a probability distribution $P(C|\mathbf{x}, m_j)$ over the set of classes $C = \{c_1, ..., c_k\}$. From this distribution, we define c_{top} as the class with the highest predicted probability and c_{sec} as the class with the second-highest predicted probability. We then generate two types of arguments for each classifier:

- **Support Argument:** An argument in favor of the most likely class. For each classifier m_j, if $c_{top} = \arg\max_i P(c_i|\mathbf{x}, m_j)$, we generate a support argument $A_{sup,j}$ with the conclusion c_{top}.
- **Attack Argument:** An argument that opposes the primary conclusion by implicitly supporting an alternative. For each classifier m_j, if there is a second-most likely class c_{sec}, we generate an attack argument $A_{att,j}$ with the conclusion c_{sec}. This argument represents the internal 'doubt' of the classifier.

This process creates a set of up to $2n$ arguments that form the basis of the debate.

3.3 Stage 2: Hybrid Argument Strength Formulation

A key innovation of our model is the formulation of an argument's initial strength, $S(A)$. We define strength as a weighted combination of two factors: the classifier's real-time confidence and its historical reliability. This ensures that arguments are not only judged on their immediate certainty but also on the past performance of their source. Formally, the initial strength of an argument A generated by classifier m_j for class c_i is:

$$S(A) = \alpha \cdot P(c_i|\mathbf{x}, m_j) + (1 - \alpha) \cdot \text{F1}(m_j, c_i) \tag{1}$$

where:

- $P(c_i|\mathbf{x}, m_j)$ is the predicted probability (confidence) of classifier m_j for class c_i.
- $\text{F1}(m_j, c_i)$ is the historical F1-score of classifier m_j for class c_i, pre-computed on a validation set. This constitutes the 'expert profile' of the classifier.
- $\alpha \in [0, 1]$ is a hyperparameter that balances the contribution of real-time confidence versus historical reliability.

For attack arguments, the probability term reflects the confidence in the second-best class, capturing the strength of the opposition.

3.4 Stage 3: Constructing the Argumentation Framework

With arguments and their initial strengths defined, we construct a formal argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is the set of all generated arguments and $\mathcal{R}$ contains attack and support relations.

- **Attack Relation:** An argument A_i attacks an argument A_j if their conclusions conflict and A_i is significantly stronger. Formally, $(A_i, A_j) \in \mathcal{R}_{attack}$ if: $\mathit{conclusion}(A_i) \neq \mathit{conclusion}(A_j)$ and $S(A_i) > S(A_j) \cdot (1 + \tau_{attack})$, where τ_{attack} is a hyperparameter representing the attack threshold. This prevents attacks between arguments of very similar strength, stabilizing the framework.
- **Support Relation:** An argument A_i supports an argument A_j if their conclusions are aligned. Formally, $(A_i, A_j) \in \mathcal{R}_{support}$ if $\mathit{conclusion}(A_i) = \mathit{conclusion}(A_j)$.

3.5 Stage 4: Argumentation Semantics and Justification

Instead of using classical extension-based semantics, we employ a graded semantics approach to calculate a final justification score, $J(A)$, for each argument. This score reflects the argument's acceptability after considering all interactions. The justification scores are computed iteratively until conver-

gence:

$$J_{t+1}(A) = S(A) \cdot \left(1 - \gamma_{\text{attack}} \cdot \max_{B:(B,A)\in\mathcal{R}_{\text{attack}}} J_t(B) + \gamma_{\text{support}} \cdot \tanh\left(\sum_{B:(B,A)\in\mathcal{R}_{\text{support}}} J_t(B)\right)\right) \tag{2}$$

where $J_t(A)$ is the justification of argument A at iteration t, and $\gamma_{attack}, \gamma_{support}$ are hyperparameters. The design of this score is a pragmatic instantiation of key principles for weighted bipolar argumentation frameworks [4,8]. The 'max' operator for attacks reflects the principle that an argument is only as strong as its weakest link. Similarly, the use of 'tanh' for support provides a dampened, cumulative effect, preventing a cascade of weak supporters from overpowering a strong attacker, thus ensuring stability. While other sophisticated scoring functions exist, our formulation provides a robust and computationally efficient method tailored for the classification context.

Illustrative Example. To make this process concrete, consider a scenario where three key arguments have been generated:

- **Argument A1 (from RF):** Concludes **Class 1** with an initial strength of $S(A1) = 0.9$.
- **Argument A2 (from DT):** Concludes **Class 0** with an initial strength of $S(A2) = 0.7$.
- **Argument A3 (from SVM):** Concludes **Class 1** with an initial strength of $S(A3) = 0.8$.
- **Relationships:** A2 attacks A1. Since they share the same conclusion, A3 supports A1, and symmetrically, A1 supports A3. However, when calculating the score for A1, we only consider the support from other arguments.
- **Hyperparameters:** Consistent with our experimental findings (see Section 5.2), we set $\gamma_{attack} = 1.0$ and $\gamma_{support} = 1.0$.

To compute the updated justification score for A1 at the first iteration ($J_1(A1)$), we apply Equation 2:

$$\begin{aligned} J_1(A1) &= S(A1) \cdot (1 - \gamma_{\text{attack}} \cdot J_0(A2) + \gamma_{\text{support}} \cdot \tanh(J_0(A3))) \\ &= 0.9 \cdot (1 - 1.0 \cdot 0.7 + 1.0 \cdot \tanh(0.8)) \\ &\approx 0.9 \cdot (1 - 0.7 + 0.664) = 0.9 \cdot (0.964) \approx 0.8676 \end{aligned}$$

In this round, the justification of A1 decreased from 0.9 to approximately 0.87. The damage from the attacker A2 was largely, but not entirely, offset by the support from A3. The process continues until all scores stabilize.

After convergence, the scores are aggregated into a bonus vector $\mathbf{b} \in \mathbb{R}^k$. For each class c_i, the corresponding bonus b_i is calculated by summing the

justification scores of all arguments concluding c_i:

$$b_i = \sum_{A \in \mathcal{A} | conclusion(A) = c_i} J(A)$$

3.6 Stage 5: Dynamic Adjudication and Final Prediction

The final stage adaptively combines the baseline soft-voting scores with the argumentation bonus. First, we compute the uncertainty of the baseline ensemble. Let $\mathbf{p}_{sv}$ be the probability vector from the weighted soft voting model, with $p_{sv,1}$ and $p_{sv,2}$ being the highest and the second-highest probabilities. The uncertainty score is:

$$U(\mathbf{p}_{sv}) = 1 - (p_{sv,1} - p_{sv,2})$$

High uncertainty (a small gap between the top two probabilities) indicates a contentious case where the argumentation module is most needed. We then compute a dynamic weight, ω_{dyn}:

$$\omega_{dyn} = \begin{cases} 0 & \text{if } p_{sv,1} \geq \tau_{gate} \\ \omega_{max} \cdot U(\mathbf{p}_{sv}) & \text{otherwise} \end{cases}$$

where ω_{max} is the maximum possible weight for the bonus, and τ_{gate} is a confidence gate. If the baseline model is already highly confident, the argumentation is bypassed to preserve efficiency and stability.

Finally, the bonus vector $\mathbf{b}$ is normalized (e.g., L1-norm) to get $\hat{\mathbf{b}}$ and combined with the soft-voting scores to produce the final score vector $\mathbf{S}_{final}$:

$$\mathbf{S}_{final} = \mathbf{p}_{sv} + \omega_{dyn} \cdot \hat{\mathbf{b}} \tag{3}$$

The final predicted class is then $\hat{y} = \arg\max_i S_{final,i}$. This dynamic mechanism ensures that A^{2}C leverages argumentative reasoning when it is most valuable, while defaulting to the efficient baseline when the case is straightforward.

4 Experimental Setup

To validate the performance of our proposed A^{2}C framework, we conducted a series of experiments on three publicly available datasets. This section details the datasets, the models used for comparison, our evaluation protocol, and the hyperparameter tuning strategy.

4.1 Datasets

We selected three well-known classification datasets from the UCI Machine Learning Repository. These datasets were chosen to represent a variety of domains and data characteristics, including a mix of numerical and categorical features, as well as differing numbers of instances and features. A summary of the datasets is presented in Table 1. For all datasets, missing values were imputed using the median for numerical features and the mode for categorical features.

Table 1
Characteristics of the Datasets Used in the Experiments.

Dataset	Instances	Features	Classes	Domain
Heart Disease (Statlog)	270	13	2	Medical
Breast Cancer Wisconsin	699	9	2	Medical
Australian Credit Approval	690	14	2	Financial

4.2 Base Models and Baseline

Our A^2C framework and the baseline model were constructed using a diverse set of seven base classifiers from the scikit-learn library. This diversity is crucial for a strong ensemble. The classifiers are Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM) with a probabilistic kernel, Random Forest (RF), LightGBM (LGBM) and Gaussian Naive Bayes (NB).

For all experiments, we used the default hyperparameters for these base models to ensure a fair and generalizable comparison.

Our primary point of comparison is a strong and widely used ensemble model: the **Weighted Soft-Voting Classifier**. This model serves as our baseline. The weights for each of the seven base classifiers were determined by their overall accuracy on the validation set, providing a robust benchmark that already accounts for the varying performance of individual models.

4.3 Evaluation Protocol

For each dataset, we followed a strict and reproducible evaluation protocol:

(i) **Data Splitting:** The dataset was first split into a training set (70%) and a final test set (30%). We used stratified splitting to maintain the original class distribution in both sets.

(ii) **Internal Splitting:** The 70% training set was further subdivided into a sub-training set (70% of the original training data) and a validation set (30% of the original training data).

(iii) **Model Training and Profiling:** The base classifiers were initially trained on the sub-training set. Their performance on the validation set was used for two purposes: (a) to determine the weights for the baseline soft-voting model, and (b) to build the 'expert profile' (class-specific F1-scores) required by the A^2C framework.

(iv) **Hyperparameter Tuning:** The validation set was also used to find the optimal set of hyperparameters for the A^2C argumentation framework via a grid search.

(v) **Final Evaluation:** Once the optimal hyperparameters were found, the base models were retrained on the full 70% training set. The final, tuned A^2C model and the baseline soft-voting model were then evaluated on the held-out 30% test set. This ensures that the test data was never used for any part of the training or tuning process.

4.4 Metrics and Implementation

The performance of the models was evaluated using four standard classification metrics:

- **Accuracy:** The proportion of correctly classified instances.
- **Weighted Precision:** The precision for each class, weighted by the number of true instances for that class.
- **Weighted Recall:** The recall for each class, weighted by the number of true instances for that class.
- **Weighted F1-Score:** The F1-score computed as the weighted average of the F1-scores for each class.

The weighted metrics are particularly important for providing a balanced view of model performance, especially in cases of class imbalance.

All experiments were implemented in Python using libraries such as scikit-learn, LightGBM, pandas, and NumPy. The experimental results were compiled and saved using the openpyxl library. A fixed random seed (42) was used throughout all experiments to ensure full reproducibility.

5 Results and Discussion

In this section, we present the empirical results of our experiments. We first provide a quantitative comparison of the A^{2}C model against the weighted soft-voting baseline. Finally, we analyze the impact of the argumentation framework.

5.1 Quantitative Performance

Table 2 presents a detailed accuracy comparison across all evaluated models, structured to illustrate the performance progression from individual classifiers to ensemble methods. The results clearly reveal a consistent hierarchy of performance.

First, the data confirms the foundational strength of the ensemble approach. On every dataset, the Soft Voting baseline either matches or exceeds the accuracy of the best-performing individual base classifier. For example, on the Heart Disease dataset, the best base model (RF, shown in *italics*) achieved an accuracy of 87.65%, which the Soft Voting baseline matched.

However, the most compelling finding is that our proposed A^{2}C model consistently establishes a new SOTA on all three datasets. More than just improving upon the baseline, A^{2}C surpasses the performance of the strongest individual component model in every scenario. This is particularly evident on the Breast Cancer Wisconsin dataset. Here, the top-performing base classifier, RF, reached a high accuracy of *96.19%*. While the standard Soft Voting ensemble failed to improve upon this (achieving 95.71%), our A^{2}C model attained an accuracy of **97.14%**, significantly outperforming the best base model by nearly one percentage point.

The performance improvement is most pronounced on the Breast Cancer

Table 2
Detailed accuracy comparison across all models on the test sets. The table shows the performance of individual base classifiers, followed by the ensemble methods. The best performance overall is highlighted in bold, while the top-performing base classifier for each dataset is shown in italics.

Model	Heart Disease	Breast Cancer	Australian Credit
RF	*87.65%*	*96.19%*	*85.51%*
NB	86.42%	95.71%	*85.51%*
LR	86.42%	95.71%	84.54%
LGBM	86.42%	94.29%	*85.51%*
SVM	85.19%	95.71%	83.57%
KNN	83.95%	95.24%	84.06%
DT	70.37%	93.81%	83.09%
Soft Voting	87.65%	95.71%	85.99%
A^2C (Ours)	**88.89%**	**97.14%**	**86.96%**

Wisconsin dataset, where A^2C achieves an accuracy of 97.14%, a significant increase of 1.43% over the baseline's 95.71%. This suggests that in domains where some classifiers may be highly specialized and accurate, the argumentation framework provides a superior mechanism for identifying and promoting the most justified conclusion, rather than simply averaging probabilistic outputs. Across all datasets, the consistent, albeit sometimes modest, gains in all metrics underscore the robustness of the adaptive argumentation approach for refining ensemble predictions.

5.2 Analysis of the Argumentation Framework's Impact

To better understand why A^2C succeeds, we analyze the behavior of the argumentation module. Figure 2 provides a visual comparison of the accuracy improvements. The consistent advantage of A^2C highlights the value added by the reasoning layer.

These findings collectively suggest that A^2C's strength lies in its aggressive yet principled intervention in cases of uncertainty, guided by an argument strength formulation that heavily weights instance-specific evidence.

5.3 Case Study: An Example of Explainable Adjudication

A core claim of our work is that A^2C provides inherent explainability by creating a transparent audit trail for its decisions. To illustrate this, we analyze a specific, challenging instance from the Heart Disease test set. This case is particularly illuminating as the baseline soft-voting model was highly conflicted and leaned towards an incorrect prediction, which our A^2C framework successfully corrected through its argumentation process.

For this instance, the true label was **Class 0** (absence of heart disease). The decision-making process unfolded in four distinct stages:

1. Initial Disagreement and Baseline Failure. The process began with significant disagreement among the base classifiers. The baseline weighted soft-

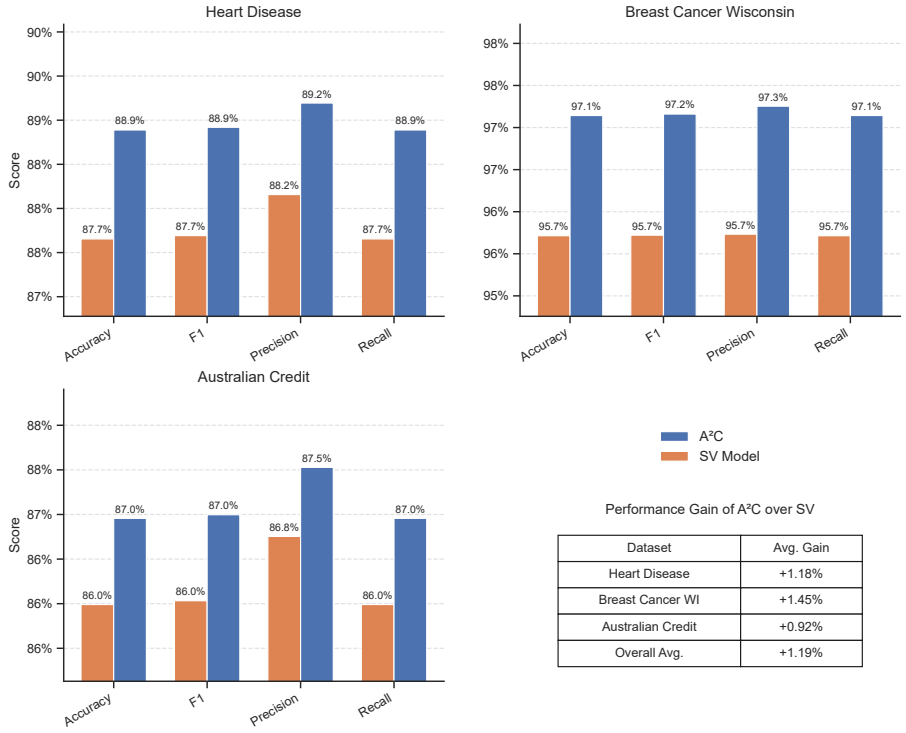


Fig. 2. Accuracy, F1 Score, Precision and Recall comparison between the baseline Soft Voting model and A^{2}C across the three datasets, demonstrating the consistent performance improvement achieved by the argumentation framework.

voting model, aggregating these conflicting signals, was deeply uncertain. It yielded a final score of 0.5057 for Class 1 (presence of disease) versus 0.4943 for Class 0. Based on this razor-thin margin, the baseline model would have made an incorrect prediction. The model's high uncertainty ($U = 0.9886$) signaled a critical disagreement, triggering a strong intervention from our argumentation module with a dynamic weight of $\omega_{dyn} \approx 1.48$.

2. Argument Generation and Key Players. In the argumentation stage, the classifiers' predictions were translated into arguments. The debate had two clear opposing factions. The primary arguments for the **correct Class 0** were championed by the Logistic Regression (LR) and Naive Bayes (NB) classifiers, with strong initial strengths of 0.78 and 0.77, respectively. On the opposing side, a strong coalition argued for the **incorrect Class 1**, led by classifiers like Random Forest (RF, strength 0.70) and Decision Tree (DT, strength 0.67).

3. Debate Outcome and Justification Scores. The core of the explanation lies in the outcome of the argumentative debate. The arguments from LR and NB proved to be overwhelmingly persuasive. Their high initial strengths

allowed them to successfully attack and dismantle the credibility of the opposing arguments. After the iterative justification process, the final scores revealed a decisive victory for the Class 0 faction. The arguments from LR and NB emerged with the highest positive justification scores (1.55 and 1.54, respectively), indicating they had 'won' the debate. Conversely, the leading arguments for the incorrect Class 1 were thoroughly defeated, ending with large negative scores (e.g., RF's argument score plummeted to -1.09, and DT's to -1.04). This signifies that their claims, despite being initially strong, could not withstand the counter-evidence presented by LR and NB.

4. Final Adjudication and Decision Reversal. The final justification scores were aggregated into a net bonus for each class. Class 0 received a large positive bonus from its winning arguments, while Class 1 received a large negative bonus (a penalty) from its defeated arguments. The dynamic adjudicator then combined these with the original, uncertain baseline scores:

- **Score for Class 0:** 0.4943 (baseline) $+ 1.4829 \times$ (Positive Bonus) $= 1.0577$
- **Score for Class 1:** 0.5057 (baseline)$+1.4829\times$(Negative Bonus) $= -0.4137$

The final score for Class 0 became strongly positive, while the score for Class 1 became negative. This decisive reversal led A^2C to correctly predict **Class 0**. This step-by-step process—from initial conflict to final, justified decision—serves as a direct and intelligible explanation for the model's behavior, demonstrating how it resolves disputes not by simple averaging, but through a structured and verifiable reasoning process.

5.4 Computational Overhead

While A^2C provides enhanced accuracy and explainability, it introduces additional computational overhead compared to a standard soft-voting ensemble. The main cost lies in two stages: the construction of the argumentation framework, which has a complexity of $O(N^2)$ where N is the number of arguments (at most twice the number of base classifiers), and the iterative calculation of justification scores. In our experiments, we observed that the justification scores typically converge within a small number of iterations (e.g., fewer than 10). For real-time applications where latency is critical, this modest overhead should be considered. However, for many offline decision-making scenarios, the benefits in performance and transparency can outweigh this additional computational cost.

6 Conclusion and Future Work

6.1 Conclusion

In this paper, we introduced the Adaptive Argumentation-based Classifier (A^2C), a novel hybrid structure that enhances ensemble learning by integrating a formal computational argumentation framework. We have demonstrated that reframing the aggregation of classifier predictions as a structured debate offers a powerful mechanism for resolving conflicts and refining final decisions. Our model's key innovations—a hybrid argument strength formulation, a dynamic

adjudication mechanism sensitive to baseline uncertainty, and an inherently transparent reasoning process—address two of the most pressing challenges in contemporary machine learning: the need for improved accuracy and the demand for greater explainability.

Our comprehensive experiments on three public datasets have shown that A^{2}C consistently outperforms a strong, weighted soft-voting baseline. The empirical results confirm that by applying a principled, argumentative reasoning layer, especially in cases of high uncertainty, we can achieve a more robust and accurate classification. Furthermore, the ability of A^{2}C to generate a clear, visual audit trail for each prediction marks a significant step away from opaque 'black box' models towards a new class of inherently interpretable systems.

6.2 Future Work

The promising results of this study open several exciting avenues for future research. We outline three potential directions:

- **Exploring Advanced Argumentation Semantics:** Our current implementation uses a graded semantics based on an iterative scoring function. Future work could investigate the use of more complex, classical argumentation semantics (e.g., preferred or ideal semantics) to select sets of winning arguments, which might provide different and potentially more robust outcomes in complex conflict scenarios.
- **End-to-End Learning of Framework Parameters:** The hyperparameters of the argumentation framework, such as the strength alpha (α) and attack threshold (τ_{attack}), were determined via a grid search on a validation set. A more sophisticated approach would be to design a differentiable version of the argumentation framework, allowing these parameters to be learned end-to-end during the training process, potentially tailoring the framework more closely to the specific dataset.
- **Application to Structured Prediction and NLP:** The concept of resolving conflicting evidence through argumentation is naturally applicable to domains beyond simple classification. Future research could adapt the A^{2}C framework to structured prediction tasks or to Natural Language Processing (NLP) problems, such as sentiment analysis or question answering, where different models or knowledge sources might provide conflicting textual evidence.

By pursuing these directions, we believe that the synthesis of argumentation and machine learning can continue to push the boundaries of what is possible in building intelligent systems that are not only powerful but also trustworthy and collaborative.

Acknowledgements. The research reported in this paper was supported by the National Natural Science Foundation of China (No. 62576309).

References

[1] Abchiche-Mimouni, N., L. Amgoud and F. Zehraoui, *Explainable ensemble classification model based on argumentation*, in: *22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, 2023, pp. 1–3.

[2] Amgoud, L., *Explaining black-box classification models with arguments*, in: *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2021, pp. 791–795.

[3] Amgoud, L., *Non-monotonic explanation functions*, in: *European Conference on Symbolic and Quantitative Approaches with Uncertainty*, Springer, 2021, pp. 19–31.

[4] Amgoud, L. and J. Ben-Naim, *Weighted bipolar argumentation graphs: Axioms and semantics*, in: *Twenty-Seventh International Joint Conference on Artificial Intelligence-IJCAI 2018*, 2018, pp. 5194–5198.

[5] Amgoud, L., C. Cayrol, M.-C. Lagasquie-Schiex and P. Livet, *On bipolarity in argumentation frameworks*, International Journal of Intelligent Systems **23** (2008), pp. 1062–1093.

[6] Amgoud, L., P. Muller and H. Trenquier, *Argument-based explanation functions*, in: *22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, International Foundation for Autonomous Agents and Multiagent Systems, 2023, pp. 2373–2375.

[7] Amgoud, L., P. Muller and H. Trenquier, *Leveraging argumentation for generating robust sample-based explanations.*, in: *IJCAI*, 2023, pp. 3104–3111.

[8] Baroni, P., A. Rago and F. Toni, *How many properties do we need for gradual argumentation?*, in: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018, pp. 1736–1743.

[9] Besnard, P. and A. Hunter, "An Introduction to Argumentation," The MIT Press, 2008.

[10] Breiman, L., *Random forests*, Machine Learning **45** (2001), pp. 5–32.

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

[12] Hastie, T., R. Tibshirani and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," Springer Science & Business Media, 2009.

[13] Leite, J. a. and J. a. Martins, *Social abstract argumentation*, in: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11 (2011), p. 2287–2292.

[14] Murray, K. and M. M. Conner, *Methods to quantify variable importance: implications for the analysis of noisy ecological data*, Ecology **90** (2009), pp. 348–355.

[15] Ribeiro, M. T., S. Singh and C. Guestrin, *"why should i trust you?": Explaining the predictions of any classifier*, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[16] Thimm, M. and K. Kersting, *Towards argumentation-based classification*, 2017.

[17] Wei, S., Y. Zhang, J. Pan and H. Liu, *A novel preprocessing approach with soft voting for hand gesture recognition with a-mode ultrasound sensing*, in: *International Conference on Intelligent Robotics and Applications*, Springer, 2022, pp. 363–374.

[18] Zhang, Y., S. Wei, Z. Wang and H. Liu, *Dual-modal gesture recognition using adaptive weight hierarchical soft voting mechanism*, IEEE Transactions on Cybernetics (2025).

[19] Zheng, S., Y. Cao and M. Yoshikawa, *Secure shapley value for cross-silo federated learning*, Proceedings of the VLDB Endowment **16** (2023), pp. 1657–1670.

A Principle-Based Robustness Analysis of Labeling-Based Bipolar Argumentation Semantics

Caren Al Anaissy[1], Chen Chen[2], Srdjan Vesic[3],
Leendert van der Torre[2,4], Liuwen Yu[5]

[1]*Sorbonne Université, France*
[2]*Zhejiang University, China*
[3]*CRIL CNRS Univ. Artois, France*
[4]*University of Luxembourg, Luxembourg*
[5]*Luxembourg Institute of Science and Technology, Luxembourg*

Abstract

Bipolar argumentation frameworks (BAFs) instantiate argumentation as balancing by modeling both attacks and supports between arguments, and are increasingly used in systems such as chatbots and debate platforms. Yet their semantics are less explored than those of Dung's abstract frameworks, especially under dynamic change. We present a principle-based robustness analysis of seven variants of labeling-based complete semantics for BAFs. The variants are grounded in three interpretations of support—deductive, necessary, and evidential. We introduce four robustness principles that assess how these semantics respond to changes, namely the addition or removal of attacks and supports. Our classification identifies, for each semantics, the change patterns that preserve labels and those that force changes, yielding a fine-grained picture of robustness across dynamic scenarios. Looking forward, these findings inform the design of efficient algorithms and enforcement procedures for dynamic BAFs (e.g., in argumentation-based chatbots), and articulate a bridge to reasoning alignment between argumentation as dialogue and argumentation as balancing.

Keywords: Artificial intelligence, knowledge representation and reasoning, bipolar argumentation, robustness principle-based approach

1 Introduction

Dung's abstract argumentation framework [14] initiates the attack–defense paradigm shift in formal argumentation [35], where the acceptability of arguments depends on their attack and defense relations rather than their internal structure. This attack-defense paradigm shift provides a unified foundation for reasoning in AI. *Bipolar argumentation frameworks (BAFs)* extend the attack–defense paradigm shift with *support* relations among arguments, thereby instantiating *argumentation as balancing* [34], which has been discussed in the first volume of Handbook of Formal Argumentation [1, Chap.3]: *where both pros and cons are in which arguments for and against alternative resolutions of the issues (options or positions) are put forward, evaluated, resolved, and balanced*. This makes BAFs particularly relevant in contexts

where decisions must be made in the presence of competing interests or values, such as legal reasoning [32], ethical deliberation, and policy-making.

Among the many proposed BAF semantics [9,31,13,25], those based on *interpretations of support* are the most discussed. Under standard interpretations—*deductive*, *necessary*—supports can be used to introduce *indirect attacks* and thus reduce BAFs to abstract attack graphs, illustrating the *universality of attack* [1, Chap. 3]. For instance, under the deductive interpretation [5], if argument A is acceptable and A supports argument B, then B must also be acceptable. In contrast, under the necessary interpretation [22], if B is acceptable and A supports B, then A must be accepted. Both interpretations treat supports as an intermediary step to introduce indirect attacks, reducing a BAF to an abstract argumentation framework. A third notion, evidential support [23], takes a different approach: support relations link arguments to pieces of evidence without enforcing acceptability, representing a qualitatively distinct form of support.

Despite growing interest, the behavior of BAF semantics is still less well understood than that of abstract argumentation frameworks, in particular in *dynamic* settings. In practical systems, agents interact through dialogue (asking, challenging, justifying), which continually *change* the underlying information: new arguments are added, attacks are discovered or retracted, and supports are introduced or withdrawn. A common architecture is therefore two–layered: *dialogue* governs the interaction between agents, while each agent maintains an *individual* reasoning state represented as a BAF that balances pros and cons for its current stance (e.g., in explainable AI assistants and argumentation-based chatbots [4,7,16,18,26,30]). As the dialogue progresses, the agent's BAF evolves—attacks and supports are added or removed, and the interpretation of support may change—requiring repeated semantic evaluation. This raises a question: *how robust are different BAF semantics to such changes (addition/removal of attacks and supports)?*

In this paper, we specifically focus on the robustness of labeling-based bipolar argumentation semantics with necessary, deductive, and evidential support. Starting with necessary and deductive interpretations, we initially follow the reduction from bipolar argumentation framework to abstract argumentation framework [14] by introducing indirect attacks. Subsequently, we define new labeling-based semantics for bipolar argumentation framework with evidential support. In alignment with the work by Rienstra et al. [28], the principles we investigate are also useful in the design of algorithms, thus bridges from formal argumentation to computational argumentation [17, Chap.13]. For example, Niskanen et al. [20] use robustness principles in the design of an algorithm for computing semantics of incomplete argumentation frameworks, where one can specify that attacks between certain arguments may or may not exist. Robustness principles are also useful in addressing enforcement problems in abstract argumentation [3]. This issue involves determining minimal sets of changes to an argumentation framework in order to enforce some result, such as the acceptance of a given set of arguments. Because robustness principles can be used to determine which changes to the attack and support relations of an argumentation framework do or do not change its evaluation, these principles can be used to guide the search for sets of changes in the enforcement problem. This idea has already been used for extension

enforcement under the grounded semantics [21].

The layout of this paper is as follows. We first present the complete semantics of BAF with necessary and deductive interpretations that is based on a reduction approach. In Section 3, we introduce the labeling-based semantics of BAF with evidential support. In Section 4, we conduct the principle-based analysis to robustness of BAF semantics. Section 5 discusses related work and Section 6 concludes and identifies some future work directions.

2 Necessary and Deductive support

This section gives the concept of indirect attack in bipolar argumentation. Dung's argumentation framework [14] consists of a set of arguments and a relation between arguments, which is called attack.

Definition 2.1 An *argumentation framework* is a pair $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ where $\mathcal{A}$ is a set of arguments and $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$ is a binary attack relation over $\mathcal{A}$. We denote by $\mathcal{AF}$ the set of all argumentation frameworks.

Given an argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ we say that an argument $a \in \mathcal{A}$ attacks an argument $b \in \mathcal{A}$ if and only if $(a,b) \in \mathcal{R}$. Given an argumentation framework $\langle \mathcal{A}, \mathcal{R} \rangle$ and an argument $x \in \mathcal{A}$ we denote by x^- the set of arguments attacking x and by x^+ the set of arguments attacked by x. Given a set $B \subseteq \mathcal{A}$ we denote by B^- the set of arguments attacking some $x \in B$ and by B^+ the set of arguments attacked by some $x \in B$.

A labeling-based semantics maps every argumentation framework to a set of labelings, which are functions that map every argument of an argumentation framework to a label. All the labeling-based semantics considered in this paper are defined using three possible labels: I indicates that the argument is accepted, O that the argument is rejected, and U that the acceptance of the argument is undecided.

Definition 2.2 [labeling [6]] A *labeling* of an argumentation framework $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ is a function L: $\mathcal{A} \to \{I, O, U\}$. We denote by $\mathcal{L}(AF)$ the set of all labelings of AF. We also denote a labeling L by the set of pairs $\{(x_1, L(x_1)), ..., (x_n, L(x_n))\}$ where $\mathcal{A} = \{x_1, ...x_n\}$.

Definition 2.3 [labeling-based semantics] A labeling-based semantics σ defines a function $\mathcal{L}_\sigma$ that associates every $AF \in \mathcal{AF}$ with a set $\mathcal{L}_\sigma(AF) \subseteq \mathcal{L}(AF)$.

Definition 2.4 [Complete labeling] Let $AF = \langle \mathcal{A}, \mathcal{R} \rangle$ be an argumentation framework. A labeling $L \in \mathcal{L}(AF)$ is complete if and only if, for all $x \in \mathcal{A}$:

- $L(x) = I$ if and only if, for all $y \in x^-$, $L(y) = O$.
- $L(x) = O$ if and only if, for some $y \in x^-$, $L(y) = I$.
- $L(x) = U$ if and only if, not for all $y \in x^-$, $L(y) = O$ and no $y \in x^-$, $L(y) = I$.

Example 2.5 [Four arguments] The argumentation framework visualized on the left hand side of Figure 1 is defined by $AF = \langle \{a,b,c,d\}, \{(a,b),(b,a),(c,d),(d,c)\} \rangle$. There are nine complete labelings: $\{(a,U),(b,U),(c,U),(d,U)\}$, $\{(a,I),(b,O),(c,U),(d,U)\}, \{(a,O),(b,I),(c,U),(d,U)\}, \{(a,U),(b,U),(c,I),(d,O)\}$,

$\{(a,U),(b,U),(c,O),(d,I)\},\{(a,I),(b,O),(c,I),(d,O)\},\{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,O),(b,I),(c,I),(d,O)\}$, $\{(a,O),(b,I),(c,O),(d,I)\}$.

Fig. 1. An argumentation framework (AF) and a bipolar argumentation framework (BAF)

A bipolar argumentation framework is an extension of Dung's framework. It is based on a binary attack relation between arguments and a binary support relation over the set of arguments.

Definition 2.6 [Bipolar argumentation framework [8]] A *bipolar argumentation framework* (BAF, for short) is a triple $\langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ consists of: a set $\mathcal{A}$ of arguments, a binary relation $\mathcal{R}$ on $\mathcal{A}$ called attack relation, and another binary relation $\mathcal{S}$ on $\mathcal{A}$ called support relation, and $\mathcal{R} \cap \mathcal{S} = \emptyset$.

An AF is a special BAF with the form $\langle \mathcal{A}, \mathcal{R}, \emptyset \rangle$. We denote by $\mathcal{BAF}$ the set of all bipolar argumentation frameworks. A *BAF* can be represented as a directed graph. Given $a,b,c \in \mathcal{A}$, $(a,b) \in \mathcal{R}$ means a attacks b, noted as $a \rightarrow b$; $(b,c) \in \mathcal{S}$ means b supports c, noted as $b \dashrightarrow c$.

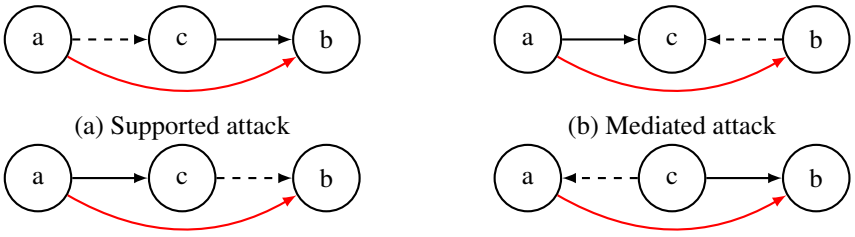
Example 2.7 [Four arguments, continued] The bipolar argumentation framework visualized at the right hand side of Figure 1 extends the argumentation framework in Example 1 so that a supports d.

Support relations only influence the semantics when there are also attacks, which leads to the study of the interactions between attack and support. In the literature, the different kinds of relations between support and attack have been studied as different notions of indirect attack.

Definition 2.8 [Four indirect attacks [24]] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework and $a,b \in \mathcal{A}$, there is:

- a supported attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from a to c and c attacks b, represented as $(a,b) \in \mathcal{R}^{sup}$.
- a mediated attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from b to c and a attacks c, represented as $(a,b) \in \mathcal{R}^{med}$.
- a secondary attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from c to b and a attacks c, $(a,b) \in \mathcal{R}^{sec}$.
- an extended attack from a to b in *BAF* iff there exists an argument c s.t. there is a sequence of supports from c to a and c attacks b, $(a,b) \in \mathcal{R}^{ext}$.

Definition 2.9 [Super-mediated attack [10]] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework and $a,b \in \mathcal{A}$, there is a super-mediated attack from a to b iff



(a) Supported attack (b) Mediated attack

(c) Secondary attack (d) Extended attack

Fig. 2. Four kinds of indirect attack

there exists an argument c such that there is a sequence of supports from b to c and a directly attacks c or supported-attacks c, represented as $(a,b) \in \mathcal{R}^{med}_{R^{sup}}$.

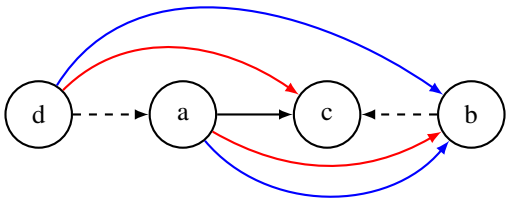


Fig. 3. Super-mediated attack

We can obtain various kinds of indirect attacks according to different interpretations of support relation. These indirect attacks were built from the combination of direct attacks and the supports. Then from the obtained indirect attacks and the support we can build additional indirect attacks and so on.

Definition 2.10 [Tiered indirect attacks [24]] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$, the tiered indirect attacks of BAF are as follows :

- $R_0^{ind} = \emptyset$
- $R_1^{ind} = \{R_\emptyset^{sup}, R_\emptyset^{sec}, R_\emptyset^{med}, R_\emptyset^{ext}\}$
- $R_i^{ind} = \{R_E^{sup}, R_E^{sec}, R_E^{med}, R_E^{ext} \mid E \subseteq R_{i-1}^{ind}\}$ for $i > 1$, where:
  - $R_E^{sup} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from a to c and $(c,b) \in R \cup \bigcup E\}$
  - $R_E^{sec} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from c to b and $(a,c) \in R \cup \bigcup E\}$
  - $R_E^{med} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from b to c and $(a,c) \in R \cup \bigcup E\}$
  - $R_E^{ext} = \{(a,b) \mid$ there exists an argument c s.t. there is a sequence of supports from c to a and $(c,b) \in R \cup \bigcup E\}$

With R^{ind} we denote the collection of all sets of indirect attacks $\bigcup_{i=0}^{\infty} R_i^{ind}$.

Definition 2.11 [Existing reductions of BAF to AF] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle, \forall a,b,c \in \mathcal{A}$:

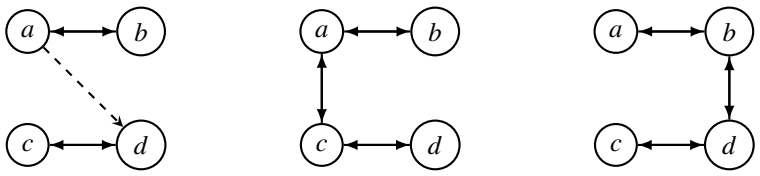
- SupportedReduction [10] (RS for short): $(a,b) \in \mathcal{R}^{sup}$ is the collection of supported attacks iff $(a,c) \in \mathcal{S}$ and $(c,b) \in \mathcal{R}$, $RS(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{sup})$.
- MediatedReduction [10] (RM for short): $(a,b) \in \mathcal{R}^{med}$ is the collection of mediated attacks iff $(b,c) \in \mathcal{S}$ and $(a,c) \in \mathcal{R}$, $RM(BAF) = \big(\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{med}\big)$.
- SecondaryReduction [10] (R2 for short): $(a,b) \in \mathcal{R}^{sec}$ is the collection of secondary attacks iff $(c,b) \in \mathcal{S}$ and $(a,c) \in \mathcal{R}$, $R2(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{sec})$.
- ExtendedReduction [10] (RE for short): $(a,b) \in \mathcal{R}^{ext}$ is the collection of extended attacks, iff $(c,a) \in \mathcal{S}$ and $(c,b) \in \mathcal{R}$, $RE(BAF) = (\mathcal{A}, \mathcal{R} \cup \mathcal{R}^{ext})$.
- DeductiveReduction [24] (RD for short): Let $\mathcal{R}' = \{\mathcal{R}^{sup}, \mathcal{R}^{med}_{\mathcal{R}^{sup}}\} \subseteq \mathcal{R}^{ind}$ be the collection of supported and super-mediated attacks in BAF, $RD(BAF) = (\mathcal{A}, \mathcal{R} \cup \bigcup \mathcal{R}')$.
- NecessaryReduction [24] (RN for short): Let $\mathcal{R}' = \{\mathcal{R}^{sec}, \mathcal{R}^{ext}\} \subseteq \mathcal{R}^{ind}$ be the collection of secondary and extended attacks in BAF, $RN(BAF) = (\mathcal{A}, \mathcal{R} \cup \bigcup \mathcal{R}')$.

Next, we define the labeling-based semantics of BAF.

Definition 2.12 [labeling of BAF] A labeling of a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ is a function L: $\mathcal{A} \to \{I, O, U\}$. We denote by $\mathcal{L}(BAF)$ the set of all labelings of BAF.

Definition 2.13 [Complete labeling of BAF] Let $\omega \in \{RS, RM, R2, RE, RD, RN\}$ be a reduction of BAF to AF. A labeling L of a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ is complete under ω, iff L is a complete labeling of $\omega(BAF)$.

Example 2.14 [Labeling-based semantics of BAF with RD and RN] Consider the bipolar argumentation framework in Figure 4.1. If the interpretation of support from a to d is deductive, a supported-attacks c, c mediated-attacks a. We have $RD(\mathcal{F}) = \langle Ar, att \cup \{(a,c),(c,a)\} \rangle$ as visualized in Figure 4.2, there are five complete labelings: $\{(a,O),(b,I),(c,U),(d,U)\}, \{(a,U),(b,U),(c,O),(d,I)\}, \{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,O),(b,I),(c,O),(d,I)\}, \{(a,O),(b,I),(c,I),(d,O)\}$. If the interpretation of support from a to d is necessary, then b secondary-attacks d, and d extended-attacks b. We have $RN(\mathcal{F}) = \langle Ar, att \cup \{(b,d),(d,b)\} \rangle$ as visualized in Figure 4.3, there are five complete labelings: $\{(a,I),(b,O),(c,U),(d,U)\}$, $\{(a,U),(b,U),(c,I),(d,O)\}$, $\{(a,I),(b,O),(c,O),(d,I)\}$, $\{(a,I),(b,O),(c,I),(d,O)\}$, $\{(a,O),(b,I),(c,I),(d,O)\}$.



4.1 A BAF $\mathcal{F}$ 4.2 RD($\mathcal{F}$) 4.3 RN($\mathcal{F}$)

Fig. 4. Deductive and necessary interpretations give different corresponding AFs

3 Evidential support

BAF with evidential support has been studied by N. Oren and T. J. Norman [23]. They analyse the importance of introducing evidential support into argumentation framework and proposed the traditional extension-based semantics of BAF with evidential support. Besides, they add moreover that elements of evidential support are unique, that support is minimal, and so on.

To keep our presentation uniform and to compare evidential support to deductive and necessary support, we only consider the fragment of bipolar argumentation frameworks where individual arguments attack or support other arguments. This also simplifies the following definitions.

Moreover, evidential support contains special arguments which do not need to be supported by other arguments. Such arguments may have to satisfy other constraints, for example that they cannot be attacked by ordinary arguments, or that they cannot attack ordinary arguments. To keep our analysis uniform, we do not explicitly distinguish such special arguments, but encode them implicitly: if an argument supports itself, then it is such a special argument. This leads to the following definition of an evidential sequence for an argument.

Definition 3.1 [Evidential sequence] Given a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$. A sequence $(a_0, \ldots, a_n)$ of elements of $\mathcal{A}$ is an evidential sequence for argument a_n iff $(a_0, a_0) \in \mathcal{S}$, and for $0 \leq i < n$ we have $(a_i, a_{i+1}) \in \mathcal{S}$.

We give our labeling-based semantics as follows.

Definition 3.2 [Complete labeling of BAF with evidential support] Let $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ be a bipolar argumentation framework. A labeling $L \in \mathcal{L}(BAF)$ is complete under evidential support iff for all $a \in \mathcal{A}$:

(i) $L(a) = I$ iff, there is an evidential sequence $(a_0, \ldots, a)$ for a, s.t. $\forall a' \in \{a_0, \ldots, a\}^-$, it holds that $L(a') = O$.

(ii) $L(a) = O$ iff, for all evidential sequence $(a_0, \ldots, a)$ for a, $\exists a' \in \mathcal{A}$ s.t. $L(a') = I$ and $a' \in \{a_0, \ldots, a\}^-$.

(iii) $L(a) = U$ iff, both of the conditions in (i) and (ii) are not satisfied.

Example 3.3 illustrates the complete semantics of BAF with evidential support.

Example 3.3 Assume a $BAF = \langle \mathcal{A}, \mathcal{R}, \mathcal{S} \rangle$ with evidential support, in which $\mathcal{A} = \{a, b, c, d\}$, $\mathcal{R} = \{(d, b)\}$, $\mathcal{S} = \{(a, a), (a, b), (b, c), (d, d)\}$, as depicted in Figure 5. The only complete labeling of BAF is $\{(a, I), (b, O), (c, O), (d, I)\}$.

4 A principle-based robustness analysis

In this section, we present a principle-based robustness analysis of bipolar argumentation. Due to space limitations, we provide some proofs of results; the remaining all other proofs are available in the supplemental material. [1] Principles 1 to 4 extend the robustness principles introduced by Baroni and Giacomin [2] and further developed

[1] https://drive.google.com/file/d/1c1B2iuAGWokm4FFIAEoM8VUSOZuWO1kt/view?usp=sharing

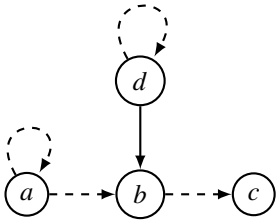


Fig. 5. A BAF with evidential support

by Rienstra et al. [28]. These principles characterize how semantics behave when an argumentation framework is modified by adding or removing an attack relation. In this work, we adapt and apply these principles to bipolar argumentation frameworks, allowing for structural changes involving both attack and support relations.

Principle 1 says that adding an attack between any two arguments does not change the original semantics of the framework.

Principle 1 (Attack addition persistence) *Let σ be a semantics and let $X,Y \in \{O,I,U\}$. We say that σ satisfies XY addition persistence if and only if for all $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ and $x,y \in \mathcal{A}$, if $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$, $L(x) = X$ and $L(y) = Y$, then $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R}\cup\{(x,y)\},\mathcal{S}))$.*

The results for Principle 1 are summarized in Table 1. It shows that all seven semantics preserve labelings when the attacked argument is labeled Out (OO, UO, IO), and all fail when an In-labeled argument is attacked or attacking (UI, IU, II). For the remaining cases: OU, UU, and OI, only RM, RD, and REv preserve the labeling across the board, making them the most robust to added attacks. In contrast, RS, RE, and RN are least robust, failing in all three of these cases.

Table 1
Attack Addition persistence for labeling-based complete semantics. Note that in Tables 1–4 each cell shows whether a semantics *persists* (✓) or *does not persist* (×) after adding or removing an attack/support (x,y). Columns are labeled by the ordered pair **XY**, where **X** is the *original* label of the source argument x and **Y** that of the target y. For example, *IO* in Table 1 denotes that we are adding the attack from an In-labeled argument to an Out-labeled argument.

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	✓	×	✓	×	×	×	✓	×	×
RM	✓	✓	✓	✓	✓	×	✓	×	×
R2	✓	✓	✓	✓	×	×	✓	×	×
RE	✓	×	✓	×	×	×	✓	×	×
RD	✓	✓	✓	✓	✓	×	✓	×	×
RN	✓	×	✓	×	×	×	✓	×	×
REv	✓	✓	✓	✓	✓	×	✓	×	×

Principle 2 says that removing an attack between any two arguments does not change the original semantics of the framework.

Principle 2 (Attack Removal persistence) *Let σ be a semantics and let $X,Y \in$*

$\{O,I,U\}$. *We say that* σ *satisfies XY removal persistence if and only if for all* $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ *and* $x,y \in \mathcal{A}$*, if* $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$*,* $L(x) = X$ *and* $L(y) = Y$*, then* $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R}\setminus\{(x,y)\},\mathcal{S}))$.

Table 2
Attack Removal persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	×	×	×	×	×	✓	✓	✓	✓
RM	✓	✓	✓	×	✓	✓	×	✓	✓
R2	✓	✓	✓	×	✓	✓	×	✓	✓
RE	×	×	×	×	✓	✓	×	✓	✓
RD	✓	×	×	×	×	✓	×	✓	✓
RN	×	×	×	×	✓	✓	×	✓	✓
REv	✓	✓	✓	×	✓	✓	×	✓	✓

The results for Principle 2 are summarized in Table 2. It shows universal satisfaction of UI, IU, and II removal persistence, and universal failure on UU. RM, R2, and REv are the most stable, preserving all other cases, while RS fails nearly all. RN, RE, and RD show mixed sensitivity, especially when an Out argument previously attacked an In or Undecided one. These differences highlight how some semantics tightly couple rejections to attack structure, while others are more relaxed.

Principle 3 says that adding a support relation between any two arguments does not change the original semantics of the framework.

Principle 3 (Support Addition persistence) *Let* σ *be a semantics and let* $X,Y \in \{O,I,U\}$. *We say that* σ *satisfies XY addition persistence if and only if for all* $\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle \in \mathcal{BAF}$ *and* $x,y \in \mathcal{A}$*, if* $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle)$*,* $L(x) = X$ *and* $L(y) = Y$*, then* $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R},\mathcal{S}\cup\{(x,y)\}))$.

The results for Principle 3 are summarized in Table 3. It shows that only the II case is universally preserved, and UO universally fails. RM, RD, and REv again stand out, preserving five of the remaining seven configurations. R2, RE, and RN are the most sensitive, only preserving IO, IU, and II. RS sits in between. The results reflect that semantics differ in how they treat added supports from non-accepted sources.

Proposition 4.1 *The complete semantics satisfy OI, UI and II support addition persistence under supported reduction.*

Proof. For any bipolar argumentation framework $\mathcal{F} = \langle \mathcal{A},\mathcal{R},\mathcal{S}\rangle$ such that there exists the set of arguments $\{l,x,y,z\} \subseteq \mathcal{A}$ where there exists a sequence of supports from l to x and an attack from y to z, let $\mathcal{F}' = \langle \mathcal{A},\mathcal{R}'\rangle$ be the argumentation framework obtained by applying the supported reduction to $\mathcal{F}$. Let $\mathcal{F}_1 = \langle \mathcal{A},\mathcal{R},\mathcal{S}\cup\{(x,y)\}\rangle$, and let $\mathcal{F}'_1 = \langle \mathcal{A},\mathcal{R}'_1\rangle$ be the argumentation framework obtained by applying the supported reduction to $\mathcal{F}_1$. We represent the four argumentation frameworks in Figure 6. Let L be a complete labeling of $\mathcal{F}'$.

- $L(x) \in \{O,I,U\}$ and $L(y) = I$: $L(z) = O$ since $L(y) = I$, $L(l) \in \{O,I,U\}$, from [28],

Table 3
Support Addition persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	×	×	×	×	✓	✓	×	×	✓
RM	✓	✓	×	✓	✓	✓	×	×	✓
R2	×	×	×	×	×	×	✓	✓	✓
RE	×	×	×	×	×	×	✓	✓	✓
RD	✓	✓	×	✓	✓	✓	×	×	✓
RN	×	×	×	×	×	×	✓	✓	✓
REv	✓	✓	×	✓	✓	✓	×	×	✓

the complete semantics satisfy OO, UO and IO attack addition persistence, which means that L is a complete labeling of $\mathcal{F}'_1$.

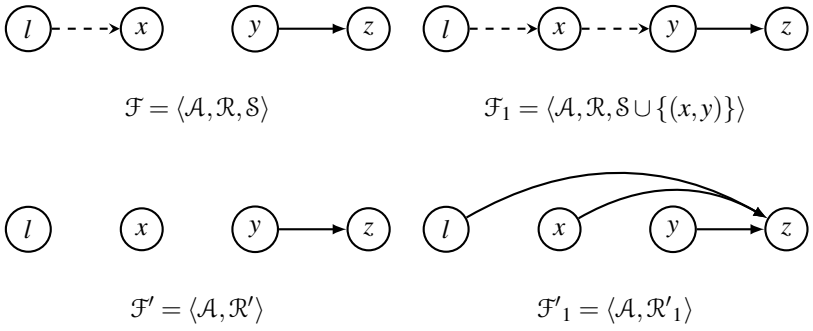


Fig. 6. The complete semantics satisfy OI, UI and II support addition persistence under supported reduction.

□

Proposition 4.2 *The complete semantics violates IO support addition persistence under supported reduction.*

Proof. Consider the following counterexample in Figure 7. The complete labeling of $\mathcal{F}'$ $\{(l,I),(x,I),(b,I),(y,O),(z,I)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(l,I),(x,I),(b,I),(y,O),(z,O)\}$. □

Proposition 4.3 *The complete semantics violates OU support addition persistence under supported reduction.*

Proof. Consider the following counterexample in Figure 8. The complete labeling of $\mathcal{F}'$ $\{(l,I),(x,O),(h,I),(b,U),(y,U),(z,U)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(l,I),(x,O),(h,I),(b,I),(y,O),(z,O)\}$. □

Proposition 4.4 *The complete semantics violates UO support addition persistence under supported reduction.*

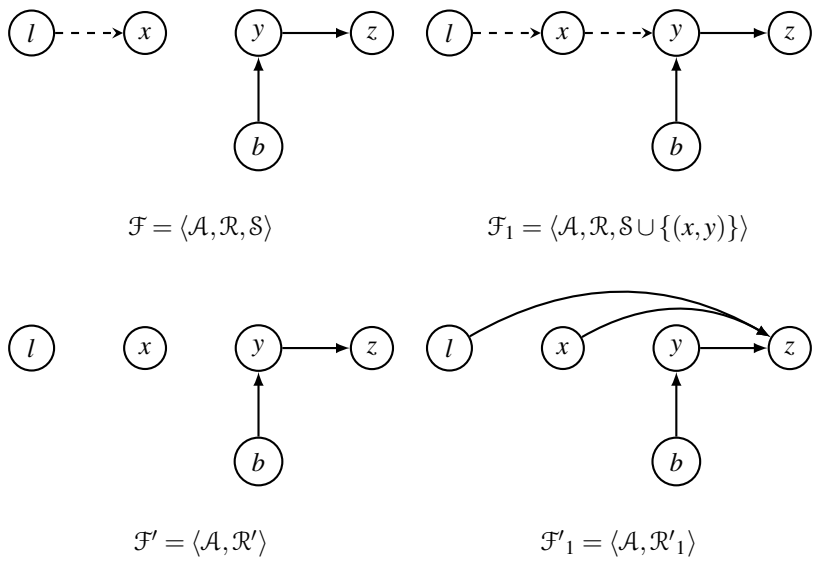


Fig. 7. The complete semantics violates IO support addition persistence under supported reduction.

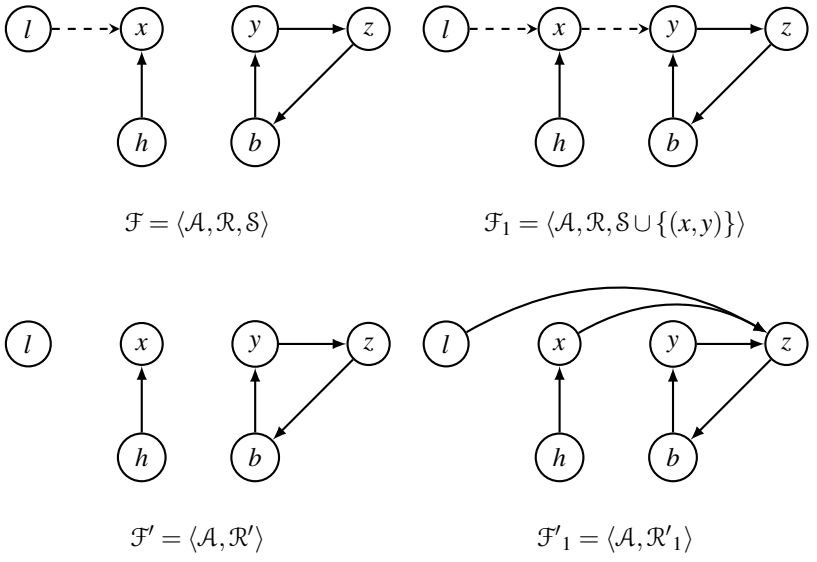


Fig. 8. The complete semantics violates OU support addition persistence under supported reduction.

Proof. Consider the following counterexample in Figure 9. The complete labeling of $\mathcal{F}'$ $\{(f,I),(l,O),(x,U),(b,I),(y,O),(z,I)\}$ is no longer a complete labeling after adding the attacks from l and x to z in $\mathcal{F}'$. The new complete labeling (of $\mathcal{F}'_1$) is $\{(f,I),(l,O),(x,U),(b,I),(y,O),(z,U)\}$. □

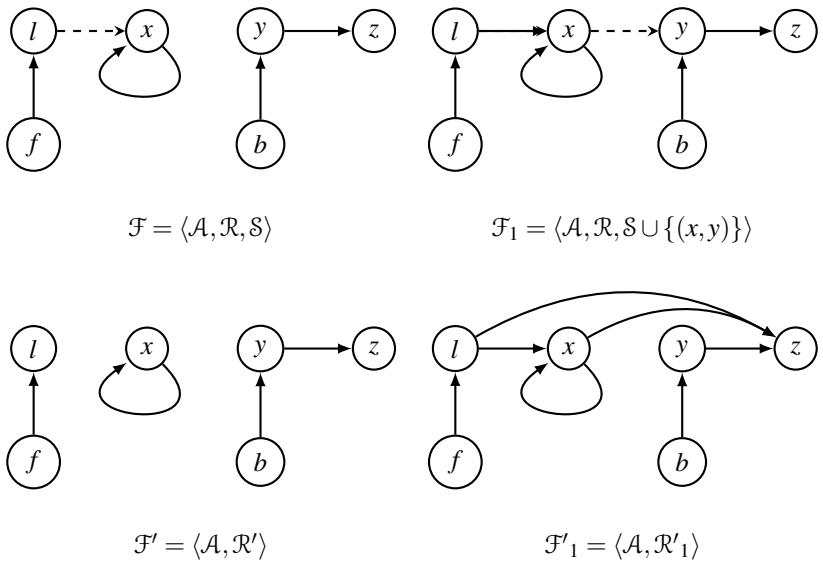


Fig. 9. The complete semantics violates UO support addition persistence under supported reduction.

Principle 4 says that removing a support between any two arguments does not change the original semantics of the framework.

Principle 4 (Support Removal persistence) *Let σ be a semantics and let $X,Y \in \{O,I,U\}$. We say that σ satisfies XY removal persistence if and only if for all $\langle \mathcal{A},\mathcal{R},\mathcal{S} \rangle \in \mathcal{BAF}$ and $x,y \in \mathcal{A}$, if $L \in \mathcal{L}_\sigma(\langle \mathcal{A},\mathcal{R},\mathcal{S} \rangle)$, $L(x) = X$ and $L(y) = Y$, then $L \in \mathcal{L}_\sigma((\mathcal{A},\mathcal{R},\mathcal{S} \setminus \{(x,y)\}))$.*

The results for Principle 4 are summarized in Table 4. It shows that all semantics preserve UI, and most preserve IO and II—with the exception of REv, which fails both due to its direct reliance on evidential chains. RM and R2 are the most robust, failing only in OO and UU. RN, RS, and RE are more fragile when supports between Out-labeled or Undecided arguments are removed, which calls for reevaluation.

Table 4
Support Removal persistence for labeling-based complete semantics

	OO	OU	UO	UU	OI	UI	IO	IU	II
RS	✓	✓	×	✓	✓	✓	×	×	✓
RM	×	✓	✓	×	✓	✓	✓	✓	✓
R2	×	✓	✓	×	✓	✓	✓	✓	✓
RE	✓	×	✓	×	×	✓	✓	✓	✓
RD	×	✓	✓	×	×	✓	✓	✓	✓
RN	×	×	✓	×	×	✓	✓	✓	✓
REv	✓	✓	✓	×	✓	✓	✓	×	×

5 Related work

Support relations, unlike the attack relation, remain controversial in the literature. The evaluation of bipolar argumentation semantics through a principle-based lens is relatively recent but growing. For example, there are studies analyzing bipolar argumentation semantics by focusing on different interpretations of support [11,33]. This line of research has been further extended to domains such as legal reasoning, where multiple interpretations of support correspond to different legal interpretations [32]. Principle-based analyses have also been applied to new semantics defined using novel notions of defense and to social choice-based approaches to argument evaluation [31]. Doder et al. [13] specifically investigate principle-based characterizations of ranking semantics for frameworks with necessities.

Applications of BAFs have been particularly studied in fields like explainable AI (XAI) and argumentation-based chatbots, providing a motivation for our research. In the context of XAI, Kampik et al. explore the changes in quantitative bipolar argumentation frameworks to provide sufficient, necessary, and counterfactual explanations in response to updates within these frameworks [18]. It underscores the importance of studying the dynamic aspects of BAFs, highlighting their crucial role in understanding operational dynamics. In the realm of chatbots, as Federico Castagna et al noted [7]: "Speaking of the underlying argumentation framework of argumentation-based chatbots, when embedding a knowledge base into an AF, the Bipolar framework (and its variants QBAF and WBAF) turns out to be the most common option. This choice is related to the additional information provided by BAFs which encompass support relations rather than just attacks, allowing for an intuitive formalisation of both endorsements and conflicts between pieces of data." For instance, the interactive recommender systems developed by Rago et al. [27,26] utilize a BAF and tripolar argumentation framework to embed their underlying knowledge bases, thereby enhancing the clarity of their recommendations through rich, argument-based explanations. In a similar vein, Cocarascu et al. describe argumentative dialogical agents that construct a quantitative bipolar argumentation framework to facilitate structured dialogues based on movie reviews [12]. The integration of BAFs with advances in generative AI and hybrid models has fostered innovations such as ArguBot [4], developed using Google DialogFlow [30]. This system employs ASPARTIX [15] to compute arguments from an underlying BAF to support (pro-bot) or challenge (con-bot) the user's opinion about the topic of dialogue. Lastly, the conversational agent designed by Fazzinga et al. incorporates BAFs to manage dialogues and argumentation effectively, showcasing the versatility and extensive applicability of BAFs in contemporary AI applications [16]. All these developments underline the significance of ongoing studies into the dynamics and robustness of bipolar argumentation semantics.

6 Summary

In this paper we analysed robustness properties for seven variants of the complete semantics for bipolar argumentation frameworks. Six variants arise from reduction-based approaches that interpret support as necessary or deductive, while the seventh is defined directly for evidential support. We use four robustness principles—the two

attack-oriented principles of Rienstra et al. [28] together with two support-oriented principles—and compare the variants exhaustively. Tables 1–4 make the impact of adding or removing attacks or supports explicit, allowing practitioners (1) to select a semantics that remains stable under the specific updates their application performs, and (2) to identify precisely the situations in which recomputation of labels is unavoidable.

Beyond these results, our analysis situates naturally within the A-BDI metamodel (Argumentation as Balancing, Dialogue, and Inference) [34]. A-BDI views the three conceptualizations as complementary rather than exclusive, and principles as a means to select among existing methods or to define new ones. This perspective aligns with the reasoning alignment view [29]: Reasoning Alignment Diagrams (RADs) are commutative reasoning representations that align a source specification with an argumentation-based explanation path; they compose an "assert (inference)" RAD with a "listen (revision)" RAD to model dialogue—agents that can say (argumentation/inference) and hear (belief revision) while preserving alignment. Our robustness classification supports this agenda by indicating when local inference within a BAF remains stable under dialogue-driven updates, and how changes to relations or to the interpretation of support affect balancing in dynamic contexts.

Future work. A natural next step is to generalise reasoning alignment between argumentation as inference and argumentation as balancing, by studying bipolar argumentation within structured argumentation [19]. In parallel, the rise of large language models foregrounds the dialogue perspective: in multi-agent settings, each agent can maintain a local BAF as its individual reasoning state while the dialogue protocol drives assert, question, and revise moves; our robustness results then indicate when edits prompted by these moves leave the agent's labels stable and when recomputation is required. On the technical side, we plan to extend the robustness analysis beyond complete semantics (e.g., grounded, preferred, stable); to investigate settings where interpretations of support vary over time or across agents and use our tables to anticipate when acceptability persists or changes; and to formalise optimisation heuristics suggested by Tables 1–4 (for instance, "no-recompute" cases under specific edit patterns) to support incremental solvers and enforcement procedures in dynamic BAF-based systems.

Acknowledgments

We thank the anonymous reviewer for their comments. This work is supported by the Luxembourg National Research Fund (FNR) through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx), Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN), and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

References

[1] Baroni, P., D. Gabbay, M. Giacomin and L. van der Torre, editors, **1**, College Publications, 2018.

[2] Baroni, P. and M. Giacomin, *On principle-based evaluation of extension-based argumentation semantics*, Artificial Intelligence **171** (2007), pp. 675–700.

[3] Baumann, R., S. Doutre, J.-G. Mailly and J. P. Wallner, *Enforcement in formal argumentation*, IfColog Journal of Logics and their Applications (FLAP) **8** (2021), pp. 1623–1678.

[4] Bistarelli, S., C. Taticchi and F. Santini, *A chatbot extended with argumentation*, in: M. D'Agostino, F. A. D'Asaro and C. 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*, CEUR Workshop Proceedings **3086** (2021).

[5] Boella, G., D. M. Gabbay, L. van der Torre and S. Villata, *Support in abstract argumentation*, in: *Proceedings of the Third International Conference on Computational Models of Argument (COMMA'10)*, Frontiers in Artificial Intelligence and Applications, IOS Press, 2010, pp. 40–51.

[6] Caminada, M., *On the issue of reinstatement in argumentation*, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2006, pp. 111–123.

[7] Castagna, F., N. Kokciyan, I. Sassoon, S. Parsons and E. Sklar, *Computational argumentation-based chatbots: a survey*, arXiv preprint arXiv:2401.03454 (2024).

[8] Cayrol, C. and M.-C. Lagasquie-Schiex, *On the acceptability of arguments in bipolar argumentation frameworks*, in: *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer, 2005, pp. 378–389.

[9] Cayrol, C. and M.-C. Lagasquie-Schiex, *Bipolar abstract argumentation systems*, in: *Argumentation in Artificial Intelligence*, Springer, 2009 pp. 65–84.

[10] Cayrol, C. and M.-C. Lagasquie-Schiex, *Bipolarity in argumentation graphs: Towards a better understanding*, International Journal of Approximate Reasoning **54** (2013), pp. 876–899.

[11] Cayrol, C. and M.-C. Lagasquie-Schiex, *An axiomatic approach to support in argumentation*, in: *International Workshop on Theory and Applications of Formal Argumentation*, Springer, 2015, pp. 74–91.

[12] Cocarascu, O., A. Rago and F. Toni, *Extracting dialogical explanations for review aggregations with argumentative dialogical agents.*, in: *AAMAS*, 2019, pp. 1261–1269.

[13] Doder, D., S. Vesic and M. Croitoru, *Ranking semantics for argumentation systems with necessities*, in: *IJCAI 2020-29th International Joint Conference on Artificial Intelligence*, International Joint Conferences on Artificial Intelligence Organization, 2021, pp. 1912–1918.

[14] Dung, P. M., *On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games*, Artificial Intelligence **77** (1995), pp. 321–357.

[15] Egly, U., S. A. Gaggl and S. Woltran, *Aspartix: Implementing argumentation frameworks using answer-set programming*, in: *International Conference on Logic Programming*, Springer, 2008, pp. 734–738.

[16] Fazzinga, B., A. Galassi and P. Torroni, *An argumentative dialogue system for covid-19 vaccine information*, in: *International Conference on Logic and Argumentation*, Springer, 2021, pp. 477–485.

[17] Gabbay, D., G. Kern-Isberner, G. Simari and M. Thimm, editors, **3**, College Publications, 2024.

[18] Kampik, T., K. Čyras and J. R. Alarcón, *Change in quantitative bipolar argumentation: Sufficient, necessary, and counterfactual explanations*, International Journal of Approximate Reasoning **164** (2024), p. 109066.

[19] Müller, M. A., S. Vesic and B. Yun, *Interpreting preferred semantics in structured bipolar argumentation* (2025).

[20] Niskanen, A., D. Neugebauer, M. Järvisalo and J. Rothe, *Deciding acceptance in incomplete argumentation frameworks*, in: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-20)* (2020), pp. 2942–2949.

[21] Niskanen, A., J. P. Wallner and M. Järvisalo, *Extension enforcement under grounded semantics in abstract argumentation*, in: *Sixteenth International Conference on Principles of Knowledge Representation and Reasoning*, 2018.

[22] Nouioua, F. and V. Risch, *Argumentation frameworks with necessities*, in: *International Conference on Scalable Uncertainty Management*, Springer, 2011, pp. 163–176.

[23] Oren, N. and T. J. Norman, *Semantics for evidence-based argumentation*, in: *Computational Models of Argument*, IOS Press, 2008 pp. 276–284.
[24] Polberg, S., *Intertranslatability of abstract argumentation frameworks*, Technical Report DBAI-TR-2017-104, Institute for Information Systems, Technical University of Vienna (2017).
[25] Potyka, N., *Continuous dynamical systems for weighted bipolar argumentation*, in: M. Thielscher, F. Toni and F. Wolter, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference, KR 2018, Tempe, Arizona, 30 October - 2 November 2018* (2018), pp. 148–157.
[26] Rago, A., O. Cocarascu, C. Bechlivanidis, D. Lagnado and F. Toni, *Argumentative explanations for interactive recommendations*, Artificial Intelligence **296** (2021), p. 103506.
[27] Rago, A., O. Cocarascu and F. Toni, *Argumentation-based recommendations: Fantastic explanations and how to find them.*, , **18**, 2018, pp. 1949–1955.
[28] Rienstra, T., C. Sakama, L. van der Torre and B. Liao, *A principle-based robustness analysis of admissibility-based argumentation semantics*, Argument & Computation (2020), pp. 1–35.
[29] Rienstra, T., L. van der Torre and L. Yu, *Reasoning alignment for Agentic AI: Argumentation, belief revision, and dialogue*, Journal of Applied Logics - IfCoLog Journal **12** (2025), pp. 1683–1712.
[30] Sabharwal, N. and A. Agrawal, *Introduction to google dialogflow*, Cognitive virtual assistants using google dialogflow: develop complex cognitive bots using the google dialogflow platform (2020), pp. 13–54.
[31] Yu, L., C. Al Anaissy, S. Vesic, X. Li and L. van der Torre, *A principle-based analysis of bipolar argumentation semantics*, in: *European Conference on Logics in Artificial Intelligence*, Springer, 2023, pp. 209–224.
[32] Yu, L., R. Markovich and L. van der Torre, *Interpretations of support among arguments*, in: *Legal Knowledge and Information Systems*, IOS Press, 2020 pp. 194–203.
[33] Yu, L. and L. van der Torre, *A principle-based approach to bipolar argumentation*, in: *NMR 2020: Non-Monotonic Reasoning Workshop Notes*, 2020, CEUR Workshop Proceedings, Vol. 2672.
[34] Yu, L. and L. van der Torre, *The A-BDI metamodel for human-level AI: Argumentation as balancing, dialogue and inference*, in: *International Conference on Logic and Argumentation*, Springer, 2025, pp. 361–379.
[35] Yu, L., L. van der Torre and R. Markovich, *Thirteen challenges of formal and computational argumentation*, in: M. Thimm and G. R. Simari, editors, *Handbook of Formal Argumentation, Volume 3*, forthcoming .

Binary Spiking Neural Networks as Causal Models

Aditya Kar

Institut de Recherche en Informatique de Toulouse (IRIT),
Centre de Recherche Cerveau et Cognition (CerCo), CNRS, France

Emiliano Lorini

Institut de Recherche en Informatique de Toulouse (IRIT), CNRS, France

Timothée Masquelier

Centre de Recherche Cerveau et Cognition (CerCo), CNRS, France

Abstract

We provide a causal analysis of Binary Spiking Neural Networks (BSNNs) to explain their behavior. We formally define a BSNN and represent its spiking activity as a binary causal model. Thanks to this causal representation, we are able to explain the output of the network by leveraging logic-based methods. In particular, we show that we can successfully use a SAT as well as a SMT solver to compute abductive explanations from this binary causal model. To illustrate our approach, we trained the BSNN on the standard MNIST dataset and applied our SAT-based and SMT-based methods to finding abductive explanations of the network's classifications based on pixel-level features. We also compared the found explanations against SHAP, a popular method used in the area of explainable AI. We show that, unlike SHAP, our approach guarantees that a found explanation does not contain completely irrelevant features.

Keywords: Logic, Causality, Explanation, Binary Spiking Neural Networks

1 Introduction

In recent times, interest in the study of binary artificial neural networks has grown, where binarization can occur at the level of the connection weights between the neural units, at the level of their activation function, or at both levels. In the field of AI, binarized neural networks (BNNs) were recently proposed by [20] and [40], while in neuroscience particular attention has been paid to binary spiking neural networks (BSNNs) [25,29]. The main difference between BNNs and BSNNs is mainly due to the presence of temporal dynamics in BSNNs over BNNs and to the fact that in BSNNs inputs are given sequentially in discrete time, while they are instantaneously presented to BNNs. Binarization obviously comes with a price in terms of the size of the network parameters relative to its learning power: a binary neural network requires a considerably higher number of neural units, compared to its non-binary counterpart, in order to achieve an acceptable level of accuracy in a given classification

task after training. Nonetheless, this disadvantage is counterbalanced by an advantage in terms of logical representability and therefore explainability. Specifically, thanks to the Boolean nature of BNNs and BSNNs, one can represent their firing dynamics as binary causal models and, consequently, explain their behaviors in an efficient way using logic-based methods.

The present paper is devoted to exploring this trade-off between accuracy and explainability in the context of BSNNs. We focus our analysis on BSNNs instead of BNNs since, from the causal point of view, the former are more general than the latter and we prefer to concentrate on the more general model first. To fully capture the causal structure of a BSNN, one has to model the firing activities of its neural units and to represent their causal dependencies over an extended time span. BNNs are less general since the presentation of the input is not sequential and, consequently, their dynamics and the resulting causal dependencies between the neural units do not extend over time. We represent the internal mechanism of a BSNN through a binary causal model and, thanks to this representation, we explain the BSNN's behavior. Different notions of explanation exist in the literature including abductive [22], contrastive [34], counterfactual [45] and alterfactual [33] explanation. In the present paper, we rely on abductive explanation (AXp) because of its simplicity and its emphasis of minimality which is a guarantee of non-redundancy. For a set of input features to be an abductive explanation of a classification by a neural network, it has to be *minimally* sufficient to ensure the classification, i.e., where minimality means that all proper subsets of features are no longer sufficient for the classification. Thus, an AXp is by definition non-redundant. The causal component will be essential to our analysis. Having an explicit representation of the BSNN's causal dependencies will enable us to formally verify that all features included in an explanation are genuinely causally relevant, unlike traditional machine learning explainability methods such as SHAP.

The paper is structured as follows. After discussing related work, we illustrate the BSNN architecture as well as the learning task we considered, namely MNIST classification, and the learning algorithm we used to train our BSNNs on the MNIST dataset. Then, we focus on the mathematical aspects of our framework. First, we introduce the mathematical model of the BSNN spiking dynamics. Then, we map it onto a binary causal model that represents the causal dependencies between the firing activities of the neural units over time. We then move to the explanation of the BSNN behavior. Specifically, we present an algorithm that combines the binary causal model with a SAT solver to compute abductive explanations of the BSNN classification, where an abductive explanation is constructed from pixel-level features at a specific time point. We also consider a variant of the abductive explanation search algorithm based on a SMT (Satisfiability Modulo Theory) encoding of the binary causal model for the BSNN architectures. We present some experimental results on computation time for the SAT-based and for the SMT-based explanation search algorithm. Finally, in Section 8 we compare our logic-based approach relying on abductive explanation with SHAP.

To the best of our knowledge, this is the first attempt i) to map a BSNN onto a binary causal model, and ii) to leverage the resulting Boolean representation of causal dependencies among its neural units to explain its behavior using both SAT and SMT

solvers.

2 Related Work

We organize the discussion of relevant literature in three parts: binary neural networks, causal models, and logic-based explanation of artificial neural networks.

Binary neural networks Binary Neural Networks (BNNs) are a class of artificial neural networks (ANNs) that have been studied extensively by researchers [39] in the deep learning community, especially by [8] and [20], who provided a viable way to train these networks using standard back-prop based optimisation methods. BNNs adopt an extreme form of quantization, by resorting to binary weights and binary activation values. [44] have shown that with back-prop based methods, it is possible to train these binarized neural networks with reasonable, near full precision accuracy. Moreover, [40] demonstrated a drastic reduction in computation time and model size with XNOR-Nets, due to the fact that the computationally expensive multiply-accumulate operations in deep learning can be replaced by faster XNOR and popcount operations when using binarized networks. Hence, for these reasons, BNNs have gained significant popularity in resource-constrained, low-power, and hardware-efficient AI applications. Binary Spiking Neural Networks (BSNNs), the subject of the present paper, are the bio-plausible counterpart of BNNs, taking inspiration from the spiking dynamics of biological neurons in the brain. The most useful feature of BSNNs is the way they process input data using spike encodings, where spikes are binary all-or-none pulses occurring at discrete time steps, as opposed to the continuous-valued representations used in conventional ANNs (including BNNs). These spike encodings are particularly convenient, as they allow us to apply our formalism to both the pixel space and the intermediate feature space. BSNNs have been trained using both temporal [25] and rate coding schemes [29].

Causal models Causal models are mathematical objects that have been extensively studied in AI [36], logic [13,15], and in the field of explainable AI [34], given the urgent need to provide formally rigorous causal explanations of AI systems. A causal model is a system of structural equations describing the causal dependencies between variables. Binary causal models (BCMs) that we use in the present work are the subclass of causal models in which variables are assumed to be Boolean. They were studied in depth in previous work [9,1,28,12]. Given their close connection with propositional logic, they offer the possibility to automate reasoning about causality with the aid of a SAT solver.

Abductive explanation of artificial neural networks The central concept in the field of logic-based explanations for artificial neural networks (ANNs)—and more broadly for machine learning models—is the *abductive explanation* (AXp) [10], which forms the foundation of the present work. This notion builds on prior theoretical research on *abduction* [32] and is grounded in the concept of the *prime implicant* (PI). For this reason, it is also referred to as a *PI-explanation* [42] or a *sufficient reason* [11]. Abductive explanations have been applied to both tractable models, such as monotone and linear classifiers[31,10,3], and intractable ones, including random forests [23], boosted trees [2], and artificial neural networks [41,22]. In [41], binary neural net-

works are compiled into Ordered Binary Decision Diagrams (OBDDs), which are then used to compute AXps for the networks' classifications. In contrast, [22] employ a *Mixed Integer Linear Programming* (MILP) formulation to derive AXps for a neural network's classifications in a three-digit MNIST task. Unlike our work and that of [41], [22] focus on neural networks with real-valued weights. Our approach differs from [22] and [41] in two key respects. First, causality plays a central role in our framework: we map a BSNN onto a binary causal model and leverage this causal representation to generate explanations. In contrast, neither [22] nor [41] incorporate any notion of causality. Second, they do not consider BSNNs, whereas BSNNs are the central focus of our analysis and the type of neural networks we aim to explain using logic and causal models.

Before concluding, it is worth mentioning the work on argumentation-based explanations of multi-layer perceptrons (MLPs) presented in [4]. This approach builds on the mathematical relationships between MLPs and *quantitative argumentation frameworks* (QAFs) established in [37]. The proposed method first sparsifies an MLP and then maps the resulting network onto an equivalent QAF, which can be used to explain and interpret the model's underlying mechanisms and decisions. Although this approach takes a different perspective, without an explicit grounding in logic or causality, we believe that a connection could be drawn between QAFs and causal models with continuous variables, and thus between the MLPs studied in [4] and causal models. We leave this question for future work.

3 Architecture, Learning and Dataset

In this section, we outline the details of the neural network models that we considered, along with the exact learning task, dataset and accuracies.

3.1 Learning task

For our training purposes, we used the MNIST classification task for hand written digit recognition. We trained networks with a single fully connected hidden layer on both tasks, 3-digit and 10-digit MNIST classification. As we will show in Table 1, we could achieve very high accuracy with binary quantized networks on the 3-digit classification task. We could also achieve a high accuracy on the 10-digit classification task with three-value quantized networks with weights ranging over $\{-1, 0, 1\}$.

3.2 Spike encoding

For our experiments, we used two different approaches to convert MNIST images into spikes. Firstly, we used a classic Poisson rate coding scheme [38] to convert images into spike trains in multiple time-steps and also a threshold-binarized scheme with just one time-step as presented in Table 1. We did not pursue temporal coding in our experiments since, as shown by [25], temporal coding requires larger time-steps for training with high accuracy. Since having more time-steps significantly increases the complexity of finding an explanation, we chose to not use temporal coding in this work. Nonetheless, the novel mapping of BSNNs to binary causal models we will present can be generalized to to other forms of spike encodings. We used a simple Integrate and Fire (IF) model for our spiking neurons, since mapping BSNNs to binary

causal models is easier in the absence of leaks.

3.3 Weight quantization

As we will demonstrate later in the paper, mapping a BSNN to a binary causal model requires the network to have weights quantized either in a binary (i.e., $\{0, 1\}$) or a three-valued (i.e., $\{-1, 0, 1\}$) way. To train our networks, the weight quantization procedure that we adopted closely follows the XNOR-Net proposal by [40], i.e., during a forward pass the network uses a binarized weight matrix $\mathcal{B}(W)$, while during the backward pass it retains a proxy full-precision weight matrix W for gradient calculation. Straight-through-estimator (STE) [8] was used without any gradient clipping for our training. The following equations represent the two variants of the quantizing functions $\mathcal{B}^{bin}$ and $\mathcal{B}^{tern}$ we used:

$$\mathcal{B}^{bin}(W_{i,j}) = \begin{cases} 0, \text{if } W_{i,j} = 0, \\ (\textit{sign}(W_{i,j}) + 1)/2, \text{if } W_{i,j} \neq 0, \end{cases} \tag{1}$$

$$\mathcal{B}^{tern}(W_{i,j}) = \textit{sign}(W_{i,j}), \tag{2}$$

with $W_{i,j}$ the (i, j)-coordinate of the weight matrix W. In order to train our networks through standard back-propagation based methods for supervised learning, we employed a surrogate gradient descent approach [35] with *arctan* as the surrogate function along with a STE for updating binary weights [8], in a way similar to [24].

4 Formal Model of Spiking Neurons

In this section, we introduce the formal model of a binary spiking neural network (BSNN) and of its integrate-fire (IF) spiking dynamics. Spiking neurons have the ability to process rich temporal dynamics in the data due to the state fullness of the neurons much like in recurrent neural networks (RNNs). We first introduce the static architecture of a BSNN.

Definition 4.1 [BSNN architecture] The architecture of a BSNN is a tuple $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \mathit{Scale}, (\tau_X)_{X\in\mathbf{L}}\big)$ where:

- $\mathbf{I}$ and $\mathbf{L}$ are two non-empty disjoint sets, respectively, the set of input neurons and the set of non-input neurons, with $\mathbf{N} = \mathbf{I} \cup \mathbf{L}$ (the set of neurons);
- $\mathcal{R} \subseteq \mathbf{L} \times \mathbf{N}$ is a connectivity relation relating each non-input neuron to its predecessors;
- $\mathcal{W} : \mathcal{R} \longrightarrow \mathit{Scale}$ is the weighting function for the connectivity relation, with Scale a finite scale of integers (i.e., $\mathit{Scale} = \{k_1, \ldots, k_n\} \subset \mathbb{Z}$ such that $k_1 < \ldots < k_n$);
- τ_X is the firing threshold for the non-input neuron $X \in \mathbf{L}$.

Given the architecture of a BSNN, we introduce the following notion of BSNN-compatible fire spiking dynamics.

Definition 4.2 [BSNN-compatible fire spiking dynamics] Let $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \mathit{Scale}, (\tau_X)_{X\in\mathbf{L}}\big)$ be the architecture of a BSNN and let $F = (\mathcal{F}_X)_{X\in\mathbf{N}}$ be a family of firing functions for S's neurons, with $\mathcal{F}_X : \mathbb{N} \longrightarrow \{0, 1\}$. We say that

F represents a possible spiking dynamics for the BSNN S up to time $\mathtt{t}_{end} \geq 0$, or simply F is S-compatible up to time $\mathtt{t}_{end}$, if and only if the following condition holds for every $X \in \mathbf{L}$ and for every $t \leq \mathtt{t}_{end}$:

$$\mathcal{F}_X(t) = \begin{cases} 0, \text{ if } t = 0, \\ \Theta\big(\mathcal{A}(X,t) - \tau_X\big), \text{ if } t > 0, \end{cases} \tag{3}$$

where

$$\mathcal{A}(X,t) = \begin{cases} 0, \text{ if } t = 0, \\ \mathcal{A}(X,t-1) \cdot \big(1 - \mathcal{F}_X(t-1)\big) \\ + \sum_{(X,X') \in \mathcal{R}} \mathcal{W}(X,X') \cdot \mathcal{F}_{X'}(t), \text{ if } t > 0, \end{cases}$$

and

$$\Theta(x) = \begin{cases} 1, \text{ if } x \geq 0, \\ 0, \text{ otherwise.} \end{cases}$$

Some explanations of the previous two definitions are in order. The weighting function $\mathcal{W}$ in Definition 4.1 specifies for each non-input neuron $X \in \mathbf{L}$ and each predecessor $X' \in \mathcal{R}(X)$ the weight of the connection from X' to X, with $\mathcal{R}(X) = \big\{X' \in \mathbf{N} : (X,X') \in \mathcal{R}\big\}$. In the general model, a weight can take any value from the set of numerical values $Scale$. In the rest of our paper we will only consider the BSNN variants of the model with $Scale = \{-1, 0, +1\}$ or $Scale = \{0, +1\}$. From a mathematical point of view, BSNNs are nothing but special cases of SNNs with either Boolean or three-valued weights.

Note that by means of the connectivity relation $\mathcal{R}$ we can specify the set of output neurons $\mathbf{O}$ as the non-input neurons that have no successors, that is,

$$\mathbf{O} = \big\{X \in \mathbf{L} : \forall X' \in \mathbf{L}, (X', X) \notin \mathcal{R}\big\}.$$

Definition 4.2 describes the possible spiking dynamics of a BSNN S. In particular, the firing function $\mathcal{F}_X$ represents a possible dynamics of the non-input neuron X in the BSNN architecture: it is the Heaviside step function of the difference between the neuron's activation value and the spiking threshold τ_X. The firing activity of the input neurons does not depend on the firing activity of other neurons, it is uniquely determined by the temporally sequential presentation of the input. This is the reason why the condition for $\mathcal{F}_X$ only applies to the case $X \in \mathbf{L}$.

The activation value of the non-input neuron X at time t depends recursively on its value at time $t-1$ and a weighted sum over the incoming stimulus at time t. Therefore, to respect the recursive nature of the activation function, we have to define that at time 0, the network is completely inactive, i.e., no node $X \in \mathbf{N}$ is firing at time $t = 0$. Moreover, the incoming stimulus gets perfectly integrated as in an Integrate-Fire (IF) model, without any leak in the neurons. But there is a hard reset term in our neuron model, which resets the activation value to zero every time it fires a spike.

Model type	Number of hidden neurons (k)	Digits	Spike encoding	Time-steps (t_{end})	Validation Accuracy (%)	Test Accuracy (%)
$\mathscr{S}_k^{bin}$	32	1,5,9	Poisson	16	92.98	94.29
	16		Poisson	16	94.68	94.62
	8		Poisson	8	95.20	95.27
	32		Thresholded	1	92.47	93.63
	16		Thresholded	1	92.09	91.66
	8		Thresholded	1	91.29	93.41
$\mathscr{S}_k^{tern}$	128	0,1,2,3,4,5,6,7,8,9	Poisson	4	92.00	92.16
	64		Poisson	4	91.82	92.03
	32		Poisson	4	90.55	91.06
	128		Thresholded	1	86.56	87.00
	64		Thresholded	1	84.97	86.10
	32		Thresholded	1	85.12	85.03

Table 1
Accuracies of different BSNN architectures trained on the MNIST classification task.

The BSNN architectures we trained for the MNIST classification task we informally described above are specific instances of Definition 4.1. Specifically, each network has 28×28 input neurons, one neuron per pixel in the image to be classified. That is, $\mathbf{I} = \{I_{x,y} : 1 \leq x, y \leq 28\}$.

Moreover, it has either 8, 16, 32, 64 or 128 hidden neurons in the hidden layer that are fully connected to the input neurons, that is, given $k \in \{8, 16, 32, 64, 128\}$: $\mathbf{H} = \{H_z : 1 \leq z \leq k\}$, and $\forall H \in \mathbf{H}, \forall I \in \mathbf{I}, (H, I) \in \mathcal{R}$.

Finally, it has 10 classification neurons in the output layer, one neuron for each digit to be recognized in the general MNIST classification task that are fully connected to the hidden neurons, that is, $\mathbf{C} = \{C_z : 1 \leq z \leq 10\}$, and $\forall H_z \in \mathbf{H}, \forall C_{z'} \in \mathbf{C}, (C_{z'}, H_z) \in \mathcal{R}$.

Thus, we considered a BSNN with a set of non-input neurons $\mathbf{L} = \mathbf{H} \cup \mathbf{C}$. Notice that in this BSNN architecture the set of classification neurons coincides with the set of output neurons, that is, $\mathbf{O} = \mathbf{C}$.

The class of BSNN architectures with binary weights are denoted by $\mathscr{S}_k^{bin}$ while those with three-valued weights are denoted by $\mathscr{S}_k^{tern}$, depending on the number k of their hidden units. We only trained and tested 12 variants of BSNN networks varying along the three dimensions: the specific spike encoding used (Poisson vs. threshold binarized), as detailed above, the weight quantization used ($\{0, 1\}$ vs. $\{-1, 0, 1\}$), and the number $k \in \{8, 16, 32, 64, 128\}$ of hidden units. For each variant, the value of $\mathcal{W}(X, X')$ for each $(X, X') \in \mathcal{R}$ was determined through learning. Specifically, we have three networks for each of the following four cases: i) binary weights, Poisson encoding and $k \in \{8, 16, 32\}$; ii) binary weights, threshold binarized encoding and $k \in \{8, 16, 32\}$; iii) three-valued weights, Poisson encoding and $k \in \{32, 64, 128\}$; iii) three-valued weights, threshold binarized encoding and $k \in \{32, 64, 128\}$.

5 Causal Model

A causal model is a mathematical object describing the causal dependencies between variables. It is a central concept of current analyses of causality in AI. A binary causal model (BCM) is nothing but a causal model in which variables are assumed to be Boolean. In a BCM causal information is expressed by means of Boolean expressions (*alias* propositional formulas), the set of Boolean expressions being generated inductively as follows: i) each Boolean variable p is a Boolean expression; ii) if ω is a

Boolean expression, so is $\neg\omega$ ("negation"); iii) if ω_1 and ω_2 are Boolean expressions, so is $\omega_1 \wedge \omega_2$ ("conjunction"). Additional Boolean constructs $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are definable as abbreviations in the usual way. In formal terms, a BCM is a triplet $\Gamma = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ where i) $\mathbf{U}$ is a set of exogenous variables, ii) $\mathbf{V}$ is a set of endogenous variables, iii) $\mathcal{E}$ is a function mapping each endogenous variable $p \in \mathbf{V}$ to a Boolean expression $\mathcal{E}(p)$ of the form $p \leftrightarrow \omega_p$, where ω_p is a Boolean expression built from $\mathbf{U} \cup \mathbf{V}$ that does not contain p. Specifically, the Boolean expression $p \leftrightarrow \omega_p$ stipulates that the endogenous variable p is true iff the condition ω_p is true. It can be seen as the compact representation of a Boolean function for the endogenous variable p. From a binary causal model $\Gamma = (\mathbf{U}, \mathbf{V}, \mathcal{E})$ it is straightforward to extract a causal graph representing the causal dependencies between the variables: the vertices of the causal graph are the variables in $\mathbf{U} \cup \mathbf{V}$, and we draw an edge from a variable q to an endogenous variable p if the Boolean expression ω_p such that $\mathcal{E}(p) = p \leftrightarrow \omega_p$ contains the variable q.

The model of the BSNN given in Definition 4.1 can be mapped onto a BCM that represents the causal dependencies between the BSNN's neural units over time. The idea of the mapping is simple: we assign a Boolean variable $p_{X,t}$ to each neuron X for each time t in $\{0, \ldots, \mathtt{t}_{end}\}$, where $\mathtt{t}_{end}$ is the final time step at which the network stops receiving incoming spike train from the image currently being presented. The variable $p_{X,t}$ is true (resp. false) if the neuron X fires (resp. does not fire) at time t. The exogenous variables are for the input neurons, while the endogenous ones are for the non-input neurons. The causal dependencies between the firing activities of the neurons are represented by the Boolean equations. Here, we only give the BCM for the variants of the BSNN with Boolean weights $\{0, 1\}$.

Definition 5.1 [BCM for BSNN with Boolean weights] Let $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \{0,1\}, (\tau_X)_{X \in \mathbf{L}}\big)$ be the architecture of a BSNN with Boolean weights in the sense of Definition 4.1. The BCM for S is the triplet $\Gamma_S = \big(\mathbf{U}_S, \mathbf{V}_S, \mathcal{E}_S\big)$ where $\mathbf{U}_S = \bigcup_{0 \leq t \leq \mathtt{t}_{end}} \mathbf{U}_S^t$, $\mathbf{V}_S = \bigcup_{0 \leq t \leq \mathtt{t}_{end}} \mathbf{V}_S^t$, $\mathbf{U}_S^t = \{p_{X,t} : X \in \mathbf{I}\}$, $\mathbf{V}_S^t = \{p_{X,t} : X \in \mathbf{L}\}$, and $\forall X \in \mathbf{L}$:

$$\mathcal{E}_S(p_{X,0}) := p_{X,0} \leftrightarrow \bot,$$

and for $t > 0$:

$$\mathcal{E}_S(p_{X,t}) := p_{X,t} \leftrightarrow \Bigg(\Big(\neg p_{X,t-1} \rightarrow \bigvee_{\substack{\Omega \subseteq \mathcal{R}^+(X): \\ \mathcal{A}(X,t-1)+|\Omega| \geq \tau_X}} \big(\bigwedge_{X' \in \Omega} p_{X',t}\big)\Big) \wedge \Big(p_{X,t-1} \rightarrow \bigvee_{\substack{\Omega \subseteq \mathcal{R}^+(X): \\ |\Omega| \geq \tau_X}} \big(\bigwedge_{X' \in \Omega} p_{X',t}\big)\Big)\Bigg),$$

with

$$\mathcal{R}^+(X) = \big\{X' \in \mathbf{N} : (X, X') \in \mathcal{R} \text{ and } \mathcal{W}(X, X') = 1\big\}.$$

We conclude this section by showing that the spiking dynamics of a BSNN are correctly represented by its BCM. Specifically, let $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \{0,1\}, (\tau_X)_{X \in \mathbf{L}}\big)$ be a BSNN with Boolean weights and $\mathcal{I}$ a Boolean interpretation for the variables in

$\mathbf{U}_S \cup \mathbf{V}_S$, i.e., $\mathcal{I} : \mathbf{U}_S \cup \mathbf{V}_S \longrightarrow \{0,1\}$, such that for every time $t \in \{0,\ldots,\mathtt{t}_{end}\}$ and for every neuron X, the function $\mathcal{F}_X$ assigns to time t the same value assigned by the interpretation $\mathcal{I}$ to the corresponding variable $p_{X,t}$. Then, the family of firing functions $F = (\mathcal{F}_X)_{X\in\mathbf{N}}$ is S-compatible up to time $\mathtt{t}_{end}$ if and only if $\mathcal{I}$ satisfies all Boolean equations of the BCM $\Gamma_S = \big(\mathbf{U}_S, \mathbf{V}_S, \mathcal{E}_S\big)$ for S. This correspondence between a BSNN and its BCM is formally expressed by the following Theorem 5.2 where, for any Boolean expression ω, $\mathcal{I} \models \omega$ denotes the fact that the Boolean interpretation $\mathcal{I}$ satisfies the Boolean expression ω. For the readers unfamiliar with Boolean (propositional) logic, we remind that $\mathcal{I} \models \omega$ iff $Val(\mathcal{I},\omega) = 1$, where $Val(\mathcal{I},\omega)$ is defined inductively, as follows: i) $Val(\mathcal{I},p) = \mathcal{I}(p)$ for $p \in (\mathbf{U}_S \cup \mathbf{V}_S)$; ii) $Val(\mathcal{I},\neg\omega) = 1 - Val(\mathcal{I},\omega)$; iii) $Val(\mathcal{I},\omega_1 \wedge \omega_2) = \min\big(Val(\mathcal{I},\omega_1), Val(\mathcal{I},\omega_2)\big)$.

Theorem 5.2 *Let $\mathcal{I}(p_{X,t}) = \mathcal{F}_X(t)$ for all $X \in \mathbf{N}$ and for all $t \leq \mathtt{t}_{end}$. Then, the following are equivalent:*

- *$(\mathcal{F}_X)_{X\in\mathbf{N}}$ is S-compatible up to time $\mathtt{t}_{end}$,*
- *$\mathcal{I} \models \bigwedge_{p_{X,t}\in\mathbf{V}_S} \mathcal{E}_S(p_{X,t})$.*

The proof of the theorem is given in the appendix A.1 at the end of the paper.

6 Explanation

In this section, we are going to show how to use binary causal models (BCMs) for formalizing and computing explanations in the context of the BSNN architectures we trained for the MNIST classification task. Following the literature on abductive explanation (AXp) [22,27], we define it to be a prime implicant that is actually true. Moreover, we define it in relation to a binary causal model. For simplicity, we assume an AXp (the *explanans*) is a term made of exogenous variables and the property to be explained (the *explanandum*) is a Boolean expression made of endogenous ones. This assumption is perfectly compatible with our application to the MNIST classification task in which we want to explain the network classification on the basis of the pixel-level features. Nonetheless, this assumption could be dropped without consequence; we would only need to assume that the explanans and the explanandum involve different variables.

Some preliminary notions are needed before defining AXp formally. We define a *term* to be a conjunction of literals in which a variable can occur at most once, a literal being a variable p or its negation $\neg p$. Terms are denoted by $\lambda, \lambda', \ldots$ Given two terms λ, λ', with a bit of abuse of notation, we write $\lambda' \subseteq \lambda$ (resp. $\lambda' \subset \lambda$) to mean that the set of literals appearing in λ' is a subset (resp. strict subset) of the set of literals appearing in λ. Given a BCM $\Gamma = \big(\mathbf{U}, \mathbf{V}, \mathcal{E}\big)$ and an arbitrary set of variables $\mathbf{X} \subseteq \mathbf{U} \cup \mathbf{V}$, $Term_{\mathbf{X}}$ denotes the set of terms built from $\mathbf{X}$.

Definition 6.1 [Abductive explanation] Let $\Gamma = \big(\mathbf{U}, \mathbf{V}, \mathcal{E}\big)$ be a BCM, $\mathcal{I}_{\mathbf{U}} : \mathbf{U} \longrightarrow \{0,1\}$ a Boolean interpretation for its exogenous variables, $\lambda \in Term_{\mathbf{U}}$ and ω_0 a Boolean expression built from $\mathbf{V}$. We say that λ is an abductive explanation (AXp) of

ω_0 with respect to Γ and $\mathcal{I}_{\mathbf{U}}$ if and only if:

$$
\begin{aligned}
&i)\ \mathcal{I}_{\mathbf{U}} \models \lambda,\\
&ii)\models \Big(\bigwedge_{p\in\mathbf{V}} \mathcal{E}(p)\wedge\lambda\Big)\rightarrow\omega_0,\\
&iii)\ \forall\lambda'\subset\lambda, \not\models \Big(\bigwedge_{p\in\mathbf{V}} \mathcal{E}(p)\wedge\lambda'\Big)\rightarrow\omega_0,
\end{aligned}
$$

where, for a given Boolean expression ω built from the set of variables $\mathbf{U}\cup\mathbf{V}$, $\models\omega$ means that ω is valid, i.e., $\mathcal{I}\models\omega$ for every Boolean interpretation $\mathcal{I}\in\{0,1\}^{\mathbf{U}\cup\mathbf{V}}$.

Let us illustrate how Definition 6.1 applies to a BSNN S with Boolean weights (i.e., $S\in\mathscr{S}_k^{bin}$) trained on the MNIST three-digit classification task. Given an input sequence $input : \{0,\ldots,\mathtt{t}_{end}\}\times\mathbf{I}\longrightarrow\{0,1\}$ and an observed output sequence $out : \{0,\ldots,\mathtt{t}_{end}\}\times\mathbf{C}\longrightarrow\{0,1\}$ for this input, we aim to abductively explain the output at a chosen time $t\in\{0,\ldots,\mathtt{t}_{end}\}$ using only variables for the input at time t. More precisely, we take the *explanandum* (i.e., ω_0) to be the Boolean expression

$$
\mathsf{out}_{S,t} =_{def} \bigwedge_{\substack{C_z\in\mathbf{C}:\\ out(t,C_z)=1}} p_{C_z,t}\wedge \bigwedge_{\substack{C_z\in\mathbf{C}:\\ out(t,C_z)=0}} \neg p_{C_z,t}.
$$

It represents the observed output of the network at time t. Then, we search for an abductive explanation $\lambda\in Term_{\mathbf{U}_S^t}$ of $\mathsf{out}_{S,t}$ with respect to the BCM Γ_S and to the Boolean interpretation $\mathcal{I}_{\mathbf{U}_S}$ encoding the input sequence *input* (i.e., $\mathcal{I}_{\mathbf{U}_S}(p_{I_{x,y},t}) = input(t,I_{x,y})$ for every $t\in\{0,\ldots,\mathtt{t}_{end}\}$ and $I_{x,y}\in\mathbf{I}$). The latter condition guarantees that the found explanation of the network's output at time t represents a portion of the actual input presented to the network at t.

The following proposition highlights an important property of a BSNN's abductive explanation: any input feature/neuron being mentioned in an abductive explanation of the output has necessarily a non-zero weight connection with the network's hidden layer. This guarantees that an abductive explanation does not contain completely irrelevant information. Later in the paper, we will contrast this result with the SHAP explanation method for which there is no guarantee that a found explanation does not contain completely irrelevant information.

Proposition 6.2 *Let $\lambda\in Term_{\mathbf{U}_S^t}$ be an abductive explanation of $\mathsf{out}_{S,t}$. Then,*

$$
\forall p_{I,t}\subseteq\lambda, \exists H\in\mathbf{H} \textit{ such that } I\in\mathcal{R}^+(H),
$$

where we recall $\mathcal{R}^+(H) = \big\{I\in\mathbf{I} : (H,I)\in\mathcal{R}$ and $\mathcal{W}(H,I)=1\big\}$.

The proof of the proposition is given in the Appendix A.2 at the end of the paper.

To compute an abductive explanation, we rely on a standard abductive explanation search algorithm, whose pseudo code is presented in Algorithm 1. The algorithm is initialized with a complete term λ_{init} over the set of exogenous variables (i.e., $\mathbf{U}_S$), which fully represents the actual input at the selected time t. Then, literals are systematically removed from λ_{init}, and at each iteration, we check whether condition (ii) in Definition 6.1 is still satisfied.

Algorithm 1 Computing Abductive Explanation

Require: Initial implicant λ_{init} and explanandum ω_0 that satisfy conditions (i) and (ii) in Def. 6.1
Ensure: Abductive explanation λ
  Set $\lambda = \lambda_{init}$
  for $l \in \lambda$ **do**
    if $\models \big(\bigwedge_{p\in\mathbf{V}} \mathcal{E}(p) \wedge \lambda\big) \rightarrow \omega_0$ **then**
      $\lambda \rightarrow \lambda \setminus l$
    end if
  end for
  return λ

At the end of the search algorithm we further verify the validity of condition (iii) in Definition 6.1 for a *prime implicant* check of the resulting abductive explanation λ. Algorithm 1 has a time complexity of $\mathcal{O}(|\mathbf{U}_{S_k^{bin}}|)$ which is the total number of exogenous variables in the model. This linear dependency guarantees the scalability of the algorithm with respect to the number of input neurons.

7 Experimental Results

In this section, we provide the experimental results on computing explanations for some of the BSNNs listed in Table 1. We implemented the AXp search Algorithm 1 using the open-source $Z3$ solver, which is an efficient and flexible theorem proving system implemented in Python developed by Microsoft Research. Since the time required to compute explanations using the SAT-based Algorithm 1 was very high (on the order of hours), we also considered a variant of this algorithm based on an SMT (Satisfiability Modulo Theories) representation of the binary causal model for the BSNNs and of the notion of abductive explanation. (See [6] for a general introduction to SMT.) In particular, SMT over Linear Integer Arithmetic (LIA) was sufficient for our purpose. We could again use Z3 given that it fully supports SMT. The exact representation can be found in the Appendix B at the end of the paper.

Number of hidden neurons (k)	Mean search time		Length of found explanation	
	SAT (hrs)	SMT (s)	(%) Total features	Mean
32	10.7	491	20.91	164
16	5.84	483	27.3	214
8	11.13	192	12.5	98

(a) Results for classes $\mathscr{S}_k^{bin}$

Number of hidden neurons (k)	Mean search time	Length of found explanation	
	SMT (hrs)	(%) Total features	Mean
128	0.27	56	437
64	0.78	55	432
32	1.0	36	280

(b) Results for classes $\mathscr{S}_k^{tern}$

Table 2
Computational analysis for searching explanation of BSNNs in the classes $\mathscr{S}_k^{bin}$ and $\mathscr{S}_k^{tern}$.

Table 2a provides a comprehensive overview of the run-times for the SAT-based

and SMT-based versions of the explanation search algorithm, along with the length of the AXp found for each BSNN in the class $\mathscr{S}_k^{bin}$ listed in Table 1, with $k \in \{8, 16, 32\}$. As the table clearly shows, the SMT-based approach is is significantly faster than the SAT-based approach. This is because, unlike the propositional logic representation of the binary causal model, the SMT representation avoids the need for universal quantification over sets of variables. We also computed explanations for BSNNs in the class $\mathscr{S}_k^{tern}$, using the SMT representation of their binary causal models and the corresponding notion of abductive explanation. Table 2b reports the run-times for computing these explanations with the SMT-based approach, along with their average length.

Figure 1 visualizes the abductive explanation found for the outputs of a network in the class $\mathscr{S}_{16}^{bin}$ at times 0 and 6.

Note that the set of input neurons/features in the explanation is a subset of the set of input neurons/features connected to the network's hidden layer. This is in line with Proposition 6.2.

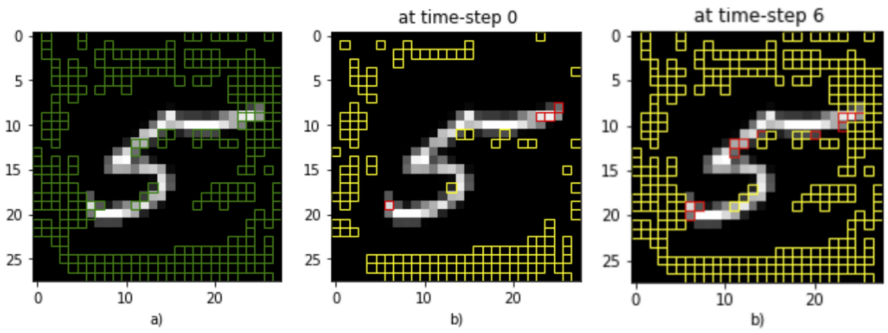


Fig. 1. Image of digit 5 (a) showing in green the input neurons/features being connected with the network's hidden layer; (b) the found AXps at times 0 and 6 showing in red the active input neurons/features (i.e., the positive literals) and in yellow the non-active input neurons/features (i.e., the negative literals) mentioned in the explanation.

8 Comparison with SHAP

In this section, we compare our logic-based explainability method with SHAP, a popular method widely used for interpreting predictions of machine learning models [30]. For our experiments, we used the pre-existing implementation of SHAP library in Python available at `https://github.com/shap/shap`. SHAP assigns relevance scores to input features based on a sample of the input space without taking into consideration the internal dynamics of the model.

Unlike our method, SHAP does not look inside the neural network and does not model the network's internal causal structure. Despite its widespread use, it has recently been shown that SHAP can provide misleading information about the relative importance of features for classification [19,18,17,26]. As discussed in [21], another limitation of SHAP is that, unlike abductive explanation, it does not take minimality of an explanation into account. To compare SHAP with our method, we fixed a

Sample size	Mean computation time (s)	Features wrongly considered relevant (%)
1000000	173.6	36.95
100000	38.3	46.34
10000	4.7	57.45

Table 3
Percentage of features wrongly considered relevant by SHAP.

threshold δ for the SHAP score and then identified the set of relevant features as those features whose SHAP score is strictly higher than δ if positive and strictly lower than $-\delta$ if negative. We observed that SHAP considered relevant some input features having zero weight connections with the network's hidden layer. This aspect is visually represented in Figure 2. This is a consequence of the model-agnostic nature of "black

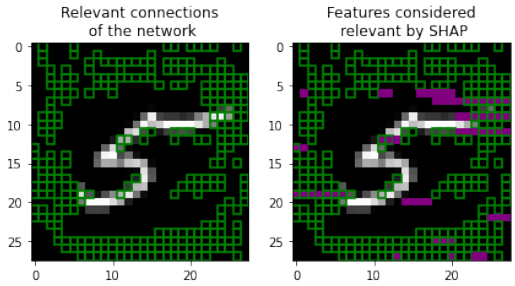


Fig. 2. Green features in the two figures are those having non-zero weight connections with the network's hidden layer. Features in purple on the right figure are considered relevant by SHAP.

box" explainability methods like SHAP.

Table 3 summarizes the results on the time required to compute the SHAP score for an input feature, as well as the percentage of features with zero-weight connections to the hidden layer that SHAP incorrectly identified as relevant, across different sample space sizes. On average, 47% of the input features deemed relevant by SHAP had zero-weight connections to the network's hidden layer. As the table shows, increasing the sample space size reduces the percentage of wrongly considered features, but at the cost of increased computation time. The performance of SHAP contrasts sharply with what is demonstrated in Proposition 6.2: our method guarantees that explanations never include input features with zero-weight connections to the hidden layer.

9 Conclusion

Let us take stock. We proposed a causal analysis of Binary Spiking Neural Networks (BSNNs) by mapping their spiking dynamics to binary causal models (BCMs). This mapping enabled the computation of abductive explanations for BSNN decisions in the context of the MNIST classification task, using both SAT-based and SMT-based approaches. Additionally, we compared our logic-based method to SHAP and demonstrated that, unlike SHAP, our approach reliably excludes causally irrelevant features from explanations. In the current work, we focused exclusively on the notion of ab-

ductive explanation (AXp). Future research will aim to extend our causal analysis of BSNNs to encompass more sophisticated concepts, including actual cause [16] and NESS (Necessary Element of a Sufficient Set) cause [7,14]. Our causal framework provides the necessary expressiveness to formally capture these concepts, and we believe that the logic-based approach we employed for computing abductive explanations can be extended to compute these notions as well. Another promising direction for future research is to develop a causal analysis of convolutional BSNNs (C-BSNNs) [43] following the approach presented here. We anticipate that incorporating convolutional layers could enhance accuracy on more complex datasets. Finally, we plan to extend our logic-based causal framework beyond simple visual classification tasks, applying it to explain BSNNs trained on language datasets [5].

Appendix

In this appendix, we present i) the proofs of the mathematical results presented in the paper (Section A), and ii) the SMT encoding of the causal model for a BSNN with Boolean weights and for a BSNN with ternary weights (Section B).

A Proofs

A.1 Proof of Theorem 5.2

Proof. ($\Rightarrow$) We first prove the left-to-right direction. Suppose i) $(\mathcal{F}_X)_{X\in\mathbf{N}}$ is S-compatible up to time $\mathtt{t}_{end}$ and ii) $\forall X \in \mathbf{N}, \forall t \leq \mathtt{t}_{end}, \mathcal{F}_X(t) = \mathcal{I}(p_{X,t})$. We are going to prove that $\mathcal{I} \models \mathcal{E}(p_{X,t})$ for every $t \in \{0,\ldots,\mathtt{t}_{end}\}$ and for every $X \in \mathbf{L}$. The case $t = 0$ is evident. In fact, $\mathcal{I}(p_{X,0}) = \mathcal{F}_X(0) = 0$ by i) and ii). Moreover, $\mathcal{I}(p_{X,0}) = 0$ iff $Val(\mathcal{I}, p_{X,0} \leftrightarrow \bot) = 1$, and $Val(\mathcal{I}, p_{X,0} \leftrightarrow \bot) = 1$ iff $\mathcal{I} \models p_{X,0} \leftrightarrow \bot$. Thus, $\mathcal{I} \models p_{X,0} \leftrightarrow \bot$ which is equivalent to $\mathcal{I} \models \mathcal{E}(p_{X,0})$. Let us prove the case $t > 0$ by reductio ad absurdum. Suppose, toward a contradiction, that $\mathcal{I} \not\models \mathcal{E}(p_{X,t})$. The latter is equivalent to $Val(\mathcal{I}, \mathcal{E}(p_{X,t})) = 0$ which is equivalent to iii) $Val(\mathcal{I}, p_{X,t}) = 0$ and $Val(\mathcal{I}, \chi) = 1$, or iv) $Val(\mathcal{I}, p_{X,t}) = 1$ and $Val(\mathcal{I}, \chi) = 0$, where χ abbreviates the following Boolean expression:

$$\chi =_{def} \Big(\neg p_{X,t-1} \rightarrow \bigvee_{\substack{\Omega\subseteq\mathcal{R}^+(X):\\ \mathcal{A}(X,t-1)+|\Omega|\geq\tau_X}} \big(\bigwedge_{X'\in\Omega} p_{X',t}\big)\Big)\wedge \Big(p_{X,t-1} \rightarrow \bigvee_{\substack{\Omega\subseteq\mathcal{R}^+(X):\\ |\Omega|\geq\tau_X}} \big(\bigwedge_{X'\in\Omega} p_{X',t}\big)\Big).$$

Suppose iii) holds. On the one hand, we have $Val(\mathcal{I}, p_{X,t}) = 0$ iff $\mathcal{I}(p_{X,t}) = 0$, and, by i) and ii), we have $\mathcal{I}(p_{X,t}) = 0$ iff $\mathcal{F}_X(t) = \Theta\big(\mathcal{A}(X,t) - \tau_X\big) = 0$. Hence, by iii), we have $\Theta\big(\mathcal{A}(X,t) - \tau_X\big) = 0$. On the other hand, by ii), it is routine mathematical exercise to verify that $Val(\mathcal{I}, \chi) = \Theta\big(\mathcal{A}(X,t) - \tau_X\big)$. Hence, by iii), we have that $\Theta\big(\mathcal{A}(X,t) - \tau_X\big) = 1$ which leads to a contradiction. In an analogous way we can prove that iv) leads to a contradiction.

($\Leftarrow$) We are going to prove the right-to-left direction. Suppose i) $\mathcal{I} \models \bigwedge_{p_{X,t}\in\mathbf{V}_S} \mathcal{E}_S(p_{X,t})$ and ii) $\forall X \in \mathbf{N}, \forall t \leq \mathtt{t}_{end}, \mathcal{F}_X(t) = \mathcal{I}(p_{X,t})$. We are going

to prove that $(\mathcal{F}_X)_{X\in\mathbf{N}}$ is S-compatible up to time $\mathtt{t}_{end}$, that is, $\mathcal{F}_X(0) = 0$ and $\mathcal{F}_X(t) = \Theta\big(\mathcal{A}(X,t) - \tau_X\big)$ for every $0 < t \leq \mathtt{t}_{end}$. The case $t = 0$ is evident. In fact, $\mathcal{I}(p_{X,0}) = 0$ iff $Val(\mathcal{I}, p_{X,0} \leftrightarrow \bot) = 1$, and $Val(\mathcal{I}, p_{X,0} \leftrightarrow \bot) = 1$ iff $\mathcal{I} \models p_{X,0} \leftrightarrow \bot$. Thus, $\mathcal{I}(p_{X,0}) = \mathcal{F}_X(0) = 0$ by i) and ii). Let us prove the case $0 < t \leq \mathtt{t}_{end}$ by reductio ad absurdum. Suppose, toward a contradiction, that $\mathcal{F}_X(t) \neq \Theta\big(\mathcal{A}(X,t) - \tau_X\big)$. By i), we have $\mathcal{I} \models \mathcal{E}_S(p_{X,t})$. The latter is equivalent to $Val(\mathcal{I}, \mathcal{E}_S(p_{X,t})) = 1$ which is equivalent to iii) $Val(\mathcal{I}, p_{X,t}) = 1$ and $Val(\mathcal{I}, \chi) = 1$, or iv) $Val(\mathcal{I}, p_{X,t}) = 0$ and $Val(\mathcal{I}, \chi) = 0$, where χ is the same abbreviation as in the proof of the $\Rightarrow$-direction. Suppose iii) holds. On the one hand, we have $Val(\mathcal{I}, p_{X,t}) = 1$ iff $\mathcal{I}(p_{X,t}) = 1$, and, by ii), we have $\mathcal{I}(p_{X,t}) = \mathcal{F}_X(t)$. Hence, by iii), we have $\mathcal{F}_X(t) = 1$. On the other hand, by ii), it is routine mathematical exercise to verify that $Val(\mathcal{I}, \chi) = \Theta\big(\mathcal{A}(X,t) - \tau_X\big)$. Hence, by iii), we have that $\Theta\big(\mathcal{A}(X,t) - \tau_X\big) = 1$ and, consequently, $\mathcal{F}_X(t) = 1$. This leads to a contradiction. In an analogous way we can prove that iv) leads to a contradiction. □

A.2 Proof of Proposition 6.2

Proof. Suppose i) the term $\lambda = p_{\mathfrak{I}_{x,y},t} \wedge \lambda'$ is an abductive explanation of $\mathsf{out}_{S_k^{bin},t}$ and, toward a contradiction, ii) $\not\exists \mathfrak{H}_z \in \mathbf{H}^k$ such that $\mathfrak{I}_{x,y} \in \mathcal{R}^+(\mathfrak{H}_z)$. By ii), we have that iii) for every $p_{X,t'} \in \mathbf{V}_{S_k^{bin}}$ the Boolean equation $\mathcal{E}(p_{X,t'})$ does not contain the variable $p_{\mathfrak{I}_{x,y},t}$. Moreover, by the definition of a term and since $p_{\mathfrak{I}_{x,y},t} \in \mathbf{U}_{S_k^{bin}}$, iv) $p_{\mathfrak{I}_{x,y},t}$ does not appear in λ' and $p_{\mathfrak{I}_{x,y},t}$ does not appear in $\mathsf{out}_{S_k^{bin},t}$. By iii) and iv), we have that v) $\models \big(\bigwedge_{p_{X,t'}\in\mathbf{V}_{S_k^{bin}}} \mathcal{E}(p_{X,t'}) \wedge p_{\mathfrak{I}_{x,y},t} \wedge \lambda'\big) \rightarrow \mathsf{out}_{S_k^{bin},t}$ iff $\models \big(\bigwedge_{p_{X,t'}\in\mathbf{V}_{S_k^{bin}}} \mathcal{E}(p_{X,t'}) \wedge \lambda'\big) \rightarrow \mathsf{out}_{S_k^{bin},t}$. Item i) implies that $\models \big(\bigwedge_{p_{X,t'}\in\mathbf{V}_{S_k^{bin}}} \mathcal{E}(p_{X,t'}) \wedge p_{\mathfrak{I}_{x,y},t} \wedge \lambda'\big) \rightarrow \mathsf{out}_{S_k^{bin},t}$ and $\not\models \big(\bigwedge_{p_{X,t'}\in\mathbf{V}_{S_k^{bin}}} \mathcal{E}(p_{X,t'}) \wedge \lambda'\big) \rightarrow \mathsf{out}_{S_k^{bin},t}$, which is in contradiction with v). □

B SMT Encodings

In this section we present the SMT representations (or encodings) of the binary causal model for a BSNN with Boolean weights and for a BSNN with ternary weights.

B.1 Boolean weights

Given the architecture of a SNN with Boolean weights $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \{0,1\}, (\tau_X)_{X\in\mathbf{L}}\big)$, the SMT representation of the binary causal model for S is a triplet $\Gamma_S = \big(\mathbf{U}_S, \mathbf{V}_S, \mathcal{E}_S^{smt}\big)$ where $\mathbf{U}_S$ and $\mathbf{V}_S$ are, respectively, the the sets of exogeneous and endogeneous variables in the sense of Definition 5.1, and $\mathcal{E}_S^{smt}$ is the function mapping each endogenous variable in $\mathbf{V}_S$ to a SMT expression such that, $\forall X \in \mathbf{L}$:

$$\mathcal{E}_S^{smt}(p_{X,0}) = p_{X,0} = 0,$$

and for $t > 0$:

$$\mathcal{E}_S^{smt}(p_{X,t}) = \Big(p_{X,t} = 1 \leftrightarrow \big((p_{X,t-1} = 0 \rightarrow \sum_{X' \in \mathcal{R}^+(X)} p_{X',t} + \mathcal{A}(X, t-1) \geq \tau_X) \wedge (p_{X,t-1} = 1 \rightarrow \sum_{X' \in \mathcal{R}^+(X)} p_{X',t} \geq \tau_X)\big)\Big).$$

In order to compute abductive explanations using the SMT encoding we use a translation of Definition 6.1 based on SAT into SMT. Specifically, let S be a SNN with Boolean weights, $\Gamma_S = \big(\mathbf{U}_S, \mathbf{V}_S, \mathcal{E}_S\big)$ its BCM, $\mathcal{I}_{\mathbf{U}} : \mathbf{U} \longrightarrow \{0,1\}$ a Boolean interpretation for the exogenous variables in $\mathbf{U}_S$, $\lambda \in \mathit{Term}_{\mathbf{U}_S}$ and ω_0 a Boolean expression built from $\mathbf{V}_S$. We can check whether λ is an abductive explanation (AXp) of ω_0 relative to Γ_S and $\mathcal{I}_{\mathbf{U}_S}$ by checking whether the following three SMT conditions are satisfied:

$$\begin{aligned}
&i)\ \mathcal{I}_{\mathbf{U}_S} \models tr(\lambda),\\
&ii)\ \models \big(\bigwedge_{p \in \mathbf{V}_S} \mathcal{E}_S^{smt}(p) \wedge tr(\lambda) \wedge \mathsf{hyp}_S\big) \rightarrow tr(\omega_0),\\
&iii)\ \forall \lambda' \subset \lambda, \not\models \big(\bigwedge_{p \in \mathbf{V}_S} \mathcal{E}_S^{smt}(p) \wedge tr(\lambda') \wedge \mathsf{hyp}_S\big) \rightarrow tr(\omega_0),
\end{aligned}$$

where

$$\begin{aligned}
tr(p_{X,t}) &= (p_{X,t} = 1),\\
tr(\neg\omega) &= \neg tr(\omega),\\
tr(\omega_1 \wedge \omega_2) &= tr(\omega_1) \wedge tr(\omega_2),
\end{aligned}$$

and

$$\mathsf{hyp}_S =_{def} \bigwedge_{p_{X,t} \in \mathbf{U}_S \cup \mathbf{V}_S} \big(p_{X,t} = 1 \vee p_{X,t} = 0\big).$$

B.2 Ternary weights

For a BSNN with ternary weights $S = \big(\mathbf{I}, \mathbf{L}, \mathcal{R}, \mathcal{W}, \{-1, 0, 1\}, (\tau_X)_{X \in \mathbf{L}}\big)$ we just need a different SMT representation of its binary causal model. In particular, in the case of ternary weights the function $\mathcal{E}_S^{smt}$ should map each endogenous variable in $\mathbf{V}_S$ to the following SMT expressions, $\forall X \in \mathbf{L}$:

$$\mathcal{E}_S^{smt}(p_{X,0}) = p_{X,0} = 0,$$

and for $t > 0$:

$$\mathcal{E}_S^{smt}(p_{X,t}) = \Big(p_{X,t} = 1 \leftrightarrow \big((p_{X,t-1} = 0 \rightarrow \sum_{X' \in \mathcal{R}^+(X)} p_{X',t} - \sum_{X'' \in \mathcal{R}^-(X)} p_{X'',t} + \mathcal{A}(X, t-1) \geq \tau_X) \wedge (p_{X,t-1} = 1 \rightarrow \sum_{X' \in \mathcal{R}^+(X)} p_{X',t} - \sum_{X'' \in \mathcal{R}^-(X)} p_{X'',t} \geq \tau_X)\big)\Big).$$

References

[1] Aleksandrowicz, G., H. Chockler, J. Y. Halpern and A. Ivrii, *The computational complexity of structure-based causality*, Journal of Artificial Intelligence Research **58** (2017), pp. 431–451.

[2] Audemard, G., S. Bellart, J. Lagniez and P. Marquis, *Computing abductive explanations for boosted regression trees*, in: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)* (2023), pp. 3432–3441.

[3] Audemard, G., F. Koriche and P. Marquis, *On tractable XAI queries based on compiled representations*, in: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)*, 2020, pp. 838–849.

[4] Ayoobi, H., N. Potyka and F. Toni, *Sparx: Sparse argumentative explanations for neural networks*, in: *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, Frontiers in Artificial Intelligence and Applications **372** (2023), pp. 149–156.

[5] Bal, M. and A. Sengupta, *Spikingbert: Distilling BERT to train spiking language models using implicit differentiation*, in: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)* (2024), pp. 10998–11006.

[6] Barrett, C., R. Sebastiani, S. Seshia and C. Tinelli, *Satisfiability modulo theories*, in: *Handbook of Satisfiability*, Frontiers in Artificial Intelligence and Applications **185**, IOS Press, 2009 pp. 825–885.

[7] Beckers, S., *The counterfactual NESS definition of causation*, in: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)* (2021), pp. 6210–6217.

[8] Bengio, Y., N. Léonard and A. Courville, *Estimating or propagating gradients through stochastic neurons for conditional computation*, CoRR **abs/1308.3432** (2013).

[9] Chockler, H. and J. Y. Halpern, *Responsibility and blame: A structural-model approach*, Journal of Artificial Intelligence Research **22** (2004), pp. 93–115.

[10] Cooper, M. C. and J. Marques-Silva, *Tractability of explaining classifier decisions*, Artificial Intelligence **316** (2023), p. 103841.

[11] Darwiche, A. and A. Hirth, *On the reasons behind decisions*, in: *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)* (2020), pp. 712–720.

[12] de Lima, T. and E. Lorini, *Model checking causality*, in: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI 2024)* (2024).

[13] Halpern, J. Y., *Axiomatizing causal reasoning*, Journal of Artificial Intelligence Research **12** (2000), pp. 317–337.

[14] Halpern, J. Y., *Defaults and normality in causal structures*, in: *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference (KR 2008)* (2008), pp. 198–208.

[15] Halpern, J. Y., "Actual causality," MIT Press, 2016.

[16] Halpern, J. Y. and J. Pearl, *Causes and explanations: a structural-model approach. Part I: Causes*, British Journal for Philosophy of Science **56** (2005), pp. 843–887.

[17] Huang, X. and J. Marques-Silva, *A refutation of Shapley values for explainability*, CoRR **abs/2309.03041** (2023).

[18] Huang, X. and J. Marques-Silva, *Refutation of Shapley values for XAI - additional evidence*, CoRR **abs/2310.00416** (2023).

[19] Huang, X. and J. Marques-Silva, *On the failings of Shapley values for explainability*, International Journal of Approximate Reasoning **171** (2024), p. 109112.

[20] Hubara, I., M. Courbariaux, D. Soudry, R. El-Yaniv and Y. Bengio, *Binarized neural networks*, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, editors, *Advances in Neural Information Processing Systems* (2016), pp. 4107–4115.

[21] Ignatiev, A., *Towards trustable explainable AI*, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (2020), pp. 5154–5158.

[22] Ignatiev, A., N. Narodytska and J. Marques-Silva, *Abduction-based explanations for machine learning models*, in: *Proceedings of the Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019, pp. 1511–1519.

[23] Izza, Y. and J. Marques-Silva, *On explaining random forests with SAT*, in: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI 2021)*, 2021, pp. 2584–2591.

[24] Jang, H., N. Skatchkovsky and O. Simeone, *Bisnn: Training spiking neural networks with binary weights via bayesian learning*, 2021 IEEE Data Science and Learning Workshop (DSLW) (2020), pp. 1–6.

[25] Kheradpisheh, S. R., M. Mirsadeghi and T. Masquelier, *Bs4nn: Binarized spiking neural networks with temporal coding and learning*, Neural Processing Letters **54** (2022).
[26] Letoffe, O., X. Huang, N. Asher and J. Marques-Silva, *From shap scores to feature importance scores*, CoRR (2024).
[27] Liu, X. and E. Lorini, *A unified logical framework for explanations in classifier systems*, Journal of Logic and Computation **33** (2023), pp. 485–515.
[28] Lorini, E., *A rule-based modal view of causal reasoning*, in: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI 2023)*, 2023, pp. 3286–3295.
[29] Lu, S. and A. Sengupta, *Exploring the connection between binary and spiking neural networks*, Frontiers in Neuroscience **14** (2020).
[30] Lundberg, S. M. and S. I. Lee, *A unified approach to interpreting model predictions*, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, editors, *Advances in Neural Information Processing Systems* (2017).
[31] Marques-Silva, J., T. Gerspacher, M. C. Cooper, A. Ignatiev and N. Narodytska, *Explaining naive bayes and other linear classifiers with polynomial time and delay*, in: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020.
[32] Marquis, P., *Extending abduction from propositional to first-order logic*, in: *Procedings of the International Workshop on Fundamentals of Artificial Intelligence Research (FAIR'91)*, LNCS (1991), pp. 141–155.
[33] Mertes, S., T. Huber, C. Karle, K. Weitz, R. Schlagowski, C. Conati and E. André, *Relevant irrelevance: Generating alterfactual explanations for image classifiers*, in: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence IJCAI 2024* (2024), pp. 467–475.
[34] Miller, T., *Contrastive explanation: a structural-model approach*, The Knowledge Engineering Review **36** (2021).
[35] Neftci, E. O., H. Mostafa and F. Zenke, *Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks*, IEEE Signal Processing Magazine **36** (2019), pp. 51–63.
[36] Pearl, J., "Causality: Models, Reasoning and Inference," Cambridge University Press, 2009.
[37] Potyka, N., *Interpreting neural networks as quantitative argumentation frameworks*, in: *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAA-21)* (2021), pp. 6463–6470.
[38] Prescott, S. A. and T. J. Sejnowski, *Spike-rate coding and spike-time coding are affected oppositely by different adaptation mechanisms*, Journal of Neuroscience **28** (2008), pp. 13649–13661.
[39] Qin, H., R. Gong, X. Liu, X. Bai, J. Song and N. Sebe, *Binary neural networks: A survey*, Pattern Recognition **105** (2020), p. 107281.
[40] Rastegari, M., V. Ordonez, J. Redmon and A. Farhadi, *Xnor-net: Imagenet classification using binary convolutional neural networks*, in: B. Leibe, J. Matas, N. Sebe and M. Welling, editors, *Computer Vision – ECCV 2016* (2016), pp. 525–542.
[41] Shi, W., A. Shih, A. Darwiche and A. Choi, *On tractable representations of binary neural networks*, in: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR 2020)*, 2020, pp. 882–892.
[42] Shih, A., A. Choi and A. Darwiche, *A symbolic approach to explaining bayesian network classifiers*, in: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, 2018, pp. 5103–5111.
[43] Srinivasan, G. and K. Roy, *Restocnet: Residual stochastic binary convolutional spiking neural network for memory-efficient neuromorphic computing*, Frontiers in Neuroscience **13** (2019).
[44] Tang, W., G. Hua and L. Wang, *How to train a compact binary neural network with high accuracy?*, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (2017), pp. 2625–2631.
[45] Verma, S., J. P. Dickerson and K. Hines, *Counterfactual explanations for machine learning: A review*, CoRR (2020).

Classifying Impact of Arguments in the Jiminy Advisor Framework

Yini Huang and Beishui Liao

Zhejiang University
Hangzhou, China

Abstract

The Jiminy Advisor framework is a multi-stakeholder ethical recommendation framework grounded in norms and argumentation. While it effectively provides conflict resolution outcomes, it leaves open the question of how individual arguments influence the final decision. In this paper, we extend Jiminy with a *relevant-argument removal semantics*. By refining the notion of relevance and then analyzing how the outcome changes when specific arguments are removed, we classify arguments into three categories: *critical*, which are indispensable for producing an obligation; *supportive*, which provide non-decisive co-support; and *null*, which have no positive effect on the result. Our approach is compatible with standard Dung semantics and with Jiminy's four-level pipeline, offering a local and modular explanation of stakeholder influence. We also establish key properties of the framework, including well-definedness, local consistency, critical persistence, and supportive demotion, and illustrate the account with a running example. This enhances both the explainability and robustness analysis of Jiminy's recommendations without altering its decision mechanism.

Keywords: Normative systems, Argumentation, Partial semantics, Explainable AI, Ethical reasoning

1 Introduction

Autonomous artificial systems are increasingly operating in ethically sensitive environments, where their decisions are influenced by a range of multiple stakeholders, including manufacturers, regulators, end-users, and bystanders [25,26]. Consequently, a central challenge in machine ethics is how to effectively integrate the divergent normative perspectives of these various groups into the autonomous system's behavior [24]. This challenge is particularly acute in cases involving moral dilemmas, necessitating an interdisciplinary approach and open dialogue among all stakeholders to achieve socially aligned moral behaviors [21].

The *Jiminy Advisor framework* [13] provides a new perspective for modeling and resolving this dilemma by combining normative systems and formal argumentation to produce moral recommendations that reflect multi-stakeholder input. The framework models each stakeholder's normative systems through a set of constitutive, regulative, and permissive norms, and uses argumentation

semantics to resolve conflicts and output a final obligation for the agent to follow.

While the Jiminy Advisor framework provides a novel way to integrate different moral inputs, it leaves open an important question: how should we evaluate the *influence* of individual arguments on the final obligation recommendation? In the current semantics, once the system has aggregated stakeholders' norms into an argumentation framework and computed an extension, the process delivers a set of obligations. However, the framework does not reveal whether the outcome critically depended on a specific argument, whether the absence of some arguments would have changed the conclusion, or whether certain arguments merely provided redundant support.

To provide a solution to this limitation, we propose a new *relevant-argument removal semantics* for the Jiminy Advisor framework. First, building on *partial semantics* [12], we refine the notion of relevance by closing not only under defeat paths but also under subarguments, thereby isolating the minimal subframework that can influence a target obligation o. This restriction ensures locality of computation while preserving all structurally necessary arguments. Second, within this restricted framework, we systematically remove arguments, together with their dependent superarguments, and recompute the outcome. The resulting analysis yields a principled classification of arguments into three categories: *critical*, *supportive* and *null*.

This paper is organized as follows. Section 2 reviews preliminaries on argumentation framework, the notion of relevant arguments, and the Jiminy Advisor framework. Section 3 introduces the proposed relevant-argument removal semantics in detail and defines the three impact categories. Section 4 establishes several basic properties of our approach, including well-definedness, local consistency, critical persistence, and supportive demotion, showing that the classification is both stable and robust under framework variations. We introduce some related work in Section 5. In Section 6 we conclude the paper and outline directions for future research.

2 Preliminaries

2.1 Argumentation Framework and Relevant Arguments

We recall the basics of the argumentation framework in the sense of Dung [7]. Note that although Dung's framework is abstract, the Jiminy Advisor framework we introduce later uses a structured argumentation formalism (a simplified version of ASPIC+). Thus, each argument has an internal structure and a well-defined set of subarguments, while defeats and extension semantics are still handled at the abstract level.

To align with the definitions in [13], we define that an *argumentation framework* (AF) is a pair $AF = (Arg, Def)$, where Arg is a finite set of arguments, and $Def \subseteq Arg \times Arg$ is a binary relation representing defeats between arguments.

Definition 2.1 *Given an AF, $A, B, C \in Arg$, let $\mathcal{E} \subseteq Arg$ be a set of argu-*

ments. We have that:

- $\mathcal{E}$ *is conflict-free iff* $\nexists A, B \in \mathcal{E}$*, s.t.* $(A, B) \in Def$*.*
- *An argument* $A \in Arg$ *is defended by* $\mathcal{E}$ *iff* $\forall (B, A) \in Def$*,* $\exists C \in \mathcal{E}$*, s.t.* $(C, B) \in Def$*.*
- $\mathcal{E}$ *is admissible iff* $\mathcal{E}$ *is conflict-free and each argument in* $\mathcal{E}$ *is defended by* $\mathcal{E}$*.*
- $\mathcal{E}$ *is a complete extension iff* $\mathcal{E}$ *is admissible and each argument in* Arg *that is defended by* $\mathcal{E}$ *is in* $\mathcal{E}$*.*
- $\mathcal{E}$ *is a preferred extension iff* $\mathcal{E}$ *is a maximal complete extension.*
- $\mathcal{E}$ *is a grounded extension iff* $\mathcal{E}$ *is the minimal complete extension.*
- $\mathcal{E}$ *is a stable extension iff* $\mathcal{E}$ *is conflict-free, and for each argument* $A \in Arg \backslash \mathcal{E}$*, there exists an argument* $B \in \mathcal{E}$*, such that* $(B, A) \in Def$*.*

We use $\sigma \in \{co, pr, gr, st\}$ to indicate the complete, preferred, grounded, and stable semantics. For each of these semantics, $\sigma(AF)$ denotes the set of σ-extensions of AF.

Partial semantics [12] identifies only those arguments that may influence a target set $\mathcal{T} \subseteq Arg$. Given $AF = (Arg, Def)$ and $\mathcal{T} \subseteq Arg$, the set of *relevant arguments* of $\mathcal{T}$ is defined as:

$$\mathrm{rlvt}_{AF}(\mathcal{T}) = \mathcal{T} \cup \bigcup_{\alpha \in \mathcal{T}} \{\beta \in Arg \setminus \mathcal{T} : \text{ there is a path from } \beta \text{ to } \alpha \text{ w.r.t. } Def\}.$$

Here, we use $\beta \rightsquigarrow \alpha$ to denote that there is a path from β to α with respect to Def. $\beta \rightsquigarrow \alpha$ iff there is a finite sequence of arguments

$$\gamma_0, \gamma_1, \ldots, \gamma_n \quad (n \geq 1),$$

such that $\gamma_0 = \beta$, $\gamma_n = \alpha$, $(\gamma_i, \gamma_{i+1}) \in Def$ for every $0 \leq i < n$.

In this paper, we adopt a modified version of this notion by additionally closing under subarguments. Following [10], for $\alpha, \beta, \gamma \in Arg$ we write $\beta \sqsubseteq \gamma$ to denote that β is a subargument of γ and correspondingly γ is a superargument of β. We write $\beta \sqsubset \gamma$ if $\beta \neq \gamma$ and $\beta \sqsubseteq \gamma$. Accordingly, we extend an argumentation framework to a triple $AF^{\mathrm{Sub}} = (Arg, Def, \mathrm{Sub})$, where Sub is a binary relation representing subargument over Arg. For each $\gamma \in Arg$, we then define the set of subarguments of γ and the set of superarguments of γ as:

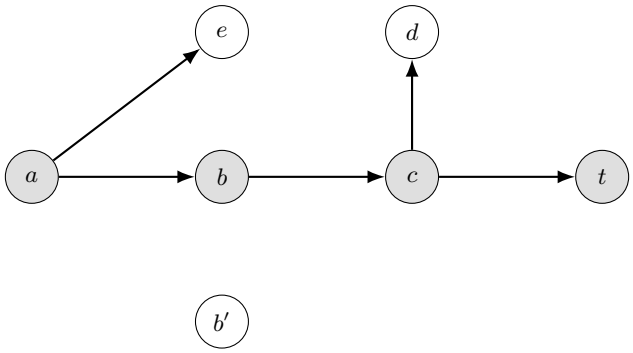
$$\mathrm{Sub}(\gamma) = \{\beta \in Arg \mid \beta \sqsubseteq \gamma\}, \ \mathrm{Sup}(\gamma) = \{\alpha \in Arg \mid \gamma \sqsubseteq \alpha\}.$$

Thus, the set of relevant arguments is:

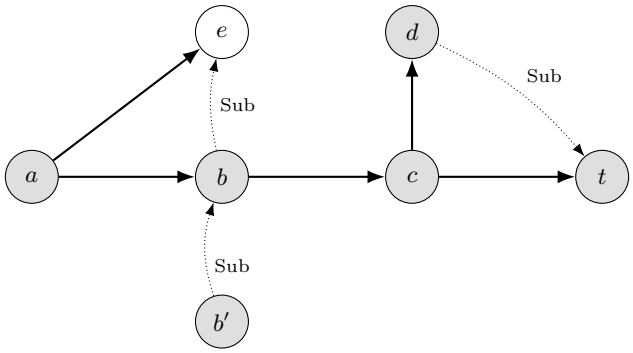
$$\mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}(B) = \mathrm{rlvt}_{AF}(B) \ \cup \bigcup_{\gamma \in \mathrm{rlvt}_{AF}(B)} \mathrm{Sub}(\gamma).$$

$\mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}(B)$ is the smallest subset of Arg that is closed both under attack chains and under subarguments.

To illustrate the difference between the standard notion of relevant arguments and our extended notion with subargument closure, consider $AF = \{a, b, b', c, d, e, t\}$ and the defeat relation is $Def^0 = \{(a, b), (b, c), (c, t), (c, d)\}$ and the subargument relation is Sub $= \{(b', b), (d, t), (b, e)\}$. According to [12], if $(\alpha, \beta) \in Def$ and $(\beta, \gamma) \in$ Sub, then we have $Def = \{(a, b), (b, c), (c, t), (c, d), (a, e)\}$. In the standard AF, the relevant set of $\{t\}$: $\{t, a, b, c\}$ is given by attack-path reachability. As shown in Figure 1, in the extended framework AF^{Sub}, if b' is a subargument of argument b (i.e., $b' \sqsubseteq b$), then b' also becomes relevant via subargument closure, even though b' has no direct defeat path to t.



(a) Standard AF (reachability): $\text{rlvt}_{AF}(\{t\}) = \{t, c, b, a\}$.



(b) Extended AF with *subargument closure only*: $\text{rlvt}^*_{AF^{\text{Sub}}}(\{t\}) = \{t, c, b, a, b', d\}$; e remains non-relevant.

Fig. 1. Relevant sets under two frameworks.

2.2 The Jiminy Advisor framework

The Jiminy Advisor [13] is a multi-stakeholder ethical advisor designed to integrate normative input from diverse stakeholders into a unified recommendation.

Definition 2.2 (Argumentation theory for a stakeholder) *Let $\mathcal{S}$ be the*

set of all stakeholders. An argumentation theory of each stakeholder $s \in \mathcal{S}$ is modeled by $\mathcal{N}_s = (\mathcal{L}, \bar{}, \mathcal{R}_s, \mathcal{K})$, where:

- *$\mathcal{L}$ is a logical language possibly containing negation.*
- *$\bar{}: \mathcal{L} \mapsto 2^{\mathcal{L}}$ is a contrariness function generalizing negation. We say $\phi = -\psi$ if and only if $\psi \in \overline{\phi}$ and $\phi \in \overline{\psi}$.*
- *$\mathcal{R}_s$ is a set of norms of the form $\varphi_1, \ldots, \varphi_n \Rightarrow_s^\tau \phi$ where $\varphi_i, \phi \in \mathcal{L}$, with $\tau \in \{r, c, p\}$. $\mathcal{R}_s^r$,$\mathcal{R}_s^c$ and $\mathcal{R}_s^p$ contain the norms in $\mathcal{R}_s$ with their corresponding superscripts, and they are called* regulative norms*,* constitutive norms*, and* permissive norms *respectively;*
- *$\mathcal{K} \subseteq \mathcal{L}$ is a set of observations called the* context*.*

$\mathcal{N} = (\mathcal{L}, \bar{}, \mathcal{R}, \mathcal{K})$ is also an argumentation theory whenever $\mathcal{R} = \bigcup_{s \in \mathcal{S}} \mathcal{R}_s$. Arguments can be of four types. We write $\mathrm{Conc}(A)$ for the conclusion of an argument A.

Definition 2.3 (Arguments) *Let $\mathcal{N} = (\mathcal{L}, \bar{}, \mathcal{R}, \mathcal{K})$ be an argumentation theory. An argument A for a conclusion $\mathrm{Conc}(A) = \phi$ is:*

- ***a brute fact argument**: $\{\phi\}$ if $\phi \in \mathcal{K}$;*
- ***an institutional fact argument**: $A_1, \ldots, A_n \Rightarrow^c \phi$ if $A_1, \ldots, A_n$ are brute or institutional fact arguments and $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow^c \phi$ is a norm in $\mathcal{R}^c$;*
- ***an obligation argument**: $A_1, \ldots, A_n \Rightarrow^r \phi$ if $A_1, \ldots, A_n$ are brute or institutional fact arguments such that there exists a norm $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow^r \phi$ that is in $\mathcal{R}^r$;*
- ***a permission argument**: $A_1, \ldots, A_n \Rightarrow^p \phi$ if $A_1, \ldots, A_n$ are brute or institutional fact arguments such that there exists a norm $\mathrm{Conc}(A_1), \ldots, \mathrm{Conc}(A_n) \Rightarrow^p \phi$ that is in $\mathcal{R}^p$.*

For each argument A, the set of its subarguments is defined recursively as

$$\mathrm{Sub}(A) = \begin{cases} \{A\}, & \text{if } A \text{ is a brute fact argument } \{\phi\}; \\ \{A\} \cup \bigcup_{i=1}^{n} \mathrm{Sub}(A_i), & \text{if } A = A_1, \ldots, A_n \Rightarrow^\tau \phi,\ \tau \in \{c, r, p\}. \end{cases}$$

For a set $S \subseteq Arg$, we write $\mathrm{Sub}(S) = \bigcup_{A \in S} \mathrm{Sub}(A)$.

Conflicts arise because different arguments may have contrary conclusions. The Jiminy Advisor framework resolves these conflicts using a built-in priority relation.

Definition 2.4 (Priority relation between arguments) *Let $\mathcal{A}$ be a set of arguments. Let $\mathcal{A}^b, \mathcal{A}^c, \mathcal{A}^r, \mathcal{A}^p \subseteq \mathcal{A}$ be the sets of brute fact arguments, institutional fact arguments, obligation arguments and permission arguments respectively. Given two arguments $A, B \in \mathcal{A}$, we use $A \succeq B$ to denote that A is non-strictly preferred to B and $A \succ B = (A \succeq$ and $B \not\succeq A)$ to denote that A is preferred to B. We have:*

- *For $A, B \in \mathcal{A}^\tau$, $A \succeq B$ for any $\tau \in \{b, c, r\}$.*

- *For* $A \in \mathcal{A}^b$ *and* $B \in \mathcal{A}^c \cup \mathcal{A}^r, A \succ B$.
- *For* $A \in \mathcal{A}^c$ *and* $B \in \mathcal{A}^r, A \succ B$.
- *For* $A \in \mathcal{A}^p$ *and* $B \in \mathcal{A}^r, A \succ B$.

We then define the attack and defeat relations of arguments.

Definition 2.5 (Attacks and defeats) *Let* $\mathcal{A}$ *be a set of arguments. A attacks B iff* $\mathrm{Conc}(A) \in \overline{\phi}$ *for some* $B' \in \mathrm{Sub}(B)$ *with* $\mathrm{Conc}(B') = \phi$. *Given a priority* $\succeq$, *the* defeat *relation* $Def = Def^{dir} \cup Def^{rev}$ *is defined by:*

- *Direct defeat* Def^{dir}*: A attacks B at* B' *and* $A \succeq B'$,
- *Reverse defeat* Def^{rev}*: B has a subargument* $B' \in \mathrm{Sub}(B)$ *that attacks A at A and* $B' \prec A$.

For a set of arguments $\mathcal{E}$, we have $Obl(\mathcal{E}) = \{\mathrm{Conc}(A) \mid A \in \mathcal{E} \cap \mathcal{A}^r\}$, where $\mathcal{A}^r$ is the set of obligation arguments. Thus, $Obl(\mathcal{E})$ collects the obligations supported in extension $\mathcal{E}$.

The Jiminy Advisor framework applies a four-level moral dilemma resolution scheme:

(i) *Individual framework*: conflicts internal to a single stakeholder's system. The individual framework is

$$AF(\mathcal{N}_s) = (Arg(\mathcal{N}_s), Def(\mathcal{N}_s), \mathrm{Sub})$$

where $s \in \mathcal{S}$.

(ii) *Combined framework*: conflicts after merging all stakeholders' arguments into one AF. The combined framework is

$$AF(\mathcal{S}) = (Arg(\mathcal{S}), Def(\mathcal{S}), \mathrm{Sub})$$

where $Arg(\mathcal{S}) = \bigcup_{s \in \mathcal{S}} Arg(\mathcal{N}_s)$ and $Def(\mathcal{S})$ is the defeat relation over this set.

(iii) *Integrated framework*: conflicts after merging stakeholders' normative systems into a single system. The integrated framework is

$$AF(\mathcal{N}_{\mathcal{S}}) = (Arg(\mathcal{N}_{\mathcal{S}}), Def(\mathcal{N}_{\mathcal{S}}), \mathrm{Sub})$$

where $\mathcal{N}_{\mathcal{S}} = (\mathcal{L}, \bar{}, \mathcal{R}_{\mathcal{S}}, \mathcal{K})$ with $\mathcal{R}_{\mathcal{S}} = \bigcup_{s \in \mathcal{S}} \mathcal{R}_s$.

(iv) *Reduced framework*: remaining conflicts are resolved by the Jiminy Advisor framework's own stakeholder-selection norms, which prioritize context-sensitive expertise. The reduced argumentation framework with respect to $\mathcal{E}$ is

$$AF^{\mathcal{E}} = (Arg, Def^{\mathcal{E}}, \mathrm{Sub})$$

where $Def^{\mathcal{E}}$ is the defeat relation induced by the revised priority relation $\succeq^{\mathcal{E}}$ defined in [13].

The Jiminy Advisor framework aims to provide the agent with a set of conflict-free obligations, thereby guaranteeing that the recommended course(s) of action are consistent and non-contradictory.

When a dilemma arises at level i (i.e., different extensions support conflicting obligations), we define the output obligation directly over the framework at that level:

$$O_\sigma(AF^{\mathrm{Sub}}) = \bigcap_{\mathcal{E}\in\sigma(AF^{\mathrm{Sub}})} Obl(\mathcal{E}),$$

where AF^{Sub} may be the argumentation framework obtained at any of the Jiminy Advisor framework's four levels. In this way, the Jiminy Advisor framework always produces a conflict-free set of obligations that are supported across all σ-extensions of the selected framework. This set can contain one or several obligations, each of which is a consistent and actionable recommendation for the agent.

Example 2.6 (Smart-speaker scenario, integrated level) *Consider the example of a smart-speaker mentioned in [13], which will serve as a running example: Law (L), Household (H), and Manufacturer (M). The context is* $\mathcal{K} = \{W_1, W_2, W_3, W_4\}$ *and the institutional fact is* i_1*:*

- W_1*: the device is made by* M*,*
- W_2*: the device collects data,*
- W_3*: the device detects a potential threat,*
- W_4*: the manufacturer is legally registered in Norway,*
- i_1*:* M *is a business in Norway.*

The conflicting moral options d_n *and morally relevant decisions* a_n *are defined*[1] *:*

- d_1*:* M *is law compliant;*
- d_2*: protect the privacy of users;*
- d_3*: report threat;*
- a_1*: comply with the GDPR;*
- a_2*: collect data without permission.*

Stakeholder norms include:

$$\begin{aligned} R_L &= \{\, W_1 \Rightarrow^r_L d_1, \quad i_1 \Rightarrow^p_L -d_2 \,\}, \\ R_H &= \{\, W_2 \Rightarrow^r_H d_2, \quad W_3 \Rightarrow^r_H d_3 \,\}, \\ R_M &= \{\, W_4 \Rightarrow^c_M i_1, \quad W_3 \Rightarrow^r_M a_2 \,\}. \end{aligned}$$

[1] They are conflicting since in the smart-speaker context, these options cannot be jointly satisfied.

From these we obtain arguments:

$$A_1 : W_1 \Rightarrow^r_L d_1, \quad A_2 : W_2 \Rightarrow^r_H d_2, \quad A_3 : W_3 \Rightarrow^r_H d_3,$$
$$A_4 : W_4 \Rightarrow^c_M i_1, \quad A_5 : W_3 \Rightarrow^r_M a_2, \quad A_6 : A_4 \Rightarrow^p -d_2.$$

$$Arg = \{ A_1, A_2, A_3, A_4, A_5, A_6,\ W_1, W_2, W_3, W_4 \}.$$
$$Def = \underbrace{\{ A_1 \leftrightarrow A_2,\ A_2 \leftrightarrow A_5 \}}_{\textit{mutual conflicts}} \cup \underbrace{\{ A_5 \rightarrow A_1,\ A_2 \rightarrow A_3,\ A_6 \rightarrow A_5,\ A_6 \rightarrow A_2 \}}_{\textit{one-way defeats}}.$$

$$\text{Sub} : \begin{cases} W_1 \mapsto \{W_1, A_1\}, & A_1 \mapsto \{A_1\}, \\ W_2 \mapsto \{W_2, A_2\}, & A_2 \mapsto \{A_2\}, \\ W_3 \mapsto \{W_3, A_3, A_5\}, & A_3 \mapsto \{A_3\},\ A_5 \mapsto \{A_5\}, \\ W_4 \mapsto \{W_4, A_4, A_6\}, & A_4 \mapsto \{A_4\},\ A_6 \mapsto \{A_6\}. \end{cases}$$

Under this variant, the integrated framework $AF(\mathcal{N}_S)$ admits a unique *preferred extension*

$$\mathcal{E} = \{W_1, W_2, W_3, W_4, A_1, A_3, A_4, A_6\},$$

with corresponding obligations

$$Obl(\mathcal{E}) = \{\ d_1,\ d_3\ \}.$$

Thus the output of the integrated framework is uniquely determined as

$$O_\sigma(AF(\mathcal{N}_S)) = \{d_1, d_3\}.$$

This version of the smart-speaker example will serve as our *running example* in the rest of the paper.

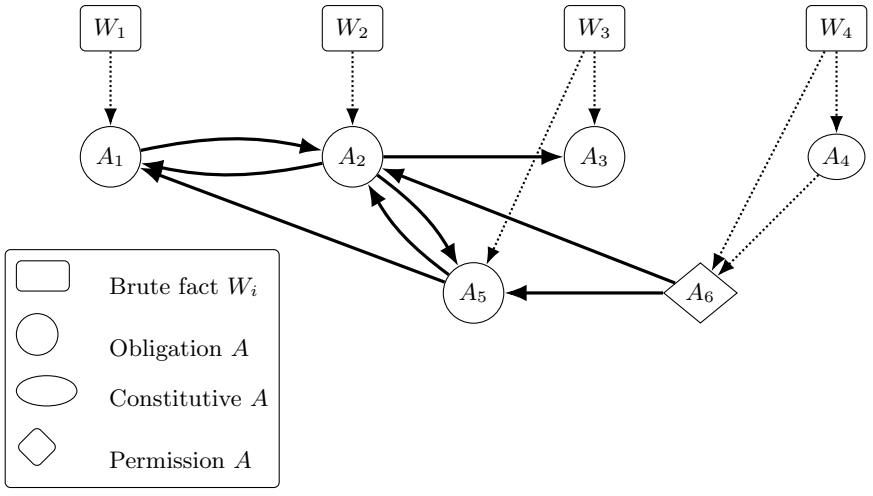


Fig. 2. The integrated argumentation framework.

3 Relevant-Argument Removal Semantics

While the Jiminy Advisor framework can provide a final recommendation, it does not reveal how sensitive this outcome is to the presence or absence of specific arguments.

In order to provide a more detailed analysis, we extend the Jiminy Advisor framework with a *relevant-argument removal semantics*. We restrict attention to arguments that can potentially affect a given obligation and then analyze how the outcome changes when each such argument is removed. This enables us to categorize arguments into three categories based on their impact on the obligation. Therefore, in our setting, we particularly focus on B_o such that B_o is an obligation argument and $\mathrm{Conc}(B_o) = o$ representing that o is a specific obligation recommendation produced by the Jiminy Advisor framework.

First, we extend the previous framework and provide a new framework:

Definition 3.1 (Restricted Framework) *Let $AF^{\mathrm{Sub}} = (Arg, Def, \mathrm{Sub})$ be an argumentation framework extended with the subargument relation, and let $o \in O_\sigma(AF^{\mathrm{Sub}})$ be an output obligation. Note that here $o = \mathrm{Conc}(B_o)$ and B_o is an obligation argument defined in Def 2.3. The restricted framework for o is the triple $AF_o^{\mathrm{Sub}} = (Arg_o, Def_o, \mathrm{Sub}_o)$, where*

$$Arg_o = \mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}(\{B_o\}), Def_o = Def \cap (Arg_o \times Arg_o), \quad \mathrm{Sub}_o = \mathrm{Sub} \cap (Arg_o \times Arg_o).$$

Note that $\mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}$ is defined to be closed under subarguments. In our impact analysis, when an argument is removed, we also remove all of its superarguments *that lie within* Arg_o. Because Sub_o is the restriction of Sub to $Arg_o \times Arg_o$, every removal is evaluated entirely within the restricted framework AF_o^{Sub}. Intuitively, AF_o^{Sub} collects arguments and defeats that can still influence the target obligation o; all subsequent removals and classifications are therefore carried out locally on this smaller framework instead of on the full Jiminy Advisor Framework.

Definition 3.2 (Removal of an argument) *For $AF_o^{\mathrm{Sub}} = (Arg_o, Def_o, \mathrm{Sub}_o)$ and $A \in Arg_o$, define $AF_o^{\mathrm{Sub}} \setminus \{A\} = (Arg'_o, Def'_o, \mathrm{Sub}'_o)$, where*

$$Arg'_o = Arg_o \setminus \{B \mid B \in \mathrm{Sup}(A)\}, \quad Def'_o = Def_o \cap (Arg'_o \times Arg'_o),$$

$$\mathrm{Sub}'_o = \mathrm{Sub}_o \cap (Arg'_o \times Arg'_o).$$

Definition 3.3 (Removal of a set of arguments) *For $AF_o^{\mathrm{Sub}} = (Arg_o, Def_o, \mathrm{Sub}_o)$ and $S \subseteq Arg_o$, define $AF_o^{\mathrm{Sub}} \setminus S = (Arg''_o, Def''_o, \mathrm{Sub}''_o)$, where*

$$Arg''_o = Arg_o \setminus \{B \mid \exists A \in S \text{ such that } B \in \mathrm{Sup}(A)\},$$

$$Def''_o = Def_o \cap (Arg''_o \times Arg''_o), \quad \mathrm{Sub}''_o = \mathrm{Sub}_o \cap (Arg''_o \times Arg''_o).$$

Here, Sub is reflexive (i.e., $A \in \mathrm{Sub}(A)$). Note that when an argument A is removed, all superarguments depending on A are also removed, and both the defeat relation and the subargument relation are restricted accordingly.

Definition 3.4 (Impact Categories) *Given AF_o^{Sub} and a semantics σ. For an obligation $o \in O_\sigma(AF^{\text{Sub}})$, we analyze the role of each argument $A \in Arg$ with respect to o:*

- *A is a **critical** argument: if $A \in Arg_o$ and $o \notin O_\sigma(AF_o^{\text{Sub}} \setminus \{A\})$.*
- *A is a **supportive** argument: if $A \in Arg_o$, $o \in O_\sigma(AF_o^{\text{Sub}} \setminus \{A\})$, and there exists $\mathcal{E} \in \sigma(AF_o^{\text{Sub}})$ and an obligation argument $\alpha \in \mathcal{E}$ with $\text{Conc}(\alpha) = o$ such that $A \in \mathcal{E}$.*
- *A is a **null** argument: if $A \notin Arg_o$; or if $A \in Arg_o$ but does not satisfy the conditions of being critical or supportive.*

Intuitively, *Critical* arguments are those without which o cannot be derived: once removed, o disappears from the output set. *Supportive* arguments contribute to sustaining o, but removing them does not eliminate it. *Null* arguments do not provide positive support for o since their presence does not improve o's derivability. We define $Category_{AF^{\text{Sub}}}(A, o) = X$ with $X \in \{Critical, Supportive, Null\}$ iff A satisfies the corresponding formal condition above.

Example 3.5 (Applying removal semantics to Example 2.6 (target $o = d_3$)) *Recall the integrated framework $AF(\mathcal{N}_S)$ and its unique output*

$$O_\sigma(AF(\mathcal{N}_S)) = \{d_1, d_3\}.$$

We now analyze the impact of each argument on the obligation $o := d_3$ ("report threat").

Relevant set. *Let $B_o := A_3$ be the obligation argument concluding d_3. We have:*

$$Arg_o \;=\; \{\, A_1, A_2, A_3, A_4, A_5, A_6 \,\} \;\cup\; \{\, W_1, W_2, W_3, W_4 \,\}.$$

The main defeats inside this set are

$$A_2 \leftrightarrow A_1, \quad A_5 \to A_1, \quad A_2 \to A_3, \quad A_2 \leftrightarrow A_5, \quad A_6 \to A_2, \quad A_6 \to A_5.$$

Restricted framework. *Let AF_o^{Sub} be the restriction of the integrated framework to Arg_o with the induced Def and* Sub *relations. We evaluate each $A \in Arg$ by removing it inside AF_o^{Sub} and checking whether o remains in O_σ (for $\sigma \in \{co, pr, gr, st\}$), and whether A co-occurs with an o-argument in some σ-extension.*

Classification for $o = d_3$:

- Critical:

  $$\{\, W_3,\ A_3,\ W_4,\ A_4,\ A_6 \,\}.$$

  W_3 and A_3 are the direct factual/obligation chain for d_3. A_6 (together with its subarguments W_4, A_4) neutralizes the privacy argument A_2, ensuring that A_3 is not defeated in every extension. Removing any of these arguments eliminates d_3 from the outcome.

- Supportive:

$$\{W_1,\ A_1,\ W_2\}.$$

A_1 (law compliance) and its factual base W_1 do not determine d_3, but in the unique extension they co-occur with A_3. Removing them leaves d_3 intact in O_σ, so they are supportive *for $o = d_3$ in the sense of our definition. W_2 is a brute-fact argument about data collection that is contained in the unique extension together with the obligation argument for d_3, and removing W_2 does not eliminate d_3 from O_σ; hence W_2 is also* supportive*.*

- Null:

$$\{A_2,\ A_5\}.$$

A_2 (privacy) and A_5 (collect without permission) are opponents of d_3: they never co-occur with A_3 in a supporting extension, and removing them cannot harm d_3.

Result. *For $o = d_3$ we obtain*

$$\text{Critical} = \{W_3, A_3, W_4, A_4, A_6\}, \qquad \text{Supportive} = \{W_1, A_1, W_2\}, \qquad \text{Null} = \{A_2, A_5\}.$$

Thus, the acceptance of d_3 relies on its direct chain and the permission chain that disables privacy opposition. Arguments for law compliance and the data-collection fact W_2 provide non-decisive co-support, whereas the privacy and data-collection chains are null.

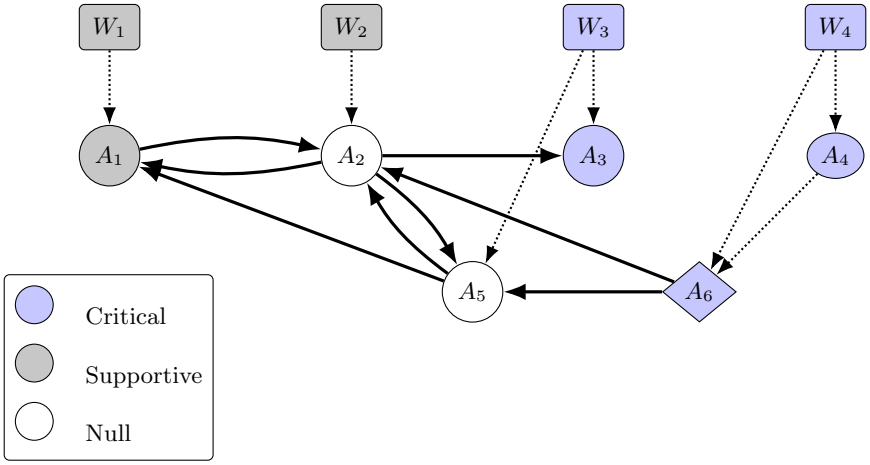


Fig. 3. The restricted framework AF_o^{Sub} for $o = d_3$, with arguments shaded by impact category (Critical, Supportive, Null).

4 Properties

Given an argumentation framework $AF^{\text{Sub}} = (Arg, Def, \text{Sub})$, an obligation argument $B_o \in Arg$, and let

$$o := \text{Conc}(B_o), \qquad Arg_o := \text{rlvt}^*_{AF^{\text{Sub}}}(\{B_o\}), \qquad AF_o^{\text{Sub}} := (Arg_o,\ Def_o,\ \text{Sub}_o),$$

where $Def_o = Def \cap (Arg_o \times Arg_o)$ and $\mathrm{Sub}_o = \mathrm{Sub} \cap (Arg_o \times Arg_o)$. For a Dung semantics $\sigma \in \{co, pr, gr, st\}$, write $O_\sigma(\cdot)$ for the outcome defined in the paper.

First, every argument receives exactly one label among *Critical*, *Supportive*, and *Null*.

Proposition 4.1 (Well-definedness) *For every $A \in Arg$ there exists a unique label*

$$Category_{AF^{\mathrm{Sub}}}(A, o) \in \{\mathrm{Critical}, \mathrm{Supportive}, \mathrm{Null}\}.$$

Proof. Given $A \in Arg$. Consider two situations.
Situation 1: $A \notin Arg_o$. By the definition of impact categories, A is labeled Null.

Situation 2: $A \in Arg_o$. Consider the outcome after removing A *inside* the restricted framework:

$$O_\sigma\big(AF_o^{\mathrm{Sub}} \setminus \{A\}\big).$$

If $o \notin O_\sigma(AF_o^{\mathrm{Sub}} \setminus \{A\})$, then A is Critical by definition.
If $o \in O_\sigma(AF_o^{\mathrm{Sub}} \setminus \{A\})$, distinguish:

$$\exists\, \mathcal{E} \in \sigma(AF_o^{\mathrm{Sub}})\ \exists\, \alpha \in \mathcal{E} \cap \mathcal{A}^r :\ \mathrm{Conc}(\alpha) = o\ \wedge\ A \in \mathcal{E}.$$

If this formula holds, A is Supportive; otherwise A is Null. Therefore, for every $A \in Arg$, A has a label.

We then prove that for every $A \in Arg$ there is only one label. Critical requires $o \notin O_\sigma(AF_o^{\mathrm{Sub}} \setminus \{A\})$, whereas both Supportive and Null require $o \in O_\sigma(AF_o^{\mathrm{Sub}} \setminus \{A\})$. Under this latter condition, the Supportive existence condition is the negation of the Null condition. Hence, exactly one label applies. □

We use the following lemma to show that arguments outside the relevant set Arg_o have no defeat or subargument edges into Arg_o, so the outside part cannot influence computations inside.

Lemma 4.2 (Boundary independence) *Let $Arg_o = \mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}(\{B_o\})$ and $U = Arg \setminus Arg_o$. Then for all $X \in U$ and $Y \in Arg_o$,*

$$(X, Y) \notin Def \quad \textit{and} \quad (X, Y) \notin \mathrm{Sub}.$$

Proof. We prove by contradiction.

If $(X, Y) \in Def$ and there is a Def-path from Y to B_o, then there is also a path from X to B_o, hence $X \in \mathrm{rlvt}_{AF}(\{B_o\}) \subseteq Arg_o$, a contradiction.

If $(X, Y) \in \mathrm{Sub}$ (i.e., $X \sqsubseteq Y$) and $Y \in Arg_o$, the closure under subarguments in rlvt^* yields $X \in Arg_o$, again a contradiction. □

With this lemma, we can then have the second proposition saying that category of any relevant argument is the same whether computed in the full framework or in the restricted framework AF_o^{Sub}.

Proposition 4.3 (Local Consistency) *Let* $o \in O_\sigma(AF^{\mathrm{Sub}})$ *and* $Arg_o = \mathrm{rlvt}^*_{AF^{\mathrm{Sub}}}(\{B_o\})$. *For every* $A \in Arg_o$,

$$Category_{AF^{\mathrm{Sub}}}(A, o) = Category_{AF_o^{\mathrm{Sub}}}(A, o).$$

Proof. Let $U := Arg \setminus Arg_o$. By the Boundary Independence Lemma, there are no Def or Sub edges from U into Arg_o. Hence, for any $S \subseteq Arg_o$,

$$Def \cap \big((Arg_o \setminus S) \times (Arg_o \setminus S)\big) = Def_o \cap \big((Arg_o \setminus S) \times (Arg_o \setminus S)\big),$$

and the same holds for Sub and Sub_o. Therefore, any reduction carried out *inside* Arg_o induces the same restricted framework whether we work in AF^{Sub} or in AF_o^{Sub}.

Critical case. For $A \in Arg_o$, let $S = \{A\}$. Then

$$o \in O_\sigma\big(AF^{\mathrm{Sub}} \setminus \{A\}\big) \quad \text{iff} \quad o \in O_\sigma\big(AF_o^{\mathrm{Sub}} \setminus \{A\}\big).$$

the Critical label is preserved.

Supportive/Null cases. Assume A is not Critical, so both frameworks agree that o survives the deletion of A.

The distinction between Supportive and Null relies on the following existence test, evaluated without removing any argument:

$$\exists\, \mathcal{E} \in \sigma(AF^{\mathrm{Sub}}) :\ A \in \mathcal{E} \ \wedge\ \exists\, \alpha \in \mathcal{E} \cap \mathcal{A}^r \text{ with } \mathrm{Conc}(\alpha) = o. \tag{1}$$

We show that this condition holds in AF^{Sub} iff it holds in the restricted framework AF_o^{Sub}.

($\Rightarrow$) If the condition holds in AF^{Sub}, then restricting any $\mathcal{E}$ to Arg_o generates $\mathcal{E}_o \in \sigma(AF_o^{\mathrm{Sub}})$.

($\Leftarrow$) Conversely, suppose $\mathcal{E}_o \in \sigma(AF_o^{\mathrm{Sub}})$ with $A \in \mathcal{E}_o$ and some obligation $\alpha \in \mathcal{E}_o$ such that $\mathrm{Conc}(\alpha) = o$. Because there are no incoming edges from U to Arg_o, the acceptability of $\mathcal{E}_o$ is unaffected by U. Remove from U all arguments attacked by $\mathcal{E}_o$, and let $\mathcal{E}_U$ be any σ-extension of the remaining subframework. Then

$$\mathcal{E} := \mathcal{E}_o \cup \mathcal{E}_U$$

is a σ-extension of AF^{Sub} with $\mathcal{E} \cap Arg_o = \mathcal{E}_o$. Hence the condition also holds in the full framework.

Thus the existence test is equivalent in AF^{Sub} and AF_o^{Sub}.

Therefore, for every $A \in Arg_o$, $Category_{AF^{\mathrm{Sub}}}(A, o) = Category_{AF_o^{\mathrm{Sub}}}(A, o)$.

□

Third, once an argument is critical in the relevant subframework, it stays critical in any larger framework that agrees with it on Arg_o and the induced relations.

Proposition 4.4 (Critical Persistence) *Let* $A \in Arg_o$ *be* Critical *in* AF_o^{Sub}. *For any* $AF'^{\mathrm{Sub}} = (Arg', Def', \mathrm{Sub}')$ *such that*

$$\mathrm{rlvt}^*_{AF'^{\mathrm{Sub}}}(\{B_o\}) = Arg_o, \quad Def' \restriction (Arg_o \times Arg_o) = Def_o, \quad \mathrm{Sub}' \restriction (Arg_o \times Arg_o) = \mathrm{Sub}_o,$$

we have $Category_{AF'^{\mathrm{Sub}}}(A, o) = \text{Critical}$.

Proof. Since A is Critical w.r.t. AF_o^{Sub},

$$o \notin O_\sigma\big(AF_o^{\mathrm{Sub}} \setminus \{A\}\big).$$

By the definitions of Def' and Sub' with Def_o and Sub_o on Arg_o, we then have $AF_o'^{\mathrm{Sub}} = AF_o^{\mathrm{Sub}}$, therefore

$$o \notin O_\sigma\big(AF_o'^{\mathrm{Sub}} \setminus \{A\}\big).$$

Since $\mathrm{rlvt}^*_{AF'^{\mathrm{Sub}}}(\{B_o\}) = Arg_o$, Local Consistency applies in AF'^{Sub} as well; hence

$$o \notin O_\sigma\big(AF'^{\mathrm{Sub}} \setminus \{A\}\big),$$

so A remains Critical in AF'^{Sub}. □

Lastly, if every path by which a supportive argument could affect o goes through nodes already labeled *Null*, then deleting those null nodes makes that argument irrelevant.

Proposition 4.5 (Supportive Demotion) *Let* A *be* Supportive *w.r.t.* AF_o^{Sub} *and define*

$$N = \{\, B \in Arg_o \mid Category_{AF^{\mathrm{Sub}}}(B, o) = \text{Null} \,\}.$$

Assume every defeat path from A *to* B_o *in* AF_o^{Sub} *contains at least one node in* N. *Let* $AF'^{\mathrm{Sub}} = AF^{\mathrm{Sub}} \setminus N$. *Then* $A \notin \mathrm{rlvt}^*_{AF'^{\mathrm{Sub}}}(\{B_o\})$ *and thus* $Category_{AF'^{\mathrm{Sub}}}(A, o) = \text{Null}$.

Proof. Let $\Pi(A \rightsquigarrow B_o)$ denote all defeat paths from A to B_o in AF_o^{Sub}, where edges are induced by Def_o. By assumption,

$$\forall \pi \in \Pi(A \rightsquigarrow B_o)\ \exists\, B_\pi \in N :\ B_\pi \in \mathrm{nodes}(\pi).$$

After deleting N, every such path is blocked; thus no defeat path from A to B_o remains in $AF_o'^{\mathrm{Sub}}$. Since node deletion cannot create new Def edges, we conclude

$$A \notin \mathrm{rlvt}^*_{AF'^{\mathrm{Sub}}}(\{B_o\}).$$

By the impact-category definition, arguments outside the relevant set are labeled Null. Hence $Category_{AF'^{\mathrm{Sub}}}(A, o) = \text{Null}$. □

5 Related Work

Foundational surveys on abstract and structured argumentation set the basis we build on [3,4]. They define arguments, defeats, and extensions, and show how structured arguments decompose into subarguments. Formal accounts of preferences explain how priority affects defeat [16,19]. These ideas ground our setting: our defeats and removals respect priorities, and our relevance closure explicitly includes subarguments because they can transmit support or attacks

that matter for the final outcome. Beyond yes/no extension outcomes, ranking-based semantics provide graded acceptability [1,2,9]. They complement our three labels: our critical/supportive/null view offers a qualitative core that future scoring schemes can refine.

In parallel, works on explanation semantics, abduction, and conditional acceptance study how to justify acceptance or rejection and how to repair missing support [5,6,14]. Our approach is complementary: we keep the original decision rule but add a local test inside a relevance-closed subframework. By removing one argument (and its superarguments) at a time, we attribute contribution and sensitivity without changing the underlying semantics.

A further line examines robustness and dynamics, that is, how changes to a framework affect outcomes and how to update efficiently [20,22]. Our relevance closure under defeat paths and subarguments provides a precise target for local updates and stability checks. This reduces recomputation and focuses analysis on the parts that can actually influence the chosen obligation, turning general principles on dynamics into a concrete per-argument procedure.

Finally, prior work links argumentation with counterfactual and causal reasoning: counterfactual tests inside abstract frameworks [23], formal accounts of actual causality [8], and responsibility analyses in law via defeasible argumentation [18]. Our removal-based classification aligns with this view by asking "what if this argument were absent?" and using the answer to explain current outcomes, while leaving open a path to connect with causal and responsibility notions in future work.

6 Conclusion and Future Work

We introduced a *relevant-argument removal semantics* for the Jiminy Advisor framework, combining partial semantics with systematic removal analysis to evaluate the impact of individual arguments. By classifying arguments into *critical*, *supportive*, or *null*, our approach goes beyond generating obligations and provides structured explanations of how stakeholder inputs shape the final outcomes. We illustrate the approach with a running example, which demonstrates its practical applicability. Overall, our contribution is to equip the Jiminy Advisor framework with a principled account of argument influence and stakeholder involvement, thereby enhancing both the explainability and robustness of the framework. More broadly, this lays the foundation for future extensions such as graded impact measures and explicit links between argument provenance and stakeholder agency, opening the way toward responsibility-aware explainable AI systems.

Building on these results, we identify several directions for further development. First, the process of determining whether an argument is critical or supportive requires repeated acceptance checks under various removals, which may become computationally demanding even within the restricted framework AF_o^{Sub}. We plan to conduct a systematic analysis of this complexity and to explore algorithmic techniques that can improve practical performance.

Second, we plan to refine the classification by distinguishing different types

of arguments, such as brute facts, institutional facts, obligations, and permissions, so that we can explain not only whether an argument is influential but also the normative route through which it shapes the final recommendation.

Third, we plan to integrate incremental [11] or ranking-based semantics [1] to move beyond discrete impact labels and develop graded or sensitivity-based measures of influence. We aim to incorporate finer notions of relevance, such as how many defeat steps separate an argument from the obligation in the underlying attack graph, so that degrees of contribution can be captured rather than only categorical roles.

Finally, we will extend the framework by linking arguments to the stakeholders who provided the norms, in line with work on AI governance and accountability that connects system behavior to stakeholder roles and input provenance [15,17], thereby enabling discussion of stakeholders' agency and tracing normative influence back to human or institutional sources.

Acknowledgements. The authors are thankful to the reviewers for their helpful comments and suggestions. The research reported in this paper was supported by the National Natural Science Foundation of China (No. 62576309).

References

[1] Amgoud, L. and J. Ben-Naim, *Ranking-based semantics for argumentation frameworks*, in: *International Conference on Scalable Uncertainty Management*, Springer, 2013, pp. 134–147.

[2] Amgoud, L., J. Ben-Naim, D. Doder and S. Vesic, *Ranking arguments with compensation-based semantics.*, KR **16** (2016), pp. 12–21.

[3] Bench-Capon, T. J. and P. E. Dunne, *Argumentation in artificial intelligence*, Artificial intelligence **171** (2007), pp. 619–641.

[4] Besnard, P., A. Garcia, A. Hunter, S. Modgil, H. Prakken, G. Simari and F. Toni, *Introduction to structured argumentation*, Argument & Computation **5** (2014), pp. 1–4.

[5] Booth, R., D. Gabbay, S. Kaci, T. Rienstra and L. van Der Torre, *Abduction and dialogical proof in argumentation and logic programming*, in: *ECAI 2014*, IOS Press, 2014 pp. 117–122.

[6] Booth, R., S. Kaci, T. Rienstra and L. Van Der Torre, *Conditional acceptance functions*, in: *Computational Models of Argument*, IOS Press, 2012 pp. 470–477.

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

[8] Halpern, J. Y., "Actual causality," MiT Press, 2016.

[9] Heyninck, J., B. Raddaoui and C. Straßer, *Ranking-based argumentation semantics applied to logical argumentation.*, in: *IJCAI*, 2023, pp. 3268–3276.

[10] Liao, B., *Layered argumentation frameworks with subargument relation and their dynamics*, Trends in Belief Revision and Argumentation Dynamics **48** (2013), pp. 247–267.

[11] Liao, B., *Toward incremental computation of argumentation semantics: A decomposition -based approach*, Annals of Mathematics and Artificial Intelligence **67** (2013), pp. 319–358.

[12] Liao, B. and H. Huang, *Partial semantics of argumentation: basic properties and empirical*, Journal of Logic and Computation **23** (2013), pp. 541–562.

[13] Liao, B., P. Pardo, M. Slavkovik and L. van der Torre, *The jiminy advisor: Moral agreements among stakeholders based on norms and argumentation*, Journal of Artificial Intelligence Research **77** (2023), pp. 737–792.

[14] Liao, B. and L. Van Der Torre, *Explanation semantics for abstract argumentation*, in: *Computational Models of Argument*, IOS Press, 2020 pp. 271–282.
[15] Miller, G. J., *Stakeholder roles in artificial intelligence projects*, Project Leadership and Society **3** (2022), p. 100068.
[16] Modgil, S. and H. Prakken, *A general account of argumentation with preferences*, Artificial Intelligence **195** (2013), pp. 361–397.
[17] Papagiannidis, E., P. Mikalef and K. Conboy, *Responsible artificial intelligence governance: A review and research framework*, The Journal of Strategic Information Systems **34** (2025), p. 101885.
[18] Pisano, G., H. Prakken, G. Sartor and R. Liepin, *Modelling cause-in-fact in legal cases through defeasible argumentation*, in: *Proceedings of the 20th International Conference on Artificial Intelligence and Law*, Association for Computing Machinery, 2025, pp. 268–277.
[19] Prakken, H., *An abstract framework for argumentation with structured arguments*, Argument & Computation **1** (2010), pp. 93–124.
[20] Rapberger, A. and M. Ulbricht, *On dynamics in structured argumentation formalisms*, Journal of Artificial Intelligence Research **77** (2023), pp. 563–643.
[21] Rhim, J., J.-H. Lee, M. Chen and A. Lim, *A deeper look at autonomous vehicle ethics: an integrative ethical decision-making framework to explain moral pluralism*, Frontiers in Robotics and AI **8** (2021), p. 632394.
[22] Rienstra, T., C. Sakama, L. van der Torre and B. Liao, *A principle-based robustness analysis of admissibility-based argumentation semantics*, Argument & Computation **11** (2020), pp. 305–339.
[23] Sakama, C., *Counterfactual reasoning in argumentation frameworks.*, in: *COMMA*, 2014, pp. 385–396.
[24] Shahriari, K. and M. Shahriari, *Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems*, in: *2017 IEEE Canada international humanitarian technology conference (IHTC)*, IEEE, 2017, pp. 197–201.
[25] Unesco, "Recommendation on the ethics of artificial intelligence," United Nations Educational, Scientific and Cultural Organization, 2022.
[26] Wallach, W. and C. Allen, "Moral machines: Teaching robots right from wrong," Oxford University Press, 2008.

DiSCo–RAD: Reasoning Alignment for Judicial Discretion

Liuwen Yu[1], Leendert van der Torre[1,2], Réka Markovich[1], Beishui Liao[2], Chenyang Cai[2]

[1] *University of Luxembourg, Luxembourg*
[2] *Zhejiang University, China*

Abstract

We analyse how judicial majorities can emerge despite internal disagreement through a model called DiSCo–RAD, developed on a real case from the Hungarian Constitutional Court. In this case, the judges agreed on the outcome but expressed distinct views through *parallel reasonings* and *dissenting opinions*, illustrating the discretionary nature of constitutional adjudication. In our model, the term *coalition* is used metaphorically to describe a convergent group of judges—a *majority configuration* that reflects agreement in outcome. DiSCo–RAD keeps two routes in view and checks whether they converge on the same majority constellation. On the *normative route*, judges are represented as a small social network built from pairwise assessments of cooperation. From these relations—friend, enemy, or mixed—we derive groups of compatible reasoning by a "friend-first with conflict tolerance" rule that allows limited disagreement when allies can compensate for it. On the *argumentation route*, possible collaborations are represented as nodes in a structured argumentation framework, with attacks expressing incompatible choices and supports expressing tolerated disagreement. When the reasoning patterns obtained from both routes coincide, the diagram explains how the majority holds together: differences in reasoning remain within the bounds of mutual support rather than full consensus.

Keywords: Artificial Intelligence, Knowledge representation and reasoning, Judicial discretion, Formal argumentation, Reasoning alignment, Conflict tolerance

1 Introduction

Constitutional courts are a special case of *discretionary judicial reasoning*. Their task is not merely to apply statutes, but to ensure that legislation and judicial decisions conform to the constitution and to the principles that sustain it—such as equality, proportionality, and the protection of fundamental rights. Because these principles may conflict, the reasoning of constitutional judges necessarily involves *discretion*: they interpret open-textured norms, weigh competing values, and justify why one interpretation better preserves constitutional coherence. Hence, *agreement on the outcome does not entail agreement on the reasoning*. Judges may attach separate opinions—*dissenting opinions* express-

ing disagreement with the decision, and *parallel reasonings* expressing agreement with the result but not with the majority's rationale [1]. These practices make constitutional courts a privileged field for studying how collective legal reasoning accommodates disagreement. We use the word "*coalition*" metaphorically to denote a convergent group of judges—a majority configuration based on compatible reasoning, not a political alliance.

Formal argumentation provides a natural lingua franca for such explanations. Rather than introduce yet another universal formalism, we adopt the "logic as a toolbox" view [23,1]: choose the mechanisms that fit the task, combine them transparently, and make their interaction auditable. In this paper, the toolbox contains three layers:

(1) a social layer that derives *friend*, *enemy*, and *frenemy* relations from pairwise collaboration scores between judges, reflecting how similar or divergent their reasoning is;

(2) an argumentation layer that encodes each judge's possible collaborations as *arguments*, where *attacks* represent incompatible choices (a judge cannot coherently adopt both), and *supports* represent tolerable disagreement (allies whose reasons partly cover a conflict);

(3) a coalition layer that extracts majorities (*Maj*) and checks individual rationality (*IR*, meaning no mutually harmful pairs).

We introduce **DiSCo–RAD**, a *Reasoning Alignment Diagram*, which keeps two reasoning routes side by side and checks that they match.

(1) Normative route. From collaboration values $\pi(\alpha, \beta)$ we classify each pair of judges as friends, enemies, or frenemies, then grow coalitions by a *friend-first + tolerance* closure. Tolerance is governed by a simple margin k: an ally can *cover* a local disagreement if $\pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k$. The result is filtered by *Maj* (majority threshold) and *IR* (no enemies inside). In legal terms, this route models how a majority can hold when minor disagreements are compensated by sufficient mutual understanding among judges.

(2) Argumentation route. We construct a *contrastive bipolar argumentation framework (BAF)* [2] whose arguments are directed collaborations (α, β). The *attack* relation is *local*: for a fixed judge α, (α, β) *attacks* (α, γ) whenever $\pi(\alpha, \gamma) < 0$, meaning that cooperation with γ would harm α's reasoning consistency. The *support* relation encodes the tolerance rule: (β, α) *supports* (α, γ) when α and β are friends, $\pi(\alpha, \gamma) < 0$ and $\pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k$. Acceptability is evaluated using the *β-semantics* for bipolar frameworks with *deductive support*, where accepted supporters propagate acceptance to what they support, and defeat requires neutralising all supporters. When support is ignored, this collapses to Dung's classic semantics [22]. In judicial terms, this route represents how compatible interpretations reinforce one another while still permitting disagreement within defensible limits.

[1] Section 66 of Act CLI of 2011 on the Constitutional Court of Hungary

We then *extract* coalitions from accepted collaborations by checking *mutual intent* (both (α, β) and (β, α) appear) and forming tolerance-aware cliques. A final *IR-pruning* step removes any residual enemies without losing the tolerance witnesses already present. The diagram itself provides the explanation: when both routes produce the same group of judges, the diagram *commutes*, meaning that the computational reasoning aligns with the normative account of discretion.

We validate DiSCo–RAD on a constitutional-court case. The numerical values behind π are illustrative—derived from judicial opinions rather than measured preferences—yet they reproduce the observed majority and explain why a partially disagreeing judge remains in it: supportive allies provide enough margin to cover the disagreement.

The remainder of this paper is structured as follows. Section 2 presents the case and inputs. Section 3 gives the methodology and the diagram. Section 4 formalises the social layer, the contrastive BAF with local attacks and deductive supports, and the coalition extractor. Section 5 works through the case. Section 6 discusses related work, and Section 7 concludes.

2 Discretionary Case Study

Constitutional courts are collegial bodies that review whether statutes—and, in some systems, certain judicial decisions—comply with the constitution. Cases typically reach them through a constitutional complaint, petition, or judicial referral; they are not ordinary appellate courts in the sense of re-examining facts or evidence. Once the case file is prepared, the bench (either a panel or the full court) deliberates *behind closed doors*: judges exchange drafts and reasons, discuss the constitutional issues, and finally vote on the operative part of the decision—such as to annul, uphold, or dismiss the challenged provision or complaint. These deliberations are confidential. What becomes public is the final decision (the operative part and the reasoning of the majority) together with any *separate opinions*.

Separate opinions come in two main forms. A *dissenting opinion* is written by a judge who disagrees with the decision's outcome or its reasoning and wishes to record the grounds of that disagreement. A *parallel reasoning* is written by a judge who agrees with the decision's outcome but disagrees on the reasoning. Both types of opinions serve transparency and doctrinal development by revealing the diversity of constitutional interpretation within the court.

Because internal exchanges are confidential, our model is *normative and explanatory*, not descriptive. We reconstruct a defensible pathway that *could* account for the published outcome and the pattern of separate opinions. Accordingly, the collaboration values in Table 1 and the induced win–lose signs in Table 2 are *illustrative encodings* derived from the published reasons; they are not measured preferences or survey data and should be read only as for explanation. The numerical entries of π and the sign profiles are therefore explicitly

for illustration and to make the explanation auditable.

Our approach sits within a line of work that treats judicial discretion as a normative reasoning problem with explicit freedom and obligations. In particular, Dik & Markovich formalise [20,19,21] a *duty of care* as the boundary of discretionary freedom in child custody cases: judges retain freedom to choose between the parents, but are constrained by obligations to determine relevant facts, to weigh them, and to reason consistently. We draw on that perspective in positioning the coalition-level postulates used later (IR, Tol(k), Maj) as boundaries on tolerable internal disagreement.

2.1 The Hungarian Constitutional Court case

To illustrate our approach, we consider Decision 3023/2016 (II. 23.) of the Hungarian Constitutional Court, which concerned maternity benefits and the interpretation of the Social Insurance Act (LXXXIII/1997). The petitioner, a mother who moved from one full-time to two part-time contracts before childbirth, was denied maternity benefit for one of the jobs. According to § 43(2) of the Act, each concurrent insurance relationship must independently satisfy a 365-day contribution period. The mother argued that this rule unfairly penalised her because she had paid social contributions for both jobs and should receive equal coverage.

The case therefore turned on how to weigh *formal equality under the law* against *substantive fairness of outcomes*. The constitutional judges were deeply divided. Fourteen judges (one absent) issued seven dissenting, four concurring, and only three purely concurring opinions. The majority upheld the statute, emphasising predictability and legal certainty, while a strong minority considered the rule discriminatory against women in non-standard employment. These divisions make the case a natural laboratory for studying *conflict-tolerant coalitions*: the court reached a majority decision even though complete agreement was impossible.

2.2 Agents and their simplified arguments

We model the judges as five representative *agents* ($Ag = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$) that capture the main interpretive positions observed in the case:

- α – a legal formalist who stresses constitutionality and uniform application of statutes.
- β – another formalist who focuses on textual fidelity and legal certainty, warning against vague moral criteria.
- γ – a moderate intermediary: accepts the law's constitutionality but finds the outcome unjust, proposing a more value-sensitive reading.
- δ – a reform-minded judge arguing that the rule produces discrimination and should be struck down.
- ε – a fairness-oriented judge supporting δ, claiming that formal categories ignore substantive equality.

Concrete arguments a_1–a_6. Let $Ar = \{a_1, \ldots, a_6\}$ with:

- a_1 (by α): "The rule (§43(2) Ebtv.) is constitutional and applies equally;

no discrimination."

- a_2 (by α): "Courts should avoid undefined moral terms (just/unjust) in constitutional review."
- a_3 (by β): "Stick to positive law and legal certainty; open moral notions undermine predictability."
- a_4 (by γ): "The rule is constitutional, but the outcome is unjust; adopt a value-sensitive interpretation (Art. 28)."
- a_5 (by δ): "The rule is discriminatory (Art. XV) against two-part-time contributors and should be struck down."
- a_6 (by ε): "Equal contributions but unequal benefit; formal classification ignores substantive equality."

Each triple (a_i, ω, ϵ) will later be used as a *reason* for agent ω to collaborate with ϵ when constructing the weighted base $\langle R, w \rangle$.

2.3 From arguments to collaboration values

Every ordered pair of judges (x, y) receives a *relative collaboration value* $\pi(x, y)$ measuring the net benefit that x expects from working with y. Positive values mean that x sees y as a helpful ally; negative values mean that x experiences y as an obstacle. The matrix in Table 1 summarises these relations.

$\pi(x,y)$	$y=\alpha$	$y=\beta$	$y=\gamma$	$y=\delta$	$y=\varepsilon$
$x=\alpha$	*	3	-2	-6	-8
$x=\beta$	2	*	-1	-3	-3
$x=\gamma$	2	1	*	-3	-3
$x=\delta$	-4	-3	-3	*	3
$x=\varepsilon$	-4	-3	-3	2	*

Table 1
Relative collaboration values $\pi(x, y)$ derived from the arguments.

In Table 1, rows indicate the agent that evaluates others; columns indicate the potential partner. For example, $\pi(\alpha, \beta) = 3$ shows that α perceives collaboration with β as strongly positive, while $\pi(\alpha, \gamma) = -2$ signals mild disagreement. Symmetry is not required: $\pi(\beta, \alpha) = 2$ is positive but not equal to $\pi(\alpha, \beta) = 3$, meaning that β values α slightly less. The first three rows correspond to the formalist camp: they cooperate internally (α and β) but have mild friction with γ. Rows for δ and ε show high mutual support ($\pi(\delta, \varepsilon) = 3$, $\pi(\varepsilon, \delta) = 2$) and strong opposition to the formalists (negative values across the first three columns).

Overall, two cohesive clusters appear: (α, β, γ) with mostly positive or small negative links, and (δ, ε) with mutual positive links but hostility to the first group. The contrastive structure of the network is already visible here.

2.4 Simplifying with win–lose signs

The sign transformation $tr(x)$ simplifies π into two values per pair: "+" if cooperation is favourable, "−" otherwise. We call the resulting tuples (v_1, v_2)

the *win–lose profile* between two agents. The outcome is shown in Table 2.

$win{-}lose(x,y)$	$y=\alpha$	$y=\beta$	$y=\gamma$	$y=\delta$	$y=\varepsilon$
$x=\alpha$	*	$(+,+)$	$(-,+)$	$(-,-)$	$(-,-)$
$x=\beta$	$(+,+)$	*	$(-,+)$	$(-,-)$	$(-,-)$
$x=\gamma$	$(+,-)$	$(+,-)$	*	$(-,-)$	$(-,-)$
$x=\delta$	$(-,-)$	$(-,-)$	$(-,-)$	*	$(+,+)$
$x=\varepsilon$	$(-,-)$	$(-,-)$	$(-,-)$	$(+,+)$	*

Table 2
Win–lose outcomes after applying the sign function $tr(x)$.

Each cell (x,y) in Table 2 records the pair of signs: the first symbol represents how x evaluates y, the second how y evaluates x. The pattern $(+,+)$ indicates a *friend relation*, $(-,-)$ an *enemy relation*, and mixed signs $(+,-)$ or $(-,+)$ a *frenemy relation*. In the table, α and β form a strong friendship ($(+,+)$ in both directions). They both have asymmetric frenemy relations with γ, and mutual enmity with δ and ε. Conversely, δ and ε form a second friendship pair, isolated from the rest by $(-,-)$ relations.

These patterns summarise the intuition of the court's deliberation: a cooperative bloc of formalists (α,β,γ) opposed to a smaller but cohesive fairness bloc (δ,ε). The slightly negative values between γ and the other formalists capture the reality that Judge γ partly agreed on the outcome but disagreed on the moral reasoning—a form of *tolerated dissent*.

The discretion case demonstrates the need for a model that can explain coalitions that are *stable but not unanimous*. Traditional game-theoretic or logic-based models require complete alignment of utilities or rules, whereas courts and other multi-agent institutions rely on *partial alignment*: members may disagree on values yet still form a workable majority. The numerical values in Tables 1–2 encapsulate this phenomenon in a minimal form. They will serve as input to the weighted social network and contrastive argumentation framework analysed in Sections 3–4.

3 Methodology: A-BDI and DiSCo–RAD

We use *formal and computational argumentation* in the sense defined in the third volume of *Handbook of Formal Argumentation* [38]. In that usage, *formal argumentation* is the representation, management and (at times) resolution of conflict; *computational argumentation* studies and implements those processes with computational methods (algorithms, complexity, data structures) and their integration with other technologies.

Dung's abstract argumentation theory [22] is marked as an attack-defense paradigm shift: acceptance depends on how arguments *attack* and *defend* one another, abstracting from internal structure. In this view, argumentation is a graph-based abstraction of nonmonotonic inference. This perspective has become a central bridge across AI subfields, because the same abstract machinery

can be specialised, extended, compared, and aligned with other reasoning formalisms.

Within this context, argumentation has been developed along three complementary *conceptualizations* that we will later align with the A–BDI metamodel: argumentation as *balancing*, as *dialogue*, and as *inference*. Over the past two decades this has yielded a family of argumentation representations:

Value-based and weighted models enrich abstract argumentation with priorities or numeric strengths to articulate trade-offs among competing reasons [13,34].

Bipolar and coalitional models distinguish support from attack and capture cooperation/opposition among agents (early contributions from the Toulouse school and collaborators) [3,17,14].

Conflict tolerant semantics relax conflict-freeness to model reasoning under inconsistency or disagreement [7,8].

Multi-agent and dialogical approaches view argumentation as interaction (debate, persuasion, negotiation), often linked to social choice and deliberation [10,9].

The Handbook of Formal Argumentation volumes survey this landscape in depth [25,13,38]. In this paper, rather than treating diversity as a problem, we adopt the *logic-as-toolbox* stance: combine components (semantics, weighting, aggregation, dialogue protocols) to suit the application, and verify the fit by *alignment*. In the remainder, we operationalise that stance with two devices. First, the *A–BDI* metamodel will organise the modelling layers. Second, *Reasoning Alignment Diagrams* (RADs) will make explicit how a normative route and an argumentation route commute—or where and why they diverge. Our case study on judicial discretion then instantiates this toolbox for coalition formation among judges and for explaining patterns of separate opinions.

3.1 The logic-as-toolbox perspective

In line with the tradition in the Knowledge Representation and Reasoning (KR) community, Gabbay [23] promotes the view of logic not as a single system but as a *toolbox of reasoning mechanisms*: each mechanism is a component that can be selected and combined for particular applications. This view motivates our modelling choices. Instead of introducing yet another new formalism, we assemble existing ideas into a transparent pipeline:

(1) A weighted reason base captures the balance of benefits and costs between agents;

(2) A social layer defines relations such as friendship, enmity, and support;

(3) An argumentation layer transforms these relations into a contrastive argumentation framework whose semantics determine accepted collaborations;

(4) A coalition layer extracts groups of agents that jointly satisfy rationality, tolerance, and majority postulates.

Each layer can be replaced or refined without affecting the others, mirroring the modular design advocated in the Handbook of Formal Argumentation volumes [25,13,38].

3.2 The A–BDI metamodel

The A-BDI metamodel [37] organizes this toolbox into three complementary perspectives:

Balancing as identified by Gordon [25] involves weighing the pros and cons of an issue in order to reach a balanced decision or judgment. In such a system, pro and con arguments for alternative resolutions of the issues (options or positions) are put forward, evaluated, resolved, and balanced [25]. The formal methods used are multi-criteria decision theory and case-based reasoning, and they are applied in the law [26], ethics [34] and decision theory [18].

Dialogue Argumentation as dialogue conceptualizes argumentation as a form of interaction aimed at resolving conflicts of opinion [31]. It focuses on the exchange among multiple agents according to defined protocols. Dialogue highlights the distributed nature of information, the selective disclosure of arguments [9], and the agents' strategies for achieving collective or competing objectives [9].

Inference covers the logical semantics that determine which arguments, or collaborations, are ultimately accepted. Here we use Dung-style semantics [22] but adapted to a *contrastive* setting where attacks come from competing collaboration options of the same source.

In this paper, we interpret *dialogue* in a broader sense than conversational or turn-taking protocols. It refers to the *relational layer* where agents can be friends, enemies, or *frenemies* [35,27]. We define the network of mutual influences within which balancing operates, governing how agents' reasoning interacts for mutual or individual benefit, when disagreements can be shielded by coalitions, and how collective acceptance emerges from inter-agent relations; this interpretation echoes work on bipolar [2,17] and social argumentation [36,33,14].

The A-BDI view emphasises that these three layers are not sequential computations but complementary descriptions: each layer can be formalised and reasoned about in its own right. Together, they support modular design and explanation.

3.3 Reasoning Alignment Diagrams (RADs)

RAD visualises how different reasoning routes—for example, a normative specification and a computational procedure—can be aligned or "commuted". In a RAD, each route transforms the same input into an output. If both produce the same result, the diagram *commutes*: the system is sound and complete with respect to its specification. If not, the diagram reveals where approximation or information loss occurs.

RADs can also serve as a general framework for connecting the three conceptualizations of argumentation introduced earlier. They instantiate the *attack–defence paradigm shift*—Dung's abstract argumentation—by showing how different forms of reasoning can be represented within a unified structure. For example, Dung showed that several forms of nonmonotonic inference can be rep-

resented within his theory of abstract argumentation. This correspondence can be depicted by the RAD in Figure 1 [38]. The diagram relates two approaches to deriving conclusions from a knowledge base. The first route (arrow 1) is the *canonical* inference route defined by the underlying defeasible logic. The second route (arrows 2–3–4) is the *argumentation route*: it begins with the translation of the knowledge base into a structured argumentation framework (arrow 2), applies semantics to determine extensions (arrow 3), and extracts the conclusions of the accepted arguments (arrow 4).

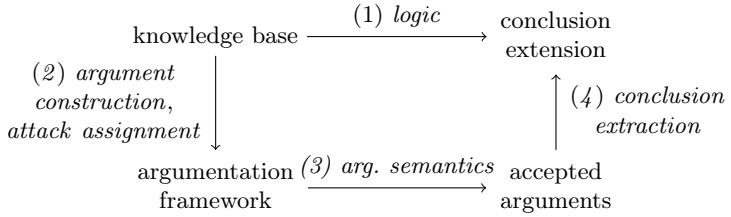


Fig. 1. The Inference-as-Argumentation RAD: the argumentation route (arrows 2–3–4) explains the inferences performed by the source formalism (arrow 1) [38].

RADs can also combine *argumentation as inference* and *argumentation as dialogue*. Consider the problem of determining the extensions of an argumentation framework under a given semantics (arrow 3 in Figure 1). This problem can be reformulated in terms of *two-player discussion games* [15], where a *proponent* and an *opponent* alternate in attacking or defending arguments. Starting from an initial claim by the proponent, the dialogue follows a fixed protocol; the existence of a winning strategy for the proponent corresponds to the argument being accepted. This correspondence is shown in the *Argumentation-as-Discussion* RAD in Figure 2. Here, arrow 1 represents the canonical evaluation of semantics, whereas arrows 2–3–4 describe its computation through a discussion game. Note that arrow 1 in Figure 2 corresponds to arrow 3 in Figure 1.

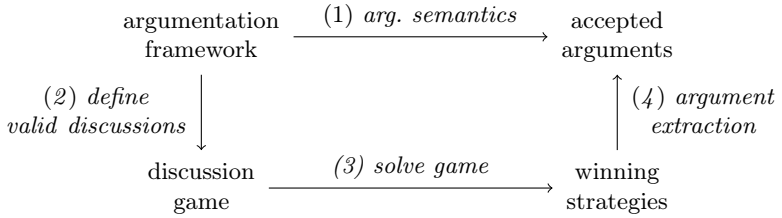


Fig. 2. The Argumentation-as-Discussion RAD: discussion games (arrows 2–3–4) explain argument acceptability (arrow 1) [1].

In this paper, we introduce the *DiSCo–RAD*. It aligns a *normative route*, defined by coalition postulates such as IR, Tol(k), and Maj, with a *computational route* expressed in a contrastive argumentation game. The diagram

commutes when both routes yield the same coalition, showing that the computational mechanism faithfully realises the normative model of discretionary reasoning.

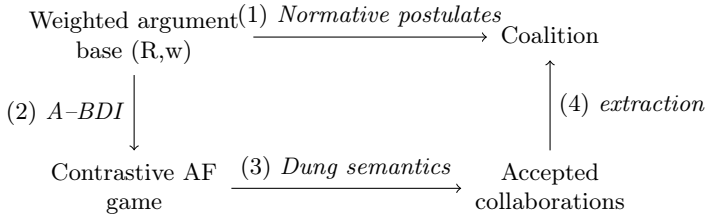


Fig. 3. DiSCo–RAD: alignment between the normative and argumentation routes.

3.4 Applying the methodology to discretion

We start from a legal case rather than from abstract definitions, because judicial discretion provides an exemplary domain where multiple reasoning models coexist and must be combined. Legal formalism relies on textual interpretation; value-based reasoning appeals to moral and social considerations; and deliberative reasoning among judges resembles multi-agent dialogue. By aligning these heterogeneous forms of reasoning, DiSCo–RAD demonstrates how the toolbox can be used in practice.

In our case study, the balancing layer captures how each judge weighs reasons of legality versus fairness, producing the dual-scale collaboration values of Table 1. The dialogue layer models interpersonal relations: formalist judges act as mutual supporters, while fairness-oriented judges form a separate alliance. The inference layer then builds a contrastive argumentation framework where each possible collaboration (e.g., "Judge α collaborates with Judge β") becomes an argument, and attacks represent incompatible choices. The semantics of this framework determine which collaborations are accepted and, through the coalition extraction function, which groups of judges form a rational and tolerant majority.

4 Formal Framework

In this section, we present two aligned routes from weighted pairwise relations to coalitions:

(1) Normative route (social network $\rightarrow$ coalition): read *friend*/*enemy*/*frenemy* from the signs of pairwise collaboration values, then grow majority coalitions by a *friend-first + tolerance* closure satisfying IR and Tol(k).

(2) Argumentation route (contrastive BAF $\rightarrow$ extensions $\rightarrow$ coalition): encode directed *collaborations* as arguments; add a *local, intention-sensitive* attack and a *deductive* support relation that implements Tol(k); compute β-extensions [2]; extract coalitions from mutual-intent edges.

4.1 Balancing and social relations

We keep the balancing layer minimal.

Definition 4.1 [Weighted base and values] Let Ag be the set of agents and Ar the set of case reasons. A reason $(a, \alpha, \beta) \in R \subseteq Ar \times Ag \times Ag$ supports collaboration "α teams with β." A weighting $w : R \to \mathbb{R}^2_{\geq 0}$ attaches (w_1, w_2) for initiator/partner. Aggregation yields a *relative collaboration value* $\pi(\alpha, \beta) \in \mathbb{R}$ with sign indicating benefit (≥ 0) or loss (< 0) to α from teaming with β.

Definition 4.2 [Friend, enemy, frenemy] For $\alpha \neq \beta$:

$$\mathsf{friend}(\alpha, \beta) \iff \pi(\alpha, \beta) \geq 0 \ \wedge \ \pi(\beta, \alpha) \geq 0;$$

$$\mathsf{enemy}(\alpha, \beta) \iff \pi(\alpha, \beta) < 0 \ \wedge \ \pi(\beta, \alpha) < 0;$$

otherwise α, β are *frenemies*.

Tolerance is enforced via a simple margin inequality.

Definition 4.3 [Support with margin k] Given $k \geq 0$, β *supports* α *against* γ iff $\mathsf{friend}(\alpha, \beta) \ \wedge \ \pi(\alpha, \gamma) < 0 \ \wedge \ \pi(\alpha, \beta) + \pi(\alpha, \gamma) \ \geq \ k$.

4.2 Normative coalition formation: friend-first + tolerance

Let F be the undirected *friend graph* on Ag with edge $\{\alpha, \beta\}$ iff $\mathsf{friend}(\alpha, \beta)$. We grow coalitions from friend components and add frenemies only when bilateral tolerance is witnessed.

Definition 4.4 [Normative closure NORM_k] For each friend component S_0 of F, define the sequence

$$S_{i+1} = S_i \ \cup \ \{\gamma\}$$

where $\gamma \notin S_i$, and for every $\beta \in S_i$, $\mathsf{friend}(\gamma, \beta)$ or $\mathsf{frenemy}(\gamma, \beta)$ (there exist $\delta_1, \delta_2 \in S_i$ such that δ_1 supports γ against β or δ_2 supports β against γ).

Let $S^\star$ be the fixpoint of this iteration. A *normative coalition at margin* k is any $S^\star$ that also satisfies Maj (majority size).

By construction, any $S^\star$ produced by NORM_k satisfies IR (no enemy pair inside) and Tol(k) (frenemy pairs are bilaterally covered by allies); the Maj filter ensures feasibility.

4.3 Argumentation route: a contrastive bipolar AF

We now move to a bipolar argumentation framework (BAF) whose *arguments* are directed collaborations and whose relations encode *local attack* (exclusive choices from one agent's perspective) and *deductive support* (our tolerance rule).

Definition 4.5 [Contrastive BAF of collaborations] Fix $k \geq 0$. Let

$A := \{(\alpha, \beta) \in Ag \times Ag \mid \alpha \neq \beta\}$. Define the bipolar AF $\mathcal{B}_k = (A, \mathrm{Att}, \mathrm{Sup})$ by: $\mathrm{Att} := \{\, ((\alpha, \beta), (\alpha, \gamma)) \mid \pi(\alpha, \gamma) < 0 \,\}$ (*local, source-α attack*); and $\mathrm{Sup} := \{\, ((\beta, \alpha), (\alpha, \gamma)) \mid \mathsf{friend}(\alpha, \beta) \wedge \pi(\alpha, \gamma) < 0 \wedge \pi(\alpha, \beta) + \pi(\alpha, \gamma) \geq k \,\}$ (*deductive support implementing Tol(k)*).

Acceptability is defined via the dual *defeat/defence* for BAFs under the deductive reading of support: a set defeats an argument iff it defeats *all* its supporters; a set defends an argument iff it defeats *all* attackers of *some* supporter. These notions generalise Dung's AAFs and yield β-admissible/β-complete/β-preferred/β-stable/β-semi-stable semantics with standard properties (existence of β-complete/β-preferred, Fundamental Lemma, labelling correspondence). We rely on these definitions and results in what follows, see [2] for further details. [2]

Definition 4.6 [Defeat, defence, β-semantics] Let Sup^* be the reflexive and transitive closure of Sup. For $S \subseteq A$ and $a \in A$:

- S *defeats* a iff $\forall u \in \mathrm{Sup}^*(a)\ \exists b \in S:\ (b,u) \in \mathrm{Att}$;
- S *defends* a iff $\exists u \in \mathrm{Sup}^*(a)\ \forall b\big((b,u) \in \mathrm{Att} \Rightarrow S \text{ defeats } b\big)$.

Write $F_{\mathcal{B}}(S) := \{a \in A \mid S \text{ defends } a\}$. Then S is *conflict-free* iff it does not defeat any of its elements; S is β*-admissible* iff it is conflict-free and $S \subseteq F_{\mathcal{B}}(S)$. β-complete/β-preferred/β-stable/β-grounded/β-semi-stable extensions are defined as in AAFs but using $F_{\mathcal{B}}$ and defeat/defence above (e.g., β-preferred = maximal β-admissible).

4.4 Coalition extraction from β-extensions

A β-extension E collects the *possible collaborations*. We read coalitions from pairs that are mutually intended and then, if desired, prune back to IR.

Definition 4.7 [Mutual intent, maximal cliques, and extraction] Let $E \subseteq A$ be a β–extension. Define the *mutual–intent graph* $H_E = (Ag, E^{\leftrightarrow})$ with

$$E^{\leftrightarrow} := \big\{\{\alpha,\beta\} \subseteq Ag \;\big|\; (\alpha,\beta) \in E \ \wedge\ (\beta,\alpha) \in E\big\}.$$

A set $C \subseteq Ag$ is E*–endorsed* iff it induces a clique in H_E, i.e., $\{\alpha,\beta\} \in E^{\leftrightarrow}$ for all distinct $\alpha,\beta \in C$. Let $\mathsf{EXTClo}_k(E)$ be the family of *inclusion–maximal* E–endorsed sets that also satisfy **Maj**. (Optionally) apply an *IR–pruning* pass: remove endpoints that form enemy pairs inside C from one side; this does not break Tol(k) witnesses because supporters are friends by construction.

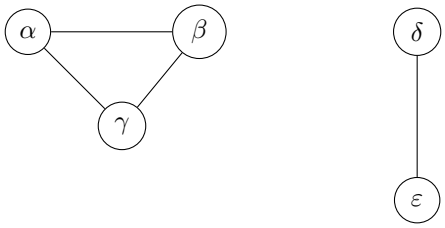


Fig. 4. Mutual–intent graph H_E: maximal cliques $\{\alpha,\beta,\gamma\}$ and $\{\delta,\varepsilon\}$. Only the triangle is a majority.

[2] We use the defeat/defence duality and β-semantics for Bipolar AFs in the sense of *deductive support*, which collapse to Dung's semantics when support is ignored [22,2].

5 Worked Illustration

We revisit the five-agent toy ($Ag = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$) with collaboration values from Tables 1–2 and set $k = 1$.

Normative route. The friend graph has two components, $\{\alpha, \beta\}$ and $\{\delta, \varepsilon\}$. Starting at $S_0 = \{\alpha, \beta\}$, judge γ is a *frenemy* to both, but has bilateral coverage:

$$\pi(\alpha, \beta) + \pi(\alpha, \gamma) = 3 + (-2) = 1 \geq 1, \quad \pi(\beta, \alpha) + \pi(\beta, \gamma) = 2 + (-1) = 1 \geq 1.$$

Thus γ is added by NORM_1 (Def. 4.4), yielding $S^\star = \{\alpha, \beta, \gamma\}$, which satisfies IR, Tol(1), and Maj. The other component $\{\delta, \varepsilon\}$ cannot reach a majority.

Argumentation route. Build $\mathcal{B}_1 = (A, \mathrm{Att}, \mathrm{Sup})$ per Def. 4.5. For source α, every (α, β) attacks $(\alpha, \gamma), (\alpha, \delta), (\alpha, \varepsilon)$ because those targets have $\pi < 0$. For source β, $(\beta, \cdot)$ analogously attacks $(\beta, \gamma), (\beta, \delta), (\beta, \varepsilon)$. Deductive supports capture tolerance: (β, α) Sup (α, γ) since $\mathsf{friend}(\alpha, \beta)$ and $3 + (-2) \geq 1$; symmetrically, (α, β) Sup (β, γ) since $\mathsf{friend}(\alpha, \beta)$ and $2 + (-1) \geq 1$.

Under the β-semantics with dual defeat/defence, support neutralises local attacks precisely when supporters stand: to defeat (α, γ), one must defeat *all* its supporters in Sup^*; here, (β, α) is unattacked, so (α, γ) can be defended and accepted together with (α, β). The same holds for (β, γ) given (α, β). Hence a β-preferred (also β-complete) extension is:

$$E = \{(\alpha, \beta), (\beta, \alpha), (\alpha, \gamma), (\gamma, \alpha), (\beta, \gamma), (\gamma, \beta)\} \ \cup \ \{(\delta, \varepsilon), (\varepsilon, \delta)\}.$$

These β-semantics generalise Dung's AAFs and ensure existence and fixpoint properties under deductive support. [2]

Extraction and alignment. With the extension E from above, H_E has edges $\{\alpha, \beta\}, \{\alpha, \gamma\}, \{\beta, \gamma\}, \{\delta, \varepsilon\}$. The inclusion–maximal cliques are $\{\alpha, \beta, \gamma\}$ and $\{\delta, \varepsilon\}$; only the former satisfies **Maj**, hence $\mathsf{EXTClo}_1(E) = \{\{\alpha, \beta, \gamma\}\}$. Optional IR-pruning has no effect (no enemy pair inside). The coalition coincides with the normative one, so the alignment diagram commutes on this instance.

Remark (IR vs. Tol(k)). Because defeat requires eliminating *all* supporters, β-preferred sets can, in other instances, contain mutually negative pairs when sufficiently shielded by friends; EXTClo_k can then be followed by IR-pruning if strict IR is desired.

6 Related Work

Our approach follows the methodological tradition of Gabbay [23], which introduced the idea of building logics from reusable components—labels for time, agents, and resources; modalities for knowledge and obligation; and nonmonotonic rules—selected and combined to suit a given application. This view, known as *logic as a toolbox*, promotes methodological pluralism with engineering discipline: choose the right tools and make design choices explicit. Extending this vision, Gabbay and Rivlin promote argumentation as the core logic of interactive and explainable reasoning for the 21st century [24]. The *A–BDI*

metamodel [37] continues this trajectory by systematising formal argumentation into three complementary conceptualisations—balancing, dialogue, and inference—and showing how each can model different aspects of a legal case such as child custody. Complementarily, the RAD framework [1] aligns distinct reasoning routes and combining conceptualisations within one system. In this paper, we apply these methodological ideas to judicial discretion, integrating the three conceptualisations within a single toolbox model rather than using them in isolation.

Dung's abstract argumentation framework [22] is not only a generalisation of nonmonotonic inference but also unifies game-theoretic concepts. In his original paper, Dung observed that the stable extensions of an argumentation framework correspond to *von Neumann–Morgenstern stable sets* in cooperative game theory and to stable matchings in matching theory. This connection validates the use of argumentation semantics for modelling rational coalitions and equilibrium behaviour [3,17,14]. Building on this insight, our *contrastive argumentation framework* transfers the same stability intuition to the level of *directed collaborations* between agents, where local attacks and deductive supports define equilibrium-like acceptability dynamics. The resulting structure bridges abstract agent reasoning [38,10,9] and coalition formation [3,17,14], grounding DiSCo–RAD in both logical and social game semantics.

Arieli's work on reasoning under inconsistency [7,8] shares a similar idea with our conflict tolerance mechanism, which preserves classical acceptance while allowing controlled internal disagreement through a *support-with-margin* condition. Our model also draws on dual-scale balancing theories [34], representing benefit and cost for each collaboration. This approach complements broader work on argument strength [32,6], weighted argumentation [13,5], preference-based argumentation [29], and value-based reasoning [11]. Related extensions—gradual [12] and ranking-based semantics [4], probabilistic [28], and decision-theoretic models—explore how degrees of acceptability evolve under uncertainty and preferences.

From an application perspective, modelling *discretionary judicial reasoning* remains a major bottleneck for knowledge-representation and reasoning approaches. Only a few recent works explicitly address this challenge. Dik and Markovich's line of research formalises discretion as a *normative reasoning problem*, centred on the interplay between judicial freedom and its boundaries. Their deontic logic of discretion introduces *nuanced permissions* to represent degrees of judicial freedom in child-custody cases [19]. Building on this, their modal logic of the *duty of care* defines the *obligations* that constrain discretion—to determine all relevant factors, to weigh them properly, and to reason consistently [21]. In a subsequent work, they implement this normative characterisation in *Answer Set Programming* [20], modelling the judicial hierarchy, the obligation to be consistent, and the declaration of violations by higher-level courts [20]. We complement this research by modeling discretionary reasoning using *formal argumentation*. We focus on a specific and paradigmatic form of discretion—constitutional adjudication—and model how majority decisions

can emerge despite internal disagreement. In contrast to the deontic logic perspective, which describes the normative duties governing a single judge's reasoning, our framework captures the collective and relational dimension of discretion among judges. *DiSCo–RAD* thus operationalises discretionary reasoning as a form of *argumentative alignment*: a structured explanation showing how individual reasoning paths converge into a normatively acceptable collective outcome within the general methodology of formal and computational argumentation.

7 Summary and Outlook

This paper presented *DiSCo–RAD*, a Reasoning Alignment Diagram designed to explain how judicial majorities can emerge despite internal disagreement. Taking the Hungarian Constitutional Court as a case study, we modelled discretionary judicial reasoning through two complementary routes: a *normative route*, capturing coalition formation under rationality and tolerance postulates (IR, Tol(k), Maj), and an *argumentation route*, formalised as a contrastive bipolar framework with local attacks and deductive supports. When the two routes converge, the diagram provides an explanatory alignment between the normative and computational levels. Methodologically, the work illustrates the *logic-as-toolbox* idea and the A–BDI metamodel within a single framework. Formally, it introduces a contrastive acceptability model and a tolerance-aware coalition extractor that together account for how partial disagreement can coexist with collective rationality.

Outlook 1: methodological and technical mutual development of RAD. The design of the RAD can be viewed through two complementary lenses: the *methodological* and the *technical*. A useful precedent is the development of the ASPIC-family. On one hand, ASPIC provided a general modelling for structured argumentation; on the other hand, it was developed and refined with technical critiques and refinements—the systematic study of rationality postulates [16]—to ASPIC+ [30]. A similar separation clarifies our own setting. The methodological or meta-level contribution (the reasoning alignment between two routes) belongs in the methodological discussion, while the technical aspects—such as the precise attack design, the explicit treatment of defence and reinstatement, and the equivalence to direct coalition construction—belong in the technical part. This work remains in progress: closer interaction between the case study and the formal machinery will refine both. Some give and take on each side is not only natural but, in our view, necessary for eventual convergence between theory and application.

Outlook 2: combining formal argumentation with large language models. With the increasing capability of large language models (LLMs), a natural next step is to combine them with formal argumentation to simulate and analyse constitutional court cases. LLMs can generate, classify, and rephrase natural-language arguments, while formal argumentation provides the structure to evaluate and align them with normative standards. Such integration would extend the RAD methodology toward a hybrid reasoning

environment—linking symbolic and subsymbolic approaches—and bring the *logic-as-toolbox* vision closer to real-world applications. From an engineering perspective, this approach aligns with the theme of the workshop for *Logic for New-Generation AI*: selecting, combining, and standardising reasoning components that connect theoretical foundations with practical, explainable systems for applications.

Acknowledgments

We thank the anonymous reviewer for their comments. This work is supported by the Luxembourg National Research Fund (FNR) through the following projects: The Epistemology of AI Systems (EAI) (C22/SC/17111440), DJ4ME – A DJ for Machine Ethics: the Dialogue Jiminy (O24/18989918/DJ4ME), Logical Methods for Deontic Explanations (LoDEx) (INTER/DFG/23/17415164/LoDEx), Symbolic and Explainable Regulatory AI for Finance Innovation (SERAFIN) (C24/19003061/SERAFIN), and the University of Luxembourg for the Marie Speyer Excellence Grant for the project Formal Analysis of Discretionary Reasoning (MSE-DISCREASON).

References

[1] *Special issue to celebrate dov gabbay's 80th birthday*, Journal of Applied Logics: The IfCoLog Journal of Logics and their Applications **12** (2025), pp. 1655–1685, to appear. Special Issue to Celebrate Dov Gabbay's 80th Birthday.

[2] Alcântara, J. and R. Cordeiro, *Bipolar argumentation frameworks with a dual relation between defeat and defence*, Journal of Logic and Computation **35** (2024).

[3] Amgoud, L., *An argumentation-based model for reasoning about coalition structures*, in: *International Workshop on Argumentation in Multi-Agent Systems*, Springer, 2005, pp. 217–228.

[4] Amgoud, L. and J. Ben-Naim, *Ranking-based semantics for argumentation frameworks*, in: *International Conference on Scalable Uncertainty Management*, Springer, 2013, pp. 134–147.

[5] Amgoud, L., J. Ben-Naim, D. Doder and S. Vesic, *Acceptability semantics for weighted argumentation frameworks*, in: *Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, International Joint Conferences on Artificial Intelligence (IJCAI), 2017.

[6] Amgoud, L., D. Doder and S. Vesic, *Evaluation of argument strength in attack graphs: Foundations and semantics*, Artificial Intelligence **302** (2022), p. 103607.

[7] Arieli, O., *Conflict-tolerant semantics for argumentation frameworks*, in: *European Workshop on Logics in Artificial Intelligence*, Springer, 2012, pp. 28–40.

[8] Arieli, O., *Conflict-free and conflict-tolerant semantics for constrained argumentation frameworks*, Journal of Applied Logic **13** (2015), pp. 582–604.

[9] Arisaka, R., J. Dauphin, K. Satoh and L. van der Torre, *Multi-agent argumentation and dialogue*, IfCoLog Journal of Logics and Their Applications **9** (2022), pp. 921–954.

[10] Arisaka, R., K. Satoh and L. van der Torre, *Anything you say may be used against you in a court of law: Abstract agent argumentation (Triple-A)*, in: *International Workshop on AI Approaches to the Complexity of Legal Systems*, Springer, 2015, pp. 427–442.

[11] Atkinson, K. and T. J. Bench-Capon, *Value-based argumentation*, IfCoLog Journal of Logics and Their Applications **8** (2021), pp. 1543–1588.

[12] Baroni, P., A. Rago and F. Toni, *From fine-grained properties to broad principles for gradual argumentation: A principled spectrum*, International Journal of Approximate Reasoning **105** (2019), pp. 252–286.

[13] Bistarelli, S., F. Santini et al., *Weighted argumentation*, Handbook of Formal Argumentation, Volume 2 (2021).

[14] Boella, G., L. Van Der Torre and S. Villata, *Social viewpoints for arguing about coalitions*, in: *Pacific Rim International Conference on Multi-Agents*, Springer, 2008, pp. 66–77.

[15] Caminada, M., *Argumentation semantics as formal discussion*, Handbook of Formal Argumentation **1** (2018), pp. 487–518.

[16] Caminada, M. and L. Amgoud, *On the evaluation of argumentation formalisms*, Artificial Intelligence **171** (2007), pp. 286–310.

[17] Cayrol, C. and M.-C. Lagasquie-Schiex, *Coalitions of arguments: A tool for handling bipolar argumentation frameworks*, International Journal of Intelligent Systems **25** (2010), pp. 83–109.

[18] Dietrich, F. and C. List, *A reason-based theory of rational choice*, Nous **47** (2013), pp. 104–134.

[19] Dik, J. and R. Markovich, *Modeling judicial discretion with nuanced permissions*, in: *JURIX* (2024), pp. 48–59.

[20] Dik, J. and R. Markovich, *Judicial discretion as normative reasoning – deontic characterization of judicial decision making with answer set programming*, in: J. Maranhao, editor, *Proceedings of the 20th International Conference on AI and Law* (2025), pp. 258–267.

[21] Dik, J. and R. Markovich, *When judges go wrong: Modeling discretion and the duty of care*, in: *Deontic Logic and Normative Systems* (2025), pp. 399–400.

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

[23] Gabbay, D. M., "Labelled deductive systems," Oxford university press, 1996.

[24] Gabbay, D. M. and L. Rivlin, *Heal2100: human effective argumentation and logic for the 21st century. the next step in the evolution of logic*, IFCoLog Journal of Logics and Their Applications (2017).

[25] Gordon, T. F., *Towards requirements analysis for formal argumentation*, in: P. Baroni, D. Gabbay, M. Giacomin and L. van der Torre, editors, *Handbook of formal argumentation, Volume 1*, College Publictions, 2018 pp. 145–156.

[26] Henkin, L., *Infallibility under law: constitutional balancing*, Colum. L. Rev. **78** (1978), p. 1022.

[27] Hoek, W. v. d., L. B. Kuijer and Y. N. Wáng, *Who should be my friends? social balance from the perspective of game theory*, Journal of Logic, Language and Information **31** (2022), pp. 189–211.

[28] Hunter, A., S. Polberg, N. Potyka, T. Rienstra and M. Thimm, *Probabilistic argumentation: A survey*, Handbook of Formal Argumentation **2** (2021), pp. 397–441.

[29] Kaci, S., L. van der Torre, S. Vesic and S. Villata, *Preference in abstract argumentation*, in: D. Gabbay, M. Giacomin, G. R. Simari and M. Thimm, editors, *Handbook of Formal Argumentation, Volume 2*, College Publications, 2021 pp. 211–248.

[30] Prakken, H., *An abstract framework for argumentation with structured arguments*, Argument & Computation **1** (2010), pp. 93–124.

[31] Prakken, H., *Historical overview of formal argumentation*, in: *Handbook of formal argumentation*, College Publications, 2018 pp. 73–141.

[32] Prakken, H., *An abstract and structured account of dialectical argument strength*, Artificial Intelligence **335** (2024), p. 104193.

[33] Qiao, L., Y. Shen, L. Yu, B. Liao et al., *Arguing coalitions in abstract argumentation*, in: *Logics for New-Generation AI 2021*, CP College Publications, 2021 pp. 93–106.

[34] Tucker, C., "The Weight of Reasons: A Framework for Ethics," Oxford University Press, 2025.

[35] Van der Hoek, W., L. Kuijer and Y. Wáng, *Logics of allies and enemies: A formal approach to the dynamics of social balance theory*, , **2020**, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 210–216.

[36] Yu, L., D. Chen, L. Qiao, Y. Shen and L. van der Torre, *A Principle-based Analysis of Abstract Agent Argumentation Semantics*, in: *Proceedings of the 18th International Conference on Principles of Knowledge Representation and Reasoning*, 2021, pp. 629–639.
URL `https://doi.org/10.24963/kr.2021/60`

[37] Yu, L. and L. van der Torre, *The a-bdi metamodel for human-level ai: Argumentation as balancing, dialogue and inference*, in: *International Conference on Logic and Argumentation (CLAR 2025)*, 2025, pp. 361–379.

[38] Yu, L., L. van der Torre and R. Markovich, *Thirteen challenges of formal and computational argumentation*, in: M. Thimm and G. R. Simari, editors, *Handbook of Formal Argumentation, Volume 3*, 2024 .

Hybrid Tense Logic of the Real Line

Zhiguang Zhao [1]

School of Mathematics and Statistics, Taishan University
Tai'an, P.R.China

Abstract

In the present paper, we give a complete axiomatization of the hybrid tense logic of the real line $\langle \mathbb{R}, < \rangle$ in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$, show its completeness by adding irrational points to the canonical model of the hybrid tense logic of the rational line $\langle \mathbb{Q}, < \rangle$, with the help of the concept of locality.

Keywords: Hybrid logic, tense logic, real line.

1 Introduction

Hybrid logics [2, Chapter 14] use languages extending the language of modal logic where a special class of propositional variables called nominals are used to refer to single states, which are true at exactly one world. In addition, some other logical connectives are used, e.g. the satisfaction operator $@_i\varphi$ which means φ is true at the state denoted by i.

In the literature, there are many existing works on the axiomatization of the modal tense logic [4]. In particular, there is the axiomatization of modal tense logic over the real line $\langle \mathbb{R}, < \rangle$ (see e.g. [3]). However, for the axiomatization of the hybrid tense logic in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ over the real line, the complete axiomatization is missing.

Since the irreflexivity condition cannot be defined in tense language, it is easier to define this condition using the satisfaction operator $@_i$. In addition, for the frame properties definable by pure hybrid tense formulas in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$, completeness result can be obtained automatically with respect to a countable canonical model, so the hybrid tense logic over the rational line $\langle \mathbb{Q}, < \rangle$ can be obtained easily. To characterize $\langle \mathbb{R}, < \rangle$, the property needed is the Dedekind completeness stating that every non-empty subset that has an upper bound has a least upper bound. It is clear that the real line satisfies this property. However, we cannot expect the same proof method for the rational line to work in the setting of the real line, since the property of Dedekind completeness cannot be defined by a pure hybrid tense formula in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$. Otherwise, suppose it is defined by a pure hybrid tense formula φ,

[1] Email: zhaozhiguang23@gmail.com. The research of Zhiguang Zhao is supported by Shandong Provincial Natural Science Foundation, China (project number: ZR2023QF021).

then by the canonical model construction for $L_{\mathbb{Q}}+\varphi$, there will be a countable canonical model which validates $L_{\mathbb{Q}}+\varphi$, which makes the underlying frame a countable unbounded dense linear order which is also Dedekind complete, which is impossible.

In this paper, we propose a complete axiomatization of the hybrid tense logic in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ over the real line. The basic proof strategy is a bit similar to [3,5]: first of all, we give a complete axiomatization of the hybrid tense logic in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ over the rational line, then add irrational points to the canonical model for the rational line, using the concept of locality which we will give later.

The structure of the paper is as follows: Section 2 gives the preliminaries on hybrid tense logic. Section 3 gives the axiomatization of the rational line $\langle \mathbb{Q}, < \rangle$ in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ and its completeness. Section 4 gives the axiomatization of the real line $\langle \mathbb{R}, < \rangle$ in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ and its completeness.

2 Preliminaries on hybrid tense logic

In this section, we give preliminaries on the hybrid tense logic in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$. For more details of hybrid logic, see [2, Chapter 14] and [6].

2.1 Language and syntax

Definition 2.1 Given a countably infinite set Prop of propositional variables and a countably infinite set Nom of nominals which are disjoint, the hybrid language $\mathcal{H}(\Diamond, \blacklozenge, @)$ is defined as follows:

$$\varphi ::= p \mid i \mid \bot \mid \top \mid \neg\varphi \mid \varphi \vee \varphi \mid \varphi \wedge \varphi \mid \varphi \to \varphi \mid \Diamond\varphi \mid \Box\varphi \mid \blacklozenge\varphi \mid \blacksquare\varphi \mid @_i\varphi,$$

where $p \in \text{Prop}$, $i \in \text{Nom}$. We define $\varphi \leftrightarrow \psi := (\varphi \to \psi) \wedge (\psi \to \varphi)$. A formula is pure if it contains no propositional variables in Prop. We use σ to denote a sorted substitution that uniformly replaces propositional variables by formulas and nominals by nominals.

2.2 Semantics

Definition 2.2 A Kripke frame is a tuple $\mathfrak{F} = \langle W, S, S' \rangle$ where $W \neq \varnothing$ is the domain of $\mathfrak{F}$, $S, S' \subseteq W \times W$ [2] are the accessibility relations. A Kripke model is a pair $\mathfrak{M} = \langle \mathfrak{F}, V \rangle$ where $V : \text{Prop} \cup \text{Nom} \to P(W)$ is a valuation on $\mathfrak{F}$ such that $V(i) \subseteq W$ is a singleton for all nominals $i \in \text{Nom}$. A Kripke frame is tense if Swv iff $S'vw$ for all $w, v \in W$. For tense Kripke frames $\mathfrak{F} = \langle W, S, S' \rangle$, sometimes we use $\mathfrak{F} = \langle W, S \rangle$ to denote it when no confusion arises.

The satisfaction relation is given as follows: for any model $\mathfrak{M} = \langle W, S, S', V \rangle$, any $w \in W$,

[2] Here we use S, S' instead of R, R' to avoid confusion with the real numbers.

$$
\begin{array}{lll}
\mathfrak{M}, w \Vdash p & \text{iff} & w \in V(p); \\
\mathfrak{M}, w \Vdash i & \text{iff} & \{w\} = V(i); \\
\mathfrak{M}, w \Vdash \bot & : & \text{never}; \\
\mathfrak{M}, w \Vdash \top & : & \text{always}; \\
\mathfrak{M}, w \Vdash \neg\varphi & \text{iff} & \mathfrak{M}, w \nVdash \varphi; \\
\mathfrak{M}, w \Vdash \varphi \vee \psi & \text{iff} & \mathfrak{M}, w \Vdash \varphi \text{ or } \mathfrak{M}, w \Vdash \psi; \\
\mathfrak{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathfrak{M}, w \Vdash \varphi \text{ and } \mathfrak{M}, w \Vdash \psi; \\
\mathfrak{M}, w \Vdash \varphi \to \psi & \text{iff} & \mathfrak{M}, w \nVdash \varphi \text{ or } \mathfrak{M}, w \Vdash \psi; \\
\mathfrak{M}, w \Vdash \Diamond\varphi & \text{iff} & \exists v(Swv \text{ and } \mathfrak{M}, v \Vdash \varphi); \\
\mathfrak{M}, w \Vdash \Box\varphi & \text{iff} & \forall v(Swv \Rightarrow \mathfrak{M}, v \Vdash \varphi); \\
\mathfrak{M}, w \Vdash \blacklozenge\varphi & \text{iff} & \exists v(S'wv \text{ and } \mathfrak{M}, v \Vdash \varphi); \\
\mathfrak{M}, w \Vdash \blacksquare\varphi & \text{iff} & \forall v(S'wv \Rightarrow \mathfrak{M}, v \Vdash \varphi); \\
\mathfrak{M}, w \Vdash @_i\varphi & \text{iff} & \mathfrak{M}, V(i) \Vdash \varphi.
\end{array}
$$

For any formula φ,

- $V(\varphi) := \{w \in W \mid \mathfrak{M}, w \Vdash \varphi\}$ denotes the truth set of φ in $\mathfrak{M}$.
- φ is globally true on $\mathfrak{M}$ (notation: $\mathfrak{M} \Vdash \varphi$) if $\mathfrak{M}, w \Vdash \varphi$ for every $w \in W$.
- φ is valid on a frame $\mathfrak{F}$ (notation: $\mathfrak{F} \Vdash \varphi$) if φ is globally true on $(\mathfrak{F}, V)$ for each valuation V.

2.3 The first-order correspondence language and the standard translation

In the first-order correspondence language, we have two binary predicate symbols S, S' corresponding to each accessibility relation, unary predicate symbols P corresponding to each propositional variable p, constant symbols i corresponding to each nominal i.

Definition 2.3 The standard translation is given as follows:

- $ST_x(p) := Px$;
- $ST_x(i) := x = i$;
- $ST_x(\bot) := x \neq x$;
- $ST_x(\top) := x = x$;
- $ST_x(\neg\varphi) := \neg ST_x(\varphi)$;
- $ST_x(\varphi \wedge \psi) := ST_x(\varphi) \wedge ST_x(\psi)$;
- $ST_x(\varphi \vee \psi) := ST_x(\varphi) \vee ST_x(\psi)$;
- $ST_x(\varphi \to \psi) := ST_x(\varphi) \to ST_x(\psi)$;
- $ST_x(\Diamond\varphi) := \exists y(Sxy \wedge ST_y(\varphi))$;
- $ST_x(\Box\varphi) := \forall y(Sxy \to ST_y(\varphi))$;
- $ST_x(\blacklozenge\varphi) := \exists y(S'xy \wedge ST_y(\varphi))$;
- $ST_x(\blacksquare\varphi) := \forall y(S'xy \to ST_y(\varphi))$;
- $ST_x(@_i\varphi) := ST_i(\varphi)$.

It is obvious that the translation is correct:

Proposition 2.4 *For any Kripke model $\mathfrak{M}$, any $w \in W$, any hybrid tense formula φ,*

$$\mathfrak{M}, w \Vdash \varphi \text{ iff } \mathfrak{M} \vDash ST_x(\varphi)[w].$$

From this proposition, we have the following corollary:

Proposition 2.5 *For any Kripke frame $\mathfrak{F}$, any pure formula φ,*

$$\mathfrak{F} \Vdash \varphi \text{ iff } \mathfrak{F} \vDash \forall x ST_x(\varphi).$$

Definition 2.6 [Correspondence] We say that a hybrid tense formula φ and a first-order sentence α correspond to each other, if for any Kripke frame $\mathfrak{F}$, we have $\mathfrak{F} \Vdash \varphi$ iff $\mathfrak{F} \vDash \alpha$.

3 The hybrid tense logic of the rational line

In this section, we give the hybrid tense logic of the rational line $\langle \mathbb{Q}, < \rangle$. The system will be a pure axiomatic extension of the basic system for the language $\mathcal{H}(\Diamond, \blacklozenge, @)$, so we get automatic completeness similar to [1, Section 7.3]. For the sake of self-containedness, we briefly sketch the completeness proof here.

3.1 The system $L_{\mathbb{Q}}$

The hybrid tense logic $L_{\mathbb{Q}}$ of the rational line in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ is defined as follows (we follow the style of system as in [6, Definition 5.1.2]):

The basic system part for $\mathcal{H}(\Diamond, \blacklozenge, @)$:

Axioms:

(CT) All classical propositional tautologies;

($\text{Dual}_{\Box}$) $\Diamond p \leftrightarrow \neg\Box\neg p$;

($\text{Dual}_{\blacksquare}$) $\blacklozenge p \leftrightarrow \neg\blacksquare\neg p$;

($\text{K}_{\Box}$) $\Box(p \to q) \to (\Box p \to \Box q)$;

($\text{K}_{\blacksquare}$) $\blacksquare(p \to q) \to (\blacksquare p \to \blacksquare q)$;

($\text{K}_{@}$) $@_i(p \to q) \to (@_i p \to @_i q)$, for all nominals $i \in \text{Nom}$;

(Selfdual) $\neg @_i p \leftrightarrow @_i \neg p$;

(Ref) $@_i i$;

(Intro) $i \wedge p \to @_i p$;

($\text{Back}_{\Box}$) $\Diamond @_i p \to @_i p$;

($\text{Back}_{\blacksquare}$) $\blacklozenge @_i p \to @_i p$;

(Agree) $@_i @_j p \to @_j p$.

Rules:

(MP) From $\varphi \to \psi$ and φ get ψ;

($\text{Nec}_{\Box}$) From φ get $\Box\varphi$;

($\text{Nec}_{\blacksquare}$) From φ get $\blacksquare\varphi$;

($\text{Nec}_{@}$) From φ get $@_i\varphi$;

(Subst) From φ get $\sigma(\varphi)$ where $\sigma(\varphi)$ is a sorted substitution which replace propositional variables by formulas and nominals by nominals;

(Name$_@$) From $@_i\varphi$ get φ, if i does not occur in φ;

(BG$_\Box$) From $@_i\Diamond j \to @_j\varphi$ get $@_i\Box\varphi$, if $i \neq j$ and j does not occur in φ;

(BG$_\blacksquare$) From $@_i\blacklozenge j \to @_j\varphi$ get $@_i\blacksquare\varphi$, if $i \neq j$ and j does not occur in φ.

The additional axioms for $\mathbb{Q}$:

(Inv) $@_i\Diamond j \leftrightarrow @_j\blacklozenge i$;

(Irref) $i \to \neg\Diamond i$;

(Tran) $\Diamond\Diamond i \to \Diamond i$;

(Lin) $\Diamond j \vee j \vee \blacklozenge j$;

(No-end$_\Box$) $\Diamond\top$;

(No-end$_\blacksquare$) $\blacklozenge\top$;

(Dense) $\Diamond i \to \Diamond\Diamond i$.

3.2 Soundness

To show the soundness of the Hilbert proof system given above, it suffices to show that the following correspondences hold:

Proposition 3.1 (i) *(Inv) corresponds to* $\forall w\forall v(Swv \leftrightarrow S'vw)$.

(ii) *(Irref) corresponds to* $\forall w(\neg Rww)$.

(iii) *(Tran) corresponds to* $\forall w\forall v\forall u(Swv \wedge Svu \to Swu)$.

(iv) *(Lin) corresponds to* $\forall w\forall v(Swv \vee w = v \vee S'wv)$.

(v) *(No-end$_\Box$) corresponds to* $\forall w\exists v(Swv)$.

(vi) *(No-end$_\blacksquare$) corresponds to* $\forall w\exists v(S'wv)$.

(vii) *(Dense) corresponds to* $\forall w\forall v(Swv \to \exists u(Swu \wedge Suv))$.

Proof. By Proposition 2.5. □

It is easy to see that the first-order conditions above characterize dense linear order without end points, so $\langle\mathbb{Q}, <, >\rangle$ (which we also denote by $\langle\mathbb{Q}, <\rangle$) validates all the pure hybrid axioms above. Therefore, $L_\mathbb{Q}$ is sound over $\langle\mathbb{Q}, <\rangle$.

3.3 The completeness proof

For the completeness proof, we essentially follow the proof strategy of [1, Section 7.3]. We only list the relevant lemmas and give the proof sketch of the main theorem for the sake of self-containedness.

Definition 3.2 A model $\mathfrak{M} := \langle U, S, S', V\rangle$ is named if for all $x \in W$, there is a nominal i such that $V(i) = \{x\}$.

Definition 3.3 Given a pure formula φ, we say that ψ is a pure instance of φ if is obtained from φ by uniformly substituting nominals for nominals.

Lemma 3.4 (Lemma 7.22 in [1]) *Given a named model* $\mathfrak{M} = \langle\mathfrak{F}, V\rangle$ *and a pure formula* φ*, if* $\mathfrak{M} \Vdash \psi$ *for all pure instances* ψ *of* φ*, then* $\mathfrak{F} \Vdash \varphi$.

Definition 3.5 We say that an $L_{\mathbb{Q}}$-maximal consistent set Γ is named if it contains a nominal, and call any nominal belonging to such a Γ a name for Γ.

Lemma 3.6 (Lemma 7.24 in [1]) *Let Γ be an $L_{\mathbb{Q}}$-maximal consistent set. For every nominal i, let $\Delta_i := \{\varphi \mid @_i\varphi \in \Gamma\}$. Then:*

(i) *For each nominal i, Δ_i is an $L_{\mathbb{Q}}$-maximal consistent set containing i.*

(ii) *For all nominals i and j, if $i \in \Delta_j$ then $\Delta_i = \Delta_j$.*

(iii) *For all nominals i and j, $@_i\varphi \in \Delta_j$ iff $@_i\varphi \in \Gamma$.*

(iv) *If k is a name for Γ, then $\Gamma = \Delta_k$.*

We call $\Delta_i = \{\varphi \mid @_i\varphi \in \Gamma\}$ a named set yielded by Γ.

Definition 3.7 We say that an $L_{\mathbb{Q}}$-maximal consistent set Γ is pasted, if

- $@_i\Diamond\varphi \in \Gamma$ implies that for some nominal j, $@_i\Diamond j \wedge @_j\varphi \in \Gamma$;
- $@_i\blacklozenge\varphi \in \Gamma$ implies that for some nominal j, $@_i\blacklozenge j \wedge @_j\varphi \in \Gamma$.

Lemma 3.8 (Extended Lindenbaum Lemma, Lemma 7.25 in [1]) *Let* Nom$'$ *be a (countably) infinite collection of nominals disjoint from* Nom*, and let $\mathcal{H}'(\Diamond, \blacklozenge, @)$ be the language obtained by adding these new nominals to $\mathcal{H}(\Diamond, \blacklozenge, @)$. Then every $L_{\mathbb{Q}}$-consistent set of formulas in language $\mathcal{H}(\Diamond, \blacklozenge, @)$ can be extended to a named and pasted $L_{\mathbb{Q}}$-maximal consistent set in language $\mathcal{H}'(\Diamond, \blacklozenge, @)$.*

The proof of this lemma is essentially the same as [1, Lemma 7.25], except that when defining the inductive steps, we use the bimodal version of the construction.

Definition 3.9 [Definition 7.26 in [1]] Let Γ be a named and pasted $L_{\mathbb{Q}}$-maximal consistent set. The named model yielded by Γ is defined as $\mathfrak{M}^\Gamma := \langle U^\Gamma, S^\Gamma, S'^\Gamma, V^\Gamma\rangle$, where

- U^Γ is the set of all named sets yielded by Γ.
- $S^\Gamma(u, v)$ iff for any formulas φ, if $\varphi \in v$ then $\Diamond\varphi \in u$.
- $S'^\Gamma(u, v)$ iff for any formulas φ, if $\varphi \in v$ then $\blacklozenge\varphi \in u$.
- $V^\Gamma(p) = \{x \in U^\Gamma \mid p \in x\}$ and $V^\Gamma(i) = \{x \in U^\Gamma \mid i \in x\}$.

By items 1 and 2 in Lemma 3.6, V^Γ assigns every nominal a singleton subset of W^Γ.

Lemma 3.10 (Existence Lemma, Lemma 7.27 in [1]) *Let Γ be a named and pasted $L_{\mathbb{Q}}$-maximal consistent set, and $\mathfrak{M} = \langle U, S, S', V\rangle$ be the named model yielded by Γ.*

- *Suppose $u \in U$ and $\Diamond\varphi \in u$, then there is a $v \in U$ such that $S(u, v)$ and $\varphi \in v$.*
- *Suppose $u \in U$ and $\blacklozenge\varphi \in u$, then there is a $v \in U$ such that $S'(u, v)$ and $\varphi \in v$.*

In this proof, we need to use the theorems $\Diamond j \wedge @_j\varphi \to \Diamond\varphi$ and $\blacklozenge j \wedge @_j\varphi \to \blacklozenge\varphi$.

Lemma 3.11 (Truth Lemma, Lemma 7.28 in [1]) *Let $\mathfrak{M} = \langle U, S, S', V\rangle$ be the named model yielded by a named and pasted $L_{\mathbb{Q}}$-maximal consistent set Γ, and let $u \in U$. Then for all formulas φ, we have $\varphi \in u$ iff $\mathfrak{M}, u \Vdash \varphi$.*

Theorem 3.12 (Completeness Theorem, Theorem 7.29 in [1]) *Every $L_{\mathbb{Q}}$-consistent set of formulas is satisfiable in a countable named model based on a frame which validates every additional pure axiom.*

Proof. Given an $L_{\mathbb{Q}}$-consistent set of formulas Σ, use the Extended Lindenbaum Lemma to expand it to a named and pasted set Γ in a countable language $\mathcal{H}'(\Diamond, \blacklozenge, @)$. Let $\mathfrak{M} = \langle U, S, S', V\rangle$ be the named model yielded by Γ. Since Γ is named, it is in U. By the Truth Lemma, formulas in Σ are satisfied at $\langle \mathfrak{M}, \Gamma\rangle$. The model is countable because each state is named by some nominal in $\mathcal{H}'(\Diamond, \blacklozenge, @)$, and there are only countably many of these.

Now to show that $\mathfrak{F} = \langle U, S, S'\rangle$ validates the additional pure axioms, since all these additional pure axioms belong to all maximal consistent sets in U, hence they are globally true in $\mathfrak{M}$. Therefore, by Lemma 3.4, $\mathfrak{F}$ validates them. □

Then since the additional axioms characterize the properties that S is the inverse of S' and S is a dense unbounded linear order, by the fact that the model is countable (indeed, it is easy to see that the model is an infinite model), it is easy to see that $\mathfrak{F} = \langle U, S, S'\rangle$ is isomorphic to $\langle \mathbb{Q}, <, >\rangle$. This concludes the completeness proof of $L_{\mathbb{Q}}$ with respect to $\langle \mathbb{Q}, <\rangle$.

4 The hybrid tense logic of the real line

In this section, we give the axiomatization of the hybrid tense logic $L_{\mathbb{R}}$ in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$ of the real line $\langle \mathbb{R}, <\rangle$.

For characterizing $\langle \mathbb{R}, <\rangle$, the property needed is the Dedekind completeness stating that every non-empty subset that has an upper bound has a least upper bound. It is clear that the real line satisfies this property.

However, we cannot expect the same proof method for the rational line to work in the setting of the real line, since the property of Dedekind completeness cannot be defined by a pure hybrid tense formula in the language $\mathcal{H}(\Diamond, \blacklozenge, @)$. Otherwise, suppose it is defined by a pure hybrid tense formula φ, then by the canonical model construction for $L_{\mathbb{Q}} + \varphi$, there will be a countable canonical model which validates $L_{\mathbb{Q}} + \varphi$, which makes the underlying frame a countable unbounded dense linear order which is also Dedekind complete, which is impossible.

It can be shown that the rational line $\langle \mathbb{Q}, <\rangle$ and the real line $\langle \mathbb{R}, <\rangle$ satisfy the same first-order sentences, using model-theoretic arguments or Ehrenfeucht-Fraïssé game. As a result, in order to define the hybrid logic for the real line, we cannot expect to use pure hybrid tense formulas, since $\langle \mathbb{Q}, <\rangle$ and $\langle \mathbb{R}, <\rangle$ validate the same pure hybrid tense formulas. We essentially need non-pure formulas.

4.1 The modal tense formula for Dedekind completeness

Indeed, we can show that there is a modal tense formula which does not contain any nominals characterize Dedekind completeness in the class of dense unbounded linear orders. We define the abbreviation $A\varphi := \Box\varphi \wedge \varphi \wedge \blacksquare\varphi$, and $E\varphi := \Diamond\varphi \vee \varphi \vee \blacklozenge\varphi$.

Lemma 4.1 *For any tense Kripke model $\mathfrak{F} = \langle W, <, >, V\rangle$ where $<$ is a linear order on W, and any $w \in W$, the following equivalences hold:*

- $\mathfrak{M}, w \Vdash A\varphi$ *iff* $\mathfrak{M}, v \Vdash \varphi$ *for all* $v \in W$*;*
- $\mathfrak{M}, w \Vdash E\varphi$ *iff* $\mathfrak{M}, v \Vdash \varphi$ *for some* $v \in W$*.*

Proof. It follows easily from the fact that $<$ is a linear order. □

Proposition 4.2 *For any tense Kripke frame $\mathfrak{F} = \langle W, <, >\rangle$ where $<$ is a dense unbounded linear order on W, the formula*

$$Ep \wedge E\neg p \wedge A(p \to \Box p) \wedge A(\neg p \to \blacksquare\neg p) \to E(\Box p \wedge \blacksquare\neg p)$$

is valid on $\mathfrak{F}$ iff $\mathfrak{F} = \langle W, <, >\rangle$ is Dedekind complete.

Proof. $\Rightarrow$: Consider any non-empty subset $U \subseteq W$ such that U has at least an upper bound. Take the valuation $V : \mathrm{Prop} \cup \mathrm{Nom} \to P(W)$ such that $w \in V(p)$ iff w is an upper bound of U.

Then it is easy to see the following:

- $V(p)$ is non-empty: since U has at least an upper bound.
- $V(\neg p)$ is non-empty: since $<$ is unbounded and there is at least a point which is not the upper bound of U.
- $V(p \to \Box p) = W$: since if $w \in V(p)$, then w is an upper bound of U, and any point v such that $w < v$ is also an upper bound of U, so $v \in V(p)$, so $w \in V(\Box p)$.
- $V(\neg p \to \blacksquare\neg p) = W$: since if $w \in V(\neg p)$, then w is not an upper bound of U, for any v such that $w > v$, v cannot be an upper bound of U, so $v \in V(\neg p)$, therefore $w \in V(\blacksquare\neg p)$.

Therefore, for any point w_0, $Ep \wedge E\neg p \wedge A(p \to \Box p) \wedge A(\neg p \to \blacksquare\neg p)$ is true at w_0. Thus $\mathfrak{F}, V, w_0 \Vdash E(\Box p \wedge \blacksquare\neg p)$, so there is a v_0 such that $\mathfrak{F}, V, v_0 \Vdash \Box p \wedge \blacksquare\neg p$. We have:

(i) For any point u such that $v_0 < u$, $u \in V(p)$, so u is an upper bound of U.

(ii) For any point u' such that $v_0 > u'$, $u' \in V(\neg p)$, so u' is not an upper bound of U.

Now we can show that v_0 is the least upper bound of U:

- If v_0 is not an upper bound of U, then there is a $w \in U$ such that $w > v$, so by item 1, w is an upper bound of U. Since $<$ is dense, there is a u such that $v < u < w$, so by item 1, u is an upper bound of U, but $u < w \in U$, a contradiction. So v_0 is an upper bound of U.

- Since for any point u' such that $v_0 > u'$, u' is not an upper bound of U, and for any u such that $v_0 < u$, u is an upper bound of U, we have that v_0 is the least upper bound of U.

Therefore, $\mathfrak{F}$ is Dedekind complete.

$\Leftarrow$: Suppose that $\mathfrak{F}$ is Dedekind complete. Then for any valuation V on $\mathfrak{F}$, any $w \in W$, suppose that $\mathfrak{F}, V, w \Vdash Ep \wedge E\neg p \wedge A(p \to \Box p) \wedge A(\neg p \to \blacksquare\neg p)$, then $V(p)$ and $V(\neg p)$ are not empty, and $V(p) \subseteq V(\Box p)$ and $V(\neg p) \subseteq V(\blacksquare\neg p)$.

Therefore, consider $V(p)$, suppose $w \in V(p) \subseteq V(\Box p)$, then for any v such that $w < v$, we have $v \in V(p)$, so $V(p)$ is upward closed. Similarly, we can show that $V(\neg p)$ is downward closed.

Now consider the set $V(\neg p)$, since $<$ is a linear order on W, for any $w \in V(\neg p)$, any $v \in V(p)$, we have that $w < v$: it is obvious that $w \neq v$, and if $w > v$ then by the upward closedness of $V(p)$ we have $w \in V(p)$, a contradiction. So $w < v$ by the linearity of $<$. Therefore by the non-emptiness of $V(p)$ and $V(\neg p)$, $V(\neg p)$ has upper bounds.

By the Dedekind completeness, we have that $V(\neg p)$ has a least upper bound w_0. Now we show that $\mathfrak{F}, V, w_0 \Vdash \Box p \wedge \blacksquare\neg p$:

- For any v such that $w_0 < v$, if $v \in V(\neg p)$ then w_0 is not an upper bound of $V(\neg p)$, a contradiction, so $v \in V(p)$, therefore $w_0 \in V(\Box p)$.
- For any v such that $w_0 > v$, if $v \in V(p)$ then v is an upper bound of $V(\neg p)$, which makes w_0 not the least upper bound, a contradiction, so $v \in V(\neg p)$, therefore $w_0 \in V(\blacksquare\neg p)$.

Therefore, $\mathfrak{F}, V, w \Vdash E(\Box p \wedge \blacksquare\neg p)$, so $Ep \wedge E\neg p \wedge A(p \to \Box p) \wedge A(\neg p \to \blacksquare\neg p) \to E(\Box p \wedge \blacksquare\neg p)$ is valid on $\mathfrak{F}$. □

Now we obtain the system $L_{\mathbb{R}}$ by adding $\varphi_{\mathbb{R}} := Ep \wedge E\neg p \wedge A(p \to \Box p) \wedge A(\neg p \to \blacksquare\neg p) \to E(\Box p \wedge \blacksquare\neg p)$ to $L_{\mathbb{Q}}$.

From the proof above, we can see that $\langle \mathbb{R}, <, > \rangle$ validates $\varphi_{\mathbb{R}}$. Since $\langle \mathbb{R}, < \rangle$ is a dense unbounded linear order, $\langle \mathbb{R}, <, > \rangle$ validates all the additional pure axioms for $L_{\mathbb{Q}}$. This concludes the soundness proof of $L_{\mathbb{R}}$ with respect to $\langle \mathbb{R}, <, > \rangle$.

4.2 Completeness proof for the real line

We will show the completeness of $L_{\mathbb{R}}$ with respect to the real line $\langle \mathbb{R}, < \rangle$.

We consider the same procedure of completeness proof for $L_{\mathbb{Q}}$, but now with the additional axiom $\varphi_{\mathbb{R}}$, as in Section 3. As we can see, the canonical frame validates all the pure additional axioms, so it is a countable frame isomorphic to $\langle \mathbb{Q}, < \rangle$ such that the axiom $\varphi_{\mathbb{R}}$ is globally true. However, $\varphi_{\mathbb{R}}$ is only globally true for the specific valuation of the canonical model, we cannot show that $\varphi_{\mathbb{R}}$ is valid on the canonical frame, since otherwise the frame will be Dedekind complete, a contradiction to it being isomorphic to $\mathbb{Q}$.

Now we identify the canonical frame for $L_{\mathbb{R}}$ with $\langle \mathbb{Q}, S, S' \rangle$, and name every node with a corresponding rational number such that $S(\Gamma_x, \Gamma_y)$ iff $x < y$ iff $S'(\Gamma_y, \Gamma_x)$.

Now our aim is to revise the canonical model $\langle \mathbb{Q}, S, S', V \rangle$ into a new canon-

ical model $\langle \mathbb{R}, T, T', V' \rangle$ such that its domain becomes the "real numbers", and the additional nodes satisfy the order relation required, and the truth lemma holds as well. Then it is clear that every $L_{\mathbb{R}}$-consistent set is satisfiable on the canonical model $\langle \mathbb{Q}, S, S', V \rangle$, and therefore on the new canonical model $\langle \mathbb{R}, T, T', V' \rangle$, which validates the additional pure axioms together with $\varphi_{\mathbb{R}}$. The basic proof strategy is similar to [3], but the proof details are significantly different.

4.2.1 The construction of the new canonical model

In what follows, we use q with indices to represent a rational number, and r to represent a real number (which might be rational as well).

For each irrational number r, we need to define a $L_{\mathbb{R}}$-maximal consistent set such that it can be added into the canonical model without affecting the value of formulas in the "rational" part.

Definition 4.3 Given the rational canonical model $\mathfrak{M} = \langle \mathbb{Q}, S, S', V \rangle$ and an irrational number r, we call a formula φ local at r, if there are $q, q' \in \mathbb{Q}$ such that $q < r < q'$ and $\varphi \in \Gamma_{q''}$ for all $q'' \in [q, q'] \cap \mathbb{Q}$.

The idea is that if φ is local at r, then there is a closed neighbourhood $[q, q']$ such that φ is in every $\Gamma_{q''}$ such that $q'' \in [q, q'] \cap \mathbb{Q}$. In principle, if all the points around r makes φ true, then φ should also be true at r in order not to "disturb" the canonical model.

Definition 4.4 For each irrational number r, we define the following three sets of formulas:

$\Lambda_r := \{\varphi \mid \varphi \text{ is local at } r\}$
$\Sigma_r^{\Diamond} := \{\Diamond\varphi \mid \text{there is a } q \text{ such that } \varphi \in \Gamma_q \text{ and } r < q\}$
$\Sigma_r^{\blacklozenge} := \{\blacklozenge\varphi \mid \text{there is a } q \text{ such that } \varphi \in \Gamma_q \text{ and } r > q\}$
$\Pi_r^{\Diamond} := \{\neg\Diamond\varphi \mid \Diamond\varphi \notin \Sigma_r^{\Diamond}\}$
$\Pi_r^{\blacklozenge} := \{\neg\blacklozenge\varphi \mid \blacklozenge\varphi \notin \Sigma_r^{\blacklozenge}\}$

Lemma 4.5 *Formulas in $\Sigma_r^{\Diamond}$, $\Sigma_r^{\blacklozenge}$ are local at r.*

Proof. We only prove for $\Sigma_r^{\Diamond}$, the other being similar. Suppose that $\Diamond\varphi \in \Sigma_r^{\Diamond}$, then there is a q such that $\varphi \in \Gamma_q$ and $r < q$. Then by denseness and unboundedness, there are q_1, q_2 such that $q_1 < r < q_2 < q$ such that for any $q_3 \in [q_1, q_2] \cap \mathbb{Q}$, there is a q such that $\varphi \in \Gamma_q$ and $q_3 < q$, so by the definition of the rational canonical model, $\Diamond\varphi \in \Gamma_{q_3}$. So $\Diamond\varphi$ is local at r. □

Lemma 4.6 *For each nominal i, $\neg i$ is local at r.*

Proof. Suppose that $i \in \Gamma_q$ for some rational number q, without loss of generality we suppose that $r < q$, then for any rational number $q' \neq q$, we have that $\neg i \in \Gamma_{q'}$. Therefore, by denseness and unboundedness, there are q_1, q_2 such that $q_1 < r < q_2 < q$ and for any $q_3 \in [q_1, q_2] \cap \mathbb{Q}$, $\neg i \in \Gamma_{q_3}$. So $\neg i$ is local at r. □

Lemma 4.7 *Λ_r is closed under taking conjunction.*

Proof. It suffices to show that if $\varphi, \psi \in \Lambda_r$, then $\varphi \wedge \psi \in \Lambda_r$. For φ, there are $q_0, q_1 \in \mathbb{Q}$ such that $q_0 < r < q_1$ and $\varphi \in \Gamma_q$ for all $q \in [q_0, q_1] \cap \mathbb{Q}$. For ψ,

there are $q_2, q_3 \in \mathbb{Q}$ such that $q_2 < r < q_3$ and $\psi \in \Gamma_q$ for all $q \in [q_2, q_3] \cap \mathbb{Q}$. Now we take q_4 to be the maximum of $\{q_0, q_2\}$ and q_5 to be the minimum of q_1, q_3, then $q_4 < r < q_5$ and $\varphi \wedge \psi \in \Gamma_q$ for all $q \in [q_4, q_5] \cap \mathbb{Q}$. □

Lemma 4.8 *Λ_r is $L_{\mathbb{R}}$-consistent.*

Proof. By the previous lemma, it suffices to see that every single formula φ in Λ_r is $L_{\mathbb{R}}$-consistent. Since every formula φ local at r, there is an interval $[q_0, q_1] \cap \mathbb{Q}$ such that $q_0 < r < q_1$ and for any q in the interval, $\varphi \in \Gamma_q$, so φ is $L_{\mathbb{R}}$-consistent. □

Lemma 4.9 *$\Lambda_r \cup \Sigma_r^{\Diamond} \cup \Sigma_r^{\blacklozenge} \cup \Pi_r^{\Diamond} \cup \Pi_r^{\blacklozenge}$ is $L_{\mathbb{R}}$-consistent with respect to the proof system. Indeed, by the locality of formulas in $\Sigma_r^{\Diamond} \cup \Sigma_r^{\blacklozenge}$, it suffices to show that $\Lambda_r \cup \Pi_r^{\Diamond} \cup \Pi_r^{\blacklozenge}$ is $L_{\mathbb{R}}$-consistent.*

Proof.

Suppose otherwise, there is a formula $\varphi \in \Lambda_r$ and formulas $\neg\Diamond\varphi_1, \ldots, \neg\Diamond\varphi_m \in \Pi_r^{\Diamond}$, formulas $\neg\blacklozenge\psi_1, \ldots, \neg\blacklozenge\psi_n \in \Pi_r^{\blacklozenge}$ such that $\varphi \wedge \neg\Diamond\varphi_1 \wedge \ldots \wedge \neg\Diamond\varphi_m \wedge \neg\blacklozenge\psi_1 \wedge \ldots \wedge \neg\blacklozenge\psi_n \rightarrow \bot$ is provable. Then $\varphi \rightarrow (\Diamond\varphi_1 \vee \ldots \vee \Diamond\varphi_m \vee \blacklozenge\psi_1 \vee \ldots \vee \blacklozenge\psi_n)$ is provable.

If both m and n are 0, then $\neg\varphi$ is provable, therefore φ cannot be local at r, a contradiction.

If $m > 0$, then consider $\varphi_1 \vee \ldots \vee \varphi_m$, we know that $\Diamond\varphi_1 \vee \ldots \vee \Diamond\varphi_m$ is provably equivalent to $\Diamond(\varphi_1 \vee \ldots \vee \varphi_m)$. Now we can show that $\neg\Diamond(\varphi_1 \vee \ldots \vee \varphi_m) \in \Pi_r^{\Diamond}$. Otherwise, $\Diamond(\varphi_1 \vee \ldots \vee \varphi_m) \in \Sigma_r^{\Diamond}$, then there is a $q > r$ such that $\varphi_1 \vee \ldots \vee \varphi_m \in \Gamma_q$, so there is a $\varphi_i \in \Gamma_q$, so $\Diamond\varphi_i \in \Sigma_r^{\Diamond}$, a contradiction to $\neg\Diamond\varphi_i \in \Pi_r^{\Diamond}$. Therefore, we have $\neg\Diamond(\varphi_1 \vee \ldots \vee \varphi_m) \in \Pi_r^{\Diamond}$. So without loss of generality we can assume that $m = 1$.

Similarly, we can assume that $n = 0$ or $n = 1$.

Since φ is local at r, we have that there is a closed interval $[q, q'] \cap \mathbb{Q}$ with $q < r < q'$ such that $\varphi \in \Gamma_{q''}$ for all $q'' \in [q, q'] \cap \mathbb{Q}$.

For the case $n = 0$, we have that $\varphi \rightarrow \Diamond\varphi_1$ is provable. Since φ is local at r, we have that $\Diamond\varphi_1$ is local at r as well, which means that $\Diamond\varphi_1 \in \Sigma_r^{\Diamond}$, a contradiction to $\neg\Diamond\varphi_1 \in \Pi_r^{\Diamond}$.

For the case $n = 1$, we have that $\varphi \rightarrow \Diamond\varphi_1 \vee \blacklozenge\psi_1$ is provable. Now for each $q'' \in [q, q'] \cap \mathbb{Q}$, either $\Diamond\varphi_1 \in \Gamma_{q''}$ or $\blacklozenge\psi_1 \in \Gamma_{q''}$.

Now we consider q. If $\blacklozenge\psi_1 \in \Gamma_q$, then there is a $q_1 < q$ such that $\psi_1 \in q_1$, so $\blacklozenge\psi_1 \in \Gamma_{q''}$ for all $q'' \in [q, q'] \cap \mathbb{Q}$, which means that $\blacklozenge\psi_1$ is local at r, a contradiction to $\blacklozenge\psi_1 \in \Pi_r^{\blacklozenge}$.

Therefore, we have that $\Diamond\varphi_1 \in \Gamma_q$. So there is a $q_2 > q$ such that $\varphi_1 \in \Gamma_{q_2}$. If $q_2 > r$, then $\Diamond\varphi_1 \in \Gamma_{q''}$ for all $q'' \in [q, q_2) \cap \mathbb{Q}$, which means that $\Diamond\varphi_1$ is local at r, a contradiction to $\Diamond\varphi_1 \in \Pi_r^{\Diamond}$. Therefore $q_2 < r$ (since r is irrational), and for all $q''' > r$, we have that $\neg\varphi_1 \in \Gamma_{q'''}$.

By a symmetric argument, consider q', if $\Diamond\varphi_1 \in \Gamma_{q'}$, we can get a similar contradiction, so $\blacklozenge\psi_1 \in \Gamma_{q'}$, and there is a $q_3 < q'$ such that $\psi_1 \in \Gamma_{q_3}$. If $q_3 < r$, we can get a similar contradiction. Therefore $q_3 > r$, and for all $q''' < r$, we have that $\neg\psi_1 \in \Gamma_{q'''}$.

Now we consider the least upper bound r_1 of $\{q_4 \mid \varphi_1 \in \Gamma_{q_4}\}$. Then by the previous argument, we have that $r_1 \leq r$. If $r_1 < r$, then for all rational number $q_5 > r_1$ we have $\neg\varphi_1 \in \Gamma_{q_5}$, therefore $\Box\neg\varphi_1 \in \Gamma_{q_5}$, which means that for any point $q_6 \in (r_1, q'] \cap \mathbb{Q}$, we have $\neg\Diamond\varphi_1 \in \Gamma_{q_6}$, so $\blacklozenge\psi_1 \in \Gamma_{q_6}$. Consider a $q_7 \in (r_1, r)$, then $\blacklozenge\psi_1 \in \Gamma_{q_7}$, since $q_7 < r$, we can get a contradiction. Therefore $r_1 = r$.

Similarly, we consider the greatest lower bound r_2 of $\{q_8 \mid \psi_1 \in \Gamma_{q_8}\}$, then we can show that $r_2 = r$.

Now $r_1 = r_2 = r$, so we have that $\Box\neg\varphi_1 \in \Gamma_{q_9}$ for all $q_9 > r$, and $\blacksquare\neg\psi_1 \in \Gamma_{q_{10}}$ for all $q_{10} < r$. Therefore, $\Box\neg\varphi_1 \leftrightarrow \neg\blacksquare\neg\psi_1$ is provable.

Now consider the axiom $Ep \wedge E\neg p \wedge A(p \rightarrow \Box p) \wedge A(\neg p \rightarrow \blacksquare\neg p) \rightarrow E(\Box p \wedge \blacksquare\neg p)$, and substitute p by $\Box\neg\varphi_1$, then it is easy to see that $E\Box\neg\varphi_1, E\neg\Box\neg\varphi_1, A(\Box\neg\varphi_1 \rightarrow \Box\Box\neg\varphi_1), A(\blacksquare\neg\psi_1 \rightarrow \blacksquare\blacksquare\neg\psi_1) \in \Gamma_{q_{11}}$ for all $q_{11} \in \mathbb{Q}$, so $E(\Box\Box\neg\varphi_1 \wedge \blacksquare\blacksquare\neg\psi_1) \in \Gamma_{q_{11}}$ for all $q_{11} \in \mathbb{Q}$, so there is a $q_{12} \in \mathbb{Q}$ such that $\Box\Box\neg\varphi_1 \wedge \blacksquare\blacksquare\neg\psi_1 \in \Gamma_{q_{12}}$. But then $\Box\neg\varphi_1 \in \Gamma_{q_{13}}$ for all $q_{13} > q_{12}$, and $\blacksquare\neg\psi_1 \in \Gamma_{q_{14}}$ for all $q_{14} < q_{12}$, which is only possible if $q_{12} = r$, a contradiction since r is irrational.

Therefore, the set is $L_{\mathbb{R}}$-consistent. □

Now we can extend $\Lambda_r \cup \Sigma_r^{\Diamond} \cup \Sigma_r^{\blacklozenge} \cup \Pi_r^{\Diamond} \cup \Pi_r^{\blacklozenge}$ to an $L_{\mathbb{R}}$-maximal consistent set Γ_r. Notice that here we do not need a named and pasted maximal consistent set.

Now we build the model $\mathfrak{M}' = \langle \mathbb{R}, T, T', V' \rangle$ as follows:

- $\mathbb{R} = \{\Gamma_r \mid r \text{ is a real number}\}$
- $T(\Gamma_{r_1}, \Gamma_{r_2})$ iff $r_1 < r_2$
- $T'(\Gamma_{r_1}, \Gamma_{r_2})$ iff $r_1 > r_2$
- $V'(p) = \{\Gamma_r \mid p \in \Gamma_r\}$

It is easy to see that when restricting the model to the rational part, it is the same as the rational canonical model $\mathfrak{M} = \langle \mathbb{Q}, S, S', V \rangle$.

Now we can prove the truth lemma for $\mathfrak{M}' = \langle \mathbb{R}, T, T', V' \rangle$:

Lemma 4.10 *For any formula φ and any real number r,*

$$\Gamma_r \Vdash \varphi \text{ iff } \varphi \in \Gamma_r.$$

Proof. The basic case and the Boolean case are easy. Notice that for nominals, they cannot be true at irrational points.

For the $\Diamond$ and $\blacklozenge$ cases, it suffices to consider the $\Diamond$ case, the other is similar.

We discuss in two cases.

- If r is a rational number,

  $\Rightarrow$: If $\Gamma_r \Vdash \Diamond\varphi$, then there is a real number r_1 such that $T(r, r_1)$ and $\Gamma_{r_1} \Vdash \varphi$. By induction hypothesis, $\varphi \in \Gamma_{r_1}$.

  Then we can find a rational number $q_1 > r$ such that $\varphi \in \Gamma_{q_1}$. Suppose otherwise, $\varphi \notin \Gamma_{q_1}$ for all rational $q_1 > r$, then $\neg\varphi \in \Gamma_{q_1}$ for all rational $q_1 > r$, so $\neg\varphi$ is local at r_1, so $\neg\varphi \in \Gamma_{r_1}$, a contradiction to $\varphi \in \Gamma_{r_1}$.

Therefore we can find a rational number $q_1 > r$ such that $\varphi \in \Gamma_{q_1}$. Therefore, by the definition of S in $\mathfrak{M} = \langle \mathbb{Q}, S, S', V \rangle$, $\Diamond\varphi \in \Gamma_r$.

$\Leftarrow$: If $\Diamond\varphi \in \Gamma_r$, then there is a rational number $q_1 \in \mathbb{Q}$ such that $\varphi \in \Gamma_{q_1}$ and $q_1 > r$. By induction hypothesis, $\Gamma_{q_1} \Vdash \varphi$, therefore from $T(r, q_1)$ we get $\Gamma_r \Vdash \Diamond\varphi$.

- If r is an irrational number,

  $\Rightarrow$: If $\Gamma_r \Vdash \Diamond\varphi$, then by the same argument as in the rational case, we can find rational number $q_1 > r$ such that $\varphi \in \Gamma_{q_1}$. By the definition of $\Sigma_r^{\Diamond}$, we have that $\Diamond\varphi \in \Sigma_r^{\Diamond} \subseteq \Gamma_r$.

  $\Leftarrow$: Suppose that $\Diamond\varphi \in \Gamma_r$, then since $\Sigma_r^{\Diamond} \cup \Pi_r^{\Diamond}$ concerns all the $\Diamond$-formulas, we have that $\Diamond\varphi \in \Sigma_r^{\Diamond}$. Otherwise $\neg\Diamond\varphi \in \Pi_r^{\Diamond} \subseteq \Gamma_r$, a contradiction to $\Diamond\varphi \in \Gamma_r$.

  Therefore, there is a $q \in \mathbb{Q}$ such that $\varphi \in \Gamma_q$ and $T(r, q)$. By induction hypothesis, $\Gamma_q \Vdash \varphi$, so $\Gamma_r \Vdash \Diamond\varphi$.

□

Now for any $L_{\mathbb{R}}$-consistent set Γ, we can satisfy Γ at a rational point in the rational canonical model $\mathfrak{M} = \langle \mathbb{Q}, S, S', V \rangle$, therefore satisfy Γ also at the same point in $\mathfrak{M}' = \langle \mathbb{R}, T, T', V' \rangle$, and $\langle \mathbb{R}, T, T' \rangle$ validate the logic of $\langle \mathbb{R}, < \rangle$. This concludes the completeness proof of $L_{\mathbb{R}}$ with respect to $\langle \mathbb{R}, < \rangle$.

References

[1] Blackburn, P., M. de Rijke and Y. Venema, "Modal Logic," Cambridge Tracts in Theoretical Computer Science **53**, Cambridge University Press, 2001.

[2] Blackburn, P., J. van Benthem and F. Wolter, "Handbook of modal logic," Studies in Logic and Practical Reasoning **3**, Elsevier, 2006.

[3] de Jongh, D., F. Veltman and R. Verbrugge, *Completeness by construction for tense logics of linear time*, Liber Amicorum for Dick de Jongh. Institute of Logic, Language and Computation, Amsterdam (2004).

[4] Goldblatt, R., "Logics of time and computation," Center for the Study of Language and Information, 1987.

[5] Gruszczynski, R. and Z. Zhao, *Hybrid logic of strict betweenness*, Submitted (2025).

[6] ten Cate, B. D., "Model theory for extended modal languages," Ph.D. thesis, University of Amsterdam, Netherlands (2005).

Kripke Semantics for MTL

Andrew Lewis-Smith

Department of Computer Science
Middlesex University
London, United Kingdom

Zhiguang Zhao [1]

School of Mathematics and Statistics
Taishan University
Tai'an, China

Abstract

We provide a generalisation of Kripke semantics for Monoidal T-Norm Logic (**MTL**), and prove adequacy of the same. In doing so, we exploit constructions found in [4], extending insights from [8,15,16,17] to **MTL**, a fuzzy logic lacking divisibility. This continues our programme of extracting generalisations of Kripke semantics from algebraic semantics for fuzzy logics.

Keywords: Monoidal T-norm logic, Kripke semantics, ordinal sum

1 Introduction

Monoidal T-Norm Logic [7] (hereon **MTL**) is an important fuzzy and substructural logic, being the so-called 'system of left-continuous t-norms' [11]. Viewed as a substructural logic, **MTL** properly extends intuitionistic affine logic [1], being therefore proof-theoretically more satisfactory [2] than fuzzy logics featuring the non-analytic condition of divisibility [5,6] - for instance, Hajek's **BL** is a proper extension of **MTL**. But **MTL** *is* a fuzzy logic, being complete for the standard unit-interval endowed with the **MTL**-algebraic structure [11].

The present paper gives relational semantics from the algebraic representation for **MTL** given in [4]. In further contrast with our earlier work, the system **MTL** is a fragment of **BL** [2] lacking divisibility, and thus cannot have

[1] Email: zhaozhiguang23@gmail.com. The research of Zhiguang Zhao is supported by Shandong Provincial Natural Science Foundation, China (project number: ZR2023QF021).

[2] Alias $\mathbf{GBL}_{ewf}$ with the axiom of Prelinearity.

the same representation.[3] This feature is reflected in the Kripke semantics we provide here, using [4]: *Zuluaga-Castiglioni*-structures defined over linearly ordered frames. Worlds and formulas are mapped to **Semihoop**-chains via sloping functions modified for the present cases, in contrast to **GBL** and **BL** in which worlds and formulas are mapped to $[0,1]_{\mathbf{MV}}$-chains – these latter come directly from the concrete characterisation via poset products of $\mathbf{GBL}_{ewf}$-algebras in terms of poset products of **MV**-chains (specifically $[0,1]_{\mathbf{MV}}$) given in [14,3].

The present paper expands our programme for extracting relational semantics for fuzzy logics from poset products to ordinal sum representations (see also [19]), while also providing relational semantics more along the lines of e.g. [16] derived from poset products given in [4]. In doing so, we extend the above-mentioned programme to systems without divisibility. This is appropriate as the majority of representation results in the algebraic literature, even around **BL** (and it's variants), concern ordinal sums as opposed to poset products. We further believe this might serve as a fruitful point of comparison for relational semantics derived from poset products [12,13,14] with ordinal sum based semantics.

Furthermore, relational semantics brings connections to modal logic and the model theory of that subject, and opens doors for tableaux and labelled calculi. Indeed, the question of suitable analytic calculi for fuzzy logics (as alternatives to the Hilbert-style presentation) poses a major open problem for fuzzy logic as a whole. By pursuing relational semantics for fuzzy logics, we aim to convert algebraic representation results into insights useful from a more general proof-theoretic, computational and semantic perspective, a unity which relational semantics traditionally suggests in modal logic.

The present paper proceeds as follows, as with [15,16]. We provide **MTL**'s Hilbert and natural deduction systems, followed by suitable definitions of algebras, validity, and our relational semantics. We prove **MTL** sound and complete for our relational semantics based on ordinal sums. We conclude with some discussion in anticipation of future research.

2 Monoidal T-Norm Logic MTL

Monoidal t-norm logic formulas are inductively defined from atomic formulas, including $\bot$, and the binary connectives $\psi \wedge \chi$, $\psi \vee \chi$, $\psi \otimes \chi$ and $\psi \to \chi$. We will refer to this language as $\mathcal{L}_{\otimes}$, since it extends the $\mathcal{L}$ of intuitionistic logic with a second form of conjunction $\psi \otimes \chi$.

Figure 1 gives a natural deduction system for monoidal t-norm logic

[3] This follows from Jipsen and Montagna's results in [14], in which the extensions of **GBL** are exhaustively classified; **MTL**-algebras lack divisibility, and so fall outside Jipsen and Montagna's classification.

$$\frac{}{\Gamma, \phi \vdash \phi}\ \mathrm{Ax} \qquad \frac{\Gamma, \phi, \psi, \Delta \vdash \chi}{\Gamma, \psi, \phi, \Delta \vdash \chi}\ \mathrm{Ex}$$

$$\frac{\Gamma, \phi \vdash \psi}{\Gamma \vdash \phi \to \psi}\ \to \mathrm{I} \qquad \frac{\Gamma \vdash \phi \qquad \Delta \vdash \phi \to \psi}{\Gamma, \Delta \vdash \psi}\ \to \mathrm{E}$$

$$\frac{\Gamma \vdash \phi \qquad \Delta \vdash \psi}{\Gamma, \Delta \vdash \phi \otimes \psi}\ \otimes \mathrm{I} \qquad \frac{\Gamma \vdash \phi \otimes \psi \qquad \Delta, \phi, \psi \vdash \chi}{\Gamma, \Delta \vdash \chi}\ \otimes \mathrm{E}$$

$$\frac{\Gamma \vdash \phi \qquad \Gamma \vdash \psi}{\Gamma \vdash \phi \wedge \psi}\ \wedge \mathrm{I} \qquad \frac{\Gamma \vdash \phi_1 \wedge \phi_2}{\Gamma \vdash \phi_i}\ \wedge \mathrm{E} \quad (i \in \{1, 2\})$$

$$\frac{\Gamma \vdash \phi_i}{\Gamma \vdash \phi_1 \vee \phi_2}\ \vee \mathrm{I} \quad (i \in \{1, 2\}) \qquad \frac{\Gamma \vdash \phi \vee \psi \qquad \Delta, \phi \vdash \chi \qquad \Delta, \psi \vdash \chi}{\Gamma, \Delta \vdash \chi}\ \vee \mathrm{E}$$

$$\frac{\Gamma \vdash \bot}{\Gamma \vdash \phi}\ \bot \mathrm{E} \qquad \frac{}{\Gamma \vdash (\phi \to \psi) \vee (\psi \to \phi)}\ \mathrm{Prelin}$$

Fig. 1. Monoidal T-Norm Logic **MTL**

MTL. When we wish to stress the precise system in which a sequent $\Gamma \vdash \phi$ is derivable we use the system as a subscript of the provability sign, e.g. $\Gamma \vdash_{\mathbf{MTL}} \phi$. Note that Γ here is a multi-set (not a set), as this logic does not have contraction, the number of occurrences of a formula in the context Γ matters. Weakening of contexts is allowed, which is captured in the axiom $\Gamma, \phi \vdash \phi$. This makes **MTL** an extension of intuitionistic affine logic.

And here we give the Hilbert system [7]:

(A0) $(\phi \to \phi)$.

(A1) $(\phi \to \psi) \to ((\psi \to \chi) \to (\phi \to \chi))$.

(A2) $(\phi \otimes \psi) \to \phi$.

(A3) $(\phi \otimes \psi) \to (\psi \otimes \phi)$.

(A4) $(\phi \wedge \psi) \to \phi$.

(A5) $(\phi \wedge \psi) \to (\psi \wedge \phi)$.

(A6) $(\phi \otimes (\phi \to \psi)) \to (\phi \wedge \psi)$.

(A7a) $(\phi \to (\psi \to \chi)) \to ((\phi \otimes \psi) \to \chi))$.

(A7b) $((\phi \otimes \psi) \to \chi)) \to (\phi \to (\psi \to \chi))$.

(A8) $((\phi \to \psi) \to \chi) \to (((\psi \to \phi) \to \chi) \to \chi)$.

(A9) $\bot \to \phi$.

($\vee$1) $(\phi \vee \psi) \to (((\phi \to \psi) \to \psi) \wedge ((\psi \to \phi) \to \phi))$.

($\vee$2) $(((\phi \to \psi) \to \psi) \wedge ((\psi \to \phi) \to \phi)) \to (\phi \vee \psi)$.

(R1) $\phi, \phi \to \psi \vdash_{\mathbf{MTL_H}} \psi$.

Proposition 2.1 *The natural deduction system* **MTL** *(Figure 1) has the same*

derivable formulas as the Hilbert-style system $\mathbf{MTL_H}$ *of [3], and hence corresponds to it in the following sense:*[4]

$$\psi_1, \ldots, \psi_n \vdash_{\mathbf{MTL}} \phi \quad \textit{iff} \quad \vdash_{\mathbf{MTL_H}} (\psi_1 \otimes \ldots \otimes \psi_n) \to \phi.$$

Proof. $\Rightarrow$: The result follows by a simple induction on the structure of the natural deduction proof once we have established that each instance of a natural deduction rule translates to a theorem of **MTL**. We translate each sequent $\phi_1, \ldots, \phi_n \vdash \chi$ to the formula $[\phi_1, \ldots, \phi_n \vdash \chi] = \phi_1 \otimes \ldots \otimes \phi_n \to \chi$, and each rule

$$\frac{\Theta_1 \ldots \Theta_m}{\Psi} \text{ to } [\Theta_1] \otimes \ldots \otimes [\Theta_m] \to [\Psi].$$

For example, (Ax) translates to $(\bigotimes \Gamma) \otimes \phi \to \phi$ (by (A1),(A2),(A3),(R1)) and ($\to$I) to $(((\bigotimes \Gamma) \otimes \phi) \to \psi) \to ((\bigotimes \Gamma) \to (\phi \to \psi))$, which is also a form of (A7b).

The analysis of many of the other rules is simplified if we introduce a relation between formulae:

$$\phi \leq \psi \text{ iff } \vdash_{\mathbf{MTL_H}} \phi \to \psi.$$

Omitting mention of use of (R1), (A0) says that this relation is reflexive, and (A1) that it is transitive. We therefore view it as generating a partial order on its equivalence classes. (A1) also implies that $\to$ is antitone in its first argument, (A3), (A7a) and (A7b) imply that $\phi \to (\psi \to \chi)$ is equivalent to $(\phi \otimes \psi) \to \chi$, $(\psi \otimes \phi) \to \chi$ and $\psi \to (\phi \to \chi)$.

Right-to-left: This follows by induction on the $\mathbf{MTL_H}$ derivation of $(\psi_1 \otimes \ldots \otimes \psi_n) \to \phi$ once we have shown that each of the axioms of $\mathbf{MTL_H}$ is a theorem of **MTL**. For example, to show $\vdash \phi \otimes \psi \to \phi$:

$$\dfrac{\dfrac{\dfrac{}{\phi \otimes \psi \vdash \phi \otimes \psi}\,\text{Ax} \qquad \dfrac{}{\phi, \psi \vdash \phi}\,\text{Ax}}{\phi \otimes \psi \vdash \phi}\,\otimes\text{E}}{\vdash \phi \otimes \psi \to \phi}\,\to\text{I}$$

□

3 Algebraic Semantics for MTL

In this section, we give the algebraic semantics for **MTL**.

Definition 3.1 *A* semihoop *is an algebra* $\mathcal{A} = (A, \otimes, \to, \wedge, \vee, 1)$ *of type* $(2, 2, 2, 2, 0)$ *such that* $(A, \wedge, \vee)$ *is a lattice with* 1 *as the greatest element,* $(A, \otimes, 1)$ *is a commutative monoid and for every* $x, y, z \in A$ *the following conditions hold:*

- *(residuation)* $x \otimes y \leq z$ *if and only if* $x \leq y \to z$.

[4] Note that we use $\psi \otimes \chi$ where Esteva and Godo in [7] use $\psi \& \chi$.

- *(prelinearity)* $(x \to y) \vee (y \to x) = 1$.

Equivalently, a semihoop is an integral, commutative and prelinear residuated lattice.

Note 1 *This makes* $(A, \otimes, \to, \wedge, \vee, 1)$ *an integral commutative residuated lattice. A semihoop* A *is* bounded *if* $(A, \wedge, \vee, 1)$ *has a least element* 0*. We often use* $\top$ *to denote* 1 *and* $\bot$ *to denote* 0*.*

Definition 3.2 *An* **MTL***-algebra* $\mathcal{A}$ *is a bounded semihoop.*

Note 2 *Hence,* **MTL***-algebras are prelinear integral bounded commutative residuated lattices, as usually defined in the literature.*

Definition 3.3 *An* **MTL***-algebra* $\mathcal{A}$ *is an MTL-chain if its semihoop reduct is totally ordered.*

Note 3 *It is known that the theory of* **MTL***-algebras forms a variety.*

Definition 3.4 (Denotation functions) *Given an* **MTL***-algebra* $\mathcal{A}$*, a mapping is from propositional variables to elements of* $\mathcal{A}$*:*

$$p \mapsto [\![p]\!] \in \mathcal{A}$$

We thus refer to the denotation of a variable p *as* $[\![p]\!]_{\mathcal{A}}$*. We can extend that mapping to all formulas in the language of* $\mathcal{L}_{\otimes}$ *in a straightforward way:*

$$\begin{aligned}
[\![\bot]\!]_{\mathcal{A}} &:= \bot \\
[\![\phi \otimes \psi]\!]_{\mathcal{A}} &:= [\![\phi]\!]_{\mathcal{A}} \otimes [\![\psi]\!]_{\mathcal{A}} \\
[\![\phi \wedge \psi]\!]_{\mathcal{A}} &:= [\![\phi]\!]_{\mathcal{A}} \wedge [\![\psi]\!]_{\mathcal{A}} \\
[\![\phi \vee \psi]\!]_{\mathcal{A}} &:= [\![\phi]\!]_{\mathcal{A}} \vee [\![\psi]\!]_{\mathcal{A}} \\
[\![\phi \to \psi]\!]_{\mathcal{A}} &:= [\![\phi]\!]_{\mathcal{A}} \to [\![\psi]\!]_{\mathcal{A}}
\end{aligned}$$

Definition 3.5 (Validity) *A sequent* $\phi_1, \ldots, \phi_n \vdash \psi$ *is said to be valid in* $\mathcal{A}$*, if* $[\![\phi_1]\!] \otimes \ldots \otimes [\![\phi_n]\!] \leq [\![\psi]\!]$ *holds in* $\mathcal{A}$ *for all denotation functions. A sequent is said to be* valid *if it is valid in* all **MTL***-algebras. It is easy to show that the valid sequents, in the sense above, are precisely the ones provable in* **MTL***.*

Proposition 3.6 (Completeness, see [7]) *A sequent* $\Gamma \vdash \psi$ *is* **MTL***-valid iff it is provable in* **MTL***.*

4 Kripke Semantics for MTL

In this section, we give the Kripke semantics for **MTL**. The Kripke semantics for **MTL** that we propose is something of a variant of our semantics introduced in [16] based on [14,3], but is more immediately derived from [4]. We first need to define a particular class of functions from the set of worlds W to **Semihoop**-chains attached with a bottom from below.

Definition 4.1 (Sloping functions) *Let* $\mathcal{W} = \langle W, \succeq \rangle$ *be a linear order with greatest element* w_G*. Let* $\{\mathcal{A}_w\}_{w \in W}$ *be a* W*-indexed family of* **Semihoop***-chains with almost disjoint domains except that they share the top element* $\top$*.*

We further require that $\mathcal{A}_{w_G}$ has a bottom element $\bot_{\mathcal{A}_{w_G}}$. We add the same new bottom element $\star$ to each $\mathcal{A}_w$ (where $w \neq w_G$) and get $\mathcal{B}_w$, and define $\mathcal{B}_{w_G} := \mathcal{A}_{w_G}$.

A function $f\colon W \to \bigcup_{w\in W} \mathcal{B}_w$ is said to be a sloping function for **MTL** *(hereon* sloping function*, or* sloping*) if it is the $\top$-constant function or there are $w_f \in W$ and $a_f \in \mathcal{A}_{w_f} - \{\top\}$ such that*

$$f(w) = \begin{cases} \top & \text{if } w \succ w_f \\ a_f & \text{if } w = w_f \\ \star & \text{if } w \prec w_f. \end{cases}$$

Notice that w_G cannot be below any element in W, so $f(w_G)$ cannot be $\star$, so the function is well-defined. Notice also that the least sloping function corresponds to $(w_G, \bot_{\mathcal{A}_{w_G}})$, i.e. it maps w_G to $\bot_{\mathcal{A}_{w_G}}$ and all other worlds to $\star$.

We can show that the set of sloping functions is closed under taking the following operations:

Lemma 4.2 *If $f, g\colon W \to \bigcup_{w\in W}(\mathcal{A}_w \cup \{\star\})$ are sloping, then so are the following functions:*

$$\begin{aligned}
(f \wedge g)(w) &:= f(w) \wedge_{\mathcal{A}_w \cup \{\star\}} g(w) \\
(f \vee g)(w) &:= f(w) \vee_{\mathcal{A}_w \cup \{\star\}} g(w) \\
(f \otimes g)(w) &:= f(w) \otimes_{\mathcal{A}_w \cup \{\star\}} g(w) \\
(f \to g)(w) &:= \begin{cases} \top & \text{if} \quad (w_f \succ w_g) \text{ or } (w_f = w_g \text{ and } a_f \leq a_g) \\ \top & \text{if} \quad g \text{ is the } \top\text{-constant function} \\ g(w) & \text{if} \quad w_f \prec w_g \text{ or } f \text{ is the } \top\text{-constant function} \\ \top & \text{if} \quad w \succ w_f = w_g \text{ and } a_f > a_g \\ a_f \to_{\mathcal{A}_w} a_g & \text{if} \quad w = w_f = w_g \text{ and } a_f > a_g \\ \star & \text{if} \quad w \prec w_f = w_g \text{ and } a_f > a_g \end{cases}
\end{aligned}$$

where $\wedge, \vee, \otimes$ are extended to $\star$ such that $\star \wedge a = \star$, $\star \vee a = a$ and $\star \otimes a = \star$ for all $a \in \mathcal{A}_w \cup \{\star\}$.

Proof. The proof is similar to its counterpart in [16]. Let f, g be sloping functions. Let us consider each case:

- $f \wedge g$. When there are $\top$-constant functions, the proof is easy. Otherwise,

$$(w_{f\wedge g}, a_{f\wedge g}) := \begin{cases} (w_f, a_f) & \text{if} \quad w_f \succ w_g \\ (w_f, a_f \wedge a_g) & \text{if} \quad w_f = w_g \\ (w_g, a_g) & \text{if} \quad w_f \prec w_g \end{cases}$$

- $f \vee g$. When there are $\top$-constant functions, the proof is easy. Otherwise,

$$(w_{f\vee g}, a_{f\vee g}) := \begin{cases} (w_f, a_f) & \text{if} \quad w_f \prec w_g \\ (w_f, a_f \vee a_g) & \text{if} \quad w_f = w_g \\ (w_g, a_g) & \text{if} \quad w_f \succ w_g \end{cases}$$

- $f \otimes g$. When there are $\top$-constant functions, the proof is easy. Otherwise,

$$(w_{f\otimes g}, a_{f\otimes g}) := \begin{cases} (w_f, a_f) & \text{if} \quad w_f \succ w_g \\ (w_f, a_f \otimes a_g) & \text{if} \quad w_f = w_g \\ (w_g, a_g) & \text{if} \quad w_f \prec w_g \end{cases}$$

- $f \to g$. When f or g is the $\top$-constant function, it is obvious. Similarly for the cases where $w_f \neq w_g$ or ($w_f = w_g$ and $a_f \leq a_g$). When $w_f = w_g$ and $a_f > a_g$, then $w_{f\to g} = w_f = w_g$ and $a_{f\to g} = a_f \to_{\mathcal{A}_w} a_g$ (it is easy to see that $a_f \to_{\mathcal{A}_w} a_g < \top$).

□

Definition 4.3 *Let $\mathcal{W} = \langle W, \succeq\rangle$ be a linear order with a greatest element w_G, and let $\{\mathcal{B}_w\}_{w\in W}$ be a W-indexed family of structures as described in Definition 4.1 such that for the greatest element w_G of W, $\mathcal{A}_{w_G}$ has a bottom element. A* Zuluaga-Castiglioni-structure *for $\{\mathcal{A}_w\}_{w\in W}$ (or* **ZC***-structure) is a pair $\mathcal{M} = \langle \mathcal{W}, \Vdash^{\mathrm{ZC}}\rangle$ where $\Vdash^{\mathrm{ZC}}$ is an infix operator (on worlds and propositional variables) taking values in $\bigcup_{w\in W} \mathcal{B}_w$, i.e. $(w \Vdash^{\mathrm{ZC}} p) \in \mathcal{B}_w$, such that for any propositional variable p the function $\lambda w.(w \Vdash^{\mathrm{ZC}} p)\colon W \to \bigcup_{w\in W} \mathcal{B}_w$ is a sloping function.*

To see that what we have defined just above (and soon to follow underneath) is indeed a Kripke semantics, we note that by restricting to the Booleans $\{\top, \bot\}$ in the above definition, one obtains a model of Gödel-Dummett logic (which logic is obtained by adjoining the structural rule of contraction to the **MTL**). The semantics introduced here therefore simply adds intermediate values to that well-known structure.

Definition 4.4 (ZC Kripke Semantics for $\mathcal{L}_\otimes$) *Given a* **ZC***-structure*

$$\mathcal{M} = \langle \mathcal{W}, \Vdash^{\mathrm{ZC}}\rangle$$

the valuation function $w \Vdash^{\mathrm{ZC}} p$ on propositional variables p can be extended to

all $\mathcal{L}_\otimes$-formulas as:

$$
\begin{aligned}
w \Vdash^{\mathrm{ZC}} \bot &:= \bot \\
w \Vdash^{\mathrm{ZC}} \phi \wedge \psi &:= (w \Vdash^{\mathrm{ZC}} \phi) \wedge_{\mathcal{A}_w \cup \{\star\}} (w \Vdash^{\mathrm{ZC}} \psi) \\
w \Vdash^{\mathrm{ZC}} \phi \vee \psi &:= (w \Vdash^{\mathrm{ZC}} \phi) \vee_{\mathcal{A}_w \cup \{\star\}} (w \Vdash^{\mathrm{ZC}} \psi) \\
w \Vdash^{\mathrm{ZC}} \phi \otimes \psi &:= (w \Vdash^{\mathrm{ZC}} \phi) \otimes_{\mathcal{A}_w \cup \{\star\}} (w \Vdash^{\mathrm{ZC}} \psi) \\
w \Vdash^{\mathrm{ZC}} \phi \to \psi &:= \begin{cases}
\top & \text{if} \quad w_\phi \succ w_\psi \text{ or } (w_f = w_g \text{ and } a_f \leq a_g) \text{ or} \\
 & \lambda w.(w \Vdash^{\mathrm{ZC}} \psi) \text{ is the } \top\text{-constant function} \\
w \Vdash^{\mathrm{ZC}} \psi & \text{if} \quad w_\phi \prec w_\psi \text{ or} \\
 & \lambda w.(w \Vdash^{\mathrm{ZC}} \phi) \text{ is the } \top\text{-constant function} \\
\top & \text{if} \quad w \succ w_\phi = w_\psi \text{ and } a_f > a_g \\
(w \Vdash^{\mathrm{ZC}} \phi) \to_{\mathcal{A}_w} (w \Vdash^{\mathrm{ZC}} \psi) & \text{if} \quad w = w_\phi = w_\psi \text{ and } a_f > a_g \\
\star & \text{if} \quad w \prec w_\phi = w_\psi \text{ and } a_f > a_g
\end{cases}
\end{aligned}
$$

We can show that each formula induces a sloping function:

Lemma 4.5 *For any formula ϕ the function $\lambda w.(w \Vdash^{\mathrm{ZC}} \phi)\colon W \to \bigcup_{w \in W}(\mathcal{A}_w \cup \{\star\})$ is a sloping function.*

Proof. By induction on ϕ. The cases for $\psi \vee \xi, \psi \wedge \xi, \psi \otimes \xi$ and $\psi \to \xi$ follow directly from Definition 4.4 and Lemma 4.2. □

We can show that the sloping functions are linearly ordered in **ZC**-structures:

Lemma 4.6 *Let $f, g : W \to \bigcup_{w \in W}(\mathcal{A}_w \cup \{\bot\})$ be sloping for* **MTL**. *Then:*

$$\forall v : (f(v) \geq g(v)) \vee \forall v : (g(v) \geq f(v)).$$

Proof. It suffices to see that $\forall v : (f(v) \geq g(v))$ iff f is the $\top$-constant function or $w_f \prec w_g$ or ($w_f = w_g$ and $a_f \geq a_g$). □

We have the following monotonicity property for the semantics:

Corollary 4.7 (Monotonicity) *The following (generalised) monotonicity property holds for all $\mathcal{L}_\otimes$-formulas ϕ, i.e.*

$$\text{if } w \preceq v \text{ then } (w \Vdash^{\mathrm{ZC}} \phi) \leq (v \Vdash^{\mathrm{ZC}} \phi).$$

Proof. This follows from the observation that the valuations are sloping functions, which are in turn monotone functions. □

5 ZC-structures and Ordinal Sums of Semihoops

In this section, we show that the algebra of sloping functions of a given **ZC**-structure is isomorphic to an ordinal sum of semihoops, which will be used in the soundness and completeness proof.

Lemma 5.1 *Given a* **ZC***-structure*

$$\mathcal{M} = \langle \mathcal{W}, \Vdash^{\mathrm{ZC}} \rangle$$

and the sloping functions $\mathsf{Slop}(\mathcal{M}) := \{f \mid f\colon W \to \bigcup_{w\in W} \mathcal{B}_w \text{ is sloping}\}$ *of* $\mathcal{M}$*, there is an isomorphism from* $\mathsf{Slop}(\mathcal{M})$ *to* $C_I := \{(w,a) : w \in W, a \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}\} \cup \{(\top_I, \top)\}$ *such that the total order on* C_I *is defined as the following order:*

$$(w,a) \leq (v,b) \quad \textit{iff} \quad (v,b) = (\top_I, \top) \quad \textit{or} \quad w \succ v \quad \textit{or} \quad (w = v \quad \textit{and} \quad a \leq_{\mathcal{A}_w} b)$$

We define on C_I *the following operations:*

$$(i,a) \wedge (j,b) = min\{(i,a),(j,b)\}$$

$$(i,a) \vee (j,b) = max\{(i,a),(j,b)\}$$

$$(i,a) \otimes (j,b) = \begin{cases} (i,a) & \textit{if} \quad i < j \\ (i, a \otimes b) & \textit{if} \quad i = j \\ (j,b) & \textit{if} \quad i > j \end{cases}$$

$$(i,a) \to (j,b) = \begin{cases} (\top_I, \top) & \textit{if} \quad (i,a) \leq (j,b) \\ (i, a \to b) & \textit{if} \quad i = j \quad \textit{and} \quad a > b \\ (j,b) & \textit{if} \quad i > j \end{cases}$$

The operations $\wedge, \vee, \otimes, \to$ *in* $\mathsf{Slop}(\mathcal{M}) := \{f\colon W \to \bigcup_{w\in W}(\mathcal{A}_w \cup \{\star\})$ *are defined as in Lemma 4.2.*

Proof. We map each $\top$-constant function to $(\top_I, \top)$ and other sloping functions to (w_f, a_f). We denote this map as h. It is easy to see that h is well-defined and is a bijection. To show that h is an isomorphism, it suffices to show that h preserves operations, which is shown as follows:

- For the case of $\wedge$, let us consider $f, g \in \mathsf{Slop}(\mathcal{M})$.
  - If both f and g are $\top$-constant functions, then $h(f) = h(g) = (\top_I, \top)$, and so $f \wedge g = g$ and $(\top_I, \top) \wedge (\top_I, \top) = (\top_I, \top \wedge \top) = (\top_I, \top)$. Therefore, $h(f \wedge g) = h(g) = (\top_I, \top) = (\top_I, \top) \wedge (\top_I, \top) = h(f) \wedge h(g)$.
  - If f is the $\top$-constant function and g is not, then $f \wedge g = g$, $h(f) = (\top_I, \top)$ and $h(g) = (w,a)$ for some $w \in W$ and $a \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$. Since $(w,a) < (\top_I, \top)$, we have that $(\top_I, \top) \wedge (w,a) = (w,a)$. Then $h(f \textit{ and } \wedge g) = h(g) = (w,a) = (\top_I, \top) \wedge (w,a) = h(f) \wedge h(g)$.
  - If g is the $\top$-constant function and f is not, the proof is similar.
  - If neither f nor g are $\top$-constant functions, then there are $w_1, w_2 \in W$ and $a_1, a_2 \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$ such that $h(f) = (w_1, a_1)$ and $h(g) = (w_2, a_2)$. Then $h(f) \wedge h(g) = (w_1, a_1) \wedge (w_2, a_2)$.

If $w_1 = w_2$, then $h(f) \wedge h(g) = (w_1, a_1 \wedge a_2)$, and

$$(f \wedge g)(w) := \begin{cases} \star & \text{if } w \prec w_1 \\ a_1 \wedge a_2 & \text{if } w = w_1 \\ \top & \text{if } w \succ w_1 \end{cases}$$

so $h(f \wedge g) = (w_1, a_1 \wedge a_2) = h(f) \wedge h(g)$.

If $w_1 \neq w_2$, without loss of generality we assume that $w_1 \prec w_2$, then $h(f) \wedge h(g) = (w_1, a_1) \wedge (w_2, a_2) = (w_2, a_2)$, and

$$(f \wedge g)(w) := \begin{cases} \star \wedge \star = \star & \text{if } w \prec w_1 \\ a_1 \wedge \star = \star & \text{if } w = w_1 \\ \top \wedge \star = \star & \text{if } w_1 \prec w \prec w_2 \\ \top \wedge a_2 = a_2 & \text{if } w = w_2 \\ \top \wedge \top = \top & \text{if } w \succ w_2 \end{cases}$$

therefore $h(f \wedge g) = (w_2, a_2) = h(f) \wedge h(g)$.

Therefore in both cases we have $h(f \wedge g) = h(f) \wedge h(g)$.

- The case of $\vee$ is similar to the case of $\wedge$.

- For the case of $\otimes$, let us consider $f, g \in \mathsf{Slop}(\mathcal{M})$.
  - If both f and g are $\top$-constant functions, then $h(f) = h(g) = (\top_I, \top)$, and it is easy to see that $f \otimes g = g$ and $(\top_I, \top) \otimes (\top_I, \top) = (\top_I, \top \otimes \top) = (\top_I, \top)$. Therefore, $h(f \otimes g) = h(g) = (\top_I, \top) = (\top_I, \top) \otimes (\top_I, \top) = h(f) \otimes h(g)$.
  - If f is the $\top$-constant function and g is not, then $f \otimes g = g$, $h(f) = (\top_I, \top)$ and $h(g) = (w, a)$ for some $w \in W$ and $a \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$. Since $(w, a) < (\top_I, \top)$, we have that $(\top_I, \top) \otimes (w, a) = (w, a)$. Then $h(f \otimes g) = h(g) = (w, a) = (\top_I, \top) \otimes (w, a) = h(f) \otimes h(g)$.
  - If g is the $\top$-constant function and f is not, the proof is similar.
  - If neither f nor g are $\top$-constant functions, then there are $w_1, w_2 \in W$ and $a_1, a_2 \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$ such that $h(f) = (w_1, a_1)$ and $h(g) = (w_2, a_2)$. Then $h(f) \otimes h(g) = (w_1, a_1) \otimes (w_2, a_2)$.

    If $w_1 = w_2$, then $h(f) \otimes h(g) = (w_1, a_1 \otimes a_2)$, and

$$(f \otimes g)(w) := \begin{cases} \star & \text{if } w \prec w_1 \\ a_1 \otimes a_2 & \text{if } w = w_1 \\ \top & \text{if } w \succ w_1 \end{cases}$$

    so $h(f \otimes g) = (w_1, a_1 \otimes a_2) = h(f) \otimes h(g)$.

    If $w_1 \neq w_2$, without loss of generality we assume that $w_1 \prec w_2$, then $h(f) \otimes h(g) = (w_1, a_1) \otimes (w_2, a_2) = (w_2, a_2)$, and

$$(f \otimes g)(w) := \begin{cases} \star \otimes \star = \star & \text{if } w \prec w_1 \\ a_1 \otimes \star = \star & \text{if } w = w_1 \\ \top \otimes \star = \star & \text{if } w_1 \prec w \prec w_2 \\ \top \otimes a_2 = a_2 & \text{if } w = w_2 \\ \top \otimes \top = \top & \text{if } w \succ w_2 \end{cases}$$

therefore $h(f \otimes g) = (w_2, a_2) = h(f) \otimes h(g)$.

Therefore in both cases we have $h(f \otimes g) = h(f) \otimes h(g)$.

- For the case of $\to$, let us consider $f, g \in \mathsf{Slop}(\mathcal{M})$.
  - If both f and g are $\top$-constant functions, then $f \to g$ is also the $\top$-constant function, so $h(f \to g) = h(f) = h(g) = (\top_I, \top)$. Since $(\top_I, \top) \to (\top_I, \top) = (\top_I, \top)$, we have $h(f \to g) = (\top_I, \top) = (\top_I, \top) \to (\top_I, \top) = h(f) \to h(g)$.
  - If f is the $\top$-constant function and g is not, then $f \to g = g$, $h(f) = (\top_I, \top)$ and $h(g) = (w, a)$ for some $w \in W$ and $a \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$. Since $(w, a) < (\top_I, \top)$, we have that $(\top_I, \top) \to (w, a) = (w, a)$. Then $h(f \to g) = h(g) = (w, a) = (\top_I, \top) \to (w, a) = h(f) \to h(g)$.
  - If g is the $\top$-constant function and f is not, then $f \to g = g$ is the $\top$-constant function, $h(g) = (\top_I, \top)$ and $h(f) = (w, a)$ for some $w \in W$ and $a \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$. Since $(w, a) < (\top_I, \top)$, we have that $(w, a) \to (\top_I, \top) = (\top_I, \top)$. Then $h(f \to g) = h(g) = (\top_I, \top) = (w, a) \to (\top_I, \top) = h(f) \to h(g)$.
  - If neither f nor g are $\top$-constant functions, then there are $w_1, w_2 \in W$ and $a_1, a_2 \in \mathcal{A}_w - \{\top_{\mathcal{A}_w}\}$ such that $h(f) = (w_1, a_1)$ and $h(g) = (w_2, a_2)$. Then $h(f) \to h(g) = (w_1, a_1) \to (w_2, a_2)$.

    If $w_1 = w_2$ and $a_1 \leq a_2$, then $h(f) \to h(g) = (w_1, a_1) \to (w_2, a_2) = (\top_I, \top)$, and $f \to g$ is the $\top$-constant function, so $h(f \to g) = (\top_I, \top) = (w_1, a_1) \to (w_2, a_2) = h(f) \to h(g)$.

    If $w_1 = w_2$ and $a_1 > a_2$, then $h(f) \to h(g) = (w_1, a_1) \to (w_2, a_2) = (w_1, a_1 \to a_2)$, and

$$(f \to g)(w) := \begin{cases} \bot & \text{if } w \prec w_1 \\ a_1 \to a_2 & \text{if } w = w_1 \\ \top & \text{if } w \succ w_1 \end{cases}$$

    so $h(f \to g) = (w_1, a_1 \to a_2) = h(f) \to (g)$.

    If $w_1 \prec w_2$, then $h(f) \to h(g) = (w_1, a_1) \to (w_2, a_2) = (w_2, a_2)$, and $f \to g = g$, therefore $h(f \to g) = h(g) = (w_2, a_2) = (w_1, a_1) \to (w_2, a_2) = h(f) \to h(g)$.

    If $w_1 \succ w_2$, then $h(f) \to h(g) = (w_1, a_1) \to (w_2, a_2) = (\top_I, \top)$, and $f \to g$ is the $\top$-constant function, so $h(f \to g) = (\top_I, \top) = (w_1, a_1) \to (w_2, a_2) = h(f) \to h(g)$.

    Therefore in all four cases we have $h(f \to g) = h(f) \to h(g)$.

□

6 Soundness and Completeness

In this section, we prove the soundness and comepleteness of the Kripke semantics for **MTL**.

Theorem 6.1 (Soundness) *If* $\vdash_{\mathbf{MTL}} \phi$ *then* $\Vdash^{\mathrm{ZC}} \phi$.

Proof. Suppose $\vdash \phi$ holds in **MTL**. By the algebraic completeness result for **MTL** algebras with respect to the Hilbert-style proof system $\mathbf{MTL_H}$ (see Proposition 3.6), it follows that for all **MTL**-algebras $\mathcal{G}$ and all mappings $h\colon Atom \to \mathcal{G}$ from atomic formulas to elements of $\mathcal{G}$, we have that $[\![\phi]\!]^{\mathcal{G}}_h = \top$.

Now for any **ZC**-structure $\mathcal{M}$, by Lemma 5.1, $\mathsf{Slop}(\mathcal{M})$ is isomorphic to some C_I. By [4, Remark 1], C_I is an **MTL**-chain, therefore we have $C_I \models \phi$, so $\mathcal{M} \Vdash^{\mathrm{ZC}} \phi$. □

Theorem 6.2 (Completeness) *If* $\Vdash^{\mathrm{ZC}} \phi$*, then* $\vdash_{\mathbf{MTL}} \phi$.

Proof. Suppose $\nvdash_{\mathbf{MTL}} \phi$. By the algebraic completeness result for **MTL**-algebras with respect to the Hilbert-style proof system $\mathbf{MTL_H}$ (see Proposition 3.6), it follows that for some **MTL**-algebra $\mathcal{G}$ and some mapping $h\colon Atom \to \mathcal{G}$ from atomic formulas to elements of $\mathcal{G}$, we have that $[\![\phi]\!]^{\mathcal{G}}_h \neq \top$. By [4, Remark 1], every **MTL**-algebra is isomorphic to an ordinal sum of semihoops, therefore we can take $\mathcal{G}$ to be some C_I as defined in Lemma 5.1. By Lemma 5.1, this C_I is isomorphic to the sloping function algebra $\mathsf{Slop}(\mathcal{M})$ of a **ZC**-structure $\mathcal{M}$ where the valuation on $\mathcal{M}$ agrees with h so that $\mathcal{M} \nVdash^{\mathrm{ZC}} \phi$. □

7 Conclusion

In the preceding, we extracted a relational semantics from the ordinal sum representation of [4] for **MTL**. By introducing generalisations of Kripke semantics adequate for **MTL** (and neighbours, see [8,15,16]), we seek a new perspective on fuzzy logics such as **MTL** as constructive or semi-constructive systems.

Relational semantics typically bring connections to proof theory, decidability and model theory. We hope in the future to exploit the semantics developed here to build semantically motivated calculi for **MTL** and similar systems. We also believe our work suggests analogies with traditional modal model theory, which we explore in forthcoming work. Furthermore, via the relational semantics developed here and in [8,15,16,17,19], one can analogise the classic correspondence theory of Intuitionistic logic (see for instance [18]).

From a proof-theoretic perspective, there is still work to be done. The present paper, together with [2] and works such as [9,10,20,22] suggest many potential pathways forward. One can build labelled calculi in the style of Gabbay [9] or Negri [20,21] over hypersequent presentations [2], or one can

utilise the notion of unsatisfiability in our relational semantics to devise a proof system, as is done in classical modal or intuitionistic logic (see [23]) or even classical Łukasiewicz logic [22]. Finally, we have the multi-type calculi of e.g. [10]. These have been successful in cases the algebraic system makes for a difficult translation into analytic rules.

References

[1] Arthan, R. and P. Oliva, *On affine logic and łukasiewicz logic* (2014).
URL `https://arxiv.org/abs/1404.0570`

[2] Baaz, M., A. Ciabattoni and F. Montagna, *Analytic calculi for monoidal t-norm based logic*, Fundam. Inform. **59** (2004), pp. 315–332.

[3] Bova, S. and F. Montagna, *The consequence relation in the logic of commutative GBL-algebras is PSPACE-complete*, Theoretical Computer Science **410** (2009), pp. 1143 – 1158.
URL `http://www.sciencedirect.com/science/article/pii/S0304397508007858`

[4] Castiglioni, J. L. and W. J. Z. Botero, *On finite MTL-algebras that are representable as poset products of Archimedean chains*, Fuzzy Sets and Systems **382** (2020), pp. 57–78.

[5] Ciabattoni, A., N. Galatos and K. Terui, *Algebraic proof theory: Hypersequents and hypercompletions*, Annals of Pure and Applied Logic **168** (2017), pp. 693 – 737.
URL `http://www.sciencedirect.com/science/article/pii/S016800721630135X`

[6] Ciabattoni, A. and F. Montagna, *Proof theory for locally finite many-valued logics: Semi-projective logics*, Theoretical Computer Science **480** (2013), pp. 26 – 42.
URL `http://www.sciencedirect.com/science/article/pii/S0304397513001126`

[7] Esteva, F. and L. Godo, *Monoidal t-norm based logic: towards a logic for left-continuous t-norms*, Fuzzy sets and systems **124** (2001), pp. 271–288.

[8] Fussner, W., *Poset products as relational models*, Studia Logica **110** (2021).

[9] Gabbay, D. M., "Labelled Deductive Systems," Oxford University Press, 1996.

[10] Greco, G., F. Liang, M. A. Moshier and A. Palmigiano, *Semi de Morgan logic properly displayed*, Studia Logica **109** (2021).

[11] Jenei, S. and F. Montagna, *A proof of standard completeness for Esteva and Godo's logic MTL*, Studia logica **70** (2002), pp. 183–192.

[12] Jipsen, P. and F. Montagna, *On the structure of generalized BL-algebras*, algebra universalis **55** (2006), pp. 227–238.
URL `https://doi.org/10.1007/s00012-006-1960-6`

[13] Jipsen, P. and F. Montagna, *The blok–ferreirim theorem for normal GBL-algebras and its application*, Algebra universalis **60** (2009), pp. 381–404.
URL `https://doi.org/10.1007/s00012-009-2106-4`

[14] Jipsen, P. and F. Montagna, *Embedding theorems for classes of GBL-algebras*, Journal of Pure and Applied Algebra **214** (2010), pp. 1559 – 1575.
URL `http://www.sciencedirect.com/science/article/pii/S0022404909002746`

[15] Lewis-Smith, A., *A Kripke Semantics for Hajek's BL*, arXiv e-prints (2023), pp. arXiv–2308.

[16] Lewis-Smith, A., P. Oliva and E. Robinson, *Kripke semantics for intuitionistic Łukasiewicz logic*, Studia Logica **109** (2021), pp. 313–339.

[17] Lewis-Smith, A. and Z. Zhao, *A Kripke Semantics for Intuitionistic Łukasiewicz Logic with Weak Excluded Middle*, Submitted (2025).

[18] Lewis-Smith, A. and Z. Zhao, *Correspondence Theory for Intuitionistic Łukasiewicz Logic*, Submitted (2025).

[19] Lewis-Smith, A. and Z. Zhao, *A Kripke Semantics for Monadic BL Chains*, Submitted (2025).

[20] Negri, S., *Proof analysis in modal logic*, Journal of Philosophical Logic **34** (2005), pp. 507–544.

[21] Negri, S. and J. V. Plato, "Proof Analysis: A Contribution to Hilbert's Last Problem," Cambridge University Press, 2011.
[22] Olivetti, N., *Tableaux for Łukasiewicz infinite-valued logic*, Studia Logica **73** (2003), pp. 81–111.
URL `https://doi.org/10.1023/A:1022989323091`
[23] Priest, G., "An Introduction to Non-Classical Logic: From If to Is," Cambridge University Press, 2008 .

Loop Formulas for Two-valued Logic Programs

Yuxin Sun [1]

Institute of Logic and Cognition, Department of Philosophy
Sun Yat-sen University, P.R.China 510275

Yuping Shen [2]

Institute of Logic and Cognition, Department of Philosophy
Sun Yat-sen University, P.R.China 510275

Abstract

Loop formulas are a bridge connecting nonmonotonic and classic logics in terms of logical equivalence. It has been discovered in the literature that a number of non-monotonic logics can be equivalently transformed into classic logic via the so-called "Completion + Loop Formula" pattern. In this paper, we introduce loop formulas for *two-valued* logic programs and show that the above pattern also applies to the translation from two-valued logic programs into propositional formulas. Our main theorem is a crucial step in understanding these programs and facilitating further practical and theoretical developments.

Keywords: Loop Formula, Logic Program, Propositional Formula, Equivalence

1 introduction

In the past decades, non-monotonic logics (NML) have been widely proposed for knowledge representation and reasoning in AI [1,2,7,23]. Particularly, *equivalent transformations* between non-monotonic and classical logics have been extensively investigated for practical and theoretical purposes. The so-called *loop formulas* [18,6] turn out to be a bridge connecting non-monotonic and classical logics in terms of logical equivalence, and play a crucial role in translation-based KR implementations [18,9] and in analyzing the abilities of boolean computational models [17,8,20].

More precisely, an interesting "NML = Completion + Loop Formulas" transformation pattern has been discovered in the literature. Intuitively, *completion* [4] simply considers non-monotonic formulas as classical ones, and loop formulas encode the dependency relations for non-monotonic reasoning. For example, Lin and Zhao [18] first showed that a propositional *normal logic program* [19,21,16] (as a non-monotonic logic) can be equivalently translated into

[1] Email: sunyx39@mail2.sysu.edu.cn

[2] Email: shyping@mail.sysu.edu.cn. Supported by NSSFC grant No. 25BZX071.

a set of classic clauses corresponding to the rules of the program, plus a set of loop formulas describing the head dependency relations. Following this research line, loop formulas for disjunctive logic programs [11], causal theories[10], circumscription [12], nested logic programs [6], description logic programs [22] and first order logic programs [3,14,13] are extensively studied.

In this work, we prove that the "Completion + Loop Formulas" pattern also applies to the so-called *two-valued logic programs* [15], which is essentially a nonmonotonic propositional logic built up with answer set semantics [19,16] and causal semantics [9,10]. As pointed out in [15,20], two-valued programs appear to be richer than their relatives in terms of computational power [5,20], yet still share the *same* expressiveness [3] and complexity with classic propositional logic. It should be noted that equivalent transformations do *not* introduce auxiliary variables. Therefore, even if a non-monotonic logic is as expressive as classic propositional logic, it may behave quite differently in describing some given boolean functions (e.g., exponentially more succinct, or vice versa). In this sense, discovering loop formulas for two-valued programs is a crucial step in understanding their essential capabilities and facilitating further practical and theoretical developments.

In the rest, we present the syntax and semantics for two-valued programs in Section 2, define completion, dependency graph and loop formulas in Section 3, prove the main theorem in Section 4 and conclude the paper in Section 5.

2 Two-valued Logic Programs

We follow the definitions and basic concepts of two-valued logic programs introduced by Lifschitz [15]. A *signature* σ is a set of atoms. A *literal* is an atom $a \in \sigma$ or its negation $\neg a$. A *formula* F (over σ) is defined as classic propositional logic.

A (two-valued logic) *program* is a finite set of rules of the form:

$$l_0 \leftarrow l_1, \ldots, l_k : F \tag{1}$$

and constraints of the form:

$$\bot \leftarrow : F \tag{2}$$

in which the *head* l_0 and the *premises* $l_1, \ldots l_k$ $(k \geq 0)$ are literals, and the *justification* F is a formula.

The rule (1) means "derive l_0 from $l_1, \ldots, l_k$ if l_0 is consistent with the justification F". The rule (2) acts as a constraint that rejects the case where F holds. By convention, a rule in form (1) can be denoted by $l \leftarrow G : F$ where $l = l_0$ and $G = \{l_1, \ldots, l_k\}$. If F in (1) is $\top$ then we may drop the colon and F as $l \leftarrow G$. If in addition $k = 0$ then the rule is a *fact* $l \leftarrow$ and we may also drop $\leftarrow$.

An *interpretation* I is an assignment from σ to $\{0, 1\}$. An interpretation I can be represented by a subset of σ or a *complete* and *consistent* set of literals

[3] In the sense that both can represent an arbitrary boolean function, see Propositiona 2.1.

over σ. By the term complete we mean for each $a \in \sigma$, either $a \in I$ or $\neg a \in I$. By consistent we mean I does not contain a and $\neg a$ at the same time for any $a \in \sigma$. In this paper, we consider I as a consistent and complete set of literals.

The *reduct* Π^I of a program Π with respect to an interpretation I is the set of rules

$$l_0 \leftarrow l_1, \ldots, l_k \tag{3}$$

corresponding to the rules (1) of Π with I *satisfies* F (i.e. $I \models F$) in the usual sense.

Clearly for any Π, the reduct Π^I is a set of rules without any justification, making it monotonic. We say that I is a *model* of Π, if the *smallest* set of literals closed under all the rules (3) equals I. Here In other words, let $\alpha(\Pi^I)$ be the smallest set of literals closed under the reduct Π^I, I is a model of Π if $I = \alpha(\Pi^I)$.

Consider the program $\Pi = \{a \leftarrow \neg b : a, \neg b \leftarrow\}$ over $\sigma = \{a, b\}$. Let $I = \{a, \neg b\}$, the reduct Π^I is:

$$a \leftarrow \neg b,$$
$$\neg b \leftarrow .$$

The smallest set of literals closed under Π^I is $\{a, \neg b\}$, which equals to I, hence $I = \{a, \neg b\}$ is a model of Π. Now consider $\{\neg a, \neg b\}$, the associated reduct is simply $\{\neg b \leftarrow\}$, whose minimal closure is $\{\neg b\}$. Clearly $\{\neg a, \neg b\}$ is not a model of Π. In fact, $\{a, \neg b\}$ is the unique model of Π. Observe that σ plays a crucial role in giving models. If Π was built on $\sigma' = \{a, b, c\}$, then it has no model at all, since Π provides no rules for deriving c or $\neg c$.

In contrast, the program $\{a \leftarrow: a, \neg a \leftarrow: \neg a, b \leftarrow: b, \neg b \leftarrow: \neg b\}$ over σ has four models $\{a, b\}, \{a, \neg b\}, \{\neg a, b\}, \{\neg a, \neg b\}$. Not hard to see, by adding suitable constraints of the form $\bot \leftarrow: l_1 \wedge l_2$ with $l_1 \in \{a, \neg a\}$ and $l_2 \in \{b, \neg b\}$, we can construct a program that possesses arbitrary models over σ. Based on these observations, it is easy to have the following results.

Proposition 2.1 (Expressiveness) *Two-valued logic programs are as expressive as classic propositional logic.*

Proof. Fix a signature σ. ($\Rightarrow$) Let F be a propositional formula built on σ. Then $\{a \leftarrow: a, \neg a \leftarrow: \neg a | a \in \sigma\} \cup \{\bot \leftarrow: \neg F\}$ is a program equivalent to F. ($\Leftarrow$) Let Π be a program over σ with models $I_1, \ldots, I_m$. Then $\bigvee_{1 \leq k \leq m}(\bigwedge_{l \in I_k} l)$ is a propositional formula equivalent to Π. □

Proposition 2.2 (Complexity) *Deciding whether a two-valued logic program has a model is NP-complete.*

Proof. For NP membership, simply note that checking whether a given interpretation is a model of a program can be done in polynomial time, since computing the reduct and its smallest closure are both easy [21]. The NP hardness can be established by the trivial translation from a formula to a program in the Proof of Proposition 2.1. □

3 Completion and Loop Formulas

Based on Clark's Completion [4], we define completion for two-valued logic programs as follows. Fix a signature σ. We denote by $Lit(\sigma)$ the set of all literals built on σ. The *completion* $Comp(\Pi)$ of a program Π over σ, is the set of the following propositional formulas:

- For each $l \in Lit(\sigma)$, let $l \leftarrow G_1 : F_1, \ldots, l \leftarrow G_n : F_n$ be all rules about l in Π, then $l \equiv (G_1 \wedge F_1) \vee \ldots \vee (G_n \wedge F_n)$ is in $Comp(\Pi)$. In particular, if $n = 0$, then the equivalence is $l \equiv \bot$, which is equivalent to $\neg l$. Here we slightly abuse G_i as the conjunction of all literals in it.
- If $\bot \leftarrow : F$ is a constraint in Π, then $\neg F$ is in $Comp(\Pi)$.

Proposition 3.1 *Let Π be a program. If an associated interpretation I is a model of Π then I is also a model of $Comp(\Pi)$.*

However, the converse of Proposition 3.1 is not true in general. Consider a program $\Pi' = \{a \leftarrow b, b \leftarrow a\}$ over $\{a, b\}$. The model of its completion $\{a \equiv b, b \equiv a, \neg a \equiv \bot, \neg b \equiv \bot\}$ is simply $\{a, b\}$, while the program itself has no model at all.

In the following, we show how to strengthen the completion such that a set is a model of a program if and only if it is a model of the strengthened theory. The key concepts in this construction are *loops* and their associated formulas.

The *dependency graph* $\mathcal{G}_\Pi$ of a Π is a pair (V, E) in which the set of nodes V consists of all literals occurring in Π, and the set of edges E contains an edge (l, l') iff there is a rule (1) in Π s.t. $l = l_0$ and $l' \in \{l_1, \ldots, l_k\}$.

A *loop* S of Π is a nonempty subset of literals occurring in Π s.t. for any l, l' in S, there is a path from l to l' of length > 0. The literals in a loop are called *loop literals*.

Given a loop S in a program Π, we define two sets of rules:

$$R^+(S, \Pi) = \{l \leftarrow G : F \in \Pi | l \in S, (\exists l')(l' \in G \wedge l' \in S)\}$$
$$R^-(S, \Pi) = \{l \leftarrow G : F \in \Pi | l \in S, \neg(\exists l')(l' \in G \wedge l' \in S)\}$$

Intuitively, $R^+(S, \Pi)$ denotes the rules whose heads and premises depend on each other in S. On the other hand, $R^-(S, \Pi)$ denotes the rules whose premises come from outside S. When the context is clear we may also omit Π.

The *loop formula* of S under Π, denoted by $LF(S, \Pi)$, or simply $LF(S)$, is the following implication:

$$\neg[\bigvee_{l \leftarrow G : F \in R^-(S,\Pi)} (G \wedge F)] \supset \bigwedge_{l \in S} \neg l \tag{4}$$

Note that the above program Π' has a loop $\{a, b\}$ and $R^-(\{a, b\}, \Pi') = \emptyset$. So the associated loop formula $LF(\{a, b\}, \Pi')$ is $\top \supset \neg a \wedge \neg b$. Now the union of $Comp(\Pi')$ and $LF(\{a, b\}, \Pi')$ has no models, which is logically equivalent to Π'. In the next section, we show that this observation holds for translating an arbitrary program into propositional formulas.

4 Main Theorem

Theorem 4.1 (Main Theorem) *Let* Π *be a program,* $Comp(\Pi)$ *be its completion, and* LF *the set of all loop formulas of* Π. *Then an associated interpretation* I *is a model of* Π *iff* I *is a model of* $Comp(\Pi) \bigcup LF$.

Proof. We follow the proof idea in [18]. It is sufficient to show the theorem for programs without constraints, and it trivially holds with constraints.

($\Leftarrow$) Let A be an interpretation that satisfies $Comp(\Pi) \bigcup LF$. We show that $A = \alpha(\Pi^A)$. Let T_0 be the following set:

$$T_0 = \{l \leftarrow G \in \Pi^A \mid l \text{ and } G \text{ are true in } A, \text{i.e., } A \models l \wedge G\}.$$

Note that for every $l \in A$ there is a rule in T_0 with head l, since A is a model of $Comp(\Pi)$. We construct T_{i+1} from T_i as follows:

- If T_i contains no loop, then let $T_{i+1} = T_i$.
- If T_i contains at least a loop, let $S_i = \{l_1, \ldots, l_n\}$ be a maximal one, and $R^+(S_i, T_i)$ be the rules: $r_1 = l_1 \leftarrow G_1 : F_1, \ldots, r_m = l_n \leftarrow G_m : F_m$, where $n \leqslant m$. Clearly S_i must be a loop in Π as well. Now since $A \models l_1$ and A is a model of $LF(S_i, \Pi)$, by (4) there exists some $1 \leqslant k \leqslant n$, and a rule $l_k \leftarrow G : F$ in $R^-(S_i, \Pi)$ such that $A \models F \wedge G$. By the definition of $R^-(S_i, \Pi)$, $G \cap S_i = \emptyset$, so the reduction of this rule, $r = l_k \leftarrow G$, is not equal to any of $r_1, \ldots, r_m$. Now let T_{i+1} be the result of deleting all those rules in $r_1, \ldots, r_m$ whose head is l_k.

We can show that T_i has the following properties:

- For some n, $T_k = T_n$ for all $k > n$, and T_n does not have any loops. This is because T_0 has only a finite number of loops, and every T_i is a subset of T_0. Let T be this T_n, i.e., $T = \bigcap_{i=1,2,\ldots} T_i$.
- For any i, if there is a loop in T_i, then the rule $r = l_k \leftarrow G$ used above is in T_{i+1}. We only need to show that it is in T_i. Clearly $r \in T_0$. Suppose that $r \in T_j$ for some $j < i$, we show that $r \in T_{j+1}$, i.e., r cannot be deleted from T_j. Suppose otherwise, there is a maximal loop S in T_j for which $r \in R^+(S, T_j)$. Now let S' be the maximal loop in T_i used in the construction of T_{i+1}. Since l_k is a element in $S \cap S'$, and $S \cup S'$ is also a loop in T_j. Hence $S' \subseteq S$, as S is maximal. Now since r is deleted from T_j, by the construction of T_{i+1}, all rules in $R^+(S, T_j)$ with head l_k are also deleted. So there are no rules in $R^+(S', T_i)$ with head l_k, a contradiction.
- For each literal $l \in A$, there is a rule in T with head l. This holds for T_0 as A is a model of $Comp(\Pi)$. Assume this also holds for T_i, we prove it for T_{i+1}: It trivially holds if no loop is in T_i. So suppose T_i has a loop. Note that T_{i+1} is the result of removing some of the rules in T_i whose heads are identical to l_k, where the rule $r = l_k \leftarrow G$ is used in the construction. So for any literal $l \in A$ that is different from l_k, there is a rule for it in T_i by inductive assumption, and the same rule is also in T_{i+1} by construction.

Now we prove by contradiction that for every $l \in A$, $l \in \alpha(T)$, i.e., l is in the smallest closure under T. Recall that we have proved: (1) for each $l \in A$, there is a rule in T with head l; (2) for each rule in T, both the head and the body are true in A; (3) T has no loops. Assume that there exists $l_0 \in A$, s.t. $l_0 \notin \alpha(T)$. Inductively construct a set of sequences Z of literals in A as follows: First, let $Z_0 = [l_0]$. Clearly, Z_0 satisfies the following properties for Z: (i) Z is a sequence of distinct literals in A; (ii) No literal in Z is in $\alpha(T)$. (iii) If l_i and l_{i+1} are in Z, then l_i depends on l_{i+1} in the sense that there is a rule $l_i \leftarrow G$ in T such that $l_{i+1} \in G$.

Suppose we have a sequence $Z_k = [l_0, \ldots, l_k]$ with the above properties, construct Z_{k+1} as follows: Let $l_k \leftarrow G_k$ be a rule in T. Note that $l_k \notin \alpha(T)$, $G_k \neq \emptyset$, and $G_k \not\subseteq \alpha(T)$. Thus there must exist $l_{k+1} \in G_k$, s.t. $l_{k+1} \notin T$. Since T is loop free, and each literal in Z_k depends on the next one, so for any $0 \leqslant i \leqslant k$, $l_{k+1} \neq l_i$. Now let $Z_{k+1} = [l_0, \ldots, l_k, l_{k+1}]$, clearly Z_{k+1} also satisfies the above properties about Z. But this is impossible as A is finite and in this way we could construct Z_k for arbitrary k.

Therefore for every $l \in A$, $l \in \alpha(\Pi^A)$ as $T \subseteq \Pi^A$. Now observe that $l \in \alpha(\Pi^A)$ implies $l \in A$ as A is a model of $Comp(\Pi)$. Consequently, if A is a model of $Comp(\Pi) \cup LF$, then A is a model of Π, i.e., $A = \alpha(\Pi^A)$.

($\Rightarrow$) Now suppose A is a model of Π, i.e., $A = \alpha(\Pi^A)$. Clearly A is a model of $Comp(\Pi)$. We show that A is also a model of LF. Let S be a loop in Π with associated loop formula $LF(S)$. Suppose $R^-(S)$ of Π contains the rules: $r_1 = l_1 \leftarrow G_1 : F_1, \ldots, r_n = l_n \leftarrow G_n : F_n$. If A does not satisfy $LF(S)$, then there must be a literal $l \in S$ such that $l \in A$, and for each rule r_i in $R^-(S)$, $A \not\models G_i \wedge F_i$. Now let the reduct of $R^-(S)$ w.r.t. A be the following rules:

$$r_{a_1} = l_{a_1} \leftarrow Ga_1, \ldots, r_{a_k} = l_{a_k} \leftarrow Ga_k,$$

$1 \leqslant a_i \leqslant n$, then for each $1 \leqslant i \leqslant k$, $A \not\models G_{a_i}$. Since $l \in A$, $l \in \alpha(\Pi^A)$. So there must be a sequence of rules in Π^A:

$$r''_1 = q_1 \leftarrow Q_1, \ldots, r''_u = q_u \leftarrow Q_u,$$

s.t. $q_u = l$, $Q_1 = \emptyset$ and for each $1 \leqslant i \leqslant u$, $Q_i \subseteq \{q_1, \ldots, q_{i-1}\}$. In this sequence, there must be a v, $1 \leqslant v \leqslant u$, s.t. $\{q_1, \ldots, q_{v-1}\} \cap S = \emptyset$ and $q_v \in S$. Since $Q_v \subseteq \{q_1, \ldots, q_{v-1}\}$, $Q_v \cap S = \emptyset$. So r''_v must be a reduct of the rule in $R^-(S)$, i.e. for some $1 \leqslant i \leqslant k$, $r''_v = r_{a_i}$. But this is a contradiction as A must satisfy the body of r''_v but not that of r_{a_i}. Consequently, if A is a model of Π, then A is a model of $Comp(\Pi) \cup LF$. □

5 Conclusion

In this paper, we introduce loop formulas for two-valued logic programs, and show that they can be equivalently translated into propositional formulas under the "Completion + Loop Formulas" pattern. In future work, we will investigate a SAT-based implementation for two-valued programs and compare their computational power w.r.t. a family of non-monotonic logics.

References

[1] Baral, C., "Knowledge Representation, Reasoning and Declarative Problem Solving," Cambridge University Press, USA, 2010, 1st edition.

[2] Brewka, G., I. Niemelä and M. Truszczyński, *Nonmonotonic reasoning*, in: F. van Harmelen, V. Lifschitz and B. Porter, editors, *Handbook of Knowledge Representation*, Foundations of Artificial Intelligence **3**, Elsevier, 2008 pp. 239–284.
URL `https://www.sciencedirect.com/science/article/pii/S1574652607030064`

[3] Chen, Y., F. Lin, Y. Wang and M. Zhang, *First-order loop formulas for normal logic programs*, in: *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'06 (2006), p. 298–307.

[4] Clark, K. L., *Negation as failure*, in: H. Gallaire and J. Minker, editors, *Logic and Data Bases* (1978), pp. 292–322.

[5] Clote, P. and E. Kranakis, "Boolean Functions and Computation Models," Springer Berlin, Heidelberg, Germany, 2002.

[6] Ferraris, P., J. Lee and V. Lifschitz, *A generalization of the lin-zhao theorem*, Annals of Mathematics and Artificial Intelligence **47** (2006), p. 79–101.
URL `https://doi.org/10.1007/s10472-006-9025-2`

[7] Gebser, M., R. Kaminski, B. Kaufmann and T. Schaub, "Answer Set Solving in Practice," Morgan & Claypool Publishers, Lexington,KY, USA, 2012.

[8] Gebser, M. and T. Schaub, *Loops: Relevant or redundant?*, in: C. Baral, G. Greco, N. Leone and G. Terracina, editors, *Logic Programming and Nonmonotonic Reasoning* (2005), pp. 53–65.

[9] Giunchiglia, E., J. Lee, V. Lifschitz, N. McCain and H. Turner, *Nonmonotonic causal theories*, Artificial Intelligence **153** (2004), pp. 49–104.
URL `https://www.sciencedirect.com/science/article/pii/S000437020300167X`

[10] Lee, J., *Nondefinite vs. definite causal theories*, in: V. Lifschitz and I. Niemelä, editors, *Logic Programming and Nonmonotonic Reasoning* (2004), pp. 141–153.

[11] Lee, J. and V. Lifschitz, *Loop formulas for disjunctive logic programs*, in: C. Palamidessi, editor, *Logic Programming* (2003), pp. 451–465.

[12] Lee, J. and F. Lin, *Loop formulas for circumscription*, Artificial Intelligence **170** (2006), pp. 160–185.
URL `https://www.sciencedirect.com/science/article/pii/S000437020500144X`

[13] Lee, J. and Y. Meng, *On loop formulas with variables*, in: *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, KR'08 (2008), p. 444–453.

[14] Lee, J. and Y. Meng, *First-order stable model semantics and first-order loop formulas*, J. Artif. Int. Res. **42** (2011), p. 125–180.

[15] Lifschitz, V., *Two-valued logic programs*, in: A. Dovier and V. Santos Costa, editors, *Technical Communications of the 28th International Conference on Logic Programming (ICLP'12)*, Leibniz International Proceedings in Informatics (LIPIcs) **17** (2012), pp. 259–266.

[16] Lifschitz, V., "Answer Set Programming," Springer, Switzerland, 2019.
URL `https://doi.org/10.1007/978-3-030-24658-7`

[17] Lifschitz, V. and A. Razborov, *Why are there so many loop formulas?*, ACM Transactions on Computational Logic **7** (2006), pp. 261–268.

[18] Lin, F. and Y. Zhao, *Assat: computing answer sets of a logic program by sat solvers*, Artif. Intell. **157** (2004), p. 115–137.

[19] Michael, G. and L. Vladimir, *The stable model semantics for logic programming*, in: *Proc. 5'h International Conference and Symposium on Logic Programming*, 1988, pp. 1070–1080.

[20] Shen, Y. and X. Zhao, *Computationally hard problems for logic programs under answer set semantics*, ACM Trans. Comput. Logic **25** (2024).

[21] Simons, P., I. Niemelä and T. Soininen, *Extending and implementing the stable model semantics.*, Artificial Intelligence **138** (2002), pp. 181–234.

[22] Wang, Y., J.-h. You, L. y. Yuan and Y.-d. Shen, *Loop formulas for description logic programs*, Theory Pract. Log. Program. **10** (2010), p. 531–545.
URL https://doi.org/10.1017/S1471068410000268

[23] Zhang, H., G. Jiang and D. Quan, *A theory of formalisms for representing knowledge*, Proceedings of the AAAI Conference on Artificial Intelligence **39** (2025), pp. 15257–15264.
URL https://ojs.aaai.org/index.php/AAAI/article/view/33674

Normative implications

Andrea De Domenico, Mattia Panettiere, Xiaolong Wang

Vrije Universiteit Amsterdam

Ali Farjami, Apostolos Tzimoulis

University of Luxembourg

Krishna Manoorkar

Institute of Computer Science of the Czech Academy of Sciences

Alessandra Palmigiano

Vrije Universiteit Amsterdam; Department of Mathematics and Applied Mathematics, University of Johannesburg

Abstract

We continue to develop a research line initiated in [6], studying I/O logic from an algebraic approach based on subordination algebras. We introduce the classes of slanted (co-)Heyting algebras, as equivalent presentations of distributive lattices with subordination relations. Interpreting subordination relations as the algebraic counterparts of input/output relations on formulas yields (slanted) modal operations with interesting deontic interpretations. We study the theory of slanted and co-slanted Heyting algebras, develop algorithmic correspondence and inverse correspondence, and present some deontically meaningful axiomatic extensions and examples.

Keywords: I/O logic, modal characterizations of normative conditions, subordination algebras, slanted Heyting algebras, (inverse) correspondence.

1 Introduction

Consider the conditional obligation $citizen \prec taxes$, which reads 'If you are a citizen then you must pay taxes'. This obligation holds under a set of *background assumptions* such as 'You earn more than a minimum threshold' ($earn$), which are often left implicit, and it is almost impossible to spell out entirely. This paper introduces a formal machinery which allows us to represent and make inferences with the sets of background assumptions of conditional norms.

This paper continues a line of investigation, initiated in [6], on the study of input/output logic [16] from an algebraic perspective based on *subordination algebras* [8]. This approach has allowed to uniformly extend input/output

logic to a large family of nonclassical logics [9], and, in this generalized context, to obtain modal characterizations of an infinite class of conditions on normative and permission systems as well as on their interaction [7]. Subordination algebras are tuples $(A, \prec)$ where A is an algebra (typically, a Boolean algebra), and $\prec \subseteq A \times A$ is a *subordination relation*, i.e. a binary relation endowed with the algebraic counterparts of the well known properties $(\top), (\bot), (\mathrm{SI}), (\mathrm{WO}), (\mathrm{AND}), (\mathrm{OR})$ of normative systems in input/output logic. These properties can be equivalently reformulated as the requirement that, for every $a \in A$, the sets $\prec[a] := \{b \in A \mid a \prec b\}$ and $\prec^{-1}[a] := \{b \in A \mid b \prec a\}$ be a filter and an ideal of A, respectively. Lindahl and Odelstad [15] have introduced structures closely related to subordination algebras in their algebraic treatment of normative systems. What distinguishes the explicit use of subordination algebras, however, is that it provides a systematic bridge between algebraic and topological perspectives, enabling general and uniform modal characterizations of classes of input/output logics.

Subordination algebras have been introduced independently from input/output logic, in the context of a research program in point-free topology, aimed at developing region-based theories of space. In this literature, subordination algebras are used as an 'umbrella' type of notion which crops up under several equivalent presentations. One of these presentations is the notion of *quasi-modal algebras* introduced by Celani [4] (see discussions in [8][10]). These are tuples (A, Δ) such that A is a Boolean algebra and Δ is a *quasi-modal operator*, i.e. a map such that $\Delta(a)$ is an ideal of A for any $a \in A$. Structures closely related to quasi modal algebras are *generalized implication lattices* [3], i.e. tuples $(A, \Rightarrow)$ such that A is a distributive lattice and $\Rightarrow$ is a *generalized implication*, i.e. a binary map such that $a \Rightarrow b$ is an ideal of A for all $a, b \in A$, and satisfying certain additional conditions. Mediated by the notion of quasi-modal operator, in [2, Lemma 5], it is shown that generalized implications and subordination relations over a given Boolean algebra bijectively correspond to each other. This observation is one of the starting points of the present paper.

In the present paper, we generalize the connection between generalized implications and subordination relations in the context of (distributive) lattices by introducing *slanted Heyting algebras* and *slanted co-Heyting algebras* (see Definitions 4.1 and 6.1). Any subordination algebra $\mathbb{S} = (A, \prec)$ induces the binary slanted operators $\Rightarrow_\prec$ and $\succcurlyeq_\prec$ on A such that each $a, b \in A$ are mapped to the following open and closed elements of its canonical extensions, respectively:

$$a \Rightarrow_\prec b := \bigvee\{c \in A \mid a \wedge c \prec b\} \quad \text{and} \quad a \succcurlyeq_\prec b := \bigwedge\{c \in A \mid b \prec a \vee c\}. \tag{1}$$

These slanted operations can be understood as *normative* counterparts of the identities defining Heyting implication and co-implication, respectively:

$$a \to b := \bigvee\{c \in A \mid a \wedge c \leq b\} \quad \text{and} \quad a \succ b := \bigwedge\{c \in A \mid b \leq a \vee c\} \tag{2}$$

indeed, the order relation in (2), encoding *logical* entailment, is replaced in (1) by the subordination relation $\prec$ which encodes *normative* entailment. We can interpret $a \Rightarrow_\prec b$ as 'the disjunction of all propositions that *normatively*

imply b when in conjunction with a' (and $a \succcurlyeq_{\prec} b$ as 'the conjunction of all propositions whose disjunction with a is *normatively* implied by b'). Hence, $a \Rightarrow_{\prec} b$ can be understood as the weakest side condition, or *context*, under which a normatively implies b.

Structure of the paper. In Section 2 we provide some examples that motivate the language we introduced and indicate its expressive power. In Section 3, we collect basic technical definitions; in Section 4, we introduce slanted Heyting algebras and show that they equivalently represent subordination algebras; in Section 5, we discuss how axioms in the language of slanted Heyting algebras capture interesting normative conditions; in Sections 6 and 7, we introduce slanted Heyting co-implication and the pseudo (co-)complements; more examples of conditions are discussed in Section 8; in Section 9, we identify the classes of axioms and normative conditions that correspond to each other; we conclude in Section 10.

2 Modelling deontic reasoning

We discussed how $a \Rightarrow_{\prec} b$ can be understood as the weakest side condition, or *context*, under which a normatively implies b, and hence, adding $\Rightarrow_{\prec}$ makes the language capable to describe situations in which obligations and permissions may change based on context, capacity, or other factors.

For example, the expression $citizen \Rightarrow_{\prec} taxes$ allows us to represent, and make inferences with, the *constellation* of background conditions which make $citizen \prec taxes$ a valid conditional obligation, purely in terms of $citizen$ and $taxes$; for example, in this case, we can represent this scenario by the inequality ('logical entailment') $earn \leq citizen \Rightarrow_{\prec} taxes$.

Also, in many real-life situations, different conditional obligations have different levels of priority based e.g. on urgency, ethics, or legal requirements [13]. For example, 'if a hospital is overcrowded, doctors should treat patients based on urgency, although, by hospital policy, patients should be visited in order of arrival.' In law, 'if a lawyer learns confidential information that could prevent a serious crime, they should report it, even though client confidentiality is generally a top priority.' We can use slanted Heyting algebras to formalize conditional obligations with different levels of priority; for instance, consider the propositions 'you are a doctor' (*doctor*), 'you visit patients according to their order of arrival' (*order*), and 'you save lives' (*save*). The fact that saving lives has higher priority for a doctor than following the order of arrival of patients can be formalized by requiring that the obligation $doctor \prec save$ hold under any context in which $doctor \prec order$ holds. That is, for any context c, if $c \wedge doctor \prec order$, then $c \wedge doctor \prec save$, which is equivalent to the inequality $doctor \Rightarrow_{\prec} order \leq doctor \Rightarrow_{\prec} save$ [1].

Another example where this framework can be applied, is to use the context of obligations to rephrase paradoxes, like the "contrary to duty" obli-

[1] If we want to explicitly model that there is a context c for which $c \wedge doctor \prec save$ holds, but $c \wedge doctor \prec order$ does not hold, then $\leq$ can be replaced by $<$.

gations. Consider the obligations '*You should never kill*' and '*if you kill, you should kill gently*'. The context of these obligations can be expressed as follows: $kill \Rightarrow_{\prec} \bot = c_1$ and $kill \Rightarrow_{\prec} gentle = c_2$. Then $c_1 \wedge c_2 = \bot$ captures situations in which both obligations are valid but under mutually exclusive contexts; for instance, the obligation '*You should never kill*' holds under usual or ordinary situations, while '*if you kill, you should kill gently*' holds in exceptional situations where killing may be permitted. Since, $c_1 \wedge c_2 = \bot$ these obligations do **not** imply $kill \prec \bot$. This formalism makes it possible to model contrary-to-duty obligations, while preventing the usual paradoxes from arising.

Finally, the operator $\Rightarrow_{\prec}$ allows us to formalize an infinite set of conditional obligations symbolically. For instance, in the context of the functioning of an autonomous vehicle, for any $0 < x \in \mathbb{R}$, consider the obligation $speed_x \wedge obst \prec brake$, which reads 'if your speed is greater than or equal to x kmph and there is an obstacle in front of you, then you must brake'. Then, the conditional obligation 'If there is an obstacle in front and your speed is *strictly greater than* 0 kmph, then you must brake' is captured by the infinite set of obligations $\{speed_x \wedge obst \prec brake \mid 0 < x \in \mathbb{R}\}$. This can be represented equivalently by the identity $obst \Rightarrow_{\prec} brake = Positive$, where $obst \Rightarrow_{\prec} brake = \bigvee\{speed_x \mid 0 < x \in \mathbb{R}\}$, and $Positive$ stands for 'speed of vehicle is positive'. Note that by requiring the identity above, we do not require the infinite join $\bigvee\{speed_x \mid 0 < x\}$ to be a proposition in our normative system (i.e. an element of the distributive lattice A); however, this join can be represented in terms of propositions $obst$ and $brake$ using the operator $\Rightarrow_{\prec}$.

3 Preliminaries

In what follows, when we say 'lattice', we mean 'bounded lattice'. Let A be a sublattice of a complete lattice A'.

(i) An element $k \in A'$ is *closed* if $k = \bigwedge F$ for some non-empty $F \subseteq A$; an element $o \in A'$ is *open* if $o = \bigvee I$ for some non-empty $I \subseteq A$;
(ii) A is *dense* in A' if every element of A is both the join of closed elements and the meet of open elements of A'.
(iii) A is *compact* in A' if, for all nonempty $F, I \subseteq A$, if $\bigwedge F \leq \bigvee I$ then $\bigwedge F' \leq \bigvee I'$ for some finite $F' \subseteq F$ and some finite $I' \subseteq I$.
(iv) The *canonical extension* of a lattice A is a complete lattice A^δ containing A as a dense and compact sublattice.

The canonical extension A^δ of any lattice A preserves only the finite meet and joins of A and destroys all other (potentially existing) infinite meets and joins. This means, for example, that the canonical extension of the powerset lattice of an infinite set, is not the powerset lattice itself, but a different, bigger in size lattice. The canonical extension always exists and is unique up to an isomorphism fixing A (cf. [11, Propositions 2.6 and 2.7]).

We let $K(A^\delta)$ (resp. $O(A^\delta)$) denote the set of the closed (resp. open) elements of A^δ. It is easy to see that $A = K(A^\delta) \cap O(A^\delta)$, which is why the elements of A are referred to as the *clopen* elements of A^δ. The following propositions collect well known facts which we will use in the remainder of the paper.

In particular, item (iv) of the next proposition is a variant of [12, Lemma 3.2].

Proposition 3.1 *(cf. [8, Proposition 2.6]) For any lattice A, all $k_1, k_2 \in K(A^\delta)$, $o_1, o_2 \in O(A^\delta)$, and $u_1, u_2 \in A^\delta$,*

(i) $k_1 \leq k_2$ *iff* $k_2 \leq b$ *implies* $k_1 \leq b$ *for all* $b \in A$.
(ii) $o_1 \leq o_2$ *iff* $b \leq o_1$ *implies* $b \leq o_2$ *for all* $b \in A$.
(iii) $u_1 \leq u_2$ *iff* $k \leq u_1$ *implies* $k \leq u_2$ *for all* $k \in K(A^\delta)$, *iff* $u_2 \leq o$ *implies* $u_1 \leq o$ *for all* $o \in O(A^\delta)$.
(iv) $k_1 \vee k_2 \in K(A^\delta)$ *and* $o_1 \wedge o_2 \in O(A^\delta)$.

Proposition 3.2 *(cf. [8, Proposition 2.7]) For any lattice A,*

(i) *for any $b \in A$, $k_1, k_2 \in K(A^\delta)$, and $o \in O(A^\delta)$,*
  (i) $k_1 \wedge k_2 \leq b$ *implies* $a_1 \wedge a_2 \leq b$ *for some* $a_1, a_2 \in A$ *s.t.* $k_i \leq a_i$;
  (ii) $k_1 \wedge k_2 \leq o$ *implies* $a_1 \wedge a_2 \leq b$ *for some* $a_1, a_2, b \in A$ *s.t.* $k_i \leq a_i$ *and* $b \leq o$;
  (iii) $\bigwedge K \in K(A^\delta)$ *for every* $K \subseteq K(A^\delta)$.
(ii) *for any $a \in A$, $o_1, o_2 \in O(A^\delta)$, and $k \in K(A^\delta)$,*
  (i) $a \leq o_1 \vee o_2$ *implies* $a \leq b_1 \vee b_2$ *for some* $b_1, b_2 \in A$ *s.t.* $b_i \leq o_i$;
  (ii) $k \leq o_1 \vee o_2$ *implies* $a \leq b_1 \vee b_2$ *for some* $a, b_1, b_2 \in A$ *s.t.* $b_i \leq o_i$ *and* $k \leq a$.
  (iii) $\bigvee O \in O(A^\delta)$ *for every* $O \subseteq O(A^\delta)$.

For the purposes of this paper, a *subordination algebra* is a tuple $\mathbb{S} = (A, \prec)$ such that A is a distributive lattice and $\prec \subseteq A \times A$ is a *subordination relation*, i.e. $\prec$ satisfies the following conditions: for all $a, b, c, d \in A$,

($\bot$-$\top$) $\bot \prec \bot$ and $\top \prec \top$; (AND) if $a \prec b$ and $a \prec c$ then $a \prec b \wedge c$;
(OR) if $a \prec c$ and $b \prec c$ then $a \vee b \prec c$; (WO-SI) if $a \leq b \prec c \leq d$ then $a \prec d$.

4 Slanted Heyting algebras and subordination algebras

Slanted Heyting algebras form a subclass of slanted DLE-algebras [10, Definition 3.2].

Definition 4.1 A *slanted Heyting algebra* is a tuple $\mathbb{A} = (A, \Rightarrow)$ s.t. A is a distributive lattice, and $\Rightarrow\colon A \times A \to A^\delta$ s.t. for all $a, b, c \in A$,

(i) $a \Rightarrow b \in O(A^\delta)$;
(ii) $a \Rightarrow (b_1 \wedge b_2) = (a \Rightarrow b_1) \wedge (a \Rightarrow b_2)$ and $a \Rightarrow \top = \top$;
(iii) $(a_1 \vee a_2) \Rightarrow b = (a_1 \Rightarrow b) \wedge (a_2 \Rightarrow b)$ and $\bot \Rightarrow b = \top$;
(iv) $c \leq a \Rightarrow b$ iff $a \wedge c \leq \top \Rightarrow b$.

The *canonical extension* of $\mathbb{A} = (A, \Rightarrow)$ (cf. [10, Definition 3.4]) is $\mathbb{A}^\delta = (A^\delta, \Rightarrow^\pi)$, where A^δ is the canonical extension of A, and $\Rightarrow^\pi\colon A^\delta \times A^\delta \to A^\delta$ is defined as follows: for every $k \in K(A^\delta)$, $o \in O(A^\delta)$ and $u, v \in A^\delta$,

$$k \Rightarrow^\pi o := \bigvee\{a \Rightarrow b \mid k \leq a, b \leq o, a, b \in A\}$$
$$u \Rightarrow^\pi v := \bigwedge\{k \Rightarrow^\pi o \mid k \in K(A^\delta), o \in O(A^\delta), k \leq u, v \leq o\}$$

The map $\Rightarrow^\pi$ extends $\Rightarrow$, distributes over arbitrary meets in its second coordinate, and distributes arbitrary joins to meets in its first coordinate (cf. [10, Lemma 3.5]). In what follows, we will omit the superscript $^\pi$, and rely on the arguments for disambiguation. Next, we discuss how slanted Heyting algebras can be understood as an equivalent presentation of subordination algebras.

Definition 4.2 The slanted Heyting algebra associated with the subordination algebra[2] $\mathbb{S} = (A, \prec)$ is the tuple $\mathbb{S}_* := (A, \Rightarrow_\prec)$, s.t. for all $a, b \in A$,

$$a \Rightarrow_\prec b := \bigvee\{c \in A \mid a \wedge c \prec b\}. \tag{3}$$

The subordination algebra associated with the slanted Heyting algebra $\mathbb{A} = (A, \Rightarrow)$ is the tuple $\mathbb{A}^* := (A, \prec_\Rightarrow)$ s.t. for all $a, b \in A$,

$$a \prec_\Rightarrow b \text{ iff } a \leq \top \Rightarrow b.$$

If $\prec$ represents a normative system, then $a \Rightarrow_\prec b$ is the disjunction of all propositions that together with a normatively imply b. That is, $a \wedge (a \Rightarrow_\prec b) \prec b$ always holds, and $a \Rightarrow_\prec b$ is the weakest element of A^δ with this property.

Proposition 4.3 *For any subordination algebra $\mathbb{S} = (A, \prec)$, any slanted Heyting algebra $\mathbb{A} = (A, \Rightarrow)$, and for all $a, b, c \in A$,*

(i) $c \leq a \Rightarrow_\prec b$ *iff* $a \wedge c \prec b$;
(ii) $\mathbb{S}_* = (A, \Rightarrow_\prec)$ *is a slanted Heyting algebra;*
(iii) $\mathbb{A}^* = (A, \prec_\Rightarrow)$ *is a subordination algebra;*
(iv) $a \prec_{\Rightarrow_\prec} b$ *iff* $a \prec b$, *and* $a \Rightarrow_{\prec_\Rightarrow} b = a \Rightarrow b$.

Proof. (i) If $a \wedge c \prec b$, then $c \leq \bigvee\{c \mid a \wedge c \prec b\} = a \Rightarrow_\prec b$. If $c \leq a \Rightarrow_\prec b = \bigvee\{c \mid a \wedge c \prec b\} \in O(A^\delta)$, by compactness, $c \leq d$ for some $d \in A$ s.t. $a \wedge d \prec b$. Hence, $a \wedge c \leq a \wedge d \prec b$, which implies $a \wedge c \prec b$ by (SI).

(ii) By definition, $a \Rightarrow_\prec b = \bigvee\{c \mid a \wedge c \prec b\} \in O(A^\delta)$ for any $a, b \in A$. Moreover, $a \Rightarrow_\prec \top = \bigvee\{c \mid a \wedge c \prec \top\} = \top$, the last identity holding because properties ($\top$) and (SI) hold for $\prec$. Likewise, $\bot \Rightarrow_\prec b = \bigvee\{c \mid \bot \wedge c \prec b\} = \top$, the last identity holding because properties ($\bot$) and (WO) hold for $\prec$. Let $a, b_1, b_2 \in A$, and let us show that

$$a \Rightarrow_\prec (b_1 \wedge b_2) = (a \Rightarrow_\prec b_1) \wedge (a \Rightarrow_\prec b_2).$$

For the left-to-right inequality, by Proposition 3.1 (ii) and item (i), this is equivalent to show that, for any $c \in A$,

$$a \wedge c \prec b_1 \wedge b_2 \text{ iff } a \wedge c \prec b_1 \text{ and } a \wedge c \prec b_2.$$

If $a \wedge c \prec b_1 \wedge b_2$, then $a \wedge c \prec b_1 \wedge b_2 \leq b_i$ for $i = 1, 2$, which implies by (WO) that $a \wedge c \prec b_i$, as required. Conversely, if $a \wedge c \prec b_i$ for $i = 1, 2$, then, by (AND), $a \wedge c \prec b_1 \wedge b_2$, as required. Let $a_1, a_2, b \in A$, and let us show that

$$(a_1 \vee a_2) \Rightarrow_\prec b = (a_1 \Rightarrow_\prec b) \wedge (a_2 \Rightarrow_\prec b).$$

by Proposition 3.1 (ii) and (i), this is equivalent to show that, for any $c \in A$,

$$(a_1 \vee a_2) \wedge c \prec b \text{ iff } a_1 \wedge c \prec b \text{ and } a_2 \wedge c \prec b.$$

If $(a_1 \vee a_2) \wedge c \prec b$, then $a_i \wedge c \leq (a_1 \vee a_2) \wedge c \prec b$, which implies, by (SI), that $a_i \wedge c \prec b$ for $i = 1, 2$, as required. Conversely, if $a_1 \wedge c \prec b$ and $a_2 \wedge c \prec b$, then

[2] Notice that, by definition, $a \Rightarrow_\prec p = a \to \blacksquare b$, where $\blacksquare b := \bigvee \prec^{-1}[b]$ and $\to$ is the Heyting algebra implication which is naturally defined on A^δ when A is a distributive lattice. However, the definition as given and some of the ensuing proofs hold in a wider setting than that of distributive lattices. In particular, Proposition 4.3 holds verbatim if $\mathbb{S}$ is a general lattice-based proto-subordination algebra with ($\top$), ($\bot$), (SI) and (WO) and $\mathbb{A}$ is a slanted algebra s.t. $\Rightarrow_\prec$ is antitone and $\bot$-reversing in the first coordinate and monotone and $\top$-preserving in the second one. In such a setting, the equivalent characterization of $a \Rightarrow_\prec p$ as $a \to \blacksquare b$ is not available anymore, since $\to$ does not exist in general.

by distributivity and (OR), $(a_1 \vee a_2) \wedge c = (a_1 \wedge c) \vee (a_2 \wedge c) \prec b$, as required. Finally, let us show that for all $a, b, c \in A$,

$$a \wedge c \leq \top \Rightarrow_{\prec} b \text{ iff } c \leq a \Rightarrow_{\prec} b.$$

By item (i), it is enough to show that $(a \wedge c) \wedge \top \prec b$ iff $a \wedge c \prec b$, which is immediately true.

(iii) As to ($\bot$), $\bot \prec_{\Rightarrow} b$ iff $\bot \leq \top \Rightarrow b$, which is clearly true. As to ($\top$), $a \prec_{\Rightarrow} \top$ iff $a \leq \top \Rightarrow \top$, which is true by Definition 4.1.2. As to (WO), $a \prec_{\Rightarrow} b \leq b'$ implies $a \leq \top \Rightarrow b \leq \top \Rightarrow b'$, hence $a \prec_{\Rightarrow} b'$, as required. As to (SI), $a' \leq a \prec_{\Rightarrow} b$ implies $a' \leq a \leq \top \Rightarrow b$, and hence $a' \prec_{\Rightarrow} b$, as required. As to (AND), if $a \prec_{\Rightarrow} b_i$ for $i = 1, 2$, then $a \leq \top \Rightarrow b_i$, therefore $a \leq (\top \Rightarrow b_1) \wedge (\top \Rightarrow b_2) = \top \Rightarrow (b_1 \wedge b_2)$, and so $a \prec_{\Rightarrow} b_1 \wedge b_2$, as required. As to (OR), if $a_i \leq b$ for $i = 1, 2$, then $a_i \leq \top \Rightarrow b$, hence $a_1 \vee a_2 \leq \top \Rightarrow b$ and so $a_1 \vee a_2 \prec_{\Rightarrow} b$, as required.

(iv) By Definition 4.2 and item (i), $a \prec_{\Rightarrow_{\prec}} b$ iff $a \leq \top \Rightarrow_{\prec} b$ iff $a = \top \wedge a \prec b$, as required. Finally, to show that $a \Rightarrow_{\prec_{\Rightarrow}} b = a \Rightarrow b$, by Proposition 3.1 (ii), it is enough to show that for all $c \in A$,

$$c \leq a \Rightarrow_{\prec_{\Rightarrow}} b \text{ iff } c \leq a \Rightarrow b.$$

By item (i), $c \leq a \Rightarrow_{\prec_{\Rightarrow}} b$ iff $a \wedge c \prec_{\Rightarrow} b$ i.e. $a \wedge c \leq \top \Rightarrow b$, which, by Definition 4.1.4, is equivalent to $c \leq a \Rightarrow b$, as required. □

Lemma 4.4 *For any slanted Heyting algebra $\mathbb{A} = (A, \Rightarrow)$, any $k, k' \in K(A^\delta)$, $o \in O(A^\delta)$, and $u, v, w \in A^\delta$,*

(i) $k \Rightarrow o \in O(A^\delta)$;
(ii) $k \leq k' \Rightarrow o$ *iff* $\exists a \exists b \exists c (a \leq c \Rightarrow b \;\&\; k \leq a \;\&\; k' \leq c \;\&\; b \leq o)$;
(iii) $k \wedge k' \leq \top \Rightarrow o$ *iff* $\exists a \exists b \exists c (a \wedge c \leq \top \Rightarrow b \;\&\; k \leq a \;\&\; k' \leq c \;\&\; b \leq o)$;
(iv) $k \leq k' \Rightarrow o$ *iff* $k \wedge k' \leq \top \Rightarrow o$;
(v) $w \leq u \Rightarrow v$ *iff* $w \wedge u \leq \top \Rightarrow v$.

Proof. (i) By definition, $k \Rightarrow o = \bigvee\{a \Rightarrow b \mid k \leq a, b \leq o, a, b \in A\}$. This implies that $k \Rightarrow o \in O(A^\delta)$ by Proposition 3.2 (ii), since $a \Rightarrow b \in O(A^\delta)$ by Definition 4.1 (i).

(ii)
$$\begin{array}{lll} & k \leq k' \Rightarrow o & \\ \text{iff} & k \leq \bigvee\{c \Rightarrow b \mid k' \leq c \;\&\; b \leq o\} & \text{Def. } \Rightarrow^\pi \\ \text{iff} & \exists a \exists b \exists c (a \leq c \Rightarrow b \;\&\; k \leq a \;\&\; k' \leq c \;\&\; b \leq o) & \text{compactness.} \end{array}$$

(iii)
$$\begin{array}{lll} & k \wedge k' \leq \top \Rightarrow o & \\ \text{iff} & \bigwedge\{a \wedge c \mid k \leq a \;\&\; k' \leq c\} \leq \bigvee\{\top \Rightarrow b \mid b \leq o\} & \text{Def. } \Rightarrow^\pi \\ \text{iff} & \exists a \exists b \exists c (a \wedge c \leq \top \Rightarrow b \;\&\; k \leq a \;\&\; k' \leq c \;\&\; b \leq o) & \text{compactness.} \end{array}$$

(iv) Immediate by items (ii) and (iii), and Definition 4.1.(iv).

(v) Since A is distributive, A^δ is completely distributive (cf. Section 3), by denseness, $w \wedge u = \bigvee\{k \wedge k' \mid k \leq w, k' \leq u\}$. Moreover, $\top \Rightarrow v = \bigwedge\{\top \Rightarrow o \mid v \leq o\}$ and $u \Rightarrow v = \bigwedge\{k \Rightarrow o \mid k \leq u, v \leq o\}$ by definition of $\Rightarrow^\pi$. Hence:

$$\begin{array}{lll} & w \leq u \Rightarrow v & \\ \text{iff} & \bigvee\{k \mid k \leq w\} \leq \bigwedge\{k' \Rightarrow o \mid k' \leq u, v \leq o\} & \\ \text{iff} & \forall k \forall k' \forall o (k \leq w \;\&\; k' \leq u \;\&\; v \leq o \implies k \leq k' \Rightarrow o) & \\ \text{iff} & \forall k \forall k' \forall o (k \leq w \;\&\; k' \leq u \;\&\; v \leq o \implies k \wedge k' \leq \top \Rightarrow o) & \text{item (iv)} \\ \text{iff} & \bigvee\{k \wedge k' \mid k \leq w, k' \leq u\} \leq \bigwedge\{\top \Rightarrow o \mid v \leq o\} & \\ \text{iff} & w \wedge u \leq \top \Rightarrow v. & \end{array}$$

□

5 Examples and discussion

In this section, we discuss how the semantic environment of slanted Heyting algebras can be used to model different properties of normative systems in a similar style to the modal characterizations of [7]. The properties of the previous section allow us to characterize well known conditions [3] of normative systems, such as (CT) and (T), in terms of axioms (i.e. algebraic inequalities which represent sequents) involving slanted Heyting implications. For the sake of enhanced readability, in what follows, all non quantified variables are quantified universally.

$$
\begin{array}{lll}
\text{(T)} & a \prec b \;\&\; b \prec c \implies a \prec c & \\
\text{iff} & a \leq \top \Rightarrow_\prec b \;\&\; b \leq \top \Rightarrow_\prec c \implies a \leq \top \Rightarrow_\prec c & \text{Prop. 4.3(i)} \\
\text{iff} & \exists b(a \leq \top \Rightarrow_\prec b \;\&\; b \leq \top \Rightarrow_\prec c) \implies a \leq \top \Rightarrow_\prec c & \\
\text{iff} & a \leq \top \Rightarrow_\prec (\top \Rightarrow_\prec c) \implies a \leq \top \Rightarrow_\prec c & \text{Lemma 4.4(ii)} \\
\text{iff} & \top \Rightarrow_\prec (\top \Rightarrow_\prec c) \leq \top \Rightarrow_\prec c & \text{Prop. 3.1(ii)}
\end{array}
$$

$$
\begin{array}{lll}
\text{(CT)} & a \prec b \;\&\; a \wedge b \prec c \Rightarrow a \prec c & \\
\text{iff} & a \leq \top \Rightarrow_\prec b \;\&\; a \leq b \Rightarrow_\prec c \implies a \leq \top \Rightarrow_\prec c & \text{Prop. 4.3(i)} \\
\text{iff} & a \leq (\top \Rightarrow_\prec b) \wedge (b \Rightarrow_\prec c) \implies a \leq \top \Rightarrow_\prec c & \\
\text{iff} & (\top \Rightarrow_\prec b) \wedge (b \Rightarrow_\prec c) \leq \top \Rightarrow_\prec c & \text{Prop. 3.1(ii)}
\end{array}
$$

Conversely, we can translate inequalities on slanted Heyting algebras into equivalent conditions on subordination algebras. For example, consider the following inequalities encoding transitivity ($\text{T}_{\Rightarrow_\prec}$) and cumulative transitivity ($\text{CT}_{\Rightarrow_\prec}$) of the implication $\Rightarrow_\prec$.

$$
\begin{array}{lll}
(\text{T}_{\Rightarrow_\prec}) & (a \Rightarrow_\prec b) \wedge (b \Rightarrow_\prec c) \leq a \Rightarrow_\prec c & \\
\text{iff} & d \leq (a \Rightarrow_\prec b) \wedge (b \Rightarrow_\prec c) \implies d \leq a \Rightarrow_\prec c & \text{Prop. 3.1(ii)} \\
\text{iff} & d \leq a \Rightarrow_\prec b \;\&\; d \leq b \Rightarrow_\prec c \implies d \leq a \Rightarrow_\prec c & \\
\text{iff} & a \wedge d \prec b \;\&\; b \wedge d \prec c \implies a \wedge d \prec c & \text{Prop. 4.3(i)}
\end{array}
$$

$$
\begin{array}{lll}
(\text{CT}_{\Rightarrow_\prec}) & (a \Rightarrow_\prec b) \wedge ((a \wedge b) \Rightarrow_\prec c) \leq a \Rightarrow_\prec c & \\
\text{iff} & d \leq (a \Rightarrow_\prec b) \wedge ((a \wedge b) \Rightarrow_\prec c) \implies d \leq a \Rightarrow_\prec c & \text{Prop. 3.1(ii)} \\
\text{iff} & d \leq a \Rightarrow_\prec b \;\&\; d \leq (a \wedge b) \Rightarrow_\prec c \implies d \leq a \Rightarrow_\prec c & \\
\text{iff} & d \wedge a \prec b \;\&\; d \wedge (a \wedge b) \prec c \implies d \wedge a \prec c & \text{Prop. 4.3(i)}
\end{array}
$$

Note that (T) and (CT) follow from the conditions equivalent to ($\text{T}_{\Rightarrow_\prec}$) and ($\text{CT}_{\Rightarrow_\prec}$), respectively, by setting $d := \top$. [4] Hence, ($\text{T}_{\Rightarrow_\prec}$) and ($\text{CT}_{\Rightarrow_\prec}$) can be seen as strengthening (T) and (CT) under any side condition or context d.

Other interesting conditions on normative systems can similarly be expressed in terms of the language of slanted Heyting algebras. For example, normative counterparts of intuitionistic tautologies such as the *Frege axiom*:

[3] Typically, the conditions we consider are expressed in terms of rules or Horn clauses, i.e. conjunction of relational atoms entails a relational atom, and the entailment relation will be represented by the symbol $\implies$.

[4] The converse implication also holds in the case of (CT), by substituting a in (CT) with $a \wedge d$ (recall that a is universally quantified).

$$
\begin{array}{lll}
& a \Rightarrow_\prec (b \Rightarrow_\prec c) \leq (a \Rightarrow_\prec b) \Rightarrow_\prec (a \Rightarrow_\prec c) & \\
\text{iff} & k \leq a \Rightarrow_\prec (b \Rightarrow_\prec c) \;\&\; k' \leq a \Rightarrow_\prec b \implies k \leq k' \Rightarrow_\prec (a \Rightarrow_\prec c) & \text{denseness} \\
\text{iff} & \exists d(k \leq d \leq a \Rightarrow_\prec (b \Rightarrow_\prec c)) \;\&\; \exists e(k' \leq e \leq a \Rightarrow_\prec b) \implies & \\
& k \leq k' \Rightarrow_\prec (a \Rightarrow_\prec c) & \text{compactness} \\
\text{iff} & d \leq a \Rightarrow_\prec (b \Rightarrow_\prec c) \;\&\; e \leq a \Rightarrow_\prec b \implies d \leq e \Rightarrow_\prec (a \Rightarrow_\prec c) & (*) \\
\text{iff} & d \leq a \Rightarrow_\prec f \;\&\; f \leq b \Rightarrow_\prec c \;\&\; e \leq a \Rightarrow_\prec b \implies & \\
& \exists g(d \leq e \Rightarrow_\prec g \;\&\; g \leq a \Rightarrow_\prec c) & \text{compactness} \\
\text{iff} & a \wedge d \prec f \;\&\; a \wedge e \prec b \;\&\; b \wedge f \prec c \implies \exists g(e \wedge d \prec g \;\&\; a \wedge g \prec c) & \text{Prop. 4.3(i)}
\end{array}
$$

As to the equivalence marked $(*)$, from bottom to top, fix $a, b, c \in A$ and $k, k' \in K(A^\delta)$ s.t. $\exists d(k \leq d \leq a \Rightarrow_\prec (b \Rightarrow_\prec c)) \;\&\; \exists e(k' \leq e \leq a \Rightarrow_\prec b)$. Then $d \leq a \Rightarrow_\prec (b \Rightarrow_\prec c) \;\&\; e \leq a \Rightarrow_\prec b$, hence by assumption, $k \leq d \leq e \Rightarrow_\prec (a \Rightarrow_\prec c) \leq k \leq k' \Rightarrow_\prec (a \Rightarrow_\prec c)$. From top to bottom, it is enough to instantiate $k := d$ and $k' := e$.

The condition above can be understood as a mode of transitive propagation of obligations under context: if a normatively implies f whenever d, and b whenever e, and b and f together normatively imply c, then c is also normatively implied by a in the context of some g that is normatively implied by $d \wedge e$. This condition can be understood as a generalization of the following principle:

$$
\begin{array}{ll}
& a \wedge d \prec f \;\&\; a \wedge e \prec b \;\&\; b \wedge f \prec c \implies a \wedge (d \wedge e) \prec c \\
\text{iff} & a \leq d \Rightarrow_\prec f \;\&\; a \leq e \Rightarrow_\prec b \;\&\; b \leq f \Rightarrow_\prec c \implies a \leq (d \wedge e) \Rightarrow_\prec c \\
\text{iff} & a \leq d \Rightarrow_\prec f \;\&\; \exists b(a \leq e \Rightarrow_\prec b \;\&\; b \leq f \Rightarrow_\prec c) \implies a \leq (d \wedge e) \Rightarrow_\prec c \\
\text{iff} & a \leq d \Rightarrow_\prec f \;\&\; a \leq e \Rightarrow_\prec (f \Rightarrow_\prec c) \implies a \leq (d \wedge e) \Rightarrow_\prec c \\
\text{iff} & a \leq (d \Rightarrow_\prec f) \wedge (e \Rightarrow_\prec (f \Rightarrow_\prec c)) \implies a \leq (d \wedge e) \Rightarrow_\prec c \\
\text{iff} & (d \Rightarrow_\prec f) \wedge (e \Rightarrow_\prec (f \Rightarrow_\prec c)) \leq (d \wedge e) \Rightarrow_\prec c
\end{array}
$$

Interesting conditions can be also captured in terms of the normative counterparts of axioms defining intermediate logics, such as the normative counterpart of the *Gödel-Dummett axiom*:

$$
\begin{array}{lll}
& \top \leq (a \Rightarrow_\prec b) \vee (b \Rightarrow_\prec a) & \\
\text{iff} & \exists e \exists f(\top \leq e \vee f \;\&\; e \leq a \Rightarrow_\prec b \;\&\; f \leq b \Rightarrow_\prec a) & \text{Prop. 3.2(ii)} \\
\text{iff} & \exists e \exists f(\top \leq e \vee f \;\&\; a \wedge e \prec b \;\&\; b \wedge f \prec a) & \text{Prop. 4.3(i)}
\end{array}
$$

The condition above requires there to be propositions e and f such that e or f is always the case and a and e normatively imply b and b and f normatively imply a. This requirement can be seen as a generalization of the following dichotomy axiom for a normative system: $\forall a \forall b(a \prec b \text{ or } b \prec a)$, i.e. for all propositions a and b, either a normatively implies b or b normatively implies a.

The following axiom encodes the distributivity of $\Rightarrow_\prec$ over disjunction in its second coordinate.

$$
\begin{array}{lll}
& a \Rightarrow_\prec (b \vee c) \leq (a \Rightarrow_\prec b) \vee (a \Rightarrow_\prec c) & \\
\text{iff} & d \leq a \Rightarrow_\prec (b \vee c) \implies d \leq (a \Rightarrow_\prec b) \vee (a \Rightarrow_\prec c) & \text{Prop. 3.1(ii)} \\
\text{iff} & d \wedge a \prec b \vee c \implies \exists e \exists f(d \leq e \vee f \,\&\, e \leq a \Rightarrow_\prec b \,\&\, f \leq a \Rightarrow_\prec c) & \text{Prop. 3.2(ii)} \\
\text{iff} & d \wedge a \prec b \vee c \implies \exists e \exists f(d \leq e \vee f \;\&\; e \wedge a \prec b \;\&\; a \wedge f \prec c) & \text{Prop. 4.3(i)}
\end{array}
$$

This axiom characterizes a form of splitting into cases for conditional obligations: if a and d normatively imply $b \vee c$ then some e and f exist s.t. d implies $e \vee f$ and a and e normatively imply b while a and f normatively imply c.

6 Slanted co-Heyting algebras

Definition 6.1 A *slanted co-Heyting algebra* is a tuple $\mathbb{A} = (A, \succcurlyeq)$ s.t. A is a distributive lattice, and $\succcurlyeq\colon A \times A \to A^\delta$ s.t. for all $a, b, c \in A$,

(i) $a \succcurlyeq b \in K(A^\delta)$;
(ii) $a \succcurlyeq (b_1 \vee b_2) = (a \succcurlyeq b_1) \vee (a \succcurlyeq b_2)$ and $a \succcurlyeq \bot = \bot$;
(iii) $(a_1 \wedge a_2) \succcurlyeq b = (a_1 \succcurlyeq b) \vee (a_2 \succcurlyeq b)$ and $\top \succcurlyeq b = \bot$;
(iv) $a \succcurlyeq b \leq c$ iff $\bot \succcurlyeq b \leq a \vee c$.

The *canonical extension* of $\mathbb{A}$ is $\mathbb{A}^\delta = (A^\delta, \succcurlyeq^\sigma)$, where A^δ is the canonical extension of A, and $\succcurlyeq^\sigma$: $A^\delta \times A^\delta \to A^\delta$ is defined as follows: for every $k \in K(A^\delta)$, $o \in O(A^\delta)$ and $u, v \in A^\delta$,

$$o \succcurlyeq^\sigma k := \bigwedge\{a \succcurlyeq b \mid a \leq o, k \leq b, a, b \in A\}$$
$$u \succcurlyeq^\sigma v := \bigwedge\{k \Rightarrow^\sigma o \mid k \in K(A^\delta), o \in O(A^\delta), k \leq u, v \leq o\}$$

The map $\succcurlyeq^\sigma$ extends $\succcurlyeq$, distributes over arbitrary joins in its second coordinate, and distributes arbitrary meets to joins in its first coordinate (cf. [10, Lemma 3.5]). In what follows, we will omit the superscript $^\sigma$, and rely on the arguments for disambiguation. Slanted co-Heyting algebras are also an equivalent presentation of subordination algebras.

Definition 6.2 The slanted co-Heyting algebra associated with a subordination algebra $\mathbb{S} = (A, \prec)$ is the tuple $\mathbb{S}_\bullet := (A, \succcurlyeq_\prec)$, s.t. for all $a, b \in S$,

$$a \succcurlyeq_\prec b := \bigwedge\{c \in A \mid b \prec a \vee c\}. \tag{4}$$

Hence, $a \succcurlyeq_\prec b$ represents the conjunction of all propositions whose disjunction with a is normatively implied by b. That is, $a \prec (a \succcurlyeq_\prec b) \vee b$, and $a \succcurlyeq_\prec b$ holds, and is the strongest element of A^δ with this property. The subordination algebra associated with $\mathbb{A} = (A, \succcurlyeq)$ is $\mathbb{A}^\bullet := (A, \prec_\succcurlyeq)$ s.t. for all $a, b \in A$,

$$a \prec_\succcurlyeq b \text{ iff } \bot \succcurlyeq a \leq b.$$

The following proposition is dual to Proposition 4.3.

Proposition 6.3 *For any subordination algebra $\mathbb{S} = (A, \prec)$, any slanted co-Heyting algebra $\mathbb{A} = (A, \succcurlyeq)$, and all $a, b, c \in A$,*
(i) $a \succcurlyeq_\prec b \leq c$ *iff* $b \prec a \vee c$;
(ii) $\mathbb{S}_\bullet = (A, \succcurlyeq_\prec)$ *is a slanted co-Heyting algebra;*
(iii) $\mathbb{A}^\bullet = (A, \prec_\succcurlyeq)$ *is a subordination algebra;*
(iv) $a \prec_{\succcurlyeq_\prec} b$ *iff* $a \prec b$, *and* $a \succcurlyeq_{\prec_\succcurlyeq} b = a \succcurlyeq b$.

The following lemma is dual to Lemma 4.4.

Lemma 6.4 *For any slanted co-Heyting algebra $\mathbb{A} = (A, \succcurlyeq)$, any $o, o' \in O(A^\delta)$, $k \in K(A^\delta)$, and $u, v, w \in A^\delta$,*
(i) $k \succcurlyeq o \in K(A^\delta)$;
(ii) $o' \succcurlyeq k \leq o$ *iff* $\exists a \exists b \exists c (a \succcurlyeq b \leq c \;\&\; a \leq o' \;\&\; k \leq b \;\&\; c \leq o)$;
(iii) $\bot \succcurlyeq k \leq o' \vee o$ *iff* $\exists a \exists b \exists c (\bot \succcurlyeq b \leq a \vee c \;\&\; a \leq o' \;\&\; k \leq b \;\&\; c \leq o)$;
(iv) $o' \succcurlyeq k \leq o$ *iff* $\bot \succcurlyeq k \leq o \vee o'$;
(v) $w \succcurlyeq u \leq v$ *iff* $\bot \succcurlyeq u \leq w \vee v$.

7 Slanted pseudo-complements and co-complements

Any slanted (co-)Heyting algebra based on a distributive lattice A induces the operation $\neg : A \to A^\delta$ (resp. $\sim : A \to A^\delta$) defined by the assignment

$a \mapsto a \Rightarrow \bot$ (resp. $a \mapsto \top \succcurlyeq a$). When $\Rightarrow = \Rightarrow_\prec$ (resp. $\succcurlyeq = \succcurlyeq_\prec$) for some $\prec$ on A, it immediately follows from (3) and (4) that, for any $a \in A$,

$$\neg a = \bigvee\{c \in A \mid a \wedge c \prec \bot\} \qquad \sim a = \bigwedge\{c \in A \mid \top \prec a \vee c\}.$$

In particular,

$$\neg\top = \bigvee\{c \in A \mid c \prec \bot\} \qquad \sim\bot = \bigwedge\{c \in A \mid \top \prec c\}.$$

The canonical extensions $\neg^\pi$ and $\sim^\sigma$ of maps $\neg$ and $\sim$ are defined as follows: for any $k \in K(A^\delta)$, $o \in O(A^\delta)$, $u \in A^\delta$,

$$\neg^\pi k := \bigvee\{\neg a \mid k \leq a, a \in A\} \text{ and } \neg^\pi u := \bigwedge\{\neg k \mid k \leq u\}$$
$$\sim^\sigma o := \bigwedge\{\sim a \mid a \leq o, a \in A\} \text{ and } \sim^\sigma u := \bigvee\{\sim o \mid u \leq o\}$$

We will typically omit the superscripts $^\sigma$ and $^\pi$, and rely on the arguments for disambiguation. Also, we omit the subscript $\prec$ even when $\neg$ and $\sim$ arise from $\Rightarrow_\prec$ and $\succcurlyeq_\prec$. The next lemma is straightforward from the definitions.

Lemma 7.1 *For any $u \in A^\delta$, $\neg u = u \Rightarrow \bot$ and $\sim u = u \succcurlyeq \top$.*

Intuitively, $\neg a$ represents the disjunction of all propositions which are normatively inconsistent with a. That is, $a \wedge \neg a \prec \bot$ and $\neg a$ is the weakest element of A^δ with this property. Likewise, $\sim a$ represents the conjunction of all propositions whose disjunction with a is an unconditional obligation. That is, $\top \prec a \vee \sim a$ and $\sim a$ is the strongest element of A^δ with this property. In particular, $\neg\top$ denotes the weakest normatively inconsistent element in A^δ, while $\sim\bot$ denotes the strongest unconditional obligation in A^δ.
The following lemma is a straightforward consequence of Lemmas 4.4 and 6.4.

Lemma 7.2 *For all $a, c \in A$, and $k \in K(A^\delta)$ and $o \in O(A^\delta)$,*
(i) $c \leq \neg a$ *iff* $a \wedge c \prec \bot$ *and* $\sim a \leq c$ *iff* $\top \prec c \vee a$;
(ii) $a \leq \neg b$ *iff* $b \leq \neg a$ *and* $\sim a \leq b$ *iff* $\sim b \leq a$;
(iii) $k' \leq \neg k$ *iff* $\exists a \exists c(k' \leq c \;\&\; k \leq a \;\&\; c \leq \neg a)$;
(iv) $\sim o \leq o'$ *iff* $\exists a \exists c(a \leq o \;\&\; c \leq o' \;\&\; \sim a \leq c)$.

8 More examples

Condition $a \prec \bot \implies a \leq \bot$ can be understood as the property of the normative system $\prec$ that any consistent (proposition) cannot yield a normative inconsistency. This condition can be axiomatically captured as follows

$$\begin{array}{lll} & a \prec \bot \implies a \leq \bot & \\ \text{iff} & a \leq \neg\top \implies a \leq \bot & \text{Lemma 7.2(i)} \\ \text{iff} & \neg\top \leq \bot & \text{Prop. 3.1 (ii)} \end{array}$$

The same condition can be also captured by $a \wedge \neg a \leq \bot$, as shown by the following computation:

$$\begin{array}{lll} & a \wedge \neg a \leq \bot & \\ \text{iff} & b \leq a \wedge \neg a \implies b \leq \bot & \text{Prop. 3.1 (ii)} \\ \text{iff} & b \leq a \,\&\, b \leq \neg a \implies b \leq \bot & \\ \text{iff} & b \leq a \,\&\, b \wedge a \prec \bot \implies b \leq \bot & \text{Lemma 7.2(i)} \\ \text{iff} & b \prec \bot \implies b \leq \bot & b \leq a \text{ iff } b = b \wedge a \end{array}$$

The next example is normative counterpart of De Morgan axiom:

$\neg(a \wedge b) \leq \neg a \vee \neg b$
iff $d \leq \neg(a \wedge b) \implies d \leq \neg a \vee \neg b$ Prop. 3.1 (ii)
iff $d \leq \neg(a \wedge b) \implies \exists e \exists f (d \leq e \vee f \,\&,\, e \leq \neg a \,\&\, f \leq \neg b)$ Prop. 3.2(ii)
iff $d \wedge a \wedge b \prec \bot \implies \exists e \exists f (d \leq e \vee f \,\&,\, e \wedge a \prec \bot \,\&\, f \wedge b \prec \bot)$ Lemma 7.2 (i)

The condition above can be interpreted as saying that if a, b, d are normatively inconsistent, then d implies $e \vee f$ s.t. e leads to normative inconsistency along with a and f leads to normative inconsistency along with b.

The next example is the normative counterpart of contraposition:

$(a \Rightarrow_\prec b) \leq (\neg b \Rightarrow_\prec \neg a)$
iff $k \leq (a \Rightarrow_\prec b)$ & $j \leq \neg b \implies (k \leq j \Rightarrow_\prec \neg a)$ denseness
iff $c \leq (a \Rightarrow_\prec b)$ & $d \leq \neg b \implies (c \leq d \Rightarrow_\prec \neg a)$ compactness
iff $c \wedge a \prec b$ & $d \wedge b \prec \bot \implies (c \leq d \Rightarrow_\prec \neg a)$ Prop. 4.3, Lem. 7.2(i)
iff $c \wedge a \prec b$ & $d \wedge b \prec \bot \implies \exists e (c \leq d \Rightarrow_\prec e \,\&\, e \leq \neg a)$ compactness
iff $c \wedge a \prec b$ & $d \wedge b \prec \bot \implies \exists e (c \wedge d \prec e \,\&\, e \wedge a \prec \bot)$ Lem. 7.2(i)

The condition above says that, if for all a, c and d some b exists s.t. $c \wedge a$ normatively imply b, and $b \wedge d$ are normatively inconsistent, then some e exists s.t. $c \wedge d$ normatively imply e and e and a together give a normative inconsistency. This condition generalizes the following principle:

$a \prec b \,\&\, d \wedge b \prec \bot \implies \exists e (d \prec e \,\&\, a \wedge e \prec \bot)$
iff $\exists b (a \leq \top \Rightarrow_\prec b \,\&\, b \leq \neg d) \implies \exists e (d \leq \top \Rightarrow_\prec e \,\&\, e \leq \neg a)$ def. of $\neg$
iff $a \leq \top \Rightarrow_\prec \neg d \implies d \leq \top \Rightarrow_\prec \neg a$ denseness
iff $a \leq \top \Rightarrow_\prec \neg d \implies d \circ \top \leq \neg a$ residuation
iff $a \leq \top \Rightarrow_\prec \neg d \implies a \leq \neg(d \circ \top)$ Lemma 7.2 (ii)
iff $\top \Rightarrow_\prec \neg d \leq \neg(d \circ \top)$ Prop. 3.1(ii)
iff $d \circ \top \leq \neg(\top \Rightarrow_\prec \neg d)$ Lemma 7.2 (ii)
iff $d \leq \top \Rightarrow_\prec \neg(\top \Rightarrow_\prec \neg d)$ residuation

where, in the computation above, $\circ$ denotes the left residual of $\Rightarrow^{\pi}_{\prec}$, which is defined as follows: $u \circ v = \bigwedge\{w \mid v \leq u \Rightarrow_\prec w\}$ for all $u, v, w \in A^\delta$. The condition above can be interpreted as the requirement that, for all a and d, if a normatively implies some b which is normatively inconsistent with d (i.e. if both d and b hold, we normatively fall into contradiction), then d normatively implies some e which is normatively inconsistent with a.

9 Correspondence and inverse correspondence

The examples discussed in Sections 5 and 8 are not isolated cases: in [7], the class of *(clopen-)analytic* axioms/inequalities is identified, each of which is shown to be equivalent to some condition on norms, permissions, or their interaction. Conversely, a class of such conditions is identified (referred to as *Kracht formulas*), each of which is shown to be equivalent to some modal axioms. In this section, we discuss how these results can be extended to the language of (distributive) lattices with slanted (co-)Heyting implications, [5] and show, via examples, that this language allows us to (algorithmically) capture conditions which cannot be captured by modal axioms of the language of [7] with the same techniques.

[5] Since the current signature is particularly simple, the definition of clopen-analytic inequality collapses to that of analytic inequality.

9.1 Correspondence

Let $\mathcal{L}$ be the language of distributive lattices expanded with one slanted Heyting (co-)implication. To characterize syntactically the class of $\mathcal{L}$-inequalities which are guaranteed to be algorithmically transformed into conditions on normative systems, we adapt the notion of *analytic $\mathcal{L}$-inequalities* of [7, Section 2.5].

The *positive* (resp. *negative*) *generation tree* of any $\mathcal{L}$-term ϕ is defined by labelling the root node of the generation tree of ϕ with the sign $+$ (resp. $-$), and then propagating the labelling on each remaining node as follows: For any node labelled with $\vee$ or $\wedge$, assign the same sign to its children nodes, and for any node labelled with $\Rightarrow$ or $\succcurlyeq$, assign the opposite sign to its first child node, and the same sign to its second child node. Nodes in signed generation trees are *positive* (resp. *negative*) if they are signed $+$ (resp. $-$). In the context of term inequalities $\varphi \leq \psi$, we consider the positive generation tree $+\varphi$ for the left side and the negative one $-\psi$ for the right side. Non-leaf nodes in signed generation trees are called Δ-*adjoints*, *syntactically left residuals (SLR)*, *syntactically right residuals (SRR)*, and *syntactically right adjoints (SRA)*, according to the specification given in the table below Nodes that are either classified as Δ-adjoints or SLR are collectively referred to as *Skeleton-nodes*, while SRA- and SRR-nodes are referred to as *PIA-nodes*. A branch in a signed generation tree $*\phi$, with $* \in \{+, -\}$, is *good* if it is the concatenation of two paths P_1 and P_2, one of which may possibly be of length 0, such that P_1 is a path from the leaf consisting (apart from variable nodes) only of PIA-nodes, and P_2 consists (apart from variable nodes) only of Skeleton-nodes. An $\mathcal{L}$-inequality $\varphi \leq \psi$ is *analytic* if every branch of $+\varphi$ and $-\psi$ is good.

Skeleton	PIA
Δ-adjoints	Syntactically Right Adjoint (SRA)
$+ \quad \vee$	$+ \quad \wedge \quad \neg$
$- \quad \wedge$	$- \quad \vee \quad \sim$
Syntactically Left Residual (SLR)	Syntactically Right Residual (SRR)
$+ \quad \wedge \quad \succcurlyeq \quad \neg$	$+ \quad \vee \quad \Rightarrow$
$- \quad \vee \quad \Rightarrow \quad \sim$	$- \quad \wedge \quad \succcurlyeq$

Based on the properties discussed in Sections 4 and 6, the algorithm of [7, Section 3] can succesfully be run also on analytic $\mathcal{L}$-inequalities; hence, the analogue of [7, Theorem 3.1] holds for analytic $\mathcal{L}$-inequalities. All inequalities discussed in Sections 5 and 8 are analytic $\mathcal{L}$-inequalities, and the chains of equivalences discussed there represent runs of the correspondence algorithm.

9.2 Inverse correspondence

The following abbreviations will be used throughout the present section:

$$
\begin{array}{rcll}
(\exists y \succ x)\varphi & \equiv & \exists y(x \prec y \ \&\ \varphi) & \text{i.e.} \quad \exists y(x \leq \top \Rightarrow_{\prec} y \ \&\ \varphi) \\
(\exists y \prec x)\varphi & \equiv & \exists y(y \prec x \ \&\ \varphi) & \text{i.e.} \quad \exists y(\bot \succcurlyeq_{\prec} y \leq x \ \&\ \varphi) \\
(\forall y \succ x)\varphi & \equiv & \forall y(x \prec y \implies \varphi) & \text{i.e.} \quad \forall y(x \leq \top \Rightarrow_{\prec} y \implies \varphi) \\
(\forall y \prec x)\varphi & \equiv & \forall y(y \prec x \implies \varphi) & \text{i.e.} \quad \forall y(\bot \succcurlyeq_{\prec} y \leq x \implies \varphi) \\
(\exists \overline{y} \leq_{\vee} x)\varphi & \equiv & \exists y_1 \exists y_2(x \leq y_1 \vee y_2 \ \&\ \varphi) & \\
(\exists \overline{y} \leq_{\wedge} x)\varphi & \equiv & \exists y_1 \exists y_2(y_1 \wedge y_2 \leq x \ \&\ \varphi) & \\
(\forall \overline{y} \leq_{\vee} x)\varphi & \equiv & \forall y_1 \forall y_2(x \leq y_1 \vee y_2 \Longrightarrow \varphi) & \\
(\forall \overline{y} \leq_{\wedge} x)\varphi & \equiv & \forall y_1 \forall y_2(y_1 \wedge y_2 \leq x \Longrightarrow \varphi). & \\
(\exists \overline{y} \leq_{\succcurlyeq_{\prec}} x)\varphi & \equiv & \exists y_1 \exists y_2(y_2 \prec y_1 \vee x \ \&\ \varphi) & \text{i.e.} \quad \exists y_1 \exists y_2(y_1 \succcurlyeq_{\prec} y_2 \leq x \ \&\ \varphi) \\
(\exists \overline{y} \geq_{\Rightarrow_{\prec}} x)\varphi & \equiv & \exists y_1 \exists y_2(x \wedge y_1 \prec y_2 \ \&\ \varphi) & \text{i.e.} \quad \exists y_1 \exists y_2(x \leq y_1 \Rightarrow_{\prec} y_2 \ \&\ \varphi) \\
(\forall \overline{y} \leq_{\succcurlyeq_{\prec}} x)\varphi & \equiv & \forall y_1 \forall y_2(y_2 \prec y_1 \vee x \implies \varphi) & \text{i.e.} \quad \forall y_1 \forall y_2(y_1 \succcurlyeq_{\prec} y_2 \leq x \implies \varphi) \\
(\forall \overline{y} \geq_{\Rightarrow_{\prec}} x)\varphi & \equiv & \forall y_1 \forall y_2(x \wedge y_1 \prec y_2 \implies \varphi) & \text{i.e.} \quad \forall y_1 \forall y_2(x \leq y_1 \Rightarrow_{\prec} y_2 \implies \varphi)
\end{array}
$$

Expressions such as $(\forall y \prec x)$ or $(\exists \overline{y} \geq_g x)$ above are referred to as *restricted quantifiers*. The variable y (resp. the variables in $\overline{y}$) in the formulas above is (resp. are) *restricted*, and the variable x is *restricting*, while the inequality occurring together with φ in the translation of the restricted quantifier is a *restricting inequality*. Throughout this section, we use the following letters to distinguish the roles of different variables ranging in the domain of an arbitrary slanted (co-)Heyting algebra. The different conditions assigned to these variables in Definition 9.1 will determine how they are introduced/eliminated:

- v variables occurring in the algebraic axiom;
- a positive variables introduced/eliminated using Proposition 3.1(ii). It must be possible to rewrite the quasi-inequality so that each such variable occurs only once on each side of the main implication;
- b negative variables introduced/eliminated using Proposition 3.1(i). The same considerations which apply to a-variables apply also to b-variables;
- c variables introduced/eliminated using Proposition 3.2 or Lemmas 4.4, 6.4, 7.2. in the antecedent of the main implication;
- d variables introduced/eliminated in the same way as c-variables in the consequent of the main implication.

The following definition adapts [7, Definition 5.2] to the present environment.

Definition 9.1 A *Kracht formula*[6] is a condition of the following shape:
$\forall \overline{v} \forall \overline{a}, \overline{b}(\forall \overline{c_m} R'_m z_m) \cdots (\forall \overline{c_1} R'_1 z_1) \big(\eta \Rightarrow (\exists \overline{d_o} R_o y_o) \cdots (\exists \overline{d_1} R_1 y_1) \zeta\big)$, where

(i) $R'_i, R_j \in \{\leq, \geq, \prec, \succ, \leq_{\succcurlyeq_{\prec}}, \geq_{\Rightarrow_{\prec}}\}$ for all $1 \leq i \leq m$ and $1 \leq j \leq o$;
(ii) variables in $\overline{z}$ are amongst those of $\overline{a}$, $\overline{b}$, and $\overline{c}$; variables in $\overline{y}$ are amongst those of $\overline{a}$, $\overline{b}$, and $\overline{d}$;
(iii) η and ζ are conjunctions of relational atoms sRt with $R \in \{\leq, \geq, \prec, \succ\}$;
(iv) all occurrences of variables in $\overline{a}$ (resp. $\overline{b}$) are positive (resp. negative) in all atoms (including those in restricting quantifiers) in which they occur;
(v) every variable in $\overline{c}$ (resp. in $\overline{d}$) occurs uniformly in η (resp. in ζ);

[6] In the modal logic literature, Kracht formulas (cf. [1, Section 3.7], [14]) are sentences in the first order language of Kripke frames which are (equivalent to) the first-order correspondents of Sahlqvist axioms. This notion has been generalized in [5,18] from classical modal logic to (distributive) LE-logics, and from a class of first order formulas targeting Sahlqvist LE-axioms to a class targeting the strictly larger class of inductive LE-axioms. The notion of Kracht formulas introduced in Definition 9.1 is different and in fact incomparable with those in [5,18], since it targets a different and incomparable class of modal axioms.

(vi) occurrences of variables in $\overline{c}$ (resp. $\overline{d}$) as restricting variables have the same polarity as their occurrences in η (resp. ζ).

(vii) all occurrences of variables in $\overline{a}$, $\overline{b}$, $\overline{c}$, and $\overline{d}$ in atoms of η and ζ are displayable;

(viii) each atom in η contains exactly one occurrence of a variable in $\overline{a}$, $\overline{b}$, or $\overline{c}$ (all other variables are in $\overline{v}$);

(ix) each atom sRt in ζ contains at most one occurrence of a variable in $\overline{d}$. Moreover, for every two different occurrences of the same variable not in $\overline{v}$ (i.e., in $\overline{a}$ or $\overline{b}$), the first common ancestor in the signed generation tree of sRt is either $+\wedge$ or $-\vee$.

For instance, the following condition (see last example of Section 5):

$$d \wedge a \prec b \vee c \implies \exists e \exists f (d \leq e \vee f \;\&\; e \wedge a \prec b \;\&\; a \wedge f \prec c)$$

can be recognized as an instance of the Kracht shape in the language of slanted Heyting algebras by assigning variables e and f the role of $\overline{d}$-variables, variable d the role of $\overline{a}$-variable, and variables a, b, c the role of $\overline{v}$-variables, and moreover, letting $d \leq e \vee f$ be the restricting inequality in the consequent, and $\eta := d \wedge a \prec b \vee c$ and $\zeta := e \wedge a \prec b \;\&\; a \wedge f \prec c$. The algorithm introduced in [7, Section 6] allows us to equivalently represent this condition as an axiom in the language of slanted Heyting algebras as follows:

$$\begin{array}{lll}
 & d \wedge a \prec b \vee c \implies \exists e \exists f (d \leq e \vee f \;\&\; e \wedge a \prec b \;\&\; a \wedge f \prec c) & \\
\text{iff} & d \leq a \Rightarrow_{\prec} (b \vee c) \implies \exists e \exists f (d \leq e \vee f \;\&\; e \leq a \Rightarrow_{\prec} b \;\&\; f \leq a \Rightarrow_{\prec} c) & \text{Prop. 4.3(i)} \\
\text{iff} & d \leq a \Rightarrow_{\prec} (b \vee c) \implies d \leq (a \Rightarrow_{\prec} b) \vee (a \Rightarrow_{\prec} c) & \text{compactness} \\
\text{iff} & a \Rightarrow_{\prec} (b \vee c) \leq (a \Rightarrow_{\prec} b) \vee (a \Rightarrow_{\prec} c) & \text{Prop. 3.1(ii)}
\end{array}$$

Notice that the condition above would *not* qualify as a Kracht formula when the intended target propositional language is the one introduced in [7]; this is because condition (vii) of Definition 9.1 requires all occurrences of $\overline{a}$-type variables in η and ζ be *displayable* when translated as inequalities (i.e. they occur in isolation on one side of the inequality). This requirement is satisfied when translating $d \wedge a \prec b \vee c$ as $d \leq a \Rightarrow_{\prec} (b \vee c)$ as we did above; however, when targeting the language of [7], the atomic formula $d \wedge a \prec b \vee c$ can be translated either as $\Diamond(d \wedge a) \leq b \vee c$ or as $d \wedge a \leq \blacksquare(b \vee d)$, and in each case, since conjunction is not in general residuated in the language of (modal) distributive lattices, that occurrence of d is not displayable, which implies that the general algorithm for computing the equivalent axiom will halt and report failure. A similar argument shows that the condition which was shown to be equivalent to the 'Gödel-Dummet axiom' in Section 5 would also violate item (vii) of Definition 9.1 when the intended target propositional language is the one introduced in [7]. The examples above show that this language contributes to widen the scope of logical/algebraic characterizations of conditions on normative systems in a principled way.

10 Conclusions and future directions

This paper introduces a framework for modeling normative relationships flexibly, where the satisfaction of an obligation depends on the specific context. The addition of slanted implications enables to model flexible dependencies that reflect real-world situations in which obligations and permissions may change based on context, capacity, and other factors.

This work suggests further directions for future research. Firstly, the current approach can be extended to permission systems [17]: it would be interesting to explore how to model contextual dependencies not only on norms but also on permissions, and on the interaction between norms and permissions. Secondly, changing the propositional base, and hence studying these 'normative implications' on (co-)Heyting algebras, Boolean algebras, and modal (e.g. epistemic, temporal) algebras could be valuable. Finally, in Footnote 2, we briefly mentioned the possibility of introducing generalized implications associated with relations $\prec$ with weaker properties than subordination relations. Exploring these settings is another interesting direction.

Appendix

A More examples of correspondence and inverse correspondence

In section, we collect some more examples of correspondence and inverse correspondence on slanted Heyting and co-Heyting algebras.

The following axiom is the normative counterpart of the weak excluded middle:

$$\top \leq \neg a \vee \neg\neg a$$

iff	$k \leq \neg a \implies \top \leq \neg a \vee \neg k$	denseness
iff	$c \leq \neg a \implies \top \leq \neg a \vee \neg c$	compactness
iff	$c \leq \neg a \implies \exists d \exists e(\top \leq d \vee e \;\&\; d \leq \neg a \;\&\; e \leq \neg c)$	compactness
iff	$c \wedge a \prec \bot \implies \exists d \exists e(\top \leq d \vee e \;\&\; d \wedge a \prec \bot \;\&\; e \wedge c \prec \bot)$	Lemma 7.2(i)

This condition can be interpreted as, for any propositions a and c, if they give normative inconsistency together, then there exist propositions d and e s.t. d or e is always the case, and both d and a and e and c together give normative inconsistencies.

The following axiom is the normative counterpart of the Kreisel-Putnam axiom:

$$\sim a \Rightarrow_\prec (b \vee c) \leq (\sim a \Rightarrow_\prec b) \vee (\sim a \Rightarrow_\prec c)$$

iff	$d \leq \sim a \Rightarrow_\prec (b \vee c) \implies d \leq (\sim a \Rightarrow_\prec b) \vee (\sim a \Rightarrow_\prec c)$	Prop. 3.1(ii)
iff	$d \leq \sim a \Rightarrow_\prec (b \vee c) \implies$ $\exists e \exists f(d \leq e \vee f \;\&\; e \leq (\sim a \Rightarrow_\prec b) \;\&\; f \leq (\sim a \Rightarrow_\prec c))$	compactness
iff	$d \leq g \Rightarrow_\prec (b \vee c) \;\&\; \sim a \leq g \implies \exists e \exists f \exists h \exists i(d \leq e \vee f \;\&\; e \leq h \Rightarrow_\prec b$ $\&\; \sim a \leq h \;\&\; f \leq i \Rightarrow_\prec c \;\&\; \sim a \leq i)$	Lemma 4.4(ii)
iff	$d \leq g \Rightarrow_\prec (b \vee c) \;\&\; \top \prec g \vee a \implies \exists e \exists f \exists h \exists i(d \leq e \vee f$ $\&\; e \leq h \Rightarrow_\prec b \;\&\; \top \prec a \vee h \;\&\; f \leq i \Rightarrow_\prec c \;\&\; \top \prec a \vee i)$	Lemma 7.2(i)
iff	$d \wedge g \prec b \vee c \;\&\; \top \prec g \vee a \implies \exists e \exists f \exists h \exists i(d \leq e \vee f$ $\&\; e \wedge h \prec b \;\&\; \top \prec a \vee h \;\&\; f \wedge i \prec c \;\&\; \top \prec a \vee i)$	Prop. 4.3(i)

This condition can be interpreted as, if for all a, b, c and d, some g exists s.t. g or a is an unconditional obligation and d and g together normatively imply b or c, then there exist e, f,h, and i s.t. d implies e or f, a or h and a or i are both unconditional obligations, and e and h together normatively imply b and f and i together normatively imply c.

References

[1] Blackburn, P., M. De Rijke and Y. Venema, *Modal logic*, Cambridge University Press **53** (2001).

[2] Calomino, I., J. Castro, S. Celani and L. Valenzuela, *A study on some classes of distributive lattices with a generalized implication*, Order (2023), pp. 1–21.

[3] Castro, J. E., S. A. Celani and R. Jansana, *Distributive lattices with a generalized implication: Topological duality*, Order **28** (2011), pp. 227–249.

[4] Celani, S., *Quasi-modal algebras*, Mathematica Bohemica **126** (2001), pp. 721–736.

[5] Conradie, W., A. De Domenico, G. Greco, A. Palmigiano, M. Panettiere and A. Tzimoulis, *Unified inverse correspondence for DLE-logics*, arXiv preprint arXiv:2203.09199 (2022).

[6] De Domenico, A., A. Farjami, K. Manoorkar, A. Palmigiano, M. Panettiere and X. Wang, *Subordination algebras as semantic environment of input/output logic*, in: *International Workshop on Logic, Language, Information, and Computation*, Springer, 2022, pp. 326–343.

[7] De Domenico, A., A. Farjami, K. Manoorkar, A. Palmigiano, M. Panettiere and X. Wang, *Correspondence and inverse correspondence for input/output logic and region-based theories of space*, arXiv preprint arXiv:2412.01722 (2024).

[8] De Domenico, A., A. Farjami, K. Manoorkar, A. Palmigiano, M. Panettiere and X. Wang, *Obligations and permissions, algebraically*, arXiv preprint arXiv:2403.03148 (2024).

[9] De Domenico, A., A. Farjami, K. Manoorkar, A. Palmigiano, M. Panettiere and X. Wang, *Obligations and permissions on selfextensional logics*, Synthese **206** (2025).

[10] De Rudder, L. and A. Palmigiano, *Slanted canonicity of analytic inductive inequalities*, ACM Transactions on Computational Logic (TOCL) **22** (2021), pp. 1–41.

[11] Dunn, J. M., M. Gehrke and A. Palmigiano, *Canonical extensions and relational completeness of some substructural logics*, J. Symb. Log. **70** (2005), pp. 713–740. URL `https://doi.org/10.2178/jsl/1122038911`

[12] Gehrke, M. and J. Harding, *Bounded lattice expansions*, Journal of Algebra **238** (2001), pp. 345–371.

[13] Horty, J. F., *Reasons as defaults*, Oxford University Press (2012).

[14] Kracht, M., *Tools and techniques in modal logic*, Elsevier Amsterdam **142** (1999).

[15] Lindahl, L. and J. Odelstad, *The theory of joining-systems*, Handbook of deontic logic and normative systems **1** (2013), pp. 545–634.

[16] Makinson, D. and L. van der Torre, *Input/output logics*, Journal of Philosophical Logic **29** (2000), pp. 383–408.

[17] Makinson, D. and L. van der Torre, *Permission from an input/output perspective*, Journal of Philosophical Logic **32** (2003), pp. 391–416.

[18] Palmigiano, A. and M. Panettiere, *Unified inverse correspondence for LE-logics*, arXiv preprint arXiv:2405.01262 (2024).

Readable Twins of Unreadable Models

Krzysztof Pancerz, Piotr Kulicki, Michał Kalisz [1]

The John Paul II Catholic University of Lublin, Poland

Andrzej Burda

VIZJA University, Warsaw, Poland

Maciej Stanisławski

University of Warmia and Mazury, Olsztyn & MakoLab S.A., Łódz, Poland

Jaromir Sarzyński

University of Rzeszów, Poland

Abstract

Creating responsible artificial intelligence (AI) systems is a key challenge in contemporary AI research. Explainability is one of their essential features. This paper focuses on explainable deep learning (XDL) systems and introduces the concept of readable twins: rough set based imprecise information flow models (IIFMs) corresponding to deep learning models (DLMs). Drawing an analogy with digital twins of physical objects, we propose a complete procedure for transforming a DLM into an IIFM. The approach is illustrated with an example based on a deep learning classifier for handwritten-digit recognition using the MNIST dataset.

Keywords: Readable twin, explainable deep learning, rough set flow graphs, hierarchical clustering, computer vision.

1 Introduction

One of the most pressing challenges in artificial intelligence (AI) is making AI tools human-readable, interpretable, and explainable, and, as a consequence, making AI systems responsible. AI tools are currently, in many cases, reinforced by deep neural networks (DNNs). Therefore, special attention in research on Explainable Artificial Intelligence (XAI) is focused on Explainable Deep Learning (XDL) (cf. [11]).

The techniques related to Deep Learning Models (DLMs) include, among others, model-agnostic techniques (MATs) and model-specific techniques

[1] {kpancerz, kulicki, mkalisz}@kul.pl

(MSTs) (cf. [2]). An explainer in MATs is capable of explaining any model (cf. the LIME technique [12]). An explainer in MSTs is correlated with a given deep learning model (cf. the DeepLIFT technique [13]). We propose a new technique that can be classified as MST: Human Readable Twin Explainer for Deep Learning Models (HuReTEx). The main idea is to transform a deep learning model (DLM) into an imprecise information flow model (IIFM) via a sequential information system (SIS). DLM is primary numerical and machine-readable, whereas IIFM is its symbolic, human-readable twin.

HuReTEx can be treated as a reference to the ideas of digital twins (cf. [7]). While digital twin models combine a physical object with its digital representation in virtual space, readable twin models combine an opaque deep learning model with its interpretable representation. Moreover, the proposed approach refers to twin-systems for XAI (cf. [5]). For a model that is unreadable to humans, its readable twin is built.

Transformation is carried out in the following main stages: (1) Building,training and testing DLM $\rightarrow$ (2) Calculating activations of entities of key layers of DLM $\rightarrow$ (3) Clustering activations of entities $\rightarrow$ (4) Creating SIS $\rightarrow$ (5) Creating IIFM for SIS $\rightarrow$ (6) Visualizing predictions paths in IIFM.

Rough set flow graphs (RSFGs) [10] (used as IIFMs) and triangular norms or co-norms together with evolutionary algorithms (EAs) can be used to mine and visualize the most confident prediction paths in RSFGs explaining decisions proposed by DLMs (see next section).

In our approach, we take into account the activation states of the outputs of individual entities of key layers of DLM, i.e., outputs of filters and neurons. Maps of such activation states are called artifacts generated by individual layers of DLM. The twin model is built on the basis of aggregated artifacts (at specific levels of abstraction) generated by individual layers of DLM for training data. In this case, the explanations given by the model are not generated for individual input cases only, but for the problem in general.

According to the taxonomy of the XAI methods given in [1], [14], our approach falls within the category of model-specific, post-hoc explainable deep learning systems, in which models are explained at a global level. The novelty of the approach lies in modelling information flow in DLMs using RSFGs, as well as mining the most confident prediction paths described by these RSFGs using EAs.

2 From Unreadable Models to Their Readable Twins

A readable twin model has the form of a rough set flow graph (RSFG). The elements of RSFG correspond to the original deep learning model (DLM), as shown in Table 1. It is worth noting that we have omitted the input layer and the flatten layer because they are not trained in deep learning models and therefore they do not acquire knowledge that could be used in the explanation process. An illustrative example is presented for a simple deep learning model built using the Keras library [3] to classify images from the MNIST database of handwritten digits [4].

Deep Learning Model (DLM)	Rough Set Flow Graph (RSFG)
A convolutional layer of DLM	A node layer of RSFG
Clusters of artifacts generated by filters of a convolutional layer of DLM	Nodes in a layer of RSFG
A dense layer of DLM	A node layer of RSFG
Clusters of artifacts generated by neurons of a dense layer of DLM	Nodes in a layer of RSFG

Table 1
DLM vs. RSFG.

2.1 Step 1: Building, training, and testing a deep learning model (DLM) on training cases (images)

That process is carried out in a standard way.

2.2 Step 2: Calculating and clustering activations generated by key layers of DLM for training cases (images)

The activations of the model entities for each image in the training dataset are computed and then clustered. Agglomerative clustering is a good choice, as it produces a hierarchical cluster structure that can be readily mapped onto a hierarchical conceptual knowledge representation of artifacts. In the example shown in Figure 1, the cluster identifiers corresponding to the artifacts generated for images from the training dataset are presented.

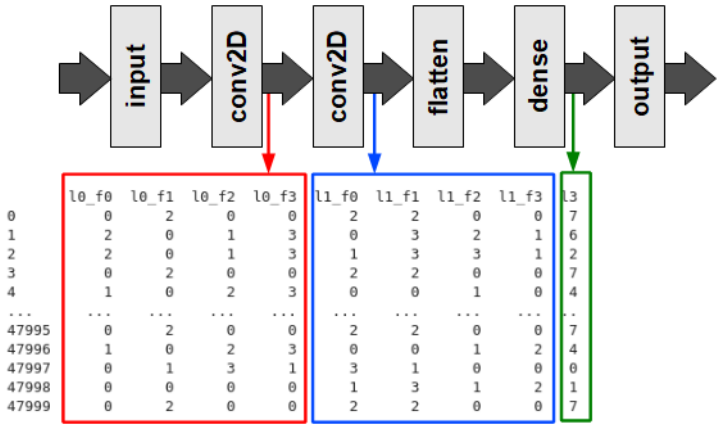
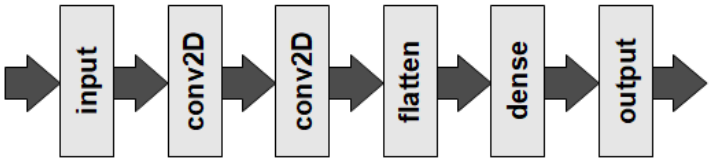


Fig. 1. Creation of a readable twin for an unreadable original model: Step 2.

2.3 Step 3: Creating a sequential information system (SIS)

The underlying information for the generation of a rough set flow graph is arranged in the form of a sequential information system (see an example in Figure 2), that is, an information system (called in [9] a knowledge representation system) with an ordered set of attributes (presented in columns). For

convolutional layers, attribute values are tuples of the identifiers of clusters, for dense layers – identifiers of clusters, and for an output layer – labels assigned to training cases (images).



	conv1	conv2	dense1	output
0	0_2_0_0	2_2_0_0	7	7
1	2_0_1_3	0_3_2_1	6	3
2	2_0_1_3	1_3_3_1	2	5
3	0_2_0_0	2_2_0_0	7	7
4	1_0_2_3	0_0_1_0	4	6
...	...	...	...	...
47995	0_2_0_0	2_2_0_0	7	7
47996	1_0_2_3	0_0_1_2	4	6
47997	0_1_3_1	3_1_0_0	0	1
47998	0_0_0_0	1_3_1_2	1	2
47999	0_2_0_0	2_2_0_0	7	7

Fig. 2. Creation of a readable twin for an unreadable original model: Step 3.

2.4 Step 4: Creating a rough set flow graph (RSFG)

RSFG has a layered structure. Each layer corresponds to one attribute (column) of SIS obtained in Step 3. The nodes in a given layer represent the values of a given SIS's attribute. An imprecise information flow between nodes is described by three coefficients (certainty, covering and strength) assigned to edges connecting nodes. The certainty of a given edge connecting nodes n and n' determines how many times a transition occurs from node n to node n' relative to all transitions from node n to other nodes. The covering of a given edge connecting nodes n and n' determines how many times there is a transition from node n to node n' relative to all transitions to node n' from other nodes. The strength of a given edge connecting nodes n and n' determines how many times there is a transition from node n to node n' out of all transitions between the layer containing n and the layer containing n'. A sample fragment of RSFG is shown in Figure 3.

2.5 Step 5: Determining confident prediction paths in RSFG

By a path in RSFG we mean a sequence of edges connecting a sequence of nodes in the graph such that a path starts at one of the nodes in the first layer and ends at one of the nodes in the last layer. An example path is marked in the RSFG shown in Figure 4. In general, we can distinguish a large number of paths in RSFG. The number of all possible paths in a particular RSFG can be estimated by $\prod_{k=1}^{m} |N_k|$, where m is the number of layers in RSFG, $|N_k|$ is the number of nodes in the k-th layer of RSFG.

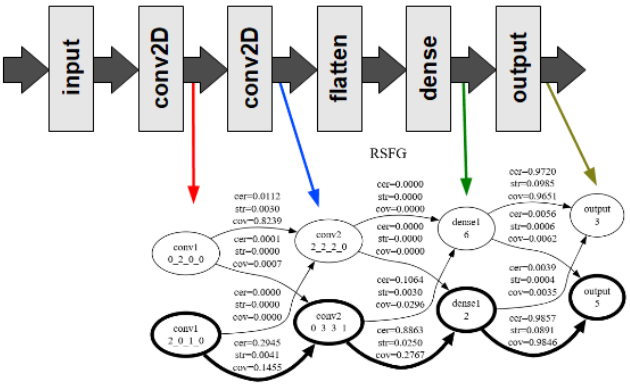
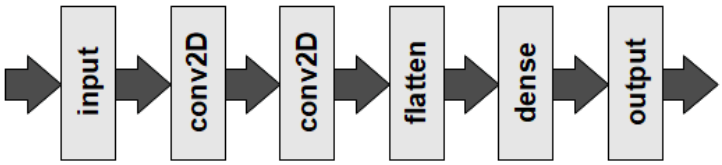


Fig. 3. Creation of a readable twin for an unreadable original model: Step 4.

In the presented approach, we propose to use the evolutionary algorithm (EA) to mine the most important (confident) paths. In EA, chromosomes are sequences of nodes in consecutive layers of RSFG. The fitness function is defined on the basis of confidence (a harmonic mean of certainty and covering) of edges between nodes in sequences represented by chromosomes. The goal is to find the sequences of nodes with the highest possible aggregated confidence value. To aggregate confidences of individual edges between nodes in sequences, triangular norms or co-norms [6] are used. A sample of the confident prediction path found by EA is shown in Figure 4.



	Confidence	Source	Target
conv1D → conv2D	0.19476	2_0_1_0	0_3_3_1
conv2D → dense	0.421719	0_3_3_1	2
dense → output	0.985136	2	5

Fig. 4. Creation of a readable twin for an unreadable original model: Step 5.

2.6 Step 6: Visualizing confident prediction paths

Visualization of confident prediction paths is the quintessence of the presented approach. A sample of the visualization of the confident prediction path is shown in Figure 5. The aggregated artifacts generated by the filters were associated with histograms indicating how many cases (images) of each class were assigned to a given cluster. The information presented in the histograms can be used to understand which prediction classes (digits) are confused by the

model. A visual summary of one of the most confident paths is presented in Figure 6. One of the most confident paths includes: aggregated artifacts generated by four filters in the first convolutional layer for input cases belonging to given clusters (top row of images, above the first red arrow), aggregated artifacts generated by four filters in the second convolutional layer for input cases belonging to given clusters (bottom row of images, under the first red arrow), aggregated artifacts generated by the dense layer for input cases belonging to given clusters (the image behind the second red arrow), and the prediction of the model for this path (digit 5).

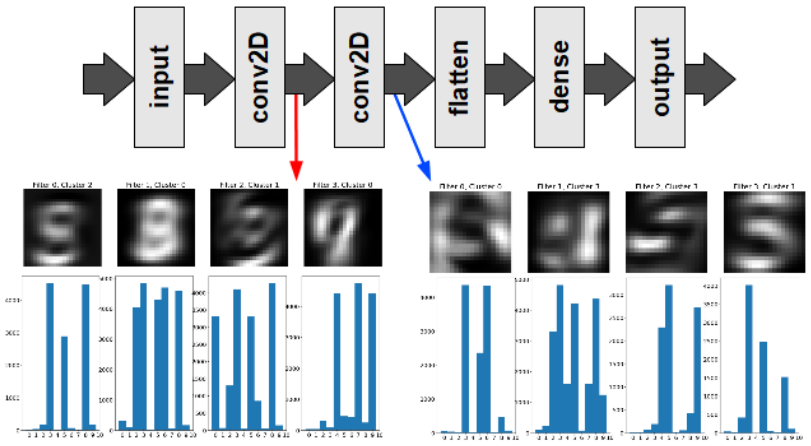


Fig. 5. Creation of a readable twin for an unreadable original model: Step 6.

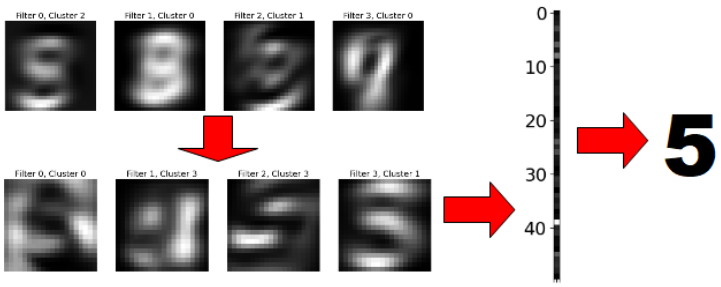


Fig. 6. Visual summarization of one of the most confident paths.

3 Conclusions

We have shown how to use rough set flow graphs to model imprecise information flow in sequential deep learning models and explain what the model has learned. This can be considered as a clear representation of the knowledge acquired by the model, as opposed to the knowledge hidden in the weights and coefficients of the original deep learning model.

Testing the approach on diverse datasets and models is needed to demonstrate the generality of the method and to develop strategies to determine the details of generating HuReTEx (e.g., alternative values of k, alternative t-norms, greedy vs. EA). In further research, we propose to incorporate ontologies of artifacts generated by model layers. Thanks to this, it will be possible to implement the following path: hierarchical clustering of artifacts, hierarchical structure of concepts describing clusters, and conceptual knowledge structure of artifacts. It is also planned to use other information flow models, such as high-level Petri nets, e.g. Petri nets over ontological graphs [8].

Acknowledgents

We would like to thank the anonymous reviewers for their valuable comments and suggestions regarding our work.

References

[1] Angelov, P. P., E. A. Soares, R. Jiang, N. I. Arnold and P. M. Atkinson, *Explainable artificial intelligence: an analytical review*, WIREs Data Mining and Knowledge Discovery **11** (2021), p. e1424.

[2] Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, *Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion **58** (2020), pp. 82–115.

[3] Chollet, F. et al., *Keras*, `https://keras.io` (2015).

[4] Deng, L., *The MNIST database of handwritten digit images for machine learning research*, IEEE Signal Processing Magazine **29** (2012), pp. 141–142.

[5] Kenny, E. M. and M. T. Keane, *Twin-Systems to Explain Artificial Neural Networks using Case-Based Reasoning: Comparative Tests of Feature-Weighting Methods in ANN-CBR Twins for XAI*, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (2019), pp. 2708–2715.

[6] Klement, E. P., R. Mesiar and E. Pap, "Triangular Norms," Springer, Dordrecht, 2000.

[7] Liu, M., S. Fang, H. Dong and C. Xu, *Review of digital twin about concepts, technologies, and industrial applications*, Journal of Manufacturing Systems **58** (2021), pp. 346–361.

[8] Pancerz, K., *A Python toolkit for dealing with Petri nets over ontological graphs*, `https://arxiv.org/abs/2504.08006` (2025).

[9] Pawlak, Z., "Rough Sets. Theoretical Aspects of Reasoning about Data," Kluwer Academic Publishers, Dordrecht, 1991.

[10] Pawlak, Z., *Flow graphs and data mining*, in: J. F. Peters and A. Skowron, editors, *Transactions on Rough Sets III*, Springer-Verlag, Berlin Heidelberg, 2005 pp. 1–36.

[11] Ras, G., N. Xie, M. van Gerven and D. Doran, *Explainable deep learning: A field guide for the uninitiated*, J. Artif. Int. Res. **73** (2022).

[12] Ribeiro, M. T., S. Singh and C. Guestrin, *"Why Should I Trust You?": Explaining the predictions of any classifier*, in: *Proceedings of the 22nd ACM SIGKDD*, KDD '16, 2016, pp. 1135–1144.

[13] Shrikumar, A., P. Greenside and A. Kundaje, *Learning important features through propagating activation differences*, `https://arxiv.org/abs/1704.02685` (2019).

[14] Speith, T., *A review of taxonomies of explainable artificial intelligence (XAI) methods*, in: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), p. 2239–2250.

Specification, Application, and Operationalization of a Metamodel of Fairness

Julian Alfredo Mendez [1]

Umeå University
Sweden

Timotheus Kampik [2]

Umeå University
Sweden

Abstract

This paper presents the AR fairness metamodel, aimed at formally representing, analyzing, and comparing fairness scenarios. The metamodel provides an abstract representation of fairness, enabling the formal definition of fairness notions. We instantiate the metamodel through several examples, with a particular focus on comparing the notions of equity and equality. We use the Tiles framework, which offers modular components that can be interconnected to represent various definitions of fairness. Its primary objective is to support the operationalization of AR-based fairness definitions in a range of scenarios, providing a robust method for defining, comparing, and evaluating fairness. Tiles has an open-source implementation for fairness modeling and evaluation.

Keywords: Metamodel of Fairness, Formalization of Fairness, Resource Distribution, Responsible Artificial Intelligence

1 Introduction

Fairness is a critical consideration in various domains, including social policy, economics, and technology. Despite its importance, defining and evaluating fairness remains a complex challenge, as assessments of what is fair may vary across contexts and stakeholders. There is no universal definition of fairness, and even seemingly purely technical decisions can have direct fairness implications [3]. Given a specific scenario, a definition of fairness can be addressed by defining a *fairness measure* that measures the fair distribution of resources among agents. Although fairness measures are subjective, they must be well defined in critical contexts. This makes it essential to systematize how these

[1] ORCID: 0000-0002-7383-0529, julian.mendez@cs.umu.se

[2] ORCID: 0000-0002-6458-2252, tkampik@cs.umu.se

measures are defined, from an abstract subjective understanding to concrete execution, and to support comparing fairness definitions.

This paper introduces the AR fairness metamodel, designed to represent and analyze fairness measures and scenarios. The metamodel extends the previous research in [26] and serves as a model of models [29], where each model is an instance of the metamodel. The specific instances of these models can then be evaluated to verify whether they comply with a given definition of fairness. The metamodel addresses the challenges of defining and evaluating fairness by offering a structured approach. It provides an abstract representation of *fairness scenarios*, incorporating key elements such as agents, resources, and attributes, which are essential components for evaluating whether a given outcome adheres to a specific definition of fairness.

We use Tiles [26], a framework designed to support the AR fairness metamodel. Tiles consists of modular blocks, called *tiles*, that can be interconnected to specify a definition of fairness. Each block is annotated to indicate how it can be connected to other blocks within the framework. The combination of the Tiles framework and the AR fairness metamodel provides a comprehensive set of tools to model fairness and evaluate fairness in various scenarios, and can be applied to real-world situations.

This paper is organized as follows. Section 2 provides an overview of computational models of fairness. Section 3 introduces the AR fairness metamodel and its components, including the definition of identifiers, measures, attributes, and auxiliary functions. Section 4 discusses the structure of the blocks and their graphical notation. Section 5 offers a discussion of the AR fairness metamodel and the Tiles framework, focusing on their capabilities and limitations. Finally, Section 6 concludes with reflections and a discussion of future work.

2 Background

The importance of fairness in machine learning and artificial intelligence (AI) systems is widely recognized. From a modeling perspective, evaluating fairness requires the ability to identify and quantify unwanted bias, which may lead to prejudice and ultimately to discrimination.

Formalizing fairness can lead to greater transparency in achieving equitable outcomes, which benefits both individuals and the groups they represent. Although operationalizing fairness is challenging, efforts to formalize it and automate fairness verification [1,2] are relevant. Several quantifiable definitions have been proposed [13,17,20,21], reflecting legal, philosophical, and social perspectives. However, different interpretations can inadvertently harm the groups they aim to protect [11] or do not account for intersectionality [21].

Two widely discussed formalizations are *individual fairness* and *group fairness*. Individual fairness requires that similar individuals, based on non-protected attributes, receive similar outcomes. Group fairness stipulates that protected groups should receive similar outcomes when non-protected factors are equal [10]. These notions can conflict [7]. For example, if two individuals with similar qualifications receive different outcomes solely because they

belong to different protected groups, then group fairness metrics such as equality of odds or equality of opportunity can be used to address the disparity. In practice, reconciling these notions and managing the associated value trade-offs remains an active research challenge [14,15,3]. Model-based methodologies, such as MBFair [27], enable the verification of software designs with respect to individual fairness.

Operational tools for fairness assessment include IBM's AI Fairness 360 [6] (AIF360), Microsoft's Fairlearn [9], and Google's What-if Tool [30] (WIT). AIF360 is a comprehensive and technical open-source Python library that includes fairness metrics and bias mitigation algorithms. It works in the three stages: pre-processing, in-processing, and post-processing. Fairlearn includes fairness metrics, bias mitigation algorithms, and also provides fairness dashboards for visual comparisons. WIT is visualization-oriented and provides a dashboard to explore counterfactuals to answer the question "What if this feature changed?". However, these tools focus on the operationalization of fairness measures rather than their definition and analysis. To address this limitation, we propose a unified metamodel that supports various perspectives of fairness, building on the frameworks ACROCPoLis [4] and AcROMAgAt [26]. We aim to integrate multiple definitions of fairness into a coherent structure, facilitating consistent evaluation and comparison between scenarios.

3 Fairness Metamodel

This section presents AR, a formal metamodel from which fairness definitions can be instantiated. Conceptually, AR focuses on *agents*, *resources*, their attributes as first-class abstractions, and the *outcomes*. The presentation of AR is accompanied by several examples that demonstrate its applicability to the assessment of fairness in specific scenarios, as well as the more abstract comparison of fairness measures.

3.1 Basic Elements

A fairness scenario provides the building blocks for relating agents, resources, and their attributes. This relation, called *outcome*, is used to evaluate whether it adheres to a defined concept of fairness. As a prerequisite, we assume two finite background sets, one of (not further specified) *agents*, denoted by $\mathcal{A}$, and one of (not further specified) *resources*, denoted by $\mathcal{R}$. We assume that the two sets are disjoint, i.e., $\mathcal{A} \cap \mathcal{R} = \emptyset$. The metamodel is defined as follows.

Definition 3.1 [Fairness Scenario] A *fairness scenario* is a tuple $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A_{at}}, \mathsf{R_{at}} \rangle$, such that:

- $\mathsf{A} \subseteq \mathcal{A}$; $\mathsf{R} \subseteq \mathcal{R}$; A and R are non-empty;
- every *agent attribute* $\mathsf{a_{AT}} \in \mathsf{A_{at}}$ is a function that takes an agent as input; every *resource attribute* $\mathsf{r_{AT}} \in \mathsf{R_{at}}$ is a function that takes a resource as input; the codomains of agent attributes and resource attributes may vary and are specified upon instantiation.

Note that in our metamodel, we exclude functions that operate on multiple

agents, multiple resources, or combinations of them.

A consistent definition of quantities is crucial for measuring fairness.

Notation 3.1 (Sets of quantities) *The set* M *is a placeholder for a set of quantities such as the set of real numbers (*$\mathbb{R}$*), rational numbers (*$\mathbb{Q}$*), integers (*$\mathbb{Z}$*), or natural numbers including 0 (*$\mathbb{N}_0$*), with its operations totally defined on* M*.*

Relevant attributes are, for example:

i) the utility function $u : \mathsf{R} \to \mathsf{M}$, which returns how much a resource is worth, and

ii) the need function $q : \mathsf{A} \to \mathsf{M}$, which returns how much of a resource is needed by an agent.

Given a fairness scenario, we can define fairness measures. Observe that we denote the *power set* of a set S by $\mathcal{P}(S)$.

Definition 3.2 [Fairness Measure] Let $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A}_{\mathsf{at}}, \mathsf{R}_{\mathsf{at}} \rangle$ be a fairness scenario. An *outcome* O is an element $O \in \mathcal{P}(\mathsf{A} \times \mathsf{R})$, and when the outcome O is clear from context, we may say that *a receives b* whenever $\langle a, b \rangle \in O$. A *fairness measure* $\tilde{f}_{\mathsf{F}}$ with respect to a fairness scenario F is a function $\tilde{f}_{\mathsf{F}} : \mathcal{P}(\mathsf{A} \times \mathsf{R}) \to [0, 1]$.

Since $\{0, 1\}$ is isomorphic to $\mathbb{B}$, which is $\{false, true\}$, we especially consider the case when $\tilde{f}_{\mathsf{F}}(O)$ returns only 0 or 1. Our interpretation is that 0 corresponds to $false$ and 1 to $true$, i.e. if $\tilde{f}_{\mathsf{F}}(O) = 1$, the outcome is fair, and if $\tilde{f}_{\mathsf{F}}(O) = 0$, it is unfair. Generally, we use $\{false, true\}$ and $\{0, 1\}$ interchangeably.

Let us take a look at how the definition of a fairness measure can be applied.

Notation 3.2 (Notation of Functions by Extension) *For a function* $f : A \to B$*, we denote* f *as a set of pairs* $\langle x, y \rangle$ *such that* x *ranges on the elements of* A *exactly once and* $y = f(x)$*. We also use the Iverson bracket notation, where* $f(x) = [P(x)]$ *denotes that* $f(x) = 1$ *if* $P(X)$*, and* $f(x) = 0$ *otherwise.*

The example below illustrates how our fairness metamodel can be instantiated.

Example 3.3 [Fairness Scenario] A group of agents A—namely Alice (A), Bob (B), Carol (C), David (D), Eve (E), and Frank (F)—apply for a subsidy. Let us assume that there are three types of resources R_1, R_2, and R_3, and their utility u is 10, 20, and 30 respectively. The agents needs are encoded in the function q, where A and D need 10, B and E need 20, and C and F need 30 each. Let us assume that it is *fair* to give everyone at least one of the two best resources. The full instantiation of the fairness scenario and fairness measure is summarized as follows:

$\mathsf{A} = \{A, B, C, D, E, F\}$, $\mathsf{R} = \{R_1, R_2, R_3\}$,

$\mathsf{A}_{\mathsf{at}} = \{q : \mathsf{A} \to \mathbb{N}_0,\ q = \{\langle A, 10\rangle,\ \langle B, 20\rangle,\ \langle C, 30\rangle,\ \langle D, 10\rangle,\ \langle E, 20\rangle,\ \langle F, 30\rangle\}\ \}$

$\mathsf{R_{at}} = \{u : \mathsf{R} \to \mathbb{N}_0,\ u = \{\langle R_1, 10\rangle,\ \langle R_2, 20\rangle,\ \langle R_3, 30\rangle\}\ \},$
$\tilde{f}_{\mathsf{F}}(O) = [\forall a \in \mathsf{A}\ (a \text{ receives } R_2 \text{ or } a \text{ receives } R_3)]$

A and R contain the agents and resources respectively, q specifies how much each agent needs, and u specifies the utility of each resource. An intuition of $\tilde{f}_{\mathsf{F}}(O)$ is that every agent receives R_2 or R_3 (or both).

Considering two different outcomes: $O_1 = \{\langle A, R_3\rangle,\ \langle B, R_3\rangle,\ \langle C, R_3\rangle,\ \langle D, R_3\rangle,\ \langle E, R_3\rangle,\ \langle F, R_3\rangle\}$ and $O_2 = \{\langle A, R_3\rangle,\ \langle B, R_2\rangle,\ \langle C, R_1\rangle,\ \langle D, R_3\rangle,\ \langle E, R_2\rangle,\ \langle F, R_1\rangle\}$, we can see that $\tilde{f}_{\mathsf{F}}(O_1) = 1$: O_1 is fair; in contrast, $\tilde{f}_{\mathsf{F}}(O_2) = 0$: O_2 is unfair.

The notion of fairness applied in the example does not consider the *needs* of the agents. We present fairness measures that consider needs further below.

3.2 An Analysis of Equity and Equality

Let us demonstrate how the fairness metamodel can be applied to formalize and compare two well-known fairness measures: *equality* and *equity*. In the case of equality, every agent receives exactly the same amount of resources.

Definition 3.4 [Equality, Equity, and Strict Equity]

Let $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A_{at}}, \mathsf{R_{at}} \rangle$ be a fairness scenario, let $u \in \mathsf{R_{at}}$ be a utility function, and O an outcome. The *accumulation of received resources* $r_O : \mathsf{A} \to \mathsf{M}$ is:

$$r_O(a) = \sum_{\langle a,b\rangle \in O} u(b)\ . \tag{1}$$

This function sums up the utility accumulated by an agent.

- The *equality* fairness measure $\tilde{f}_{\mathsf{F}_{eqa}}$ is

$$\tilde{f}_{\mathsf{F}_{eqa}}(O) = [\forall a, a' \in \mathsf{A} \text{ it holds that } r_O(a) = r_O(a')].$$

- If $q \in \mathsf{A_{at}}$ is the need function, the *equity* fairness measure $\tilde{f}_{\mathsf{F}_{eqi}}$ is

$$\tilde{f}_{\mathsf{F}_{eqi}}(O) = [\forall a \in \mathsf{A} \text{ it holds that } r_O(a) \geq q(a)].$$

- The *strict equity* fairness measure $\tilde{f}_{\mathsf{F}_{seqi}}$ is

$$\tilde{f}_{\mathsf{F}_{seqi}}(O) = [\forall a \in \mathsf{A} \text{ it holds that } r_O(a) = q(a)].$$

Note in Definition 3.4, equity stipulates that each agent receives at least as much as they need. One can strengthen the notion of *equity* so that it is violated if an agent receives more than they need. We call this alternative notion of equity *strict equity*; it constitutes a special case of equity.

Proposition 3.5 (Strict Equity Implies Equity) *For a fairness scenario* F*, for every outcome* O*, the following implication holds:*

$$\tilde{f}_{\mathsf{F}_{seqi}}(O) = 1 \Rightarrow \tilde{f}_{\mathsf{F}_{eqi}}(O) = 1.$$

Proof. Let $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A_{at}}, \mathsf{R_{at}} \rangle$ be a fairness scenario, and let $q : \mathsf{A} \to \mathsf{M}$ be the need function $q \in \mathsf{A_{at}}$. If $\tilde{f}_{\mathsf{F}seqi}(O) = 1$, then by definition $\forall a \in \mathsf{A}$ $(r_O(a) = q(a))$. Consequently, $\forall a \in \mathsf{A}$ $(r_O(a) \geq q(a))$ must hold as well, and therefore $\tilde{f}_{\mathsf{F}eqi}(O) = 1$. □

More interestingly, we can show that equality can be reduced to strict equity by stipulating that the equally distributed available resources are sufficient to satisfy the needs. In other words, if identical amounts are distributed to every agent and each one requires the same amount, then it is possible to *reduce* a fairness scenario to another one containing a need function to satisfy the needs of all the agents. As a prerequisite, we fix background sets of fairness scenarios $\mathcal{F}$ and outcomes $\mathcal{O}$.

Proposition 3.6 (Equality Reduced to Strict Equity) *For every fairness scenario* F *and outcome* O*, there exists a function* $\tau : \mathcal{F} \times \mathcal{O} \to \mathcal{F}$*, such that*

$$\tilde{f}_{\mathsf{F}eqa}(O) = 1 \iff \tilde{f}_{\tau(\mathsf{F},O)seqi}(O) = 1.$$

Proof. Let $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A_{at}}, \mathsf{R_{at}} \rangle$ be a fairness scenario and O an outcome. Choose an arbitrary element $a_0 \in \mathsf{A}$, which is well-defined because A is non-empty and define the function $q : \mathsf{A} \to \mathsf{M}$ as follows: $q(a) := r_O(a_0)$ (i.e., q is fixed, independently of input a). Also, let $\tau(\mathsf{F}, O) = \langle \mathsf{A}, \mathsf{A_{at}} \cup \{q\}, \mathsf{R}, \mathsf{R_{at}} \rangle$.

($\Rightarrow$) Assume that $\tilde{f}_{\mathsf{F}eqa}(O) = 1$ and choose $a, a' \in \mathsf{A}$. Then, by definition it holds that $r_O(a) = r_O(a')$, specifically $r_O(a) = r_O(a_0)$. By definition of q, it holds that $q(a) = r_O(a_0)$. Since a is arbitrary, we have $\forall a \in \mathsf{A}$, $r_O(a) = q(a)$ and therefore $\tilde{f}_{\tau(\mathsf{F},O)seqi}(O) = 1$.

($\Leftarrow$) Assume that $\tilde{f}_{\tau(\mathsf{F},O)seqi}(O) = 1$ and choose $a \in \mathsf{A}$. Then, by definition it holds that $r_O(a) = q(a)$. By definition of q, it holds that $q(a) = r_O(a_0)$. Since a is arbitrary, $\forall a, a' \in \mathsf{A}$ we have $r_O(a) = r_O(a')$, and therefore $\tilde{f}_{\mathsf{F}eqa}(O) = 1$.□

All agents need what an arbitrary agent receives. Only if all agents receive the same, then their needs are strictly met.

3.3 Preferences

In the examples above, we use quantitative functions to evaluate fairness in the distribution of resources among agents. However, we can consider qualitative functions for that purpose as well. In fact, attributes can be used to determine preferences and thus define fairness measures. Note that qualitative functions can also be used to group agents into categories.

Definition 3.7 [Ordinal Preference Function] An *ordinal preference function* v is a function $v : \mathsf{A} \to \text{Perm}(\mathsf{R})$, where $\text{Perm}(\mathsf{R}) = \{(b_1, b_2, \ldots, b_n) \mid \{b_1, b_2, \ldots, b_n\} = \mathsf{R}\}$, such that for each agent, it defines a ranking (a strict total preference order) over all resources, ordered from the most preferred to the least preferred. The notation $b_1 \succ_a b_2$ indicates that b_1 precedes b_2 in $v(a)$, and is read as "a prefers b_1 over b_2". Note that this particular modeling does not allow for *ties*, i.e., situations in which an agent is indifferent between two resources. To accommodate ties, the definition of ranking can be relaxed to a

weak order, where every pair is comparable but some pairs may be considered equally good.

Example 3.8 [Ordinal Preferences] Assume that A is a set of agents $\mathsf{A} = \{A, B, C, D, E\}$, and each agent can receive a jacket that is *large* (L) or *small* (S), i.e., $\mathsf{R} = \{L, S\}$. We allow each agent to choose different jackets in some order of preference. To model each agent's preference, we define the attribute $v : \mathsf{A} \rightarrow \text{Perm}(\mathsf{R})$, $v = \{\langle A, (S, L)\rangle, \langle B, (S, L)\rangle, \langle C, (L, S)\rangle, \langle D, (L, S)\rangle, \langle E, (L, S)\rangle\}$.

In this case, B would be *satisfied* with a large jacket but would *prefer* a small one, C prefers a large jacket but would accept a small one. The fairness measure can be that every agent receives at least one of the resources in the preferences, which can be written as

$$\tilde{f}_{\mathsf{F}}(O) = [\forall a \in \mathsf{A}, \exists b \in \mathsf{R} \text{ s.t. } a \text{ receives } b].$$

The agents can be satisfied with the following outcome: $O = \{\langle A, S\rangle, \langle B, L\rangle, \langle C, S\rangle, \langle D, L\rangle, \langle E, L\rangle\}$, but this does not prevent that C envies B's jacket and that B envies C's jacket. If they exchange their jackets: $O = \{\langle A, S\rangle, \langle B, S\rangle, \langle C, L\rangle, \langle D, L\rangle, \langle E, L\rangle\}$, no agent envies another agent's jacket. The concept in which no agent envies the outcome of another agent is called *envy-freeness* [16,5,28,23]. We write a new fairness measure $\tilde{f}_{\mathsf{F}}(O)$ considering what we call a *weak* envy-freeness, where each agent receives at least one resource that it most preferred than any other resource received by any other agent.

$\mathsf{A} = \{A, B, C, D, E\}$, $\mathsf{R} = \{L, S\}$,

$\mathsf{A}_{\mathsf{at}} = \{v : \mathsf{A} \rightarrow \text{Perm}(\mathsf{R}),\ v = \{\langle A, (S, L)\rangle, \langle B, (S, L)\rangle, \langle C, (L, S)\rangle, \langle D, (L, S)\rangle, \langle E, (L, S)\rangle\}$, $\mathsf{R}_{\mathsf{at}} = \emptyset$,

$\tilde{f}_{\mathsf{F}}(O) = [\ \nexists a, a' \in \mathsf{A}, b' \in \mathsf{R}$, s.t. a' receives b', a does not receive b', and $\forall b \in \mathsf{R}$ s.t. a receives b it holds that $b' \succ_a b$].

Here, v represents the resource preference of each agent and no resource attribute is needed.

The change in outcomes in Example 3.8 is a *Pareto improvement*, because at least one agent is better off without leaving anyone else worse off. An outcome is *Pareto optimal* when no Pareto improvement can be applied.

3.4 Group and Individual Fairness

Group fairness ensures that different demographic groups, which may have protected attributes, such as race and gender, receive similar outcomes. It focuses on statistical parity across groups. *Individual fairness* ensures that similar individuals receive similar outcomes. It emphasizes consistency in treatment based on relevant features, regardless of group membership. Group fairness and individual fairness may seem to be in conflict [8], but we can think of group and individual fairness as complementary. Let us assume that an attribute can be considered either *relevant* or *irrelevant* to determine the distribution of a resource. According to group fairness, if an attribute p is irrelevant, the

groups with attribute p should receive the same amount as the groups without attribute p. According to individual fairness, if an attribute q is relevant, the individuals with a similar value of attribute q should be treated similarly.

We illustrate this with the following example.

Example 3.9 [Group and Individual Fairness] Assume that a group of agents $\mathsf{A} = \{A, B, C, D, E, F\}$ apply for a loan (L), and that there is a protected demographic attribute p, which should be irrelevant for the loan application. Assuming that only D, E, and F have that attribute creates two demographic groups: $G_{\neg p} = \{A, B, C\}$ and $G_p = \{D, E, F\}$.

Following group fairness, those belonging to different demographic groups should be treated similarly, regardless of the group to which they belong. Assuming that 2 out of 3 loan applications are accepted, that relation should appear in $G_{\neg p}$ and also in G_p. For example, if the applications of A, B, E, and F are accepted and the rest (C and D) rejected, group fairness holds in this case.

Following individual fairness, if two applicants have nearly identical values for a critically relevant attribute, they should receive similar treatment. Assume that q is an essential attribute to determine whether to give a loan, and that only B, C, E, and F have this attribute. If D gets the loan and F does not, individual fairness is not observed in this outcome.

We formalize this example as follows:

$$\mathsf{A} = \{A, B, C, D, E, F\},\ \mathsf{R} = \{L\},$$

$$\mathsf{A}_{\mathsf{at}} = \{p : \mathsf{A} \to \mathbb{B}, p = \{\langle A, false\rangle, \langle B, false\rangle, \langle C, false\rangle, \langle D, true\rangle, \langle E, true\rangle, \langle F, true\rangle\},\quad q : \mathsf{A} \to \mathbb{B},\ q = \{\langle A, false\rangle, \langle B, true\rangle, \langle C, true\rangle, \langle D, false\rangle, \langle E, true\rangle, \langle F, true\rangle\}\ \},\ \mathsf{R}_{\mathsf{at}} = \emptyset,$$

$$\varepsilon = 10^{-2},\ \cdot \simeq_\varepsilon \cdot : \mathsf{M} \times \mathsf{M} \to \mathbb{B},$$

$$a \simeq_\varepsilon b = \begin{cases} true, & \text{if } a = b \text{ or } (a \neq b \text{ and } \dfrac{|a-b|}{\max(|a|,|b|)} < \varepsilon) \\ false, & \text{otherwise} \end{cases}$$

$$r_{p^+} = \frac{|\{a \in \mathsf{A} \mid p(a) \text{ and } a \text{ receives } L\}|}{|\{a \in \mathsf{A} \mid p(a)\}|},$$

$$r_{p^-} = \frac{|\{a \in \mathsf{A} \mid \neg p(a) \text{ and } a \text{ receives } L\}|}{|\{a \in \mathsf{A} \mid \neg p(a)\}|},$$

$$G\tilde{f}_{\mathsf{F}}(O) = [r_{p^+} \simeq_\varepsilon r_{p^-}]\ ,$$

$$I\tilde{f}_{\mathsf{F}}(O) = {}_{1:}\ [\forall a, a' \in \mathsf{A}, a \neq a' \text{ and } q(a) = q(a') \Rightarrow$$
$${}_{2:}\quad (a \text{ receives } L \text{ and } a' \text{ receives } L) \text{ or}$$
$${}_{3:}\quad (a \text{ does not receive } L \text{ and } a' \text{ does not receive } L)]\ .$$

Here, p is an irrelevant protected attribute for group fairness, q is an essential attribute for individual fairness, $\cdot \simeq_\varepsilon \cdot$ determines whether two quantities are similar up to a value ε, r_{p^+} and r_{p^-} are the ratios between the number of agents with/without the protected attribute p that receive the loan compared to all that have/do not have p respectively.

We provide two different fairness measures: $G\tilde{f}_{\mathsf{F}}(O)$ for group fairness and $I\tilde{f}_{\mathsf{F}}(O)$ for individual fairness. Group fairness compares the relation between the number of agents with attribute p that receive the loan compared to all the agents having p is similar to the relation between those who receive the loan for the other group. Individual fairness requires that:

1. for every two different agents with the same value of the essential attribute q,
2. either both receive the loan,
3. or neither receives the loan.

3.5 Continuous Fairness Measures

In Definition 3.2, we define $\tilde{f}_{\mathsf{F}}$ to range in the continuous interval $[0, 1]$, but we only have presented discrete examples up to this point. However, we can model a continuous example such as Jain's fairness index [18], which is a quantitative measure for assessing how evenly a resource is allocated among n agents.

Definition 3.10 [Jain's Fairness Index] Given a set of agents indexed from 1 to n, such that each agent receives $x_1, x_2, \ldots, x_n$ respectively, the index is defined (on the left) and rewritten (on the right) as:

$$J(x_1, \ldots, x_n) = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n \cdot \sum_{i=1}^{n} x_i^2} \quad , \quad \tilde{f}_{\mathsf{F}}(O) = \frac{\left(\sum_{a \in \mathsf{A}} r_O(a)\right)^2}{|\mathsf{A}| \cdot \sum_{a \in \mathsf{A}} r_O(a)^2} \ . \tag{2}$$

Jain's fairness index is used to measure fairness in network resource allocation, to evaluate load balancing schemes in distributed systems, and to balance throughput in congestion control protocols. In particular, the index provides a single scalar score that can be used to compare different allocation strategies and to tune parameters for the desired level of fairness.

In the following example, we apply this fairness measure.

Example 3.11 [Continuous Fairness Measure]

Let us assume that agents need to access resources that represent different bandwidth values in a computer network.

$\mathsf{A} = \{A, B, C, D\}$, $\mathsf{R} = \{M_0, M_{10}, M_{20}, M_{50}\}$, $\mathsf{A}_{\mathsf{at}} = \emptyset$,
$\mathsf{R}_{\mathsf{at}} = \{u : \mathsf{R} \to \mathsf{M}, u = \{\langle M_0, 0\rangle, \langle M_{10}, 10\rangle, \langle M_{20}, 20\rangle, \langle M_{50}, 50\rangle\}\ \}$,
$\tilde{f}_{\mathsf{F}}(O) =$ as in Equation (2)

In this case, u represents the utility in megabits per second (Mbps) of each resource.

Considering the following outcomes: $O_1 = \{\langle A, M_{20}\rangle, \langle B, M_{20}\rangle, \langle C, M_{20}\rangle, \langle D, M_{20}\rangle\}$, $O_2 = \{\langle A, M_{20}\rangle, \langle B, M_{20}\rangle, \langle C, M_{20}\rangle, \langle D, M_0\rangle\}$, and $O_3 = \{\langle A, M_0\rangle, \langle B, M_0\rangle, \langle C, M_0\rangle, \langle D, M_{10}\rangle\}$, we obtain that $\tilde{f}_{\mathsf{F}}(O_1) = 1$, $\tilde{f}_{\mathsf{F}}(O_2) = 0.75$, and $\tilde{f}_{\mathsf{F}}(O_3) = 0.25$. We could interpret in words that O_1 is perfectly fair, O_2 is moderately fair, and O_3 is clearly unfair, but in practice the nuances of how fair these values are strongly dependent on context.

A continuous fairness measure that is more closely aligned with *social* (in contrast to *technical*) applications is the Gini index, which we cover in the example below.

Example 3.12 [Complement of the Gini Index] The *Gini index* is a statistical measure of inequality often used to quantify the inequality of income or wealth within a population. The index has also been applied in other contexts, such as decision tree algorithms in machine learning. In the interpretation of the Gini index, the closer the Gini index is to 0, the more equal the distribution, and the closer to 1, the more unequal. This works exactly opposite to Jain's index and our definition for a fairness measure. Because of this, we define the *complement of the Gini index* by inverting the output of the Gini index. The Gini index is defined (on the left) and rewritten (on the right) as:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2 \cdot n \cdot \sum_{i=1}^{n} x_i}, \quad \tilde{f}_{\mathsf{F}}(O) = 1 - \frac{\sum_{a_1 \in \mathsf{A}}\sum_{a_2 \in \mathsf{A}}|r_O(a_1) - r_O(a_2)|}{2 \cdot |\mathsf{A}| \cdot \sum_{a \in \mathsf{A}} r_O(a)} \tag{3}$$

where n is the number of agents (or households) and x_i is the income (or wealth) of agent i.

As in the case of Jain's index in Definition 3.10, we can plug in the agents of set A, such that n is $|\mathsf{A}|$, and $r_O(a)$ is the amount received by each agent a. The agents are the households, and the resource is the wealth to be distributed.

$\mathsf{A} = \{A, B, C, D, E, F\}$, $\mathsf{R} = \{R_5, R_{10}, R_{15}, R_{20}, R_{50}, R_{100}\}$,
$\mathsf{A}_{\mathsf{at}} = \emptyset$, $\mathsf{R}_{\mathsf{at}} = \{u : \mathsf{R} \to \mathsf{M}, u = \{\langle R_5, 5\rangle, \langle R_{10}, 10\rangle, \langle R_{15}, 15\rangle, \langle R_{20}, 20\rangle, \langle R_{50}, 50\rangle, \langle R_{100}, 100\rangle\}\ \}$,
$\tilde{f}_{\mathsf{F}}(O) =$ as in Equation (3)

Here, u is the utility (for example, in thousands of euros) of each resource.

Considering the following outcomes: $O_1 = \{\langle A, R_{20}\rangle, \langle B, R_{20}\rangle, \langle C, R_{20}\rangle, \langle D, R_{20}\rangle, \langle E, R_{20}\rangle, \langle F, R_{20}\rangle\}$, $O_2 = \{\langle A, R_5\rangle, \langle B, R_{10}\rangle, \langle C, R_{15}\rangle, \langle D, R_{20}\rangle, \langle E, R_{50}\rangle, \langle F, R_{100}\rangle\}$, $O_3 = \{\langle A, R_5\rangle, \langle B, R_{10}\rangle\}$, we obtain that $\tilde{f}_{\mathsf{F}}(O_1) = 1 - \frac{0}{1440} = 1$, $\tilde{f}_{\mathsf{F}}(O_2) = 1 - \frac{1200}{2400} = 0.5$, and $\tilde{f}_{\mathsf{F}}(O_3) = 1 - \frac{130}{180} \approx 0.28$. We could interpret in words that O_1 is perfectly fair, O_2 is rather unfair, and O_3 is clearly unfair.

Example 3.13 [Fairness Measure for Equalized Odds] The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system [22] is a well-known example of a risk assessment tool used in the U.S. criminal justice system to predict the likelihood of recidivism (re-offending after a prior arrest). Studies, most notably by ProPublica in 2016 [3], found that COMPAS

[3] `https://github.com/propublica/compas-analysis`

exhibited racial bias: it was more likely to falsely flag Black defendants as future criminals (higher false positive rate) and more likely to misclassify White defendants as low risk (higher false negative rate).

Based on that example, we designed a fairness measure to detect bias in a prediction system. The system gives two possible scores (resources) with their analogous score in COMPAS:

- R_{low} : Low Risk (scores 1–4)
- R_{high} : Medium Risk (scores 5–7) and High Risk (scores 8–10)

The prediction system is defined as follows.

$\mathsf{A} = \{A, B, C, D, E, F\}$, $\mathsf{R} = \{R_{\text{low}}, R_{\text{high}}\}$,
$\mathsf{A}_{\mathsf{at}} = \{p : \mathsf{A} \to \mathbb{B},\ p = \{\langle A, false\rangle, \langle B, false\rangle, \langle C, false\rangle, \langle D, true\rangle, \langle E, true\rangle, \langle F, true\rangle\},\ res : \mathsf{A} \to \mathsf{R},\ res = \{\langle A, R_{\text{low}}\rangle, \langle B, R_{\text{high}}\rangle, \langle C, R_{\text{high}}\rangle, \langle D, R_{\text{low}}\rangle, \langle E, R_{\text{low}}\rangle, \langle F, R_{\text{high}}\rangle\}$,
$\mathsf{R}_{\mathsf{at}} = \{u : \mathsf{R} \to \mathsf{M}, u = \{\langle R_{\text{low}}, 0\rangle, \langle R_{\text{high}}, 1\rangle\}\ \}$,
$\overline{(x_i)_{i=1}^n}$ is the arithmetic mean of the sequence $(x_1, x_2, \ldots, x_n)$

$$corr : \bigcup_{n=1}^{\infty} (\mathsf{M} \times \mathsf{M}) \to [-1, 1],$$

$$corr((x_i)_{i=1}^n, (y_i)_{i=1}^n) = \frac{\sum_{k=1}^{n} (x_k - \overline{(x_i)_{i=1}^n})\,(y_k - \overline{(y_i)_{i=1}^n})}{\sqrt{\sum_{k=1}^{n} (x_k - \overline{(x_i)_{i=1}^n})^2}\ \sqrt{\sum_{k=1}^{n} (y_k - \overline{(y_i)_{i=1}^n})^2}}$$

$$\tilde{f}_{\mathsf{F}}(O) = |corr(\ (\ [\langle x_i, R_{\text{high}}\rangle \in O \wedge res(x_i) \neq R_{\text{high}}]\ )_{i=1}^n, (\ [p(x_i)]\ )_{i=1}^n\ )|$$

Here, p is the protected attribute, such as skin color in COMPAS, res is the ground truth based on facts, and u is the associated risk of each score. We use the Pearson correlation coefficient, but another correlation coefficient can also be used. Recall that the square brackets [ ] follow the Iverson notation and that the fairness measure should only return values in $[0, 1]$.

4 Operationalization

In this section, we show how we operationalize the concepts presented above, that is, how we translate abstract ideas into concrete, formal representations that can be implemented and measured. Essentially, an instance of the AR metamodel provides a formal representation of a fairness scenario $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A}_{\mathsf{at}}, \mathsf{R}_{\mathsf{at}} \rangle$. Assuming that the agent and resource attributes are given, the fairness measure still needs to be defined. As we can see from the examples above, defining a fairness measure can be both complex and error-prone. To address this, we use the Tiles framework [26] as a way of defining the fairness measure for a given fairness scenario. The purpose of this framework is to improve readability and reliability of the fairness measures. To achieve this, we decompose the fairness measure into smaller pieces of composable blocks.

4.1 Design of the Blocks

The design of blocks in Tiles is based on the concept of a *module* in software engineering. Recall that a fairness scenario is a tuple $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A}_{\mathsf{at}}, \mathsf{R}_{\mathsf{at}} \rangle$. We assume that R_{at} contains a function $u : \mathsf{R} \to \mathsf{M}$ *(utility)*, A_{at} contains a function $q : \mathsf{A} \to \mathsf{M}$ *(needs)*, and we define the auxiliary function $r_O : \mathsf{A} \to \mathsf{M}$ *(accumulates)* as in Definition 3.4, which depends on u.

We represent the agents A as a sorted sequence of identifiers $A_{\prec}$, denoted by all-agent. We define accumulates and needs by applying r_O and q, respectively, to each element of the sequence. Additionally, we define all-equal as the result of checking whether all the elements in the sequence are equal, and all-at-least as the result of checking, for each pair in the sequence, if its first component is greater than or equal to the second component.

We define equality with the following pipeline:

$$\tilde{f}_{\mathsf{F}_{eqa}}(O) = \textsf{all-equal}(\textsf{accumulates}(\textsf{all-agent})) \tag{4}$$

$\tilde{f}_{\mathsf{F}_{eqa}}$ in Definition 3.4 is equivalent to $\tilde{f}_{\mathsf{F}_{eqa}}$ given in Equation 4.

Proposition 4.1 (Correctness of the Equality Pipeline) *Let* $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A}_{\mathsf{at}}, \mathsf{R}_{\mathsf{at}} \rangle$ *be a fairness scenario,* $\tilde{f}_{\mathsf{F}_{eqa}}$ *defined as in Definition 3.4, and* $\tilde{f}_{\mathsf{F}_{eqa'}}$ *defined as in* (4)*. Then, for every outcome* O*,*

$$\tilde{f}_{\mathsf{F}_{eqa}}(O) = 1 \iff \tilde{f}_{\mathsf{F}_{eqa'}}(O) = 1 \ .$$

Proof. Applying the expansion of $\tilde{f}_{\mathsf{F}_{eqa}}(O)$ in (4) yields:

$$[\forall m, m' \in (r_O(a))_{a \in \mathsf{A}_{\prec}} \ (m = m') \ ].$$

Notice that A is non-empty and finite, and that $a \in \mathsf{A}$ if and only if $a \in \mathsf{A}_{\prec}$. In particular, the sequence $(r_O(a))_{a \in \mathsf{A}_{\prec}}$ contains, for each agent a, the value $r_O(a)$. In other words, for each value m, $m \in (r_O(a))_{a \in \mathsf{A}_{\prec}}$ if and only if $m \in \{r_O(a) \mid a \in \mathsf{A}\}$.

Then, $\tilde{f}_{\mathsf{F}_{eqa'}}(O)$ holds if and only if (by definition)

$$\forall m, m' \in (r_O(a))_{a \in \mathsf{A}_{\prec}} \ (m = m') \ .$$

In turn, $\tilde{f}_{\mathsf{F}_{eqa'}}(O)$ holds if and only if

$$\forall m, m' \in \{r_O(a) \mid a \in \mathsf{A}\} \ (m = m') \ . \tag{5}$$

Since $m \in \{r_O(a) \mid a \in \mathsf{A}\}$ if and only if there is an $a \in \mathsf{A}$, such that $m = r_O(a)$, we can re-write (5) as

$$\forall a, a' \in \mathsf{A} \ (r_O(a) = r_O(a')),$$

which is the definition of $\tilde{f}_{\mathsf{F}_{eqa}}(O)$. □

We define equity with the following pipeline

$$\tilde{f}_{\mathsf{F}_{eqi}}(O) = \textsf{all-at-least}(\textsf{accumulates}(\textsf{all-agent}), \textsf{needs}(\textsf{all-agent})) \tag{6}$$

Observe that $\tilde{f}_{\mathsf{F}_{eqi}}$ in Definition 3.4 is equivalent to $\tilde{f}_{\mathsf{F}_{eqi}}$ given in (6).

Proposition 4.2 (Correctness of the Equity Pipeline) *Let* $\mathsf{F} = \langle \mathsf{A}, \mathsf{R}, \mathsf{A}_{\mathsf{at}}, \mathsf{R}_{\mathsf{at}} \rangle$ *be a fairness scenario,* $\tilde{f}_{\mathsf{F}_{eqi}}$ *defined as in Definition 3.4,* $\tilde{f}_{\mathsf{F}_{eqi'}}$ *defined as in* (6)*. Then, for every outcome* O*,*

$$\tilde{f}_{\mathsf{F}_{eqi}}(O) = 1 \iff \tilde{f}_{\mathsf{F}_{eqi'}}(O) = 1 \ .$$

Proof. For the case of equity, consider: $X = (r_O(a))_{a \in \mathsf{A}_\prec}$ and $Y = (q(a))_{a \in \mathsf{A}_\prec}$.

Notice that, by construction, X and Y have as many elements as A. Considering that $|X| = |Y|$, $\tilde{f}_{\mathsf{F}_{eqi'}}(O)$ can be expanded as:

$$[\forall i \in \mathbb{N}_0, 1 \leq i \leq |X| \Rightarrow \text{the } i\text{-th element of } (r_O(a))_{a \in \mathsf{A}_\prec} \geq \text{ the } i\text{-th element of } (q(a))_{a \in \mathsf{A}_\prec}] \ .$$

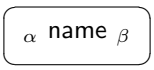
This is 1 if and only if the following holds: $\forall a \in \mathsf{A}\ (r_O(a) \geq q(a))$, which is how $\tilde{f}_{\mathsf{F}_{eqi}}(O)$ is defined in Definition 3.4. □

The functional notation for Tiles can be represented using a graphical notation.

4.2 Graphical Notation

One of the key aspects of Tiles is the ability to represent configurations in a graphical notation that clarifies how a configuration works. The whole configuration is a formal representation of how the blocks are interconnected. Each block, called *tile*, can connect with others to produce a specific definition.

Concept 4.1 (Tile) *A* tile *is a construct that contains a name, a function, an input type, an output type, and contextual information, which includes the fairness scenario, constants, and auxiliary functions. The tile is represented as*



where α *and* β *are type annotations for the input type and the output type respectively, and* name *is the function name. The input type is omitted if the tile represents a constant. A type in* Tiles *can be*

- *atomic:* a *(agent),* r *(resource),* m *(quantity or measure), and* b *(Boolean);*
- *a tuple composed of other types:* $\langle \alpha_1, \ldots, \alpha_n \rangle$*; or*
- *a sequence of a type:* (α)*.*

The type annotation in Tiles can be used not only to specify the connection between the output of a tile and the input of another type, but also to denote an input variable name when the tile has parameters. Parametric tiles are also useful for defining customized tiles.

4.3 Implementation

The Tiles framework is implemented as an open-source project [4] written in the Soda language [24], an open-source functional language [5]. The code in

[4] https://github.com/julianmendez/tiles

[5] https://github.com/julianmendez/soda

Soda can be formally proved using the Lean [12] proof assistant, and seamlessly integrated with the Java Virtual Machine ecosystem, allowing efficient execution [25]. The Tiles implementation written in Soda aims to follow the notation used for the pipelines as accurately as possible, where each tile of the implementation has its source code directly accessible.

The framework includes detailed components that ensure the correct construction of pipelines. These components are particularly focused on zipping and unzipping sequences, as well as creating and projecting tuples. Since Soda is statically typed and Tiles is a typed framework, the type consistency of the entire pipeline can be verified at compile time.

5 Discussion

This section informally discusses the capabilities and limitations of the AR metamodel and the Tiles framework. The framework is designed around the concept of *flow*, where data moves through a pipeline. This pipeline connects tiles, forming a directed acyclic graph with a single start and end point. We do not provide an effective algorithm for constructing pipelines from the first-order formula given by the fairness measure, since it is not possible in the general case.

5.1 Expressiveness

The AR metamodel can model a wide range of fairness scenarios, though not all possible ones. We specifically focus on scenarios that involve attributes of agents and resources. In particular, the metamodel supports the combination of multiple agents and resources, as demonstrated in Example 3.8, where an agent or resource serves as the first parameter of a function that returns a sequence.

5.2 Decidability and Time Complexity

Pipelines built within the Tiles framework are decidable provided that their tiles are decidable. When contracts between tiles are respected, undefined values cannot arise.

With respect to time complexity, the pipeline structure guarantees the absence of loops in the diagram, ensuring that no tile is evaluated more than once. The worst-case complexity of a pipeline is determined by the maximum complexity of its individual tiles.

5.3 Applicability

The Tiles framework has broad applicability beyond the fairness domain. As a software engineering tool, it can handle any scenario that involves finite, iterable sets of identifiers and attributes. The type system in Tiles is flexible and includes sets of identifiers, quantities, Boolean values, tuples, and sequences.

That said, the Tiles framework is not intended to be applied at all levels of a scenario. The framework describes connections and processes to provide a better explanation of complex formulas. However, if the notation of the formulas is clear enough, they should be used instead of the pipelines.

The Tiles framework is primarily intended for modeling, but the generated pipelines are naturally simple to execute in parallel. This is because pipelines are usually designed to process several agents, resources, or quantities at the same time. However, this requires expertise in modeling these pipelines, since parallel execution can be effortful to design.

5.4 Limitations

A clear limitation of our approach is the generic nature of the metamodel. To address this, we introduce Tiles as a layer of modular building blocks. Although Tiles is conceptually elegant, it may pose, similar to other declarative notations, challenges in terms of readability. Studying (and potentially improving) the readability of Tiles can therefore be considered important future work. For example, one could instantiate fairness measures and scenarios in several languages and systematically compare understandability, or one could conduct perceived usefulness studies analogous to [19] (which empirically compares several declarative modeling languages in this regard).

Furthermore, this work does not address modeling *specific* real-world problems: we either use *toy examples* that facilitate a better understanding, or work with generally applicable measures, e.g., to showcase broad applicability of the conceptual level, or to make fundamental observations about their relationship (as in the case of strict equity vs. equality). Future work could evaluate our metamodel and its operationalization in specific case studies, in which real-world users are faced with fairness modeling and analysis challenges.

6 Conclusion

In this paper, we have introduced a metamodel of fairness and demonstrated its application through concrete examples. The metamodel offers a structured approach to understanding and evaluating fairness in various scenarios. It begins by identifying agents, resources, and their relevant attributes, and concludes by the construction of complex fairness measures.

We have provided examples of an application of the metamodel of central fairness concepts, such as equality, equity, group fairness, and individual fairness. We have explored the use of the metamodel spanning from economics and game theory, like preferences and envy-freeness, to a concept vastly applied in computer networks, like Jain's fairness index.

In addition to the multiple examples, we have provided a method for operationalizing the metamodel, which simplifies the formal notation. The graphical and conceptual representation is supported by a modular design. This formal and practical notation for defining functions can also be applied across various domains. Moreover, we provide an implementation of this representation.

Looking ahead, the fairness framework can be extended and revised when exploring its use in real-world scenarios. In this context, it is essential to provide further tooling that simplifies practical modeling.

Acknowledgements
This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

[1] Albarghouthi, A., L. D'Antoni, S. Drews and A. V. Nori, *Fairsquare: Probabilistic verification of program fairness*, Proc. ACM Program. Lang. **1** (2017), https://doi.org/10.1145/3133904.
URL https://doi.org/10.1145/3133904

[2] Albarghouthi, A. and S. Vinitsky, *Fairness-aware programming*, in: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19 (2019), p. 211–219, https://doi.org/10.1145/3287560.3287588.
URL https://doi.org/10.1145/3287560.3287588

[3] Aler Tubella, A., F. Barsotti, R. G. Koçer and J. A. Mendez, *Ethical implications of fairness interventions: what might be hidden behind engineering choices?*, Ethics and Information Technology **24** (2022), p. 12.
URL https://doi.org/10.1007/s10676-022-09636-z

[4] Aler Tubella, A., D. Coelho Mollo, A. Dahlgren Lindström, H. Devinney, V. Dignum, P. Ericson, A. Jonsson, T. Kampik, T. Lenaerts, J. A. Mendez and J. C. Nieves, *ACROCPoLis: A Descriptive Framework for Making Sense of Fairness*, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (2023), p. 1014–1025.
URL https://doi.org/10.1145/3593013.3594059

[5] Amanatidis, G., G. Birmpas and E. Markakis, *Comparing approximate relaxations of envy-freeness* (2018).
URL https://arxiv.org/abs/1806.03114

[6] Bellamy, R. K. E., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney and Y. Zhang, *AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias*, IBM J. Res. Dev. **63** (2019), pp. 4:1–4:15.
URL https://doi.org/10.1147/JRD.2019.2942287

[7] Binns, R., *On the apparent conflict between individual and group fairness* (2019).
URL https://arxiv.org/abs/1912.06883

[8] Binns, R., *On the apparent conflict between individual and group fairness*, in: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (2020), p. 514–524.
URL https://doi.org/10.1145/3351095.3372864

[9] Bird, S., M. Dudík, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach and K. Walker, *Fairlearn: A toolkit for assessing and improving fairness in AI*, Microsoft, Tech. Rep. MSR-TR-2020-32 (2020).

[10] Chouldechova, A., *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, Big Data **5** (2017), pp. 153–163, pMID: 28632438.
URL https://doi.org/10.1089/big.2016.0047

[11] Corbett-Davies, S. and S. Goel, *The measure and mismeasure of fairness: A critical review of fair machine learning*, CoRR **abs/1808.00023** (2018).
URL http://arxiv.org/abs/1808.00023

[12] de Moura, L., *Lean* (2013).
URL https://github.com/leanprover/lean4

[13] Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. Zemel, *Fairness through awareness*, in: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*,

ITCS '12 (2012), p. 214–226, https://doi.org/10.1145/2090236.2090255.
URL https://doi.org/10.1145/2090236.2090255

[14] Dwork, C., M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel, *Fairness through awareness*, in: S. Goldwasser, editor, *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012* (2012), pp. 214–226.
URL https://doi.org/10.1145/2090236.2090255

[15] Fleisher, W., *What's Fair about Individual Fairness?*, in: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21 (2021), p. 480–490.
URL https://doi.org/10.1145/3461702.3462621

[16] Foley, D. K., "Resource allocation and the public sector," Yale Economics Essays, 1966.

[17] Hardt, M., E. Price and N. Srebro, *Equality of opportunity in supervised learning*, in: *Advances in Neural Information Processing Systems*, 2016, pp. 3323–3331.

[18] Jain, R. K., D.-M. W. Chiu and W. R. Hawe, *A Quantitative Measure of Fairness and Discrimination for Resource Allocation in Shared Computer Systems* (1984).

[19] Jalali, A., *Evaluating user acceptance of knowledge-intensive business process modeling languages*, Softw. Syst. Model. **22** (2023), pp. 1803–1826.
URL https://doi.org/10.1007/s10270-023-01120-6

[20] Joseph, M., M. Kearns, J. H. Morgenstern and A. Roth, *Fairness in Learning: Classic and Contextual Bandits*, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon and R. Garnett, editors, *Advances in Neural Information Processing Systems*, NIPS **29** (2016), pp. 325–333.
URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf

[21] Kearns, M., S. Neel, A. Roth and Z. S. Wu, *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*, in: *International Conference on Machine Learning*, PMLR, 2018, pp. 2564–2572.

[22] Larson, J., S. Mattu, L. Kirchner and J. Angwin, *How We Analyzed the COMPAS Recidivism Algorithm* (2016).
URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[23] Li, Z., S. Liu, X. Lu, B. Tao and Y. Tao, *A complete landscape for the price of envy-freeness* (2024).
URL https://arxiv.org/abs/2401.01516

[24] Mendez, J. A., *Soda: An Object-Oriented Functional Language for Specifying Human-Centered Problems* (2023), https://doi.org/10.48550/arXiv.2310.01961.
URL https://doi.org/10.48550/arXiv.2310.01961

[25] Mendez, J. A. and T. Kampik, *Can Proof Assistants Verify Multi-agent Systems?*, in: R. Collier, A. Ricci, V. Nallur, S. Burattini and A. Omicini, editors, *Multi-Agent Systems* (2025), pp. 323–339, https://doi.org/10.1007/978-3-031-93930-3_19.
URL https://doi.org/10.1007/978-3-031-93930-3_19

[26] Mendez, J. A., T. Kampik, A. Aler Tubella and V. Dignum, *A Clearer View on Fairness: Visual and Formal Representation for Comparative Analysis*, in: F. Westphal, E. Peretz-Andersson, M. Riveiro, K. Bach and F. Heintz, editors, *14th Scandinavian Conference on Artificial Intelligence, SCAI 2024*, Swedish Artificial Intelligence Society, 2024, pp. 112–120, https://doi.org/10.3384/ecp208013.
URL https://ecp.ep.liu.se/index.php/sais/article/view/1005/913

[27] Ramadan, Q., M. Konersmann, A. S. Ahmadian, J. Jürjens and S. Staab, *Mbfair: a model-based verification methodology for detecting violations of individual fairness*, Software and Systems Modeling **24** (2025), pp. 111–136.
URL https://doi.org/10.1007/s10270-024-01184-y

[28] Richter, M. and A. Rubinstein, *The permissible and the forbidden*, Journal of Economic Theory **188** (2020), p. 105042.
URL https://www.sciencedirect.com/science/article/pii/S0022053120300405

[29] Weske, M., "Business Process Management - Concepts, Languages, Architectures," Springer Berlin, Heidelberg, 2019, 3 edition.
URL https://doi.org/10.1007/978-3-662-59432-2

[30] Wexler, J., M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. B. Viégas and J. Wilson, *The What-If Tool: Interactive Probing of Machine Learning Models*, IEEE Trans. Vis. Comput. Graph. **26** (2020), pp. 56–65.
URL https://doi.org/10.1109/TVCG.2019.2934619

The Algebraic Semantics and Proof Theory of Intuitionistic Epistemic Logic

Lifei Wang [1]

Department of Philosophy,Xiamen University

Zhe Yu [2]

Insitute of Logic and Cognition, Sun Yat-sen University

Zhe Lin [3]

Department of Philosophy,Xiamen University

Abstract

This paper focuses on the intuitionistic epistemic logic IEL proposed by S. Artemov and T. Protopopescu. First, the algebraic semantics of IEL is investigated and it is proven that IEL satisfies the property of the algebraic completeness and finite model property (FMP). Second, from a proof-theoretic perspective, a labeled sequent system for IEL is designed; this system is verified to be sound, complete, and terminating. Finally, a natural extension of IEL is considered, which is obtained by adding a diamond operator ($\overline{\mathbf{K}}$) to IEL such that $\overline{\mathbf{K}}$ and the box operator $\mathbf{K}$ in IEL satisfy the adjoint property. For this extended logic, the properties of algebraic completeness and FMP are further proven.

Keywords: Intuitionistic epistemic Logic, Sequent Calculus, Algebraic Semantics, Finite Model Property

1 Introduction

Artemov and Protopopescu [1] introduced the intuitionistic epistemic logics (IEL and IEL$^-$) to investigate knowledge from an intuitionistic standpoint, with the interpretation of knowledge rooted in the Brouwer–Heyting–Kolmogorov (BHK) semantics—the intended semantics for intuitionistic logic [3],[4]. Within this framework, a proposition is deemed true if it is proved; accordingly, intuitionistic knowledge and belief are conceptualized as products of verification. A core tenet of these systems is that an intuitionistic proof, as

[1] wanglifei.wlf@outlook.com

[2] yuzhe28@mail.sysu.edu.cn

[3] pennyshaq@163.com

a form of verification, inherently implies intuitionistic knowledge (denoted by the modality K), which validates the coreflexion principle (or constructivity of proof) $A \to KA$ for both belief and knowledge. Notably, the classical factivity axiom $KA \to A$ is not intuitionistically valid, as verification of a statement does not guarantee the existence of its proof in the intuitionistic perspective. Instead, IEL adopts the weaker intuitionistic reflection principle $KA \to \neg\neg A$, which formally expresses that if a statement A has a verification, then A cannot be false—though the proof of A is not necessarily derivable from the verification process itself. This reflection principle is equivalent to the consistency of knowledge ($\neg K\bot$). Additionally, the classical factivity principle can be recuperated in IEL through its double negation form $\neg\neg(KA \to A)$. Overall, this approach to extending BHK semantics to epistemic logic offers a constructive resolution to the knowability paradox.

Intuitionistic Epistemic Logic (IEL) and its variants have been studied via multiple methods: Krupski and Yatmanov [9] proposed a cut-free sequent calculus, establishing its PSPACE-completeness; Protopopescu [12] embedded it into S4V for expressive extension; Hagemeier and Kirst [7] and Su et al. [14] explored Coq type theory application and multi-agent distributed knowledge respectively; Su and Sano [13] investigated sequent calculi for first-order extensions QIEL/QIEL ; Fiorino [6] developed calculi with the subformula property and linear-depth termination. Beyond these contributions, Balbiani [2] extended the box-based propositional languages of the intuitionistic doxastic logic IEL and intuitionistic epistemic logic IEL with a diamond operator. He subsequently proved the completeness of the resulting logics with respect to their respective relational semantics. Analogously to Balbiani's work [2], we also consider an extension of Intuitionistic Epistemic Logic (IEL) via the addition of a diamond operator $\overline{\mathbf{K}}$.We also consider an extension of Intuitionistic Epistemic Logic (IEL) by adding a diamond operator $\overline{\mathbf{K}}$. In particular, we interpret $\overline{\mathbf{K}}$ as the adjoint dual of the box operator $\mathbf{K}$, following the foundational framework of IntGC [5]. The algebraic semantics of IntGC, together with its frame completeness, finite model property, and duality results, are established in that paper. In the classical counterparts of IntGC studied by Järvinen, Kondo, and Kortelainen[8], the Information Logic of Galois Connections (ILGC) is introduced as a system suitable for approximate reasoning about knowledge.

Herein, we continue this line of research from both algebraic and proof-theoretic perspectives. Specifically, we investigate the algebraic semantics of IEL and establish its algebraic finite model property. We also demonstrate that this extended logic (hereafter referred to as IEL*) constitutes a conservative extension of the original IEL by algebraic meathods. From the proof-theoretic point of view, we following Marin et al. [10]'s methodological framework of labeled sequent system for intuitionistic modal logics to develop a labeled sequent system for IEL. We further show this system is sound, complete, and terminated.

The structure of this article is as follows. In Section 2, we will review Intu-

itionistic Epistemic Logic (IEL) and its frame semantics, while simultaneously discussing its algebraic semantics and completeness. In Section 3, we will focus on IEL* we first describe its axiomatic system and completeness with respect to its corresponding algebra, and finally prove that IEL* is a conservative extension of IEL by means of algebraic methods. In Section 4, we establish the algebraic finite model property for both logics. In Section 5, we present the labelled sequent system GIEL for IEL. Finally, we conclude the article and discuss potential issues for future research.

2 Intuitionistic epistemic logic

In this section, we present the semantics and syntax related to the intuitionistic epistemic logic IEL [1].

Let **Prop** be a denumerable set of variables. The language $\mathcal{L}_{\mathrm{IEL}}[1]$ for intuitionistic epistemic logic is defined inductively as follows:

$$\mathcal{L}_{\mathrm{IEL}} \ni A ::= p \mid \bot \mid (A \to A) \mid (A \wedge A) \mid (A \vee A) \mid \mathbf{K}A$$

Where $p \in \mathbf{Prop}$. We define $\neg A := A \to \bot$, $\top := (\bot \to \bot)$, and $A \leftrightarrow B := (A \to B) \wedge (B \to A)$.

An IEL-model is a quadruple $\langle W, R, V, E\rangle$ such that:

(i) $\langle W, R, V\rangle$ is an intuitionistic model;
- $\langle W, R\rangle$ is a nonempty partial order (i.e., R is a binary 'cognition' relation on W satisfying reflexivity, transitivity, and anti-symmetry);
- V is a monotonic evaluation of propositional letters in W.

(ii) E is a binary 'knowledge' relation on W coordinated with the 'cognition' relation R:
- For any $u \in W$, $E(u) \subseteq R(u)$;
- If uRv, then $E(v) \subseteq E(u)$;
- For all $u \in W$, $E(u) \neq \emptyset$.

Let $\mathcal{F} = \langle W, R, E\rangle$ be a Kripke frame of IEL. Let $\mathcal{M} = \langle \mathcal{F}, V\rangle$ be a Kripke model based on $\mathcal{F}$, and let $w \in W$. The notion of a formula A being *true* at w in $\mathcal{M}$ (denoted by $\mathcal{M}, w \Vdash A$) is defined inductively as follows:

$$\mathcal{M}, w \Vdash p \iff w \in V(p) \tag{1}$$
$$\mathcal{M}, w \nVdash \bot, \forall w \in W \tag{2}$$
$$\mathcal{M}, w \Vdash A \wedge B \iff \mathcal{M}, w \Vdash A \text{ and } \mathcal{M}, w \Vdash B \tag{3}$$
$$\mathcal{M}, w \Vdash A \vee B \iff \mathcal{M}, w \Vdash A \text{ or } \mathcal{M}, w \Vdash B \tag{4}$$
$$\mathcal{M}, w \Vdash A \to B \iff \forall v \in W, (wRv \Rightarrow (\mathcal{M}, v \Vdash A \Rightarrow \mathcal{M}, v \Vdash B)) \tag{5}$$
$$\mathcal{M}, w \Vdash \mathbf{K}A \iff \forall v \in E(w), (\mathcal{M}, v \Vdash A) \tag{6}$$

A formula A is said to be *globally true* in an epistemic model (denoted by $\mathcal{M} \Vdash A$) if $\mathcal{M}, w \Vdash A$ for all $w \in W$. A formula A is *valid* in an epistemic frame $\mathcal{F}$ if $\mathcal{F}, V \Vdash A$ for all valuations V.

The Hilbert axiom system IEL for intuitionistic epistemic logic is as follows:

- Ax.1 All intuitionistic propositional logic axioms.
- Ax.2 $\mathbf{K}(A \to B) \to (\mathbf{K}A \to \mathbf{K}B)$ (Distribution)
- Ax.3 $A \to \mathbf{K}A$ (Coreflection)
- Ax.4 $\mathbf{K}A \to ((A \to \bot) \to \bot)$ (intuitionistic reflection)
- (MP) From A and $A \to B$, infer B

Theorem 2.1 *(Theorem 2, Theorem 3 in [1]) The Hilbert system of* IEL *is sound and complete with respect to* IEL *frames.*

Definition 2.2 An algebra structure $\mathfrak{A} = (A, \wedge, \vee, \to, 0, 1)$ is called a *Heyting algebra* (HA for short) if $(A, \wedge, \vee, 0, 1)$ is a bounded distributive lattice and $\neg$ is a binary operator on A satisfying the following conditions: for any $a, b, c \in A$,

$$\text{(res) } a \wedge b \leq c \text{ iff } a \leq b \to c$$

A complete Heyting algebra is a Heyting algebra that is complete as a lattice. A complete lattice is a partially ordered set $(L, \leq)$ such that every subset X of L has both a greatest lower bound (the infimum, or meet) and a least upper bound (the supremum, or join) in $(L, \leq)$. The meet is denoted by $\bigwedge X$, and the join by $\bigvee X$.

Definition 2.3 An IEL algebra is structure $\mathfrak{A} = (A, \wedge, \vee, \to, 0, 1, K)$ such that $(A, \wedge, \vee, \to, 0, 1)$ is a Heyting algebras and K is an unary operator on A satisfying the following conditions: for any $a, b \in A$,

- $K(a \to b) \leq Ka \to Kb$
- $a \to Ka$
- $K\bot = \bot$

We denote the class of IEL algebras by $\mathbf{A_{IEL}}$. A complete IEL algebra is an IEL algebra that is complete as a lattice.

Given an IEL algebra $\mathrm{X} = (X, \wedge, \vee, \to, \mathbf{K}, 0, 1)$, an *assignment* in X is a function $\theta : \mathbf{Var} \hookrightarrow X$. Every assignment σ in X can be extended homomorphically. Let $\hat{\sigma}(A)$ be the element in X denoted by A. An *algebraic model* is a pair (X, σ) where X is an algebraic structure, and σ is an assignment in X. A formula A is *true in an algebraic model* (X, σ), notation $\models_{\mathrm{X},\sigma} A$, if $\hat{\sigma}(A) = 1$. A formula A is *valid in an algebra* X, notation $\models_{\mathrm{x}} A$, if it is true under any assignments in X. A formula A is *valid in a class of algebras* $\mathbf{X}$, notation $\models_{\mathbf{X}} A$, if $\models_{\mathrm{X}} A$ for any $\mathrm{X} \in \mathbf{X}$.

A system S is called *sound* with respect to a class of algebras $\mathbf{X}$, if for any formula A, $\vdash_{\mathrm{IEL}} A$ implies $\models_{\mathbf{X}} A$. A system S is called *complete* with respect to $\mathbf{X}$, if for any formula A, $\models_{\mathbf{K}} A$ implies $\vdash_{\mathbf{X}} A$.

Theorem 2.4 (Soundness and Completeness) *IEL is sound and complete with respect to the class of IEL algebras.*

Proof. The soundness proceeds by the induction on the height of derivation. For completeness result, it suffices to show that for any A, if $\nvdash_{\mathbf{IEL}} A$, then $\not\models_{\mathbf{A_{IEL}}} A$. It can be proved by standard construction. Let $[A] = \{B| \vdash_{\mathbf{IEL}} A \leftrightarrow B\}$. Let X be the set of all $[A]$. One defines $\{\wedge', \vee', \rightarrow', \mathbf{K}', \top', \bot'\}$ on X as follows: $[A]\#'[B] = [A\#B]$ where $\# \in \{\wedge, \vee, \rightarrow\}$ and $\mathbf{K}'[A] = [\mathbf{K}A]$. Clearly these are well-defined. And it is easy to check that $(X, \wedge', \vee', \rightarrow', \mathbf{K}', \top', \bot')$ is an IEL algbera. The order is defined as $[A] \leq' [B]$ iff $[A] \wedge' [B] = [A]$.

Define an assignment $\sigma : \mathbf{Var} \hookrightarrow X$ such that $\sigma(p) = [p]$. By induction on the complexity of the formula A, one obtains $\widehat{\sigma}(A) = [A]$ for any formula A. Suppose that $\models_{\mathbf{X}} \top \rightarrow A$. Then $\widehat{\sigma}(\top) \leq \widehat{\sigma}(A)$. Hence $\vdash_{\mathbf{IEL}} \top \rightarrow A$, which yields a contradiction. This completes the proof. □

3 Intuitionistic Epistemic Logic with Diamond IEL*

In this section we consider IEL enriched with a Diamond $\overline{\mathbf{K}}$ which is a adjoint of $\mathbf{K}$. Let **Prop** be a denumerable set of variables. The language $\mathcal{L}_{\text{IEL}^*}$ for intuitionistic epistemic logic with diamond is defined inductively as follows:

$$\mathcal{L}_{\text{IEL}^*} \ni A ::= p \mid \bot \mid (A \rightarrow A) \mid (A \wedge A) \mid (A \vee A) \mid \mathbf{K}A \mid \overline{\mathbf{K}}A$$

Where $p \in \mathbf{Prop}$. $\neg, \top$ and $\leftrightarrow$ are defined as above. The Hilbert axiom system for IEL^* is obtained from IEL by enriching the following rules

$$\overline{\mathbf{K}}A \rightarrow B \text{ iff } A \rightarrow \mathbf{K}B$$

The Kripke semantics for IEL* remains identical to that of IEL. The truth condition for formulas in this extended logic is also analogous to that of IEL, except for the inclusion of the following new clause (governing the diamond operator)

$$\mathcal{M}, w \Vdash \overline{K}A \iff \exists v, w \in E(v), (\mathcal{M}, v \Vdash A)$$

Definition 3.1 An IEL$^+$ algebra is structure $\mathbf{X} = (X, \wedge, \vee, \rightarrow, 0, 1, K, \overline{K})$ such that $(A, \wedge, \vee, \rightarrow, 0, 1, K)$ is an IEL algebra and $\overline{K}$ is an unary operator on X satisfying the following adjoint condition: for any $a, b \in X$,

$$\overline{K}a \leq b \text{ iff } a \leq Kb$$

Lemma 3.2 *In any* $A_{\mathbf{IEL}^*}$ $\mathbf{X} = (X, \wedge, \vee, \rightarrow, 0, 1, K, \overline{K})$*, let* $a, b \in X$*. then the following holds:*

(1) $\overline{\mathbf{K}}(a \vee b) \leq \overline{\mathbf{K}}a \vee \overline{\mathbf{K}}b$

(2) $\overline{K}\bot \leq \bot$

(3) if $a \leq b$ *then* $\overline{\mathbf{K}}a \leq \overline{\mathbf{K}}b$.

A *basic sequent* is an expression of the form $A \Rightarrow B$, where $A, B \in \mathcal{L}_{\text{IEL}}$. We use s, t, etc. to denote basic sequents. We write $A \Leftrightarrow B$ if we have both $A \Rightarrow B$ and $B \Rightarrow A$. A *basic sequent rule* is an expression of the form as follows:

$$\frac{s_1 \dots s_n}{s_0}(\text{R})$$

where $s_1, \dots, s_n$ are *premisses* and s_0 is the *conclusion* of (R).

Definition 3.3 The algebraic system AIEL* is a set of basic sequents that satisfies the following conditions: for $i = \{1, 2\}$,

(1) AIEL* contains all instances of the following basic sequents:

(ID) $A \Rightarrow A$ (D) $A \wedge (B \vee C) \Rightarrow (A \wedge B) \vee (A \wedge C)$ (IR) $\top \Rightarrow \neg\mathbf{K}\bot$

($\overline{\mathbf{K}}n1$) $\overline{\mathbf{K}}(A \vee B) \Rightarrow \overline{\mathbf{K}}A \vee \overline{\mathbf{K}}B$ ($\overline{\mathbf{K}}n2$) $\overline{\mathbf{K}}\bot \Rightarrow \bot$ (CR) $A \Rightarrow \mathbf{K}A$

(2) AIEL* is closed under the following rules:

$$\frac{A_i \Rightarrow B}{A_1 \wedge A_2 \Rightarrow B}(\wedge_l) \quad \frac{B \Rightarrow A_1 \quad B \Rightarrow A_2}{B \Rightarrow A_1 \wedge A_2}(\wedge_r)$$

$$\frac{A_1 \Rightarrow B \quad A_2 \Rightarrow B}{A_1 \vee A_2 \Rightarrow B}(\vee_l) \quad \frac{B \Rightarrow A_i}{B \Rightarrow A_1 \vee A_2}(\vee_r)$$

$$\frac{A \Rightarrow B}{\overline{\mathbf{K}}A \Rightarrow \overline{\mathbf{K}}B}(\overline{\mathbf{K}}m) \quad \frac{A \Rightarrow B \quad B \Rightarrow C}{A \Rightarrow C}(\text{Tran})$$

$$\frac{A \Rightarrow B}{\overline{\mathbf{K}}\mathbf{K}A \Rightarrow B}(\mathbf{K}_l) \quad \frac{\overline{\mathbf{K}}A \Rightarrow B}{A \Rightarrow \mathbf{K}B}(\mathbf{K}_r)$$

$$\frac{A \wedge B \Rightarrow C}{A \Rightarrow B \to C}\,(\text{Res}) \quad \frac{P\alpha \Rightarrow \beta}{\alpha \Rightarrow G\beta}\,(\text{rPG})$$

(3) AIEL* is closed under uniform substitution: if $A \Rightarrow B \in \mathbf{AIEL}^*$, then $\theta(A) \Rightarrow \theta(B) \in \mathbf{AIEL}^*$ for any substitution θ.

The algebraic system for IEL denoted AIEL is obtained from AIEL* by excluding all axioms and rules for $\overline{\mathbf{K}}$ and replace ($\mathbf{K}$) rules by axiom: $\mathbf{K}(A \to B) \Rightarrow \mathbf{K}A \to \mathbf{K}B$ ($\mathbf{K}_d$).

Lemma 3.4 *Let* $A, B \in \mathcal{L}_{\text{IEL}^*}$, $\vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}}A \Rightarrow B$ *iff* $\vdash_{\mathbf{AIEL}^*} A \Rightarrow \mathbf{K}B$.

The definitions of algebraic models and the validity of a formula herein align with those presented in the previous section. A basic sequent $A \Rightarrow B$ is valid in an algebra $\mathbf{X}$ denoted $\models_{\mathbf{X}} A \Rightarrow B$ if and only if A holds true under every assignment function then B holds too.

Theorem 3.5 *The following three statements are equivalent.*

(1) $\vdash_{\mathbf{IEL}^*} A \to B$

(2) $\vdash_{\mathbf{AIEL}^*} A \Rightarrow B$

(3) $\models_{\mathbf{A_{IEL}}^*} A \Rightarrow B$

Proof. From (1) to (3), one can simply check with Definition 3.2. From (2) to (1), it is obvious that all the axioms in IEL are valid in AIEL*. The (Adj) follows from Lemma 3.4. From (3) to (2), let us assume that $\nvdash_{\mathbf{AIEL}^*} A \Rightarrow B$. Then by Lemma 4.13 there is an IEL* algebra $\mathbf{X}$ such that $\not\models_{\mathbf{X}} A \Rightarrow B$, whence $\not\models_{\mathbf{A_IEL}^*} A \Rightarrow B$. □

Below, we will prove that the IEL* under consideration is a conservative extension of IEL. The specific proof strategy is as follows: First, we construct

a complex algebra based on the frame model of IEL and demonstrate that this algebra is a complete Heyting algebra, thereby establishing that IEL is sound and complete with respect to complete IEL algebras. Next, we define a diamond operator on this complete Heyting algebra and prove that every complete IEL algebra can be expanded into an IEL* algebra. Given that IEL* is complete with respect to IEL* algebras, it follows that the corresponding IEL* is a conservative extension of IEL. This completes the outline of our proof, and the details are as follows.

Let $F = (W, R, E)$ be an IEL-frame. The complex algebra of F (notation: $Cm(F)$), is the expansion of the power set algebra $P(W)$ with operations $m(R)$ and $m(E)$ for $\rightarrow$ and K defined as follows respectively

$$m(R)(X, Y) = \{u|\ \forall v.uRv \text{ if } v \in X \text{ then } v \in Y\}$$

$$m(E)(X) = \{u \in W|\ \forall v.uEv\ v \in X\}$$

Let (F, V) be an IEL model where V be a valuation in standard way. We denote with $V(A)$ the set of states where A is true. Obviously, $V(B \rightarrow C) = m(R)(V(B), V(C))$ when $A = B \rightarrow C$ and $V(KB) = m(E)(V(B))$ when $A = KB$. Therefore, it is easy to prove that $Cm(F) = (P(W), \cap, \cup, m(R), m(E), \emptyset, W)$ is a complete IEL algebra. Define assignment $\sigma : \text{Var} \hookrightarrow P(W)$ such that $\sigma(p) = V(p)$. Clearly σ can be extended to formulas naturally. Therefore $\models_{(F,V)} A$ iff $\models_{(cm(F),\sigma)} \sigma(A) = \sigma(\top)$, whence $\models_F A$ iff $\models_{cm(F)} A$. Hence IEL is sound and complete with respect to the class of complete IEL algebras.

Theorem 3.6 (Soundness and Completeness) *IEL is sound and complete with respect to the class of complete IEL algebras.*

Theorem 3.7 *Let B, C be IEL formulas. Then $\vdash_{\mathbf{AIEL}} B \Rightarrow C$ iff $\vdash_{\mathbf{AIEL}^*} B \Rightarrow C$*

Proof. The 'if' part is obvious. Let us consider the other direction. Suppose that $\nvdash B \Rightarrow C$. Then by Theorem there is a complete IEL algebra $\mathbf{A} = (A, \wedge, \vee, \rightarrow, 0, 1, K)$ which is sound and complete with respect to IEL. Hence $\not\models_{\mathbf{A}} B \Rightarrow C$. We define $\overline{K}$ as follows.

$$\overline{K}a = \bigwedge\{b|a \leq Kb\}$$

Then the adjoint property for $\overline{K}$ and K is valid. Suppose that $a \leq Kb$. Then, by the definition of $\overline{K}$, $\overline{K}a \leq b$. In contrast, let $\overline{K}a \leq b$ and $\overline{K}a = \bigwedge\{c|a \leq Kc\}$. Thus $\bigwedge c \leq b$. Then $K \bigwedge c \leq Kb$, so $\bigwedge Kc \leq Kb$. Subsequently, $a \leq Kb$. The resulting algebra denoted by $\mathbf{X}^* = (X, \wedge, \vee, \rightarrow, 0, 1, K, \overline{K})$ is a IEL* algebra and $\not\models_{\mathbf{X}^*} B \Rightarrow C$. Hence by Theorem $\nvdash_{\mathbf{AIEL}^*} B \Rightarrow C$, which completes the proof.

□

4 Finite Model Property

In this section we prove the algebraic finite model property for IEL and IEL*.

A variety V is said to have a decidable equational theory if there is a computer algorithm that determines whether an equation is true or not in this variety in finite time. Let V be any finitely based variety. If V has the finite model property (FMP), then it has a decidable equational theory.

Definition 4.1 Let VA be the variety of **A**-algebras. VA is said to have the finite model property (FMP) if for any $s, t \in T$ there is a A and a μ satisfying that $\mu(s) \leq \mu(t)$ implies $\pi(s) \leq \pi(t)$ for some finite A$'$ and π .

The FMP can be equivalently defined by the Gentzen-style algebraic systems.

Definition 4.2 Let VA be a variety of **A**-algebras and $\mathbf{GS_A}$ be its corresponding system. VA is said to have finite model property if for any $A, B \in F$ $\nvdash_{\mathbf{GS_A}} A \Rightarrow B$, then there is a A and a μ such that $\mu(A) \leq \mu(B)$ does not hold in A.

Let T be a set of formula containing $\{\bot, \overline{\mathbf{K}}\bot, \mathbf{K}\bot, \neg\bot, \neg\overline{\mathbf{K}}\bot, \neg\mathbf{K}\bot, \top, \overline{\mathbf{K}}\top, \mathbf{K}\top, \neg\top, \overline{\mathbf{K}}\top, \mathbf{K}\top\}$ and closed under $\wedge, \vee$. Let T^* be the $\wedge$ and $\mathbf{K}$ closure of T.

Definition 4.3 We define $\leq_T$ on T^* as follows: for $A, B \in T^*$, $A \leq_T B$ iff for formula $C \in T$, $\vdash_{\mathbf{AIEL}^*} B \Rightarrow C$ implies $\vdash_{\mathbf{AIEL}^*} A \Rightarrow C$

Let $A \sim_T B$ be $A \leq_T B$ and $B \leq_T A$, then $\sim_T$ is an equivalence relation. Let $[A]_T = \{B \mid B \sim_T A \ \& B \in T^*\}$ for any $A \in T$. Let $[T] = \{[A]_T \mid A \in T\}$.

Definition 4.4 A formula A is called a disjunction normal form with respect to the set of terms T if it is the disjunction of conjunction of some formula in T.

Remark 4.5 It is important to note that the notion of DNF used here differs from the standard definition. Our DNF is defined relative to a given set T, where every formula in T is treated as a literal. Consequently, any formula that lies in the conjunctive–disjunctive closure generated by T has a corresponding DNF formula, and the two are equivalent with respect to provability. This result holds for any logic whose semantics is based on a distributive lattice. In contrast, the traditional definition of DNF enjoys this property only in classical logic.

Since distributive of lattices are always assumed here, for any $A \in T$ there is a formula $dnf(A)$ in disjunction normal form such that $A \sim_T dnf(A)$. Clearly $\{dnf(A) | A \in T\}$ is finite since T is finite. Due to $[A]_T = [dnf(A)]_T$ and the number of $[dnf(A)]_T$ is finite, $[T]$ is finite.

Lemma 4.6 *For any $A \in T^*$, there is a $B \in dnf(T)$ such that $A \sim_T B$.*

Proof. Consider the set $X = \{C \in dnf(T) | \vdash_{\mathbf{IEL}^*} A \Rightarrow C\}$. Clearly it is not empty since $\top \in X$. Let $B = \bigwedge X$. Clearly $\vdash_{\mathbf{IEL}^*} B \Rightarrow C$ for any $C \in X$.

Therefore $B \leq_T A$. Moreover $\vdash_{\mathbf{IEL}^*} A \Rightarrow B$. Hence for any $D \in T$ such that $\vdash_{\mathbf{IEL}^*} B \Rightarrow D$, $\vdash_{\mathbf{IEL}^*} A \Rightarrow D$. Thus $A \leq_T B$. Hence $A \sim_T B$. □

Definition 4.7 Let $\mathcal{Q} = ([T], \wedge^*, \vee^*, \rightarrow^*, 0^*, 1^*, \mathbf{K}^*, \overline{\mathbf{K}}^*)$ be the quotient algebra of $[T]$ where all operations are defined as follows: for any $[A]_T, [B]_T \in [T]$,

(1) $1^* = [\top]_T$;

(2) $0^* = [\bot]_T$;

(3) $[A]_T \wedge^* [B]_T = [A \wedge B]_T$;

(4) $[A]_T \vee^* [B]_T = [A \vee B]_T$;

(6) $[A]_T \rightarrow * [B]_T = \bigvee^* \{[C] \in [dnf(T)] \mid A \wedge C \leq_T B\}$;

(7) $\overline{\mathbf{K}}^* [A]_T = [B]$ s.t. $B \sim_T A$

(8) $\mathbf{K}[A]_T = \bigvee^* \{[B]_T \in [dnf(T)] \mid \overline{\mathbf{K}} B \leq_T A\}$;

We define $[A]_T \leq^* [B]_T$ as $[A]_T \wedge^* [B]_T = [A]_T$.

Lemma 4.8 *The following conditions are equivalent for all $s, q \in T^*$:*

(1) $A \leq_T B$;

(2) $\vdash_{\mathbf{AIEL}^*} A \Rightarrow B$;

(3) $[A]_T \leq^* [B]_T$

The proof can be checkedregularly.

Lemma 4.9 *All the operations defined in Definition 4.7 are well-defined.*

Proof. $\wedge, \vee$ are well defined since T is closed under these operations. We provide the proof for operation $\overline{\mathbf{K}}$. Then by the functional definition, $\rightarrow, \mathbf{K}$ are all well defined. Let $[A]_T = [B]_T$. It suffice to show $\overline{\mathbf{K}}[A] = \overline{\mathbf{K}}[B]$ which equals to $\overline{\mathbf{K}} A \sim_T \overline{\mathbf{K}} B$. Suppose that $\vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}} A \Rightarrow C$ for some $C \in T$. Clearly $\vdash_{\mathbf{AIEL}^*} B \Rightarrow A$. Thus $\vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}} B \Rightarrow \overline{\mathbf{K}} A$. Hence $\vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}} B \Rightarrow C$. Thus $\overline{\mathbf{K}} B \leq_T \overline{\mathbf{K}} A$. By similar argument, $\overline{\mathbf{K}} A \leq_T \overline{\mathbf{K}} B$,

□

Lemma 4.10 *For any $[A]_T, [B]_T, [C]_T \in [T]$, in $\mathcal{Q}$ according to Definition 4.7 $[A]_T \wedge^* [B]_T \leq^* [C]_T$ iff $[A]_T \leq^* [B]_T \rightarrow^* [C]_T$*

Proof. Let us show $[A]_T \wedge^* [B]_T \leq^* [C]_T$ implies $[A]_T \leq^* [B]_T \rightarrow^* [C]_T$. Let $[B]_T \rightarrow^* [C]_T = [D]_T = [\bigvee \{dnf_T(D_i) | [B]_T \wedge^* [D_i]_T \leq [C]_T\}]_T$. Assume that $[A]_T \cdot^* [B]_T \leq^* [C]_T$. Then there is a D_i such that $D_i = A$. Then $\vdash_{\mathbf{AIEL}^*} A \Rightarrow D$. Thus by Lemma 4.8 $[A]_T \leq^* [D]_T$.

Let us show the opposite direction. Assume that $[A]_T \leq^* [B]_T \rightarrow^* [C]_T$. By Lemma 4.8 $\vdash_{\mathbf{AIEL}^*} A \Rightarrow D$. Thus $\vdash_{\mathbf{AIEL}^*} B \wedge D_i \Rightarrow C$. Clearly $\vdash_{\mathbf{AIEL}^*} B \wedge D \Rightarrow \bigvee \{B \vee D_i\}$. So $\vdash_{\mathbf{AIEL}^*} B \wedge D \Rightarrow C$. Thus $\vdash_{\mathbf{AIEL}^*} A \wedge B \Rightarrow C$. Hence $[A]_T \wedge^* [B]_T \leq^* [C]_T$.

□

Lemma 4.11 *For any $[A]_T, [B]_T \in [T]$ in Q according to Definition 4.7 $\overline{\mathbf{K}}^* [A]_T \leq^* [B]_T$ iff $[A]_T \leq^* \mathbf{K}[B]_T$.*

The proof is quite similar to Lemma 4.10.

Lemma 4.12 *For any $[A]_T \in [T]$ in $\mathcal{Q}$ according to Definition 4.7*

(1) $[A]_T \leq^* \mathbf{K}[A]_T$;

(2) $[\top]_T \leq^* \neg^*\mathbf{K}^*[\bot]_T$.

Proof. (2) is easy. Let us show (1). Let $\mathbf{K}[A]_T = [B]_T$. So $B \sim_T \mathbf{K}A$. Hence for any $C \in T$ s.t. $\vdash_{\mathbf{AIEL}^*} A \Rightarrow C$. So $\vdash_{\mathbf{AIEL}^*} \mathbf{K}A \Rightarrow C$. Hence $\vdash_{\mathbf{AIEL}^*} B \Rightarrow C$. Therefore $[A]_T \leq^* \mathbf{K}^*[A]_T$. □

Lemma 4.13 *The following conditions hold for $\mathcal{Q}$: for any $A, B \in T$,*

(1) *If $\overline{K}A \in T$, then $\overline{\mathbf{K}^*}[A]_T = [\overline{\mathbf{K}}A]_T$;*

(2) *$f\,\mathbf{K}A \in T$, then $\mathbf{K}^*[A]_T = [\mathbf{K}A]_T$;*

(3) *If $A \to B \in T$, then $[A]_T \to^* [B]_T = [A \to B]_T$.*

Proof. Let us show (1). Let $\overline{\mathbf{K}^*}[A]_T = [B]_T$. Then $\overline{\mathbf{K}}A \sim_T B$. Since $\overline{\mathbf{K}}A \in T$, $\vdash_{\mathbf{LG}} B \Leftrightarrow \overline{\mathbf{K}}A$. By Lemma 4.8, $[B]_T = [\overline{\mathbf{K}}A]_T$.

Let us show (2). Let $\mathbf{K}^*[A]_T = [B]_T$. Let $B = \bigvee\{B_i \mid \vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}}B_i \Rightarrow A\}$. Thus $\vdash_{\mathbf{AIEL}^*} B_i \Rightarrow \mathbf{K}A$. Then $\vdash_{\mathbf{AIEL}^*} B \Rightarrow \mathbf{K}A$. It is obviously $\vdash_{\mathbf{AIEL}^*} \overline{\mathbf{K}}\mathbf{K}A \Rightarrow A$. Thus $\vdash_{\mathbf{AIEL}^*} \mathbf{K}A \Rightarrow B$. Therefore (2) holds.

□

Lemma 4.14 *Let T be set of formula containing all formula in A, B and generated as above. If $\nvdash_{\mathbf{AIEL}^*} A \Rightarrow B$, then there is an IEL* algebra $\mathcal{Q}$ and an assignment σ satisfying that $\not\models_{\mathcal{Q}} [A]_T \leq^* [B]_T$.*

Proof. Recall that T be the smallest set containing all the formula in $Sub(A) \cup Sub(B) \cup \{0, 1\}$ and defined as above. Clearly T is finitely based. Assume that $\nvdash_{\mathbf{AIGL}^*} A \Rightarrow B$. Construct $\mathcal{Q}$ as in Definition 4.7 with respect to **AIEL***. Let $\sigma : \mathbf{Var}(T) \mapsto [T]$ such that $\sigma(p) = [p]_T$. σ can be simply extended to $\hat{\sigma}$ satisfying that $\hat{\sigma}(C) = [C]_T$ for any $C \in T$. Assume that $\sigma\hat{(}A) \leq^* \sigma\hat{(}B)$. Then $[A]_T \leq^* [B]_T$. By Lemma 4.8 and Lemma 4.13, one obtains $\vdash_{\mathbf{AIEL}^*} A \Rightarrow B$ which yields contradiction.

□

Theorem 4.15 (FMP and Decidability) AIEL* *has FMP and thus is decidable.*

Proof. Suppose $\nvdash_{\mathbf{AIEL}^*} A \Rightarrow B$ By Lemma 4.14, there is a finite counter-algebra $\mathcal{Q}$ such that $\not\models_{\mathcal{Q}} A \Rightarrow B$. Thus, $\not\models_{\mathbf{A}_{\mathbf{AIEL}^*}} A \Rightarrow B$. □

Theorem 4.16 (FMP and Decidability) AIEL *has FMP and thus is decidable.*

5 labelled sequent calculus GIEL

This section builds upon Negri's framework [11] to construct a labelled sequent calculus suitable for intuitionistic epistemic logic. The system extends the relational atomic formulas to include both xRy and xEy, while preserving all other concepts—including labels, labelled formulas, sequent rules, and the

height-preserving admissible and height-preserving invertibility of rules. We subsequently introduce the labelled sequent calculus GIEL for IEL.

Table 1
The system GIEL.

1. Initial rules:

$x : p, xRy, \Gamma \Rightarrow \Delta, y : p$

2. Propositional rules:

$$\frac{x : A, x : B, \Gamma \Rightarrow \Delta}{x : A \wedge B, \Gamma \Rightarrow \Delta} L\wedge \qquad \frac{\Gamma \Rightarrow \Delta, x : A \quad \Gamma \Rightarrow \Delta, x : B}{\Gamma \Rightarrow \Delta, x : A \wedge B} R\wedge$$

$$\frac{x : A, \Gamma \Rightarrow \Delta \quad x : B, \Gamma \Rightarrow \Delta}{x : A \vee B, \Gamma \Rightarrow \Delta} L\vee \qquad \frac{\Gamma \Rightarrow \Delta, x : A, x : B}{\Gamma \Rightarrow \Delta, x : A \vee B} R\vee$$

$$\frac{x : A \to B, xRy, \Gamma \Rightarrow \Delta, y : A \qquad y : B, x : A \to B, xRy, \Gamma \Rightarrow \Delta}{x : A \to B, xRy, \Gamma \Rightarrow \Delta} L\to$$

$$\frac{xRy, y : A, \Gamma \Rightarrow \Delta, y : B}{\Gamma \Rightarrow \Delta, x : A \to B} R\to \qquad \frac{}{x : \bot, \Gamma \Rightarrow \Delta} L\bot$$

3. Modal rules:

$$\frac{y : A, x : \mathbf{K}A, xEy, \Gamma \Rightarrow \Delta}{x : \mathbf{K}A, xEy, \Gamma \Rightarrow \Delta} L\mathbf{K} \qquad \frac{xEy, \Gamma \Rightarrow \Delta, y : A}{\Gamma \Rightarrow \Delta, x : \mathbf{K}A} R\mathbf{K}$$

4. Relational rules:

$$\frac{xRz, xRy, yRz, \Gamma \Rightarrow \Delta}{xRy, yRz, \Gamma \Rightarrow \Delta}\text{Trans} \qquad \frac{xRx, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\text{Ref}$$

$$\frac{xRy, xEy, \Gamma \Rightarrow \Delta}{xEy, \Gamma \Rightarrow \Delta}\mathbf{KR1} \qquad \frac{xEz, xRy, yEz, \Gamma \Rightarrow \Delta}{xRy, yEz, \Gamma \Rightarrow \Delta}\mathbf{KR2}$$

$$\frac{xEy, \Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}\text{Ser}$$

In $R\to, R\mathbf{K}$, Ref, y does not appear in the conclusion.

Definition 5.1 (Principal Formulas and Active Formulas) For a sequent rule, the formulas in its conclusion that are the target of the rule's operation—specifically, any logical formulas with principal connectives $\wedge$, $\vee$, $\to$, or $\mathbf{K}$, and any required relational atomic formulas (xRy or xEy), are collectively called the *principal formulas*. Any formulas in the premises that do not appear in the conclusion and lie outside the contexts Γ and Δ, are called *active formulas*.

Definition 5.2 (Eigenvariable) During the application of a rule, an *eigenvariable* is a variable that appears only in the premises of the rule application but does not appear in its conclusion.

In the rules $R\rightarrow$, $R\mathbf{K}$, and Ser, the variable y is an eigenvariable.

Definition 5.3 (Complexity of Formulas) The *complexity* of a labelled formula $x : A$ is inductively defined as follows: $c(x : p) = 0$, $c(x : \bot) = 0$, $c(x : A \circ B) = \max\{c(x : A), c(x : B)\} + 1$ where $\circ \in \{\wedge, \vee, \rightarrow\}$, $c(x : \mathbf{K}A) = c(x : A) + 1$

Lemma 5.4 GIEL $\vdash x : A, xRy, \Gamma \Rightarrow \Delta, y : A$

Proof. By the induction on the complexity of $x : A$.

(i) $x : A = x : p$, by the initial rule, $x : p, xRy, \Gamma \Rightarrow \Delta, y : p$ is derivable.

(ii) $x : A = x : \bot$, $x : \bot, xRy, \Gamma \Rightarrow \Delta, y : \bot$ is clearly the conclusion of the rule $L\bot$, and is thus derivable.

(iii) $x : A = B \wedge C$

$$\dfrac{\dfrac{x : B, x : C, xRy, \Gamma \Rightarrow \Delta, y : B \qquad x : B, x : C, xRy, \Gamma \Rightarrow \Delta, y : C}{x : B, x : C, xRy, \Gamma \Rightarrow \Delta, y : B \wedge C}\, R\wedge}{x : B \wedge C, xRy, \Gamma \Rightarrow \Delta, y : B \wedge C}\, L\wedge$$

(iv) $x : A = B \vee C$

$$\dfrac{\dfrac{x : B, xRy, \Gamma \Rightarrow \Delta, y : B, y : C}{x : B, xRy, \Gamma \Rightarrow \Delta, y : B \vee C}\, R\vee \qquad \dfrac{x : C, xRy, \Gamma \Rightarrow \Delta, y : B, y : C}{x : C, xRy, \Gamma \Rightarrow \Delta, y : B \vee C}\, R\vee}{x : B \vee C, xRy, \Gamma \Rightarrow \Delta, y : B \vee C}\, L\vee$$

(v) $x : A = B \rightarrow C$

$$\dfrac{\dfrac{\dfrac{\dfrac{zRz, \ldots, z : B, \Gamma \Rightarrow \Delta, z : C, z : B \qquad z : C, \ldots, zRz, \Gamma \Rightarrow \Delta, z : C}{zRz, xRz, yRz, z : B, x : B \rightarrow C, xRy, \Gamma \Rightarrow \Delta, z : C}\, L\rightarrow}{xRz, yRz, z : B, x : B \rightarrow C, xRy, \Gamma \Rightarrow \Delta, z : C}\, \text{Ref}}{yRz, z : B, x : B \rightarrow C, xRy, \Gamma \Rightarrow \Delta, z : C}\, \text{Trans}}{x : B \rightarrow C, xRy, \Gamma \Rightarrow \Delta, y : B \rightarrow C}\, R\rightarrow$$

(vi) $x : A = \mathbf{K}B$

$$\dfrac{\dfrac{\dfrac{z : B, xEz, yEz, x : \mathbf{K}B, xRy, \Gamma \Rightarrow \Delta, z : B}{xEz, yEz, x : \mathbf{K}B, xRy, \Gamma \Rightarrow \Delta, z : B}\, L\mathbf{K}}{yEz, x : \mathbf{K}B, xRy, \Gamma \Rightarrow \Delta, z : B}\, \mathbf{KR2}}{x : \mathbf{K}B, xRy, \Gamma \Rightarrow \Delta, y : \mathbf{K}B}\, R\mathbf{K}$$

□

Definition 5.5 (Substitution) Given a labelled formula or a relational atomic formula t, $t(z/x)$ denotes the result of replacing every occurrence of x in t with z. Given a set of labelled formulas and relational atoms Γ, we define $\Gamma(z/x) = \{t(z/x) \mid t \in \Gamma\}$.

Lemma 5.6 *(Substitution Lemma) If* GIEL $\vdash_n \Gamma \Rightarrow \Delta$, *then* GIEL $\vdash_n \Gamma(z/x) \Rightarrow \Delta(z/x)$.

Proof. The proof proceeds by induction on the derivation height n of $\Gamma \Rightarrow \Delta$. For $n = 0$, if the derivation is an instance of initial sequent or the conclusion of $L\bot$, the result follows trivially. For $n > 0$, assume the statement holds for all derivations of height $m < n$. Let $\mathfrak{R}$ be the last rule applied in the derivation of $\Gamma \Rightarrow \Delta$, with premises of lower height. By the induction hypothesis, the substitution z/x can be performed in all premises, yielding derivations of $\Gamma_i(z/x) \Rightarrow \Delta_i(z/x)$. If $\mathfrak{R}$ is a rule other than $R\rightarrow$, $R\mathbf{K}$, and *Ser*, substitution commutes with rule application. For $R\rightarrow$, $R\mathbf{K}$, and *Ser*, if z is not an eigenvariable, substitution proceeds as above; If z is an eigenvariable, avoid clashes by introducing a fresh variable, substituting, and applying the rules. □

Lemma 5.7 *(Height-Preserving Admissibility of Weakening Rules).*

$$\frac{\Gamma \Rightarrow \Delta}{x : A, \Gamma \Rightarrow \Delta} LW_1 \qquad \frac{\Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, x : A} RW_1$$

$$\frac{\Gamma \Rightarrow \Delta}{xRy, \Gamma \Rightarrow \Delta} LW_2 \qquad \frac{\Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, xRy} RW_2$$

$$\frac{\Gamma \Rightarrow \Delta}{xEy, \Gamma \Rightarrow \Delta} LW_3 \qquad \frac{\Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Delta, xEy} RW_3$$

The above rules are height-preserving admissible in GIEL.

Proof. The proof proceeds by induction on the derivation height n of $\Gamma \Rightarrow \Delta$. For the base case ($n = 0$), the sequent is an instance of an initial rule or derived by $L\bot$, and the claim holds immediately after applying weakening. For the inductive step ($n > 0$), assume the claim holds for all derivation heights $m < n$. We analyze the last rule $\mathfrak{R}$ applied. For most rules, the result follows by direct application of the inductive hypothesis and the rule. For the rules $R\rightarrow$, $R\mathbf{K}$, and Ser, which introduce eigenvariables, we apply the Substitution Lemma 5.6 to to avoid variable clashes by substituting a fresh variable, then reapply the inductive hypothesis and the rule.

□

Lemma 5.8 *(Invertibility Lemma)*

(i) *If* $\vdash_n x : A \wedge B, \Gamma \Rightarrow \Delta$, *then* $\vdash_n x : A, x : B, \Gamma \Rightarrow \Delta$.

(ii) *If* $\vdash_n \Gamma \Rightarrow \Delta, x : A \wedge B$, *then* $\vdash_n \Gamma \Rightarrow \Delta, x : A$ *and* $\vdash_n \Gamma \Rightarrow \Delta, x : B$.

(iii) *If* $\vdash_n x : A \vee B, \Gamma \Rightarrow \Delta$, *then* $\vdash_n x : A, \Gamma \Rightarrow \Delta$ *and* $\vdash_n x : B, \Gamma \Rightarrow \Delta$.

(iv) *If* $\vdash_n \Gamma \Rightarrow \Delta, x : A \vee B$, *then* $\vdash_n \Gamma \Rightarrow \Delta, x : A, x : B$.

(v) *If* $\vdash_n x : A \rightarrow B, xRy, \Gamma \Rightarrow \Delta$, *then* $\vdash_n x : A \rightarrow B, xRy, \Gamma \Rightarrow \Delta, y : A$ *and* $\vdash_n y : B, x : A \rightarrow B, xRy, \Gamma \Rightarrow \Delta$.

(vi) *If* $\vdash_n \Gamma \Rightarrow \Delta, x : A \rightarrow B$, *then* $\vdash_n xRy, y : A, \Gamma \Rightarrow \Delta, y : B$, *where* y *does not occur in the premises.*

(vii) *If* $\vdash_n x : \mathbf{K}A, xEy, \Gamma \Rightarrow \Delta$, *then* $\vdash_n y : A, x : \mathbf{K}A, xEy, \Gamma \Rightarrow \Delta$.

(viii) *If* $\vdash_n \Gamma \Rightarrow \Delta, x : \mathbf{K}A$, *then* $\vdash_n xEy, \Gamma \Rightarrow \Delta, y : A$, *where* y *does not occur in the premises.*

(ix) *If* $\vdash_n xRy, yRz, \Gamma \Rightarrow \Delta$, *then* $\vdash_n xRz, xRy, yRz, \Gamma \Rightarrow \Delta$.

(x) *If* $\vdash_n \Gamma \Rightarrow \Delta$, *then* $\vdash_n xRx, \Gamma \Rightarrow \Delta$.

(xi) *If* $\vdash_n xEy, \Gamma \Rightarrow \Delta$, *then* $\vdash_n xRy, xEy, \Gamma \Rightarrow \Delta$.

(xii) *If* $\vdash_n xRy, yEz, \Gamma \Rightarrow \Delta$, *then* $\vdash_n xEz, xRy, yEz, \Gamma \Rightarrow \Delta$.

(xiii) *If* $\vdash_n \Gamma \Rightarrow \Delta$, *then* $\vdash_n xEy, \Gamma \Rightarrow \Delta$, *where* y *does not occur in the premises.*

Proof. The proof proceeds by induction on the derivation height n.

(i)-(iv) The height-preserving invertibility proofs for $L\wedge$, $R\wedge$, $L\vee$, and $R\vee$ are similar to [15], which can be consulted for reference.

(v) The height-preserving invertibility of $L\rightarrow$ follows directly from the

height-preserving admissibility of weakening rules (LW_1) and (RW_1).

(vi) To prove the height-preserving invertibility of $R\rightarrow$: (a) When $n = 0$, the sequent is either an instance of initial rule or the conclusion of $L\bot$, and the invertibility holds. (b) When $n > 0$, the induction hypothesis applies. Discuss separately whether the final step is $R\rightarrow$ or another rule.

(vii) The height-preserving invertibility of $L\mathbf{K}$ follows directly from the height-preserving admissibility of weakening rule (LW_1).

(viii) To prove the height-preserving invertibility of $R\mathbf{K}$: (a) When $n = 0$, the sequent is either an instance of initial rule or the conclusion of $L\bot$, and the invertibility holds. (b) When $n > 0$, the induction hypothesis applies. Discuss separately whether the final step is $R\mathbf{K}$ or another rule.

(ix) The height-preserving invertibility of Trans follows directly from the height-preserving admissibility of LW_2.

(x) The height-preserving invertibility of Ref follows directly from the height-preserving admissibility of LW_2.

(xi) The height-preserving invertibility of **KR1** follows directly from the height-preserving admissibility of LW_2.

(xii) The height-preserving invertibility of **KR2** follows directly from the height-preserving admissibility of LW_3.

(xiii) The height-preserving invertibility of Ser follows directly from the height-preserving admissibility of LW_3. □

Theorem 5.9 *(Height-preserving Admissibility of the Contraction Rule)*

$$\frac{x:A, x:A, \Gamma \Rightarrow \Delta}{x:A, \Gamma \Rightarrow \Delta} LC_1 \qquad \frac{\Gamma \Rightarrow \Delta, x:A, x:A}{\Gamma \Rightarrow \Delta, x:A} RC_1$$

$$\frac{xRy, xRy, \Gamma \Rightarrow \Delta}{xRy, \Gamma \Rightarrow \Delta} LC_2 \qquad \frac{\Gamma \Rightarrow \Delta, xRy, xRy}{\Gamma \Rightarrow \Delta, xRy} RC_2$$

$$\frac{xEy, xEy, \Gamma \Rightarrow \Delta}{xEy, \Gamma \Rightarrow \Delta} LC_3 \qquad \frac{\Gamma \Rightarrow \Delta, xEy, xEy}{\Gamma \Rightarrow \Delta, xEy} RC_3$$

The above rules are height-preserving admissible in GIEL.

Proof. We prove LC_1, RC_1, LC_2, RC_2, LC_3, and RC_3 simultaneously by induction on the derivation height n.

(1) When $n = 0$, the sequent is either an initial axiom or the conclusion of $L\bot$. After contraction, it remains an initial axiom or $L\bot$, so the claim holds.

(2) For $n > 0$, assume inductively that contraction is admissible for derivations of height less than n. Consider two cases: (a) If the contracted formula is not principal, then in the last inference rule $\mathfrak{R}$, all premises contain two identical occurrences t. By the induction hypothesis, these can be contracted to one t, and applying $\mathfrak{R}$ yields the conclusion. (b) If the contracted formula is principal, consider the relevant rule: for $L\wedge$, $R\wedge$, $L\vee$, $R\vee$, $L\rightarrow$, $R\rightarrow$, $L\mathbf{K}$, $R\mathbf{K}$, **KR1**, **KR2**, Trans, etc., the induction hypothesis and invertibility allow contraction of repeated formulas in the premises. Applying the corresponding rule then yields the desired conclusion. □

Lemma 5.10 *(Admissibility of Monotonicity Rules)*

$$\frac{xRy, x : A, y : A, \Gamma \Rightarrow \Delta}{xRy, x : A, \Gamma \Rightarrow \Delta} LM \qquad \frac{xRy, \Gamma \Rightarrow \Delta, x : A, y : A}{xRy, \Gamma \Rightarrow \Delta, y : A} RM$$

are admissible in GIEL.

Proof. We prove the admissibility of both LM and RM simultaneously by induction on the derivation height n of the premise. When $n = 0$, the premise is either an instance of an initial rule or a conclusion of $L\bot$.

For the LM rule:

- If the premise is an initial rule of the form $xRy, x : p, y : p, \Gamma \Rightarrow \Delta, y : p$, then $xRy, x : p, \Gamma \Rightarrow \Delta, y : p$ is also an initial rule.
- If the premise is an initial rule of the form $xRy, yRz, x : p, y : p, \Gamma \Rightarrow \Delta, z : p$, then applying Trans yields:

$$\frac{xRz, xRy, yRz, \Gamma, x : p \Rightarrow \Delta, z : p}{xRy, yRz, x : p, \Gamma \Rightarrow \Delta, z : p} \text{Trans}$$

- If the premise is from $L\bot$ with $x : \bot$ eliminated, of form $xRy, x : \bot, y : \bot, \Gamma \Rightarrow \Delta$, then $xRy, x : \bot, \Gamma \Rightarrow \Delta$ remains a $L\bot$ conclusion.
- If the premise is from $L\bot$ with some $z : \bot \in \Gamma$ $(z \neq x)$ eliminated, then $xRy, x : A, \Gamma \Rightarrow \Delta$ remains a $L\bot$ conclusion.

For the RM rule: The premise is either initial or from $L\bot$, and the argument proceeds symmetrically through three analogous subcases.

Now, let $n > 0$ and assume as the induction hypothesis (IH) that both rules are admissible for all derivations of height less than n. Consider the last rule $\mathfrak{R}$ applied in the derivation of the premise. We examine two main cases based on the role of the eliminated formula in $\mathfrak{R}$:

- The eliminated formula is not principal in $\mathfrak{R}$. In this case, we apply IH first and apply $\mathfrak{R}$ again.
- The eliminated formula is principal in $\mathfrak{R}$. This case requires a analysis on $\mathfrak{R}$, considering rules such as $L\wedge$, $R\vee$, $L\rightarrow$, and $R\mathbf{K}$. For more complex rules like $R\rightarrow$ and $R\mathbf{K}$, the proof leverages:
  · the height-preserving invertibility of the rule $\mathfrak{R}$, and
  · the admissibility of relevant structural rules (weakening and contraction).

□

Definition 5.11 (Cut rule)

$$\frac{\Gamma \Rightarrow \Delta, t \quad t, \Gamma' \Rightarrow \Delta'}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'} Cut$$

The term t can be xRy, xEy, or $x : A$.

Theorem 5.12 *(Cut Elimination Theorem) The cut rule is admissible in the* GIEL *system.*

Proof. See Appendix A □

Theorem 5.13 *GIEL is soundness and completeness with respect to IEL-frames.*

Proof. See Appendix B □

Theorem 5.14 *GIEL is terminated*

Proof. See Appendix C □

6 Conclusion

In this paper, we investigate the proof theory and algebraic semantics of Intuitionistic Epistemic Logic (IEL), as introduced by Artemov and Protopopescu. Specifically, we examine IEL algebras and establish both algebraic completeness and the finite model property (FMP) for this logic. From a proof-theoretic perspective, we further study IEL by constructing a labelled sequent calculus for the system; we prove that this sequent calculus is sound and complete with respect to IEL, and additionally demonstrate its termination property.

Following the tradition of W. Díaz, we extend IEL by incorporating a diamond operator $\overline{\mathbf{K}}$. For this extended logic, we introduce the corresponding algebraic structures and establish results for algebraic soundness, completeness, and the FMP. The development of a labelled sequent calculus for this extended logic is reserved for future work.

In subsequent research, it would be valuable to explore extensions of IEL within the well-established framework of intuitionistic modal logic, particularly following the tradition of Fischer Servi (FS). A key question for such extensions is whether properties such as decidability and the FMP (or other structural results, e.g., admissibility of rules) continue to hold.

Acknowledgements We would like to express our special acknowledgement to Jinsheng Chen (Sun Yat-sen University) for many stimulating discussions about the subject of this article.

References

[1] Artëmov, S. N. and T. Protopopescu, *Intuitionistic epistemic logic*, The Review of Symbolic Logic **9** (2014), pp. 266 – 298.
URL `https://api.semanticscholar.org/CorpusID:6117764`

[2] Balbiani, P., *Iuitionistic epistemic logic with two modal operators*, Synthese **206:169** (2025).

[3] Brouwer, L. E. J., "Over de grondslagen der wiskunde," Maas & Van Suchtelen, Amsterdam, 1907.

[4] Brouwer, L. E. J., "Brouwer's Cambridge Lectures on Intuitionism," Cambridge University Press, New York, 1981.

[5] Dzik, W., J. Järvinen and M. Kondo, *Intuitionistic propositional logic with galois connections*, Logic Journal of the IGPL **18** (2010), pp. 837–858.

[6] Fiorino, G., *Linear depth deduction with subformula property for intuitionistic epistemic logic*, Journal of Automated Reasoning **67:3** (2023).

[7] Hagemeier, C. and D. Kirst, *Constructive and mechanised meta-theory of iel and similar modal*, Journal of Logic and Computation **32** (2022), pp. 1585–1610.

[8] Järvinen, J., M. Kondo and J. Kortelainen, *Logics from galois connections*, International Journal of Approximate Reasoning **49** (2008), pp. 595–606.

[9] Krupski, V. N. and A. Yatmanov, *Sequent calculus for intuitionistic epistemic logic iel*, in: S. Artemov and A. Nerode, editors, *Logical Foundations of Computer Science* (2016), pp. 187–201.

[10] Marin, S., M. Morales and L. Straßburger, *A fully labelled proof system for intuitionistic modal logics*, Journal of Logic and Computation **31** (2021), pp. 998–1022.

[11] Negri, S., *Proof analysis in modal logic*, Journal of Philosophical Logic **34** (2005), pp. 507–544.

[12] Protopopescu, T., *Intuitionistic epistemology and modal logics of verification*, in: W. van der Hoek, W. H. Holliday and W.-f. Wang, editors, *Logic, Rationality, and Interaction* (2015), pp. 295–307.

[13] Su, Y., R. Murai and K. Sano, *On Artemov and Protopopescu's intuitionistic epistemic logic expanded with distributed knowledge*, in: *Logic, Rationality, and Interaction*, 2021, pp. 216–231.

[14] Su, Y. and K. Sano, *First-order intuitionistic epistemic logic*, in: P. Blackburn, E. Lorini and M. Guo, editors, *Logic, Rationality, and Interaction* (2019), pp. 326–339.

[15] Troelstra, A. S. and H. Schwichtenberg, "Basic Proof Theory," Cambridge Tracts in Theoretical Computer Science, Cambridge University Press, 2000, 2 edition.

Appendix

A Cut Elimination of GIEL

Theorem A.1 *(Cut Elimination Theorem) The cut rule is admissible in the* GIEL *system.*

Proof.
1. If the cut term is a relational atom xRy (similarly for xEy), such atoms can only be generated by left-side rules; their appearance on the right must pass through the right weakening rule LW_2. Therefore, removing xRy from the premise $\Gamma \Rightarrow \Delta, xRy$ yields $\Gamma \Rightarrow \Delta$. By applying weakening rules, we directly obtain the conclusion $\Gamma, \Gamma' \Rightarrow \Delta, \Delta'$.
2. If the cut term is a labeled formula $x : A$, we proceed by main induction on the complexity of the formula and sub-induction on the sum of the derivation heights of the premises. To avoid such conflicts, appropriate variable substitutions must be made based on Lemma 5.6 before permutation.
2.1 At least one of the cut premises is an instance of initial rule or an $L\bot$ conclusion. We discuss the left and right premises separately:
2.1.1 If the left premise is an instance of initial rule or an $L\bot$ conclusion, there are three subcases:
2.1.1.1 Suppose the left premise is an instance of initial rule of the form $x : p, xRy, \Gamma'' \Rightarrow \Delta'', y : p, t$, and $\Gamma = x : p, xRy, \Gamma'', \Delta = \Delta'', y : p$. In this case, the cut term t does not participate in the specific instance of an instance of initial rule. Therefore, even after removing t, the left premise still satisfies the conditions of an instance of initial rule (the left side contains $x : p, xRy$, and the right side contains $y : p$), so $\Gamma \Rightarrow \Delta$ remains derivable. By repeatedly applying weakening rules, we directly obtain the desired conclusion $\Gamma, \Gamma' \Rightarrow \Delta, \Delta'$.
2.1.1.2 Suppose the left premise is an instance of initial rule of the form $x : p, xRy, \Gamma'' \Rightarrow \Delta, y : p$ and the cut term is $y : p$. Then the right premise is

$y : p, \Gamma' \Rightarrow \Delta'$. First, apply weakening rules to the right premise to yield $xRy, x : p, y : p, \Gamma' \Rightarrow \Delta'$. Next, apply the (LM) rule to obtain:

$$\frac{xRy, x : p, y : p, \Gamma' \Rightarrow \Delta'}{xRy, x : p, \Gamma' \Rightarrow \Delta'} \text{ (LM)}$$

Repeatedly apply weakening rules to obtain $x : p,\ xRy,\ \Gamma'', \Gamma' \Rightarrow \Delta,\ \Delta'$.

2.1.1.3 If the left premise $\Gamma \Rightarrow \Delta, x : A$ is an $L\bot$ conclusion, then Γ must contain a labeled formula of the form $x : \bot$. Therefore, $\Gamma, \Gamma' \Rightarrow \Delta, \Delta'$ remains an $L\bot$ conclusion.

2.1.2 If the right premise $x : A, \Gamma' \Rightarrow \Delta'$ is an instance of initial rule or an $L\bot$ conclusion, there are four subcases:

2.1.2.1 Suppose the right premise is an instance of initial rule of the form $t, x : p, xRy, \Gamma'' \Rightarrow \Delta'', y : p$, and $\Gamma' = x : p, xRy, \Gamma'', \quad \Delta' = \Delta'', y : p$. The cut term t does not participate in the specific instance of an instance of initial rule. Therefore, even after removing t, the premise still satisfies the conditions of an instance of initial rule (the left side contains $x : p, xRy$, and the right side contains $y : p$), so $\Gamma' \Rightarrow \Delta'$ remains derivable. By applying weakening rules we directly obtain the desired conclusion $\Gamma, \Gamma' \Rightarrow \Delta, \Delta'$.

2.1.2.2 Suppose the right premise is an instance of initial rule of the form $x : p, xRy, \Gamma'' \Rightarrow \Delta'', y : p$ and the cut term is $x : p$. Then the left premise is $\Gamma \Rightarrow \Delta, x : p$. First, apply weakening rules to the left premise to yield $xRy, \Gamma \Rightarrow \Delta, x : p, y : p$. Next, apply the (RM) rule to obtain:

$$\frac{xRy, \Gamma \Rightarrow \Delta, x : p, y : p}{xRy, \Gamma \Rightarrow \Delta, y : p}(RM)$$

Repeatedly apply weakening rules to obtain $xRy, \Gamma, \Gamma'' \Rightarrow \Delta, \Delta'', y : p$.

2.1.2.3 If the right premise is an $L\bot$ conclusion and the formula $x : \bot$ is contained in Γ', then $\Gamma, \Gamma' \Rightarrow \Delta, \Delta'$ remains an $L\bot$ conclusion.

2.1.2.4 If the right premise is an $L\bot$ conclusion of the form $x : \bot, \Gamma' \Rightarrow \Delta'$, where the cut term is $x : \bot$. Since $x : \bot$ is not a principal formula, the right premise $x : \bot, \Gamma' \Rightarrow \Delta'$ must be derived through some rule. This case belongs to subsequent discussions and is not expanded here.

2.2 If the cut premises are neither instance of an instance of initial rules nor $L\bot$ conclusions, we discuss three cases:

2.2.1 If the cut term is not principal in the left premise, we verify based on the rule used in the left premise. It is easily verifiable that in all cases, the cut height in the new derivation is reduced.

2.2.2 If the cut term is not principal in the right premise, we verify based on the rule used in the right premise. It is easily verifiable that in all cases, the cut height in the new derivation is reduced.

2.2.3 If the cut term is principal in both premises, we consider the following four cases:

2.2.3.1 The cut term is $x : A \wedge B$

$$\dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A^{(1)} \quad \Gamma \Rightarrow \Delta, x : B^{(2)}}{\Gamma \Rightarrow \Delta, x : A \wedge B}\,(R\wedge) \qquad \dfrac{x : A, x : B, \Gamma' \Rightarrow \Delta'^{(3)}}{x : A \wedge B, \Gamma' \Rightarrow \Delta'}\,(L\wedge)}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)$$

$\curvearrowright$

$$\dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A^{(1)} \qquad \dfrac{\Gamma \Rightarrow \Delta, x : B^{(2)} \qquad x : A, x : B, \Gamma' \Rightarrow \Delta'^{(3)}}{x : A, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)}{\Gamma, \Gamma, \Gamma' \Rightarrow \Delta, \Delta, \Delta'}\,(Cut)}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Con^*)$$

Here, Con^* denotes multiple applications of contraction rules.

2.2.3.2 The cut term is $x : A \vee B$

$$\dfrac{\dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A, x : B^{(1)}}{\Gamma \Rightarrow \Delta, x : A \vee B}\,(R\vee) \qquad \dfrac{x : A, \Gamma' \Rightarrow \Delta'^{(2)} \qquad x : B, \Gamma' \Rightarrow \Delta'^{(3)}}{x : A \vee B, \Gamma' \Rightarrow \Delta'}\,(L\vee)}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)$$

$\curvearrowright$

$$\dfrac{\dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A, x : B^{(1)} \qquad x : B, \Gamma' \Rightarrow \Delta'^{(3)}}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta', x : A}\,(L\vee) \qquad x : A, \Gamma' \Rightarrow \Delta'^{(2)}}{\Gamma, \Gamma, \Gamma' \Rightarrow \Delta, \Delta, \Delta'}\,(Cut)}{\Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Con^*)$$

2.2.3.3 The cut term is $x : A \to B$

$$\dfrac{\dfrac{\dfrac{xRz, z : A, \Gamma \Rightarrow \Delta, z : B^{(1)}}{\Gamma \Rightarrow \Delta, x : A \to B^{(2)}}\,(R\to) \qquad \dfrac{x : A \to B, xRy, \Gamma' \Rightarrow \Delta', y : A^{(3)} \qquad y : B, \ldots^{(4)}}{x : A \to B, xRy, \Gamma' \Rightarrow \Delta'}\,(L\to)}{xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)$$

Assume m and n are the derivation heights of the cut premises. Substituting y/z in ${}^{(1)}xRz, z : A, \Gamma \Rightarrow \Delta, z : B$ (since z is an eigenvariable, Γ and Δ remain unchanged), Lemma 5.6 yields ${}^{(1)'}xRy, y : A, \Gamma \Rightarrow \Delta, y : B$ at the same height. The derivation is then transformed as follows:

$$\dfrac{\dfrac{\dfrac{xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta', y : B^{(5)} \qquad \dfrac{\Gamma \Rightarrow \Delta, x : A \to B^{(2)} \qquad y : B, x : A \to B, xRy, \Gamma' \Rightarrow \Delta'^{(4)}}{y : B, xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)}{xRy, xRy, \Gamma, \Gamma', \Gamma' \Rightarrow \Delta, \Delta, \Delta', \Delta'}\,(Cut)}{xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Con^*)$$

Here, ${}^{(5)}$ is derived as follows:

$$\dfrac{\dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A \to B^{(2)} \qquad x : A \to B, xRy, \Gamma' \Rightarrow \Delta', y : A^{(3)}}{xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta', y : A}\,(Cut) \qquad xRy, y : A, \Gamma \Rightarrow \Delta, y : B^{(1)'}}{xRy, xRy, \Gamma, \Gamma, \Gamma' \Rightarrow \Delta, \Delta, \Delta', y : B}\,(Cut)}{xRy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta', y : B^{(5)}}\,(Con^*)$$

2.2.3.4 The cut term is $x : \mathbf{K}A$

$$\dfrac{\dfrac{xEz, \Gamma \Rightarrow \Delta, z : A}{\Gamma \Rightarrow \Delta, x : \mathbf{K}A}\,(R\mathbf{K}) \qquad \dfrac{y : A, x : \mathbf{K}A, xEy, \Gamma' \Rightarrow \Delta'}{x : \mathbf{K}A, xEy, \Gamma' \Rightarrow \Delta'}\,(L\mathbf{K})}{xEy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)$$

By substituting y/z in $xEz, \Gamma \Rightarrow \Delta, z : A$ — noting that z is an eigenvariable, so Γ and Δ remain unchanged — Lemma 5.6 yields $xEy, \Gamma \Rightarrow \Delta, y : A$ at the same derivation height. The derivation is then transformed as follows:

$$\dfrac{\dfrac{\dfrac{xEy, \Gamma \Rightarrow \Delta, y : A \qquad \dfrac{\Gamma \Rightarrow \Delta, x : \mathbf{K}A \qquad y : A, x : \mathbf{K}A, xEy, \Gamma' \Rightarrow \Delta'}{y : A, xEy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Cut)}{xEy, xEy, \Gamma, \Gamma, \Gamma' \Rightarrow \Delta, \Delta, \Delta'}\,(Cut)}{xEy, \Gamma, \Gamma' \Rightarrow \Delta, \Delta'}\,(Con^*)$$

□

B Sound and completeness of GIEL

This Appendix establishes the soundness and completeness of GIEL.

Definition B.1 (Interpretation Function and Satisfiability) Given an intuitionistic epistemic model $\mathfrak{M} = \langle W, R, V, E\rangle$ and a set of labels Lab, an *interpretation function* is a mapping $\tau : \mathsf{Lab} \to W$. The pair $(\mathfrak{M}, \tau)$ is called a *labelled extension* of the model.

The satisfiability of a term t in the labelled model $(\mathfrak{M}, \tau)$ (denoted $\mathfrak{M}, \tau \Vdash t$) is defined inductively as follows:

$$\begin{array}{lcl} \mathfrak{M}, \tau \Vdash x : A & \iff & \mathfrak{M}, \tau(x) \Vdash A \\ \mathfrak{M}, \tau \Vdash xRy & \iff & (\tau(x), \tau(y)) \in R \\ \mathfrak{M}, \tau \Vdash xEy & \iff & (\tau(x), \tau(y)) \in E \end{array}$$

Definition B.2 (Satisfiability of Sequents) Let $\mathfrak{M}$ be a model and τ an interpretation function:

(i) For a labelled sequent $\Gamma \Rightarrow \Delta$, the satisfaction is defined as:

$$\mathfrak{M}, \tau \Vdash \Gamma \Rightarrow \Delta \iff (\forall A \in \Gamma,\, \mathfrak{M}, \tau \Vdash A) \implies (\exists B \in \Delta,\, \mathfrak{M}, \tau \Vdash B)$$

That is, if the model $\mathfrak{M}$ under the interpretation function τ satisfies all formulas in Γ, then it must satisfy at least one formula in Δ.

(ii) For a set of sequents $\{\Gamma_i \Rightarrow \Delta_i\}_{1 \le i \le n}$, the satisfaction is defined as:

$$\mathfrak{M}, \tau \Vdash \{\Gamma_i \Rightarrow \Delta_i\}_{1 \le i \le n} \iff \forall i \in \{1, \ldots, n\},\, \mathfrak{M}, \tau \Vdash \Gamma_i \Rightarrow \Delta_i$$

That is, the model $\mathfrak{M}$ under the interpretation function τ satisfies all sequents in the set.

Definition B.3 (Validity of Sequents) A labelled sequent $\Gamma \Rightarrow \Delta$ is said to be valid if for any interpretation function τ and model $\mathfrak{M}, \tau$, $\mathfrak{M}, \tau \Vdash \Gamma \Rightarrow \Delta$.

Definition B.4 (Validity Preservation of Sequent Rules) A sequent rule $\dfrac{S_1 \ldots S_n}{S}$ is said to be *validity-preserving* if it satisfies:

$$\bigwedge_{i=1}^{n} \forall \mathfrak{M}, \forall \tau, (\mathfrak{M}, \tau \Vdash S_i) \implies \forall \mathfrak{M}, \forall \tau, (\mathfrak{M}, \tau \Vdash S)\,.$$

That is, if the premises $S_1, \ldots, S_n$ are valid under all models $\mathfrak{M}$ and assignments τ, then the conclusion S must also hold under all models and assignments.

Lemma B.5 $\Vdash_{\mathrm{GIEL}} \Rightarrow x : A$ *if and only if* $\Vdash_{\mathrm{IEL}} A$.

Proof. This follows directly from the definition of validity. □

Theorem B.6 *(Soundness) If* $\vdash_{\mathrm{GIEL}} \Gamma \Rightarrow \Delta$, *then* $\Vdash \Gamma \Rightarrow \Delta$.

Proof. It suffices to show that all axioms of system $GIEL$ are valid and all rules preserve validity. The detailed proof is omitted. □

Proposition B.7 $\vdash_{\text{GIEL}} \Rightarrow x : \mathbf{K}(A \to B) \to (\mathbf{K}A \to \mathbf{K}B)$

Proof.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{wRw, w : A, w : A \to B, yEw, zEw, yRz, xRy, z : \mathbf{K}A \Rightarrow w : B}{w : A, w : A \to B, yEw, zEw, yRz, xRy, z : \mathbf{K}A \Rightarrow w : B}\,Ref}{w : A \to B, yEw, zEw, yRz, xRy, z : \mathbf{K}A \Rightarrow w : B}\,L\mathbf{K}}{yEw, zEw, yRz, xRy, y : \mathbf{K}(A \to B), z : \mathbf{K}A \Rightarrow w : B}\,L\mathbf{K}}{zEw, yRz, xRy, y : \mathbf{K}(A \to B), z : \mathbf{K}A \Rightarrow w : B}\,\mathbf{KR2}}{yRz, xRy, y : \mathbf{K}(A \to B), z : \mathbf{K}A \Rightarrow z : \mathbf{K}B}\,R\mathbf{K}}{xRy, y : \mathbf{K}(A \to B) \Rightarrow y : (\mathbf{K}A \to \mathbf{K}B)}\,R \to}{\Rightarrow x : \mathbf{K}(A \to B) \to (\mathbf{K}A \to \mathbf{K}B)}\,R \to$$

By Lemma 5.4, the topmost sequent is derivable. □

Proposition B.8 $\vdash_{\text{GIEL}}, \Rightarrow x : A \to \mathbf{K}A$

Proof.

$$\dfrac{\dfrac{\dfrac{yRz, yEz, xRy, y : A \Rightarrow z : A}{yEz, xRy, y : A \Rightarrow z : A}\,\mathbf{KR1}}{xRy, y : A \Rightarrow y : \mathbf{K}A}\,R \to}{\Rightarrow x : A \to \mathbf{K}A}\,R \to$$

By Lemma 5.4, this sequent is derivable. □

Proposition B.9 $\vdash_{\text{GIEL}} \Rightarrow x : \mathbf{K}A \to ((A \to \bot) \to \bot)$

Proof. The topmost sequents $zRz', zEz', z' : A, yEz', xRy, yRz, y : \mathbf{K}A \Rightarrow z : \bot, z' : A$ and $z' : \bot, zRz', zEz', z' : A, yEz', xRy, yRz, y : \mathbf{K}A \Rightarrow z : \bot$ instances of initial sequent, then

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\vdots}{}\,L \to}{zRz', z' : A, yEz', zEz', yRz, xRy, y : \mathbf{K}A, z : A \to \bot \Rightarrow z : \bot}\,\mathbf{KR1}}{z' : A, yEz', zEz', yRz, xRy, y : \mathbf{K}A, z : A \to \bot \Rightarrow z : \bot}\,L\mathbf{K}}{yEz', zEz', yRz, xRy, y : \mathbf{K}A, z : A \to \bot \Rightarrow z : \bot}\,\mathbf{KR2}}{zEz', yRz, xRy, y : \mathbf{K}A, z : A \to \bot \Rightarrow z : \bot}\,\text{Ser}}{yRz, xRy, y : \mathbf{K}A, z : A \to \bot \Rightarrow z : \bot}\,R \to}{xRy, y : \mathbf{K}A \Rightarrow y : (A \to \bot) \to \bot}\,R \to}{\Rightarrow x : \mathbf{K}A \to ((A \to \bot) \to \bot)}\,R \to$$

□

Theorem B.10 *(The MP rule) Suppose the sequents* $\Rightarrow x : A$ *and* $\Rightarrow x : A \to B$ *are derivable in* GIEL*, then the sequent* $\Rightarrow x : B$ *is also derivable.*

Proof.

$$\dfrac{\Rightarrow x : A \qquad \dfrac{\dfrac{\Gamma \Rightarrow \Delta, x : A \qquad x : B, \Gamma \Rightarrow \Delta}{xRx, x : A \to B, x : A \Rightarrow x : B}\,L \to}{x : A \to B, x : A \Rightarrow x : B}\,\text{Ref} \qquad \Rightarrow x : A \to B}{x : A \Rightarrow x : B}\,Cut$$

$\Gamma = xRx, x : A \to B, x : A, \Delta = x : B$ □

Theorem B.11 *If* $\vdash_{\text{IEL}} A$*, then* $\vdash_{\text{GIEL}} \Rightarrow x : A$*.*

Proof. Induct on the complexity of the proof of A. The result follows from Propositions B.7, B.8, B.9, and Theorem B.10. □

Theorem B.12 *(Completeness) For any variable* x *and formula* A*, if* $\Vdash_{\text{GIEL}} \Rightarrow x : A$*, then* $\vdash_{\text{GIEL}} \Rightarrow x : A$*.*

Proof. Suppose $\Vdash_{\text{GIEL}} \Rightarrow x : A$. By Lemma B.5, $\Vdash_{\text{IEL}} A$. By the completeness of IEL, $\vdash_{\text{IEL}} A$. Then by Theorem B.11, $\vdash_{\text{GIEL}} A$. □

Theorem B.13 *The following sequents are derivable in* GIEL*:*

(i) $\Rightarrow x : \mathbf{K}\bot \to \bot$

(ii) $\Rightarrow x : ((\mathbf{K}A \to ((A \to \bot) \to \bot)) \to \bot) \to \bot$

(iii) $\Rightarrow x : (A \to \bot) \to (\mathbf{K}A \to \bot)$

(iv) $\Rightarrow x : ((\mathbf{K}A \to A) \to \bot) \to \bot$

(v) $\Rightarrow x : \mathbf{K}(A \wedge B) \leftrightarrow (\mathbf{K}A \wedge \mathbf{K}B)$

Proof. Proof omitted. □

Proposition B.14 *The following propositions hold:*

(i) *The* $\mathbf{K}$*-necessitation rule, that is, if* $\Rightarrow x : A$*, then* $\Rightarrow x : \mathbf{K}A$*, is derivable in* GIEL*.*

(ii) *The deduction theorem, that is, if* $x : A \Rightarrow x : B$*, then* $\Rightarrow x : A \to B$*, is derivable in* GIEL*.*

(iii) GIEL *is normal intuitionistic modal logics.*

(iv) *Positive and negative introspection, that is,* $\Rightarrow x : \mathbf{K}A \to \mathbf{K}\mathbf{K}A$ *and* $\Rightarrow x : (\mathbf{K}A \to \bot) \to \mathbf{K}(\mathbf{K}A \to \bot)$*, are derivable in* GIEL*.*

Proof.

(i) Suppose $\Rightarrow x : A$ is derivable. Then, by the substitution rule, $\Rightarrow y : A$ is also derivable. By the weakening rule LW_2 get $xEy \Rightarrow y : A$. Then

$$\dfrac{xEy \Rightarrow y : A}{\Rightarrow x : \mathbf{K}A}\ R\mathbf{K}$$

(ii) Suppose $x : A \Rightarrow x : B$ is derivable. By the weakening rule LW_2, $xRx, x : A \Rightarrow x : B$ is derivable. The remaining derivation is as follows:

$$\dfrac{xRx, x : A \Rightarrow x : B}{\Rightarrow x : A \to B}\ R \to$$

(iii) This follows from Proposition B.7 and the $\mathbf{K}$-necessitation rule.

(iv) We derive from the root sequent $\Rightarrow x : \mathbf{K}A \to \mathbf{K}\mathbf{K}A$ as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{yRz, yEz, xRy, y : \mathbf{K}A \Rightarrow z : \mathbf{K}A}{yEz, xRy, y : \mathbf{K}A \Rightarrow z : \mathbf{K}A}\ \mathbf{KR1}}{xRy, y : \mathbf{K}A \Rightarrow y : \mathbf{K}\mathbf{K}A}\ R\mathbf{K}}{\Rightarrow x : \mathbf{K}A \to \mathbf{K}\mathbf{K}A}\ R \to$$

For the root sequent $\Rightarrow x : (\mathbf{K}A \to \bot) \to \mathbf{K}(\mathbf{K}A \to \bot)$, the derivation is as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{yRu, yRz, zRu, u : \mathbf{K}A, yEz, xRy \Rightarrow u : \bot, u : \mathbf{K}A \quad u : \bot, yRu, yRz, zRu, u : \mathbf{K}A, yEz, xRy, \Rightarrow u : \bot}{yRu, yRz, zRu, u : \mathbf{K}A, yEz, xRy, y : \mathbf{K}A \to \bot \Rightarrow u : \bot}\ L \to}{yRz, zRu, u : \mathbf{K}A, yEz, xRy, y : \mathbf{K}A \to \bot \Rightarrow u : \bot}\ \text{Trans}}{zRu, u : \mathbf{K}A, yEz, xRy, y : \mathbf{K}A \to \bot \Rightarrow u : \bot}\ \mathbf{KR1}}{yEz, xRy, y : \mathbf{K}A \to \bot \Rightarrow z : \mathbf{K}A \to \bot}\ R \to}{xRy, y : \mathbf{K}A \to \bot \Rightarrow y : \mathbf{K}(\mathbf{K}A \to \bot)}\ R\mathbf{K}}{\Rightarrow x : (\mathbf{K}A \to \bot) \to \mathbf{K}(\mathbf{K}A \to \bot)}\ R \to$$

□

C Decidability

Hereafter, we prove that the GIEL system supports terminating proof search. Below we provide a direct proof, where the main ideas and proof structure are based on the work of [11].

We first prove the weak subformula property of L_{IEL}-formulas, which serves as the foundation for subsequent analysis.

Theorem C.1 (Weak subformula property) *In any derivation of* GIEL*, every formula that appears satisfies one of the following two conditions: either it is a subformula of some formula in the end sequent, or it is a relational atom generated by the system's rules.*

Proof. It can be directly verified. □

According to the definition of formula complexity, the complexity of any subformula of a formula $x : A$ is strictly less than the complexity of $x : A$ itself. Every derivation in GIEL satisfies the weak subformula property.

During root-first search, the potential risks that may cause proof search to fail to terminate primarily stem from the following following five cases:

(1) The $(L \to)$,$(L\mathbf{K})$,$(\mathbf{KR1})$,$(\mathbf{KR2})$, (Trans) rule can be applied infinitely on the same principal formula.

(2) The (Ref) rule can introduce new relational atoms infinitely.

(3) Through the iterative use of $(L\mathbf{K})$,$(L \to)$, or other rules, an infinite chain of accessible worlds can be constructed.

(4) The (Ser) rule can be applied infinitely on different principal formulas.

For cases (1)-(3), discussions can be found in [11], with similar methods. Here we only briefly explain:

To prove that the space of root-first proof search is finite, we consider minimal derivations, i.e., derivations that cannot be shortened further. For case (1), lemma about rule permutation properties such repeated applications should be excluded when searching for minimal derivations. can be applied at most once on the same pair of principal formulas.

For case (2), we need the following lemma:

Lemma C.2 *In* GIEL*, for any minimal derivation of a sequent $\Gamma \Rightarrow \Delta$, all variables appearing in relational atoms of the form xRx that are introduced by the* (Ref) *rule and subsequently removed already occur in Γ or Δ.*

Proof. It can be proven by tracing the origin of the relational atom xRx upward along the derivation tree. □

For case (3), we need to prove the following key proposition:

Proposition C.3 *In a minimal derivation of a sequent $\Gamma \Rightarrow \Delta$ in* GIEL*, for each formula $x : \mathbf{K}A$ appearing in its positive part, the number of iterative applications of the $(R\mathbf{K})$ rule with $x_i : \mathbf{K}A$ as the principal formula on a chain of accessible worlds of the form $xRx_1, x_1Rx_2, \ldots$ is at most $n(\mathbf{K})$, where $n(\mathbf{K})$ represents the number of occurrences of the $\mathbf{K}$ operator in the negative part of $\Gamma \Rightarrow \Delta$.*

Proposition C.4 *In a minimal derivation of a sequent* $\Gamma \Rightarrow \Delta$ *in* GIEL*, for each formula* $x : A \to B$ *appearing in its positive part, the number of iterative applications of the* $(L \to)$ *rule with* $x_i : A \to B$ *as the principal formula on a chain of accessible worlds of the form* $xRx_1, x_1Rx_2, \ldots$ *is at most* $n(\to)$*, where* $n(\to)$ *represents the number of occurrences of the* $\to$ *operator in the negative part of* $\Gamma \Rightarrow \Delta$*.*

For case (4), we need to prove the following key proposition:

Proposition C.5 *In a minimal derivation of a sequent* $\Gamma \Rightarrow \Delta$ *in* GIEL*, the number of applications of the Ser rule is at most* M*, where* $M = 2^{2|F|}$*, and* F *is the set of all subformulas of all formulas in* $\Gamma \Rightarrow \Delta$ *(including all subformulas and atomic formulas).*

Proof. Let S be the set of all formulas in $\Gamma \Rightarrow \Delta$, and define F as the set of all subformulas of S (including all subformulas and atomic formulas). Since S is finite, F is also finite. By the weak subformula property, all formulas appearing in the proof search belong to F. For each label x, define its state as the tuple $(\text{Left}(x), \text{Right}(x))$, where $\text{Left}(x) = \{A \in F \mid x : A \in \Gamma\}$ and $\text{Right}(x) = \{A \in F \mid x : A \in \Delta\}$. Since F is finite, the number of possible states is at most $2^{2|F|}$.

The Ser rule requires introducing a new label y and the relation xEy. During proof search, when applying the Ser rule, we check whether there exists an existing label z such that:

- $\text{Left}(z) = \text{Left}(y)$ and $\text{Right}(z) = \text{Right}(y)$ (same state),
- For all other labels w, the relations satisfy zRw if and only if yRw, wRz if and only if wRy, zEw if and only if yEw, and wEz if and only if wEy (consistent relational pattern).

If such a z exists, then reuse z instead of introducing a new label y. Since the number of states is finite, the total number of labels is at most $2^{2|F|}$. In a minimal derivation, if the Ser rule is applied multiple times for the same x but the introduced labels have the same state, then the derivation is not minimal. Therefore, the number of applications of the Ser rule is bounded by $M = 2^{2|F|}$. □

The above analysis shows that all cases that may cause non-termination in the GIEL system can be controlled within the framework of minimal derivations, thus ensuring that the proof search process necessarily terminates.

The Identity Rule Reconsidered: from Logic Programming to Formal Theories of Truth

Shuwen Wu [1]

Fudan University
Shanghai 200433, P. R. China

Abstract

This paper compares three systems that restrict the use of the identity rule or related principles: Kreuger's restriction in definitional reflection (1994), Schroeder-Heister's LI system (2016), and French's substructural system LKR (2016). Although these systems were developed in different contexts—logic programming, proof theory, and theories of truth—they share a common insight: identity is not merely a trivial background axiom; rather, it is a structural principle that requires further examination and clarification. Kreuger initiated a systematic reflection on reflexivity by restricting the form of initial sequents in logic programs. Schroeder-Heister's system demonstrates that by limiting identity, contraction and cut can remain admissible, thereby revealing the structural trade-offs among these rules. By contrast, the development of formal substructural theories of truth has been slower and remains incomplete. French proposed a non-reflexive substructural framework for truth by eliminating structural reflexivity to block paradoxes; however, compared with the other two approaches, this framework is relatively weak, with all inferences existing only at the metasequent level. Overall, the comparison highlights a common methodological theme: reflexivity is not entirely harmless. By weakening or abandoning the reflexive principle, one can sacrifice some derivability in exchange for desirable properties such as consistency. This suggests that, just like contraction and cut, reflexivity deserves systematic investigation within substructural theories of truth.

Keywords: identity rule, sequent calculus, structural restrictions, semantic paradox, truth theories

1 Introduction

Theories of truth and the study of semantic paradoxes have long motivated modifications of classical logic. A central line of research has focused on the role of structural rules, especially contraction and transitivity. For example, non-contractive approaches have been developed by Petersen (2000, [1]) and Zardini (2011, [2]), while Ripley (2013, [3]) has proposed a non-transitive approach. The restriction or elimination of contraction, for instance, is at the heart of

[1] wwushuwen@163.com

substructural truth theories that seek to block paradoxical derivations without abandoning transparency. By contrast, the role of the identity rule has received comparatively little systematic attention.

The identity rule—also referred to as reflexivity—is the initial sequent $A \vdash A$, assumed for any formula A. Yet this seemingly trivial assumption encodes important proof-theoretic commitments. To accept it unconditionally is to assume that any formula can be taken as derivable from itself, regardless of whether its content is reducible or defined in terms of other expressions. Once this assumption is questioned, new possibilities arise for both proof theory and the theory of truth.

The central motivation of this paper is to reconsider the identity rule by comparing three contexts in which it, or closely related principles, have been explicitly restricted. The first is logic programming with definitional reflection, where Kreuger (1994, [4]) proposed a restricted form of identity, allowing $A \vdash A$ only when A cannot be further reduced. The second is Schroeder-Heister's LI system (2016, [5]), which systematically investigates the consequences of restricting initial sequents, highlighting the trade-offs between identity, contraction, and cut. The third is French's LKR system (2016, [6]), a substructural calculus that removes structural reflexivity from Gentzen's sequent framework and introduces a truth predicate.

This paper has three aims: (i) to survey and compare these restrictions (Kreuger, Schroeder-Heister, French); (ii) to identify a common methodological theme, namely that reflexivity, though often regarded as an innocuous rule, turns out to be more problematic than it is usually assumed to be and requires further examination; and (iii) to explore the implications of this theme for theories of truth and the handling of semantic paradoxes.

2 Preliminaries

2.1 Substructural logic

Let us first briefly introduce what is meant by substructural logic. Substructural logics are so called because they arise from systematic modifications of the structural rules in Gentzen-style sequent calculi. By restricting or removing rules such as weakening, contraction, exchange, or cut, one obtains a family of logics that depart from the classical structural background while retaining much of the original inferential framework.

Sequent calculus was originally developed by Gerhard Gentzen(1935, [8]) as a tool for formal reasoning. Gentzen believed that the axiomatic systems developed by Frege, Russell, and Hilbert failed to reflect the actual process of human reasoning. In response, he proposed the method of natural deduction, aiming to construct a formal system that more closely mirrors our real inferential practices.

Within this framework, inference rules are divided into operational rules and structural rules. Operational rules are those that introduce a logical connective in the principal formula (as Paoli (2002, [7]) states, "operational rules are those which introduce a connective in their principal formula"), while all other rules

are classified as structural. In contrast to operational rules, structural rules do not introduce logical symbols; rather, they operate on the structure of sequents themselves. Apart from the cut rule, they typically appear in left/right pairs within Gentzen's LK system. Structural rules have traditionally been regarded as essential for the derivation systems of classical and intuitionistic logic alike and have played a crucial role in the development of logic as a methodology of rigorous scientific reasoning.

2.2 Sequent Calculus and Structural Rules

We will use this background as the basis for comparing three different systems, each representing a distinct approach to restricting the identity rule or related reflexive principles.

We begin with the sequent calculus of Classical Propositional Logic (CL) as a basis theory for our introduction. Let Γ and Δ be finite multisets of formulas, and let ϕ and ψ denote individual formulas. A sequent takes the form $\Gamma\Rightarrow \Delta$, where the arrow $\Rightarrow$ indicates the sequent relation, expressing a consequence relation. The operational and structural rules of CL are defined as follows:

Structural rules

$$\text{(Reflexivity)Id}\quad \frac{}{\phi \Rightarrow \phi}$$

$$\text{(Weakening)LK}\quad \frac{\Gamma \Rightarrow \Delta}{\Gamma, \Sigma \Rightarrow \Delta} \qquad \text{RK}\quad \frac{\Gamma \Rightarrow \Delta}{\Gamma \Rightarrow \Pi, \Delta}$$

$$\text{(Contraction)LW}\quad \frac{\Gamma, \phi, \phi \Rightarrow \Delta}{\Gamma, \phi \Rightarrow \Delta} \qquad \text{RW}\quad \frac{\Gamma \Rightarrow \phi, \phi, \Delta}{\Gamma \Rightarrow \phi, \Delta}$$

$$\text{Cut}\quad \frac{\Gamma \Rightarrow \phi, \Delta \quad \Gamma, \phi \Rightarrow \Delta}{\Gamma \Rightarrow \Delta}$$

Operational rules

$$\text{L}\neg\quad \frac{\Gamma \Rightarrow \phi, \Delta}{\neg\phi, \Gamma \Rightarrow \Delta} \qquad \text{R}\neg\quad \frac{\Gamma, \phi \Rightarrow \Delta}{\Gamma \Rightarrow \neg\phi, \Delta}$$

$$\text{L}\wedge\quad \frac{\Gamma, \phi \Rightarrow \Delta \quad \Gamma, \psi \Rightarrow \Delta}{\Gamma, \phi \wedge \psi \Rightarrow \Delta} \qquad \text{R}\wedge\quad \frac{\Gamma \Rightarrow \phi, \Delta \quad \Gamma \Rightarrow \psi, \Delta}{\Gamma \Rightarrow \phi \wedge \psi, \Delta}$$

$$\text{L}\vee\quad \frac{\Gamma, \phi \Rightarrow \Delta}{\Gamma, \phi \vee \psi \Rightarrow \Delta} \qquad \text{R}\vee\quad \frac{\Gamma \Rightarrow \phi, \Delta}{\Gamma \Rightarrow \phi \vee \psi, \Delta}$$

To extend CL with a truth predicate, we introduce the following rule:

$$LTr\ \frac{\Gamma, \phi \Rightarrow \Delta}{\Gamma, Tr(\ulcorner\phi\urcorner) \Rightarrow \Delta} \qquad RTr\ \frac{\Gamma \Rightarrow \phi, \Delta}{\Gamma \Rightarrow Tr(\ulcorner\phi\urcorner), \Delta}$$

Where $\ulcorner\phi\urcorner$ is the name of the sentence ϕ, and Tr$\ulcorner\phi\urcorner$ means that the sentence is true. The double horizontal lines indicate that this is a rule that works both top-down and bottom-up.

It is known that the liar paradox leads to a contradiction in classical logic. Let λ be the liar sentence, that is, λ is a sentence referring to $\neg Tr\ulcorner\lambda\urcorner$. We perform the derivation using the above sequent method as follows:

$$\dfrac{\dfrac{\dfrac{\dfrac{}{Tr(\ulcorner\lambda\urcorner) \Rightarrow Tr(\ulcorner\lambda\urcorner)}\,\mathrm{Id}}{\Rightarrow \neg Tr(\ulcorner\lambda\urcorner), Tr(\ulcorner\lambda\urcorner)}\,\mathrm{R}\neg}{\Rightarrow Tr(\ulcorner\lambda\urcorner), Tr(\ulcorner\lambda\urcorner)}\,\mathrm{R}Tr}{\Rightarrow Tr(\ulcorner\lambda\urcorner)}\,\mathrm{RW} \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{Tr(\ulcorner\lambda\urcorner) \Rightarrow Tr(\ulcorner\lambda\urcorner)}\,\mathrm{Id}}{\neg Tr(\ulcorner\lambda\urcorner), Tr(\ulcorner\lambda\urcorner) \Rightarrow}\,\mathrm{L}\neg}{Tr(\ulcorner\lambda\urcorner), Tr(\ulcorner\lambda\urcorner) \Rightarrow}\,\mathrm{L}Tr}{Tr(\ulcorner\lambda\urcorner) \Rightarrow}\,\mathrm{LW}$$

$$\dfrac{\Rightarrow Tr(\ulcorner\lambda\urcorner) \qquad Tr(\ulcorner\lambda\urcorner) \Rightarrow}{\Rightarrow}\,\mathrm{Cut}$$

As shown above, the derivation ends with the empty sequent. For some logicians, an empty sequent signifies a contradiction; even if it does not, obtaining an empty sequent is bad—in the presence of weakening, it means that any sequent can be derived. Therefore, substructural logicians have argued that, beyond investigations into negation, as in Kripke's fixed-point approach (1975, [9]), or into the truth predicate, as in Tarski's way (1933, [10]), structural rules themselves deserve careful scrutiny. Most attention has traditionally been given to contraction, cut, exchange, and weakening, whose restriction can indeed prevent paradoxical derivations of the empty sequent. By contrast, the identity rule has often been taken for granted and has not received comparable attention. In sequent calculi, the identity rule provides a "reflexive anchor" for any formula A, allowing the immediate derivation of $A \vdash A$. This seemingly trivial principle, however, embodies significant proof-theoretic commitments. If identity is applied without restriction to all formulas, it may also facilitate circular reasoning or non-terminating derivations in certain contexts.

This observation motivates a further line of inquiry: can restrictions on the identity rule, analogous to those on contraction or cut, yield desirable meta-properties such as consistency or normalizability? It is precisely along these lines that several systems have been developed — from Kreuger's restriction of identity in the context of definitional reflection, to Schroeder-Heister's LI system, and French's LKR system. Taken together, these approaches reveal that even the most basic structural rule, identity, merits renewed investigation.

3 Approach

We focus on three different systems in which either the identity rule or closely related principles are restricted: Kreuger's proposal in the context of logic programming, Schroeder-Heister's system LI, and French's system LKR. Although these systems arise from different motivations—logic programming, proof theory, and the philosophy of truth—they converge on a shared theme: reflexivity is not as innocuous as it is typically assumed to be; on the contrary, by weakening or abandoning the principle of reflexivity, one can obtain desirable properties such as consistency. In particular, the relatively successful results obtained in Kreuger's and Schroeder-Heister's systems may provide instructive insights for the design of substructural theories of truth.

3.1 Kreuger (1994): Restricted Identity in Definitional Reflection

Kreuger's contribution arose within the framework of logic programming with definitional reflection. In this setting, atomic formulas are not entirely irreducible; rather, they are associated with definitional rules that govern their

meaning. The unrestricted use of the identity axiom, $A \vdash A$, risks trivializing the role of these definitions. Kreuger therefore proposed restricting initial sequents to irreducible formulas only, i.e., to those that cannot be further reduced via definitional rules.

The effect of this modification is twofold. On the computational side, it secures determinism in the operational semantics of logic programming languages such as GCLA II. On the proof-theoretic side, it enables cut elimination, since cycles introduced by definitional reflection are blocked when reducible formulas are denied an identity anchor. Kreuger's system thus exemplifies how limiting identity yields both technical and conceptual clarity: derivations proceed only when anchored in genuine atomic data, not in definitional artifacts.

3.2 Schroeder-Heister (2016): The LI System

Schroeder-Heister's LI system represents a systematic exploration of restricting initial sequents in sequent calculus. The central observation is that the unrestricted identity axiom interacts with contraction and cut in non-trivial ways. By limiting identity—for example, to atomic or otherwise restricted formulas—one can reestablish cut elimination while avoiding certain undesirable derivations.

The distinctive feature of LI is its explicit articulation of trade-offs:

- With full identity, contraction and cut may undermine normalization.
- With restricted identity, cut elimination can be recovered, but some derivability is lost.
- Different versions of identity restriction yield different balances between expressive strength and proof-theoretic discipline.

The LI system demonstrates that identity is not a trivial background axiom but a central structural rule whose regulation reshapes the meta-theory of sequent calculus. Schroeder-Heister thereby highlights a neglected dimension in the design of substructural systems, extending the methodological options for handling paradox and circularity.

3.3 French (2016): The LKR System

French's system LKR addresses the problem of semantic paradoxes by removing structural reflexivity from Gentzen's sequent calculus and introducing a truth predicate. The removal of reflexivity blocks paradox-generating derivations that would otherwise render the system inconsistent once a transparent truth predicate is added.

The motivation behind LKR is to avoid the triviality that arises when a transparent truth predicate is added to Gentzen's multiple-conclusion sequent calculus. By removing structural reflexivity and allowing derivability only at the level of metasequents, French shows that the paradox-generating patterns responsible for collapse can be blocked while preserving the classical operational rules. LKR therefore provides a non-reflexive substructural setting in which transparency and consistency are jointly maintained.

3.4 Comparative Strategy

Our method is to analyze each system in terms of two dimensions:

- Restricted principle:
  - Kreuger: In the system of definitional reflection within the logic programming language GCLA, the identity rule is restricted to irreducible atoms. The motivation for this restriction is to impose an operational direction on definitions, thereby avoiding indeterminacy in definitional reasoning.
  - Schroeder-Heister: Within an intuitionistic framework, he further develops this approach by systematically examining the behavior of the system when initial sequents are limited to uratoms.
  - French: Within the traditional multi-conclusion sequent calculus (LK), he removes structural reflexivity and, by introducing a truth predicate, constructs a non-reflexive substructural framework for truth.
- Gained property:
  - Kreuger: By restricting the initial identity rule, the system achieves operational determinism while preserving its soundness and completeness.
  - Schroeder-Heister: Reveals a controlled trade-off among identity, contraction, and cut: If the system contains (explicit or implicit) contraction and admits cut, it becomes inconsistent; Without contraction, cut is admissible; When identity is restricted to uratoms, both contraction and cut remain admissible.
  - French: Avoiding reflexivity, the system achieves consistency in a truth-theoretic setting, although derivability is confined to the metasequent level.

Taken together, these three systems illustrate a shared methodological theme: reflexivity is not as natural or harmless as it appears. Weakening or abandoning the reflexive rule may reduce derivability, yet it restores desirable proof-theoretic properties such as determinism and consistency. This offers a new perspective on the role of structural rules in formal theories of truth and indicates that non-reflexive substructural approaches still require further semantic and inferential development.

4 Final Remarks

Our findings support three main claims. First, the identity rule plays a far more substantial role than its traditional presentation suggests. Second, restricting reflexivity yields a clear methodological trade-off: some inferential strength is sacrificed, yet desirable properties such as consistency are restored. Third, these insights carry philosophical significance for the theory of truth: by blocking the circularity exploited in the Liar and Curry paradoxes, identity restrictions provide a structural means of developing consistent and transparent truth theories. Moreover, the application of reflexivity restrictions in other areas of logic provides additional support for this claim, as well as concrete methods that can be explored within a truth-theoretic framework.

References

[1] Petersen, H. (2000). A paraconsistent approach to semantic paradoxes. *Studia Logica*, 65(3), 317–342.
[2] Zardini, E. (2011). Truth without contra(di)ction. *Review of Symbolic Logic*, 4(4), 498–535.
[3] Ripley, D. (2013). Paradoxes and failures of cut. *Australasian Journal of Philosophy*, 91(1), 139–164.
[4] Kreuger, P. (1994). Axioms in definitional calculi. In R. Dyckhoff (Ed.), *Extensions of Logic Programming* (pp. 196–205). Springer, Berlin, Heidelberg.
[5] Schroeder-Heister, P. (2016). Restricting Initial Sequents: The Trade-Offs Between Identity, Contraction and Cut. In *The Nature of Logic* (pp. 339–351). Springer. https://doi.org/10.1007/978-3-319-29198-7_10
[6] French, R. (2016). Structural Reflexivity and the Paradoxes of Self-Reference. *Ergo: An Open Access Journal of Philosophy*, 3.
[7] Paoli, F. (2002). *Substructural Logics: A Primer*. Springer, Dordrecht.
[8] Gentzen, G. (1935). Investigations into Logical Deduction. In M. E. Szabo (Ed.), *The Collected Papers of Gerhard Gentzen*. North-Holland, 1969.
[9] Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72(19), 690–716.
[10] Tarski, A. (1933). The concept of truth in formalized languages. In J. Corcoran (Ed.), *Logic, Semantics, Metamathematics* (pp. 152–278). Indianapolis: Hackett Publishing Company, 1983. (Originally published in Polish, 1933.)

Toward learning and reasoning in first-order justification logic

Stipe Pandžić [1]

LUCI Lab, University of Milan
Via Festa del Perdono 7, Milan, Italy

Abstract

In his seminal 1980 paper, Reiter introduced default logic to formalize reasoning with incomplete information. During the 1990s, researchers explored default logic as a qualitative counterpart to inductive-statistical reasoning. A key limitation of Reiter's original framework is its inability to eliminate unintended extensions, where defeated conclusions, such as Tweety the bird flying despite being a penguin, cannot be retracted. We propose first-order default justification logic as a novel formal system to qualitatively represent uncertainty resulting from learning and generalization. We demonstrate how to encode default reasoning schemas in this system, avoiding the pitfall of unintended extensions. Using (quantifier-free) first-order justification logic, we efficiently formalize rules with exceptions directly in the object language, rendering extra-logical solutions, such as rule prioritization, unnecessary. We establish that the basic system is well-behaved and argue that it offers an intuitive logical foundation for integrating learning and reasoning in artificial intelligence.

Keywords: Default logic, Justification logic, Learning and reasoning, Undercutting attacks, Exclusionary reasons.

1 Introduction

In this paper, we revisit default logic, one of the foundational system for commonsense reasoning in AI. We explore how introducing default *schemas* with first-order justification logic formulas brings us close to the desiderata that motivated standard Reiter's default logics [18]. Reiter's logic was already investigated as a reinvention of inductive, statistical reasoning [21]. The added value of first-order justification logic [2] in this context is that it is expressive enough to formalize rules with exceptions in the object-language. This, in turn, enables us to formalize useful forms of higher-order evidence, such as, undercutting attacks. We show how the resulting first order justification logic default

[1] Email: stipepandzic@gmail.com
The author gratefully acknowledges support from the Italian Ministry of University and Research through the "Reasoning with Data" project (ReDa, G53C23000510001, awarded to Prof. Hykel Hosni under the FIS1 Advanced Grant scheme).

theories are well-behaved and, finally, contextualize this logic within the larger project of establishing justification logic syntax for integration of learning and reasoning.

The paper starts by exploring quantifier-free first-order justification logic, which is the main formal component of the nonmonotonic logic presented in Section 4. We revisit the famous 'Tweety' example to showcase the possibilities that merging default reasoning and first-order justification logic offers. In Section 5, we present a concrete application of the system's capability to integrate learning and reasoning, illustrated by a simple example involving a traffic sign database. Finally, we conclude the paper and indicate the current state of the larger project behind this logical system.

2 Quantifier-free first-order logic of reasons

Standard justification logic (JL) formulas $t{:}F$ are interpreted as assertions of the type 't is a reason for F'. In the language of first-order JL, the general form of justification assertions is $t{:}_X F$, where t is a proof term, X is a finite set of individual variables, and F is a formula. This formula can be read as stating that t represents a reason for F such that the variables in X can be substituted for as placeholders in the reasoning encoded by t. The idea of a placeholder variable can be illustrated as follows. Given a proof of $F(x)$ and a constant a, we can replace all the occurrences of x in the proof of $F(x)$ and obtain a proof of $F(a)$ [7, p. 226]. We use a quantifier-free fragment of the first-order justification logic language L, which is a first-order logic language with individual variables and n-ary predicate symbols, but without equality and extended with reason variables and constants and operations on justification terms ('$+$' and '$\cdot$'). The formal definition of the qfFOJT is given by the following two-layer grammar of reason terms (or 'reason polynomials') and formulas:

$$t ::= p \mid c \mid (t \cdot t) \mid (t + t),$$

where $p_1, p_2, \ldots p_j$ are *reason variables* and $c_1, c_2, \ldots c_k$ are *reason (or proof) constants* and

$$F ::= \top \mid P_{x_1,\ldots,x_n} \mid (F \to F) \mid (F \vee F) \mid (F \wedge F) \mid \neg F \mid t{:}_X F,$$

where $P_1, P_2, \ldots, P_n$ are *n-ary predicates* and X is a finite set of *individual variables* $x_1, x_2, \ldots, x_m$.[2]

We will work with a minimal set of axioms and inference rules of the **quantifier-free first-order variant of the logic of factive** or truth-inducing **justifications** JT, that is, qfFOJT.[3]

Axioms:

[2] Note the difference between the reason variables $p_1, p_2, \ldots p_j$ representing arbitrary or unknown reasons and the individual variables $x_1, x_2, \ldots, x_m$ representing arbitrary objects from the domain.

[3] See, e.g., [7, p. 227] for an axiomatization of first-order logic of proofs. Except for omitting the generalization rule, we do not consider the positive introspection axiom $t{:}_X F \to$

A1 Classical QF-FOL Axioms
A2 $t{:}_{Xy}F \to t{:}_X F$ (where y does not occur in F)
A3 $t{:}_X F \to t{:}_{Xy} F$
B1 $t{:}_X F \to F$
B2 $s{:}_X(F \to G) \to (t{:}_X F \to (s \cdot t){:}_X G)$
B3 $t{:}_X F \to (t + s){:}_X F$
$s{:}_X F \to (t + s){:}_X F$

Rules:
R1 If $\vdash F, F \to G$ then $\vdash G$
R3 $\vdash c{:}_\emptyset A$ (A is an axiom, c is a proof constant)

The axioms and rules of qfFOJT are inherently monotonic. We use them to describe certain information and fully specified states of affairs. The monotonic component of the system determines the behavior of the initial set of facts W in default logic.

The semantics for qfFOJT is given by Mkrtychev models, which are essentially one-world Fitting models for first-order justification logics.

Definition 2.1 [Mkrtychev model] A *Mkrtychev* qfFOJT *model* is defined as $\mathcal{M} = \langle \mathcal{D}, \mathcal{I}, \mathcal{E} \rangle$, where $\mathcal{D}$ is the domain of the model and $\mathcal{I}$ an interpretation assigning each n-ary predicate symbol with a set of n-tuples of elements of $\mathcal{D}$, and $\mathcal{E}$ is an evidence function mapping each justification term t and formula F to a truth value, $\mathcal{E}(t, F)$. The evidence function needs to meet the following conditions:

$$\begin{aligned} (\cdot) &\quad \text{if } \mathcal{E}(t, F \to G) \;\&\; \mathcal{E}(u, G), \text{ then } \mathcal{E}((t \cdot u), G); \\ (+) &\quad \text{if } \mathcal{E}(t, F) \text{ or } \mathcal{E}(u, F), \text{ then } \mathcal{E}((t + u), F); \\ \text{(Instantiation)} &\quad \text{if } a \in \mathcal{D}, \text{then, if } \mathcal{E}(t, F(x)) \text{ then } \mathcal{E}(t, F(a)). \end{aligned}$$

The truth relation for Mkrtychev models, $\mathcal{M} \models F$, is defined as usual for connectives, with the special case of

$$\mathcal{M} \models t{:}_X F(\boldsymbol{x}) \;\text{ iff } \mathcal{E}(t, F(\boldsymbol{x})) \text{ and } \mathcal{M} \models F(\boldsymbol{a}) \text{ for all } \boldsymbol{a} \in \mathcal{D}.$$

We assume *reflexivity* for qfFOJT Mkrtychev models, that is, we assume that if $\mathcal{E}(t, F)$, then $\mathcal{M} \models F$. This is important to regulate certainty of starting premises for default reasoning. For a set of qfFOJT formulas $S \subseteq L$, we define the closure operation $\mathrm{Th}_L(S)$ as a set of all the qfFOJT formulas F such that $S \models F$. In constructing qfFOJT default theories, we will use the $\mathrm{Th}_L^-(S)$ closure that does not meet the $(+)$ condition.

Based on the grammar of first-order justification logic, we introduce the

---

$!t{:}_X(t{:}_X F)$ because this property is not essential for our system. Quantifier-free versions of first-order justification logics have not been thoroughly studied, with a brief mention of quantifier-free logic of proofs in [1, pp. 499-502]

following default rule logic schemas with justifications:

$$\frac{t{:}_{\{x\}}F :: (u \cdot t){:}_{\{x\}}G}{(u \cdot t){:}_{\{x\}}G},$$ [4]

where x in $t{:}_{\{x\}}F$ is a free variable throughout the derivation t. Such rules could be read as "If t is a reason for x being F and if it is consistent to assume that $(u \cdot t)$ is a reason for x being G, then conclude that $(u \cdot t)$ is a reason for x being G. Notice how the default schema inherits the (monotonic) application step in the B2 axiom, that is, $u{:}_{\{x\}}(F \to G) \to (t{:}_{\{x\}}F \to (u \cdot t){:}_{\{x\}}G)$, but without the assumption that the justified conditional $u{:}_{\{x\}}(F \to G)$ has been established.

The proposed way of thinking about default generalizations in qfFOJL is analogous to the way in which such generalizations are handled in Reiter's logic. From the standard default schema

$$\frac{F(x) : G(x)}{G(x)},$$

we can reconstruct the monotonic implication formula

$$[(F(x) \to G(x)) \wedge F(x)] \to G(x).$$

As in the case of the qfFOJL default, the first conditional of this first-order formula is not established to hold without exceptions. The difference is that the latter formula does not track justifications and it is purely truth-functional.

Later, we argue that justification-enriched rules provide a more powerful framework for encoding statistical regularities of the kind that originally motivated Reiter's default rules. In the next section, we juxtapose Reiter's defaults with first-order JL defaults. We show how default theories based on first-order JL default schemas solve the problem of 'unintended' extensions while retaining desirable properties of Reiter's default logic such as, e.g., existence of extensions.

3 Tweety example and higher-order evidence

The paradigmatic example of a default generalization that birds (usually) fly is given by the following Reiter's schema:

$$\frac{\mathit{bird}(x) : \mathit{flies}(x)}{\mathit{flies}(x)},$$

Substituting for the free variable x, we get the default rule $\frac{\mathit{bird}(tw):\mathit{flies}(tw)}{\mathit{flies}(tw)}$, where t is a domain object representing 'Tweety'. Moreover, assume that

[4] The use of the double-colon '::' is simply to make it distinct from the colon ':' operator used in defining the qfFOJT syntax. In the Reiter's format of defaults we present below, a single colon ':' delimits the prerequisite and the consistency condition formula.

the following schema formalizes the intuition that penguins do not fly: $\frac{penguin(x):\neg flies(x)}{\neg flies(x)}$.[5] Together with the starting facts $\{bird(tw), penguin(tw)\}$. Given the two default schemas, the following two sets define the resulting Reiter's default theory:

$$\{bird(tw), penguin(tw)\}, \left\{\frac{bird(tw) : flies(tw)}{flies(tw)}, \frac{penguin(tw) : \neg flies(tw)}{\neg flies(tw)}\right\}.$$

Already at the level of the 'Tweety example', Reiter's default theories cannot accommodate for the induced non-monotonicity. The default theory above has two logically indistinguishable extensions, namely, $\{bird(tw), penguin(tw), flies(tw)\}$ and $\{bird(tw), penguin(tw), \neg flies(tw)\}$. Of course, only the second one can be intuitively endorsed as a correct extension, but standard default logic does not provide formal resources to distinguish the two. That is, there are no available logical means to express that the reason justifying the formula *flies*(tw) has been *excluded* on the grounds of the fact that *penguin*(tw).

Example 3.1 In the qfFOJT syntax, the following schemas represent the required gneralizations:

$$\frac{r{:}_{\{x\}}B :: (s \cdot r){:}_{\{x\}}F}{(s \cdot r){:}_{\{x\}}F} \text{ and } \frac{t{:}_{\{x\}}P :: (u \cdot t){:}_{\{x\}}\neg F}{(u \cdot t){:}_{\{x\}}\neg F},$$

namely that birds usually fly and penguins do not fly, respectively. Given the starting set of facts $\{r{:}_{\{tw\}}B, t{:}_{\{tw\}}P\}$ for $tw \in \mathcal{D}$, we get the following two default rule instantiations: $\frac{r:_{\{tw\}}B::(s\cdot r):_{\{tw\}}F}{(s\cdot r):_{\{tw\}}F}$ and $\frac{t:_{\{tw\}}P::(u\cdot t):_{\{tw\}}\neg F}{(u\cdot t):_{\{tw\}}\neg F}$. With the two schemas above, the qfFOJT generates two inconsistent extensions, analogously to Reiter's theory. However, with qfFOJT defaults, we can go a step further in representing the status of an excluded reason by representing reasoning about higher-order evidence in default reasoning. The following default schema formalizes the reasoning pattern needed to exclude the reason for the conclusion that Tweety flies:

$$\frac{t{:}_{\{x\}}P :: (v \cdot t){:}_{\{x\}}\neg s{:}_{\{x\}}(B \to F)}{(v \cdot t){:}_{\{x\}}\neg s{:}_{\{x\}}(B \to F)}.$$

The consequent of the rule

$$\frac{t{:}_{\{tw\}}P :: (v \cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B \to F)}{(v \cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B \to F)},$$

namely, $(v \cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B \to F)$, is added to both extensions of the theory resulting from the facts and schemas above. Intuitively, the formula says that

[5] It is well-known that the original suggestion of encoding this regularity via material implication $\forall x(penguin(x) \to \neg flies(x))$ does not scale if we were to allow exceptions to exceptions, as we should in the domain of commonsense reasoning. For example, one such exception to this generalization is that a genetically altered penguin might be able to fly [17, p. 294].

the default reasoning step codified in the term 's', allowing for the inference that the object *tw* is such that *tw* 'flies', was not justified in the context of the new information provided by P. The latter, higher-order type of evidential reasoning is itself codified in the term $(v \cdot t)$.

The type of conflict induced by the term $(v \cdot t)$ has been elusive for the logics of default reasoning, and corresponds to Pollock's [15] notion of **undercut**.[6] The above rule exemplifies the typical formal pattern of the undercut, where a reason u undercuts $(s \cdot t)$ as a reason for a being F because u compromises the default rule $\frac{t:_{\{a\}}E::(s\cdot t):_{\{a\}}F}{(s\cdot t):_{\{a\}}F}$ in light of the justified conclusion that $u:_{\{a\}}\neg[s:_{\{a\}}(E \rightarrow F)]$. Importantly, notice that the undercut does not affect the generalization expressed in the default schema $\frac{t:_{\{x\}}P::(v\cdot t):_{\{x\}}\neg s:_{\{x\}}(B\rightarrow F)}{(v\cdot t):_{\{x\}}\neg s:_{\{x\}}(B\rightarrow F)}$, as expected from a rule here applied to a specific case of the object *tw*. In the next section, we capture this pattern in a definition that underpins formal notions of *conflict-free* sets and *admissible* sets.

As we show in the following section, the non-monotonic system based on the qfFOJT is able to represent defeasibility by its unique treatment of reason terms. In specific, we develop default theories where two terms may 'defeat' each other by simply supporting opposing conclusions, e.g., $s:_{\{a\}}F$ and $t:_{\{a\}}\neg F$. Delving deeper into defeasibility, the following section accounts for those reason terms that exclude the applicability of other (default) reason terms, thus providing a logical theory of *undercutting* defeat.

4 qfFOJT default theories

We informally introduced qfFOJT default theories in Example 3.1. The following definition formalizes the intuition:

Definition 4.1 [qfFOJT Default Theory] A qfFOJT default theory Δ is a pair $\langle W, D\rangle$, where the set W is a finite set of qfFOJT formulas and D is a countable set of default rules with qfFOJT formulas.

Notice that the rules are not default schemas, but rather a result of assigning values to free variables through a ground substitution.

To formalize the central notion of a default extension, we first define *reason extension* sets via a quasi-inductive characterization due to [18, p. 89].

Definition 4.2 [Reason Extension] For a qfFOJT default theory $\Delta = \langle W, D\rangle$,

[6] The standard way to deal with undercut in default logic is by imposing priority orderings on default rules [5,9].

we define

$$R_0 = \{W\}, \quad \textit{and for } i \geq 0$$
$$R_{i+1} = \left\{ \mathrm{Th}_L\left(R_i \cup \{(u \cdot t){:}_{\{a\}}G\}\right) \;\middle|\; \begin{array}{l} \left(\frac{t{:}_{\{a\}}F \quad :: \quad (u\cdot t){:}_{\{a\}}G}{(u\cdot t){:}_{\{a\}}G}\right) \in D, \\ t{:}_{\{a\}}F \in R_i, \\ \neg(u \cdot t){:}_{\{a\}}G \notin R \end{array} \right\}. \textit{Then}$$
$$R = \cup_{i=0}^{\infty} R_i \;\; \textit{is a reason extension for } \Delta, \; \textit{where } \mathrm{Th}_L(R) \subseteq L.$$

Notice that the definition is referred to as 'quasi-inductive', rather than simply 'inductive', because we use R to define R_{i+1}. For a qfFOJT default theory Δ, we define the reasoning space set as follows: $\mathcal{R}(\Delta) = \{R \subseteq L \mid R \text{ is a reason extension of } \Delta\}$.

Against the background of reasoning space, we now specify how reasons interact with each other to form 'defensible' sets of reasons. The fundamental type of conflict is known as *undercut*.

Definition 4.3 [Undercut] For some qfFOJT formulas $t{:}_{\{a\}}F$ and $u{:}_{\{a\}}\neg[v{:}_{\{a\}}(G \rightarrow H)]$ in R, where t, u and v are some specific reason polynomials and v is a subterm of t, a reason u undercuts t as a reason for $a \in \mathcal{D}$ being F iff for a reason variable p such that $p{:}_{\{a\}}G$ is in R, it holds that $(v \cdot p){:}_{\{a\}}H$ is also in R.

Notice the basic undercut case where $F = H$ and $t = (v \cdot p)$. To see what makes this relation an *undercutting* relation in specific, consider the following set of qfFOJT formulas $\{s{:}_{\{b\}}A, (r \cdot s){:}_{\{b\}}B, u{:}_{\{b\}}\neg[r{:}_{\{b\}}(A \rightarrow B)]\}$. This set is not inconsistent according to the qfFOJT consequence relation. However, there is an asymmetric conflict where the reason u is a reason to deny that the reason r justifies the reasoning step from A to B. Thus, the reason u attacks the applicability of the conditional in the circumstances captured by the default reason r, but u does not attack B. Definition 4.3 generalizes the relation to include attacks to the reason subterms for nested default inferences. A similar account of the undercut as an attack on the evidential support is discussed in [16, p. 176], although more informally.

We say that a set of qfFOJT formulas S is **conflict-free** if $Th_L(S)$ does not contain an undercutting reason for any justified formula $t{:}F \in Th_L(S)$. Definition 4.3 suffices to fully formalize the default theories of qfFOJT and their extensions. Based on a reasoning space $\mathcal{R}(\Delta)$, we define the basics of admissibility semantics.

Definition 4.4 [Admissible Extension] A conflict-free set of qfFOJT formulas $\Gamma \subseteq R$ is **admissible** iff for any undercutting reason u for a formula $t{:}_{\{a\}}F \in \Gamma$ such that $u{:}_{\{a\}}\neg[v{:}_{\{a\}}(G \rightarrow H)] \in R$ and $u{:}_{\{a\}}\neg[v{:}_{\{a\}}(G \rightarrow H)] \notin \Gamma$, $Th_L^{-}(\Gamma)$ contains an undercutter for u.

The above iterative process of undercuts and admissibility suffice to define all the familiar notions of extensions that are given for argumentation frameworks.

We are ready to define the notion of a *preferred extension* for a theory Δ.

Definition 4.5 [Preferred Extension] For a reasoning space $\mathcal{R}(\Delta)$ of the default theory $\Delta = \langle W, D\rangle$, the closure $\mathrm{Th}_L^-(\Gamma)$ of an admissible set of qfFOJT formulas $\Gamma \subseteq \mathcal{R}(\Delta)$ is a **qfFOJT-preferred extension** of Δ iff $Th_L(\Gamma)$ is a maximal element (w.r.t. set inclusion) among admissible sets in $\mathcal{R}(\Delta)$.

Other admissibility-based extensions and stable extensions for qfFOJT default theories are also easily definable, but we omit them here for the sake of conciseness. The existence of preferred extensions is guaranteed, as it can be assumed based on the format of *normal* qfFOJT default rules.

Theorem 4.6 *Every qfFOJT default theory $\Delta = \langle W, D\rangle$ has at least one qfFOJT-preferred extension.*

To prove the existence of preferred extensions, we notice that the admissible extensions for a reason extension R form a complete partial order with respect to $\subseteq$. For the case of an infinite set of rules D, maximality is guaranteed by Zorn's lemma [23].

The notion of qfFOJT-preferred extension is analogous to the notion of preferred extensions in argumentation theory. Multiple qfFOJT-preferred extensions can occur when there is a symmetric attack based on inconsistent consequents with default reasons added to different reason extensions of a theory, and their conflict is not resolved by an undercut to either of the two reasons. That is, multiple qfFOJT-preferred extensions are typically a result of two reasons 'rebutting' each other. Although we have not explicitly discussed rebuttal, this type of attack is implicit because of a potential multiplicity of reason extensions for qfFOJT default theory.

Example 4.7 [Continuation of Example 3.1] Let the qfFOJT default theory $\Delta_1 = \langle W_1, D_1\rangle$ be defined by the set of facts $W_1 = \{r{:}_{\{tw\}}B, t{:}_{\{tw\}}P\}$ and the set of defaults

$$D_1 = \Big\{\frac{r{:}_{\{tw\}}B :: (s\cdot r){:}_{\{tw\}}F}{(s\cdot r){:}_{\{tw\}}F}, \frac{t{:}_{\{tw\}}P :: (u\cdot t){:}_{\{tw\}}\neg F}{(u\cdot t){:}_{\{tw\}}\neg F}, \frac{t{:}_{\{tw\}}P :: (v\cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B\to F)}{(v\cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B\to F)}\Big\}.$$

We can establish that the only qfFOJT-preferred extension is $Th_L(\Gamma_1)$, where $\Gamma_1 = \{r{:}_{\{tw\}}B, t{:}_{\{tw\}}P, (u\cdot t){:}_{\{tw\}}\neg F, (v\cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B\to F)\}$. This output is intuitively acceptable. The system does not allow contamination by 'unintended extensions' and it relies exclusively on *logical* means to handle default reasoning type of non-monotonicity. Notice that the reason extension $\{r{:}_{\{tw\}}B, t{:}_{\{tw\}}P, (s\cdot r){:}_{\{tw\}}F, (v\cdot t){:}_{\{tw\}}\neg s{:}_{\{tw\}}(B\to F)\}$ does not generate an alternative qfFOJT-preferred extension. Any admissible extensions defined based on this reason extension is a subset of Γ_1.

5 Combining learning and default reasoning in first-order JL

The project of default reasoning with first-order justification logic targets barriers to effective neurosymbolic integration that lie in the currently used logical frameworks. Propositional and first-order logics (including modal) that dominate existing "neuro-symbolic" approaches [10] face key limitations:

(i) propositional representation is poorly suited to capturing relations in data, which undermines the capacity for extrapolation;

(ii) first-order and modal logic extend expressivity but still fail to represent exceptions to rules in a robust way, limiting their ability to model error recovery in systems that revise conclusions given new evidence;

(iii) they all lack the expressive resources to trace the provenance of inferences, which is a precondition for interpretability.

To overcome these limitations, the representational core of an architecture combining learning and reasoning needs to be flexible enough to provide a robust qualitative representation of *exceptions*. Existing neurosymbolic approaches have merged ANNs with logic programming [20], formal argumentation [6], and input/output logic [3]. These systems made significant progress in merging learning and reasoning, but they still treat exceptions indirectly and provide limited traceability.

To illustrate the methodological pipeline of such an architecture, consider a toy dataset of traffic signs labeled "Stop" and "Wrong Way." A neural component first processes the raw image and extracts simple numerical features, such as the proportion of red pixels (color ratio) and the number of detected edges corresponding to the sign's shape. For a given image s, the neural network produces the following outputs, that is, images are preprocessed into features (e.g., convolutional layers extract octagon, red).

As an output, the following is an example probability distribution over candidate features obtained from the Shape and Color Convolutional Neural Networks (CNNs).

Shape CNN $P(shape = octagon) = 0.95$;

Color CNN $P(color = red) = 0.85$.

The next step is to translate these numerical outputs into symbolic predicates expressed in first-order justification logic. The features extracted from CNNs are translated into interpretable properties:

s is 'Red' if, for image 's', $color = red$;

s is 'Octagon' if, for image 's', $shape = octagon$.

In justification logic, each of these predicates is accompanied by an explicit reason term that records both its origin and weight. For the example above, we obtain:

- $t_{color}{:}_{\{s\}} Red$ with weight 0.85 assigned to t_{color};

- $t_{shape}{:}_{\{s\}}\,Octagon$ with weight 0.95 assigned to t_{shape}.

Based on reasons for *Red* and *Octagon*, the following two default rules are triggered:

$$\frac{t_{color}{:}_{\{s\}}\,Red :: (u_1 \cdot t_{color}){:}_{\{s\}}\,Stop}{(u_1 \cdot t_{color}){:}_{\{s\}}\,Stop} \quad \text{and}$$

$$\frac{t_{shape}{:}_{\{s\}}\,Octagon :: (u_2 \cdot t_{shape}){:}_{\{s\}}\,Stop}{(u_2 \cdot t_{shape}){:}_{\{s\}}\,Stop},$$

to derive defeasible conclusions $(u_1 \cdot t_{color}){:}_{\{s\}}\,Stop$ and $(u_2 \cdot t_{shape}){:}_{\{s\}}\,Stop$.

The weights provide a neural-style confidence measure, while the justification terms preserve a symbolic trace of how the decision was reached. If a feature is perturbed (e.g., a sticker covers part of the stop sign, reducing the red ratio) or if the context makes the inference uncertain (e.g., an octagon-shaped orange "HAZMAT" sign warning for trucks before a tunnel or a "Yield" sign that is dominantly red in Romania, both making questionable the two patterns of reasoning that led the system to infer that 's' is a "Stop" sign), justification logic can represent these as an exception to a rule and record the alternative reasoning path. For example, if it was the case that the starting information on the shape of the sign 's', that is, $t_{shape}{:}_{\{s\}}\,Octagon$, was combined with the information that s is $t_{color'}{:}_{\{s\}}\,Orange$, the formula $((d \cdot t_{shape}) \cdot t_{color'}){:}_{\{s\}})(Octagon \wedge Orange)$, where d is some reason constant, would trigger the rule

$$\frac{((d\cdot t_{shape})\cdot t_{color'})){:}_{\{s\}}(Octagon \wedge Orange) :: (u_3\cdot((d\cdot t_{shape})\cdot t_{color'}))){:}_{\{s\}}\,Hazmat}{(u_3\cdot((d\cdot t_{shape})\cdot t_{color'}))){:}_{\{s\}}\,Hazmat},$$

which then provides an undercutting reason to the reason u_2 via the rule:

$$\frac{(u_3\cdot((d\cdot t_{shape})\cdot t_{color'}))){:}_{\{s\}}\,Hazmat :: (u_4\cdot(u_3\cdot((d\cdot t_{shape})\cdot t_{color'})))){:}_{\{s\}}\neg[u_2{:}_{\{s\}}(Octagon \rightarrow Stop)]}{(u_4\cdot(u_3\cdot((d\cdot t_{shape})\cdot t_{color'})))){:}_{\{s\}}\neg[u_2{:}_{\{s\}}(Octagon \rightarrow Stop)]}.$$

This undercut would in effect reduce the weight of the initial conclusion $(u_2 \cdot t_{shape}){:}_{\{s\}}\,Stop$ based on the shape of the sign 's'. Whether the undercutting attack would defeat the reason $(u_2 \cdot t_{shape})$ depends on any available counterattacks and the threshold below which the original conclusion is not any more supported by $(u_2 \cdot t_{shape})$.

The logical structure for this seemingly simple inference can become even richer. Firstly, the symbolic subsystem could combines justifications into a structured inference:

$$((u_1 \cdot t_{color}) + (u_2 \cdot t_{shape})){:}_{\{s\}}\,Stop.$$

By concatenating reasons, system would infer with greater confidence that the sign is a "Stop" sign based on merging two explicitly tracked reasons $(u_1 \cdot t_{color})$ and $(u_2 \cdot t_{shape})$.

Secondly, there could be synergistic effects of combining reasons beyond the simple concatenation of reasons above. This is the case in the current domain

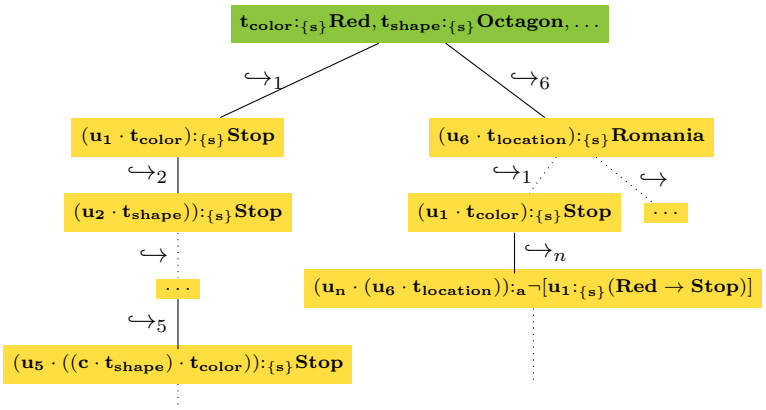


Fig. 1. Illustration of a network-style integration of learning and reasoning using quantifier-free first-order JL logic.

of application, where the combination of features gives rise to a stronger default conclusion that the sign 's' is a "Stop" sign. That is, a rule might be based on the justified formula $((c \cdot t_{shape}) \cdot t_{color}):_{\{s\}}(Octagon \wedge Red)$ with some reason constant c to derive a stronger conclusion $(u_5 \cdot ((c \cdot t_{shape}) \cdot t_{color})):_{\{s\}}Stop$. By adding such a rule to a default theory based on Figure 1, we would again obtain a unique qfFOJT-preferred extension accepting the justification that s is *Stop*. The strength of the conclusion could, in turn, be attenuated by the undercutter based on the location information $(u_6 \cdot t_{location}):_{\{s\}}Romania$, which may reduce confidence in default reasoning from the predicate *Red*.

This example demonstrates how systematically translating raw numerical inputs into symbolic terms enriched with reasons and weights facilitates both inductive learning capacity and interpretability. In future work, this pipeline will be scaled from simple toy domains to structured relational datasets and real-world benchmarks.

6 Concluding remarks and future research

We presented the idea of a logical system that successfully deals with the issue of 'unintended' extensions in default reasoning. The system has yet to be fully explored in its full potential. The line of research started by Reiter's original work [18] on default logic is too comprehensive to be systematically covered here. What is worth mentioning here are related solutions that focus on default logic in specific. Some of the most popular solutions to the problem of interaction between the defaults that we started with need to rely on non-logical means to address the issue. For example, one such method is prioritization of default rules [5,9] that relies on encoding extra-logical orderings on rules to represent their interaction. Another well-known approach is exploiting subclass relations to substantiate prioritization choices [19,4]. This content-sourced solution works for most of the examples discussed in the early work on

default logic, but, as noted by [9, p. 19], priority relations among defaults, and their corresponding reasons, can have different sources, where specificity is one of the possible sources.

Horty's work, comprehensively presented in [9], comes closest to ours in terms of the ability to represent undercutting defeat. Horty's approach is nested in reasoning about prioritization of defaults and the possibility to refer to defaults in the object-level language. Thus, expressions such as $d_1 \succ d_2$ that says the rule d_1 is preferred over the rule d_2 or $Out(d_1)$ that says the rule d_1 is undercut are allowed, for some constants d_1 and d_2 naming the corresponding default rules. This approach has been venerable and flexible enough to accommodate a variety of exclusionary reasons. Without attempting to find downsides of prioritization in general, and Horty's approach in specific, we emphasize that our project is to avoid introducing meta-level content into the object level language. We seek to define the logical structure of undercutting as attacks on *reasons* as object-level entities, rather than reducing undercutting to statements about excluded rules, such as $Out(d)$. What justification logic adds is a fully syntactic treatment of undercut and interactions among defaults, thereby focusing on providing a qualitative account of uncertainty induced by default reasons.

In future work, we will explore the connections of the framework to the argumentation frameworks. This connection becomes apparent through the use of admissibility in Definition 4.4. Across existing formal argumentation frameworks, there are systems that explicitly model undercut (e.g., ASPIC+ [11]) and systems whose functionality allows for modeling exceptions in the way of implicit undercut (e.g., DeLP [8] and ABA [22]).

Moving forward, the system proposed here will be developed as a candidate logical syntax for the 'neuro-symbolic integration' project. The choice of a logical syntax remains one of the open questions of neuro-symbolic integration [10]. The state-of-the-art hybrid architectures are divided between the "propositional" nature of neural networks and the need to introduce more sophisticated languages with the internal structure that refers to domain objects, such as first-order or relational logic. Both propositional logic and first-order logic are not suitable for learning relations in a database or handling exceptions to rules in a robust way. These limitations make it difficult to include features such as extrapolation and error recovery as a result of learning in the fashion of neural networks. Moreover, propositional and first-order languages are not expressive enough to *trace* where the inferences in these systems come from. The ability to trace the origin of inferences matters in each domain with a requirement of transparent decision-making.

We conclude the paper with an overview of the current state of the larger project behind this logic. The guiding idea behind the project is that a combination of numerical reasoning, default reasoning, and first-order JL offers a unique formal toolbox for qualitative representation of *uncertainty* that results from inductive learning and generalization. Default reasoning and justification logic have already been explored as logical systems for structured arguments

[12,13]. In [14], default reasoning and numerical reasoning were combined with propositional justification logic to foster explainability of neural network-type of inference. Introducing default *schemas* in this paper brings us closer to the original motivation behind the standard default logics [18] and to the potential of modeling inductive, statistical reasoning [21]. We have left open the question of how reason terms precisely change their weights through the type of reasoning illustrated in Figure 1. This area presents ample opportunities for future research.

References

[1] Artemov, S. N., *The logic of justification*, The Review of Symbolic Logic **1** (2008), pp. 477–513.

[2] Artemov, S. N. and T. L. Yavorskaya, *On first order logic of proofs*, Moscow Mathematical Journal **1** (2001), pp. 475–490.

[3] Besold, T. R., A. d'Avila Garcez, K. Stenning, L. van der Torre and M. van Lambalgen, *Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples*, Minds and Machines **27** (2017), pp. 37–77.

[4] Brewka, G., *Adding priorities and specificity to default logic*, in: C. MacNish, D. Pearce and L. M. Pereira, editors, *Logics in Artificial Intelligence (JELIA1994)* (1994), pp. 247–260.

[5] Brewka, G., *Reasoning about priorities in default logic*, in: *Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI'94*, AAAI National Conference Proceedings Series **2**, AAAI Press/The MIT Press, Seattle, WA, USA, 1994, pp. 940–945.

[6] d'Avila Garcez, A., D. M. Gabbay and L. C. Lamb, *Argumentation neural networks*, in: N. R. Pal, N. Kasabov, R. K. Mudi, S. Pal and S. K. Parui, editors, *Neural Information Processing (ICONIP 2004)*, Lecture Notes in Computer Science **3316** (2004), pp. 606–612.

[7] Fitting, M., *Possible world semantics for first-order logic of proofs*, Annals of Pure and Applied Logic **165** (2014), pp. 225–240.

[8] García, A. J. and G. R. Simari, *Defeasible logic programming: An argumentative approach*, Theory and Practice of Logic Programming **4** (2004), pp. 95–138.

[9] Horty, J. F., "Reasons as Defaults," Oxford University Press, 2012.

[10] Marra, G., S. Dumančić, R. Manhaeve and L. De Raedt, *From statistical relational to neurosymbolic artificial intelligence: A survey*, Artificial Intelligence (2024), p. 104062.

[11] Modgil, S. and H. Prakken, *The ASPIC+ framework for structured argumentation: A tutorial*, Argument & Computation **5** (2014), pp. 31–62.

[12] Pandžić, S., *A logic of defeasible argumentation: Constructing arguments in justification logic*, Argument & Computation **13** (2022), pp. 3–47.

[13] Pandžić, S., *Structured argumentation dynamics: Undermining attacks in default justification logic*, Annals of Mathematics and Artificial Intelligence **90** (2022), p. 297–337.

[14] Pandžić, S. and J. Graff, *A logic of weighted reasons for explainable inference in AI*, in: L. Longo, S. Lapuschkin and C. Seifert, editors, *World Conference on Explainable Artificial Intelligence, xAI 2024*, Springer, 2024, pp. 243–267.

[15] Pollock, J. L., *Defeasible reasoning*, Cognitive Science **11** (1987), pp. 481–518.

[16] Pollock, J. L., *A recursive semantics for defeasible reasoning*, in: I. Rahwan and G. R. Simari, editors, *Argumentation in Artificial Intelligence*, Springer, 2009 pp. 173–197.

[17] Prakken, H. and G. Vreeswijk, "Logics for defeasible argumentation," Springer Netherlands, Dordrecht, 2002 pp. 219–318.

[18] Reiter, R., *A logic for default reasoning*, Artificial intelligence **13** (1980), pp. 81–132.

[19] Rintanen, J., *On specificity in default logic*, in: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, IJCAI (1995), pp. 1474–1479.
[20] Saldanha, E. D., S. Hölldobler, C. D. P. K. Ramli and L. P. Medinacelli, *A core method for the weak completion semantics with skeptical abduction*, Journal of Artificial Intelligence Research **63** (2018), pp. 51–86.
[21] Tan, Y.-H., *Is default logic a reinvention of inductive-statistical reasoning?*, Synthese **110** (1997), pp. 357–379.
[22] Toni, F., *A tutorial on assumption-based argumentation*, Argument & Computation **5** (2014), pp. 89–117.
[23] Zorn, M., *A remark on method in transfinite algebra*, Bulletin of the American Mathematical Society **41** (1935), pp. 667–670.