Towards an Argumentative Foundation for Evaluative AI

Xiang Yin[1,*], Tim Miller[2], Nico Potyka[3], Antonio Rago[4] and Francesca Toni[1]

[1]Imperial College London, Exhibition Road, London, SW7 2AZ, UK
[2]The University of Queensland, St Lucia, Brisbane, Queensland 4072, Australia
[3]Cardiff University, Main Building, Park Place, Cardiff, CF10 3AT, UK
[4]King's College London, Strand, London, WC2R 2LS, UK

Abstract

Evaluative AI (EAI) has been recently proposed as a way to support human decision-making, not by producing a single recommendation, but by presenting competing hypotheses together with evidence for and against each. In this ongoing work, we investigate how computational argumentation can provide a formal and computable foundation for explainable and contestable forms of EAI. We present a ranking-based formulation of EAI and show how weighted Quantitative Bipolar Argumentation Frameworks can be used to realise and analyse such systems.

Keywords

Evaluative AI, Argumentation, Explainability, Contestability

1. Introduction

Explainable AI (XAI) aims to enhance the transparency and trustworthiness of AI systems by elucidating their decision-making processes [1], which is crucial in high-stakes domains such as healthcare, law, finance, and safety-critical applications such as aviation [2, 3]. Recently, *Evaluative AI (EAI)* [4] has been proposed as a form of XAI that supports human decision-making in a *hypothesis-*, rather than *recommendation-*, driven manner. In the dominant recommend-and-explain paradigm, an AI model first produces an output and XAI methods then justify it [5, 6, 7]. However, this can induce cognitive fixation, leading users to over-rely on or dismiss AI recommendations without sufficient deliberation [8, 9, 10]. EAI instead presents multiple plausible hypotheses with structured evidence for and against each, helping users actively evaluate competing options while preserving human agency. It therefore strengthens human-in-the-loop AI by repositioning AI systems as facilitators rather than decision-makers.

In this ongoing work, we take initial steps towards distributed, human-centred EAI with formal and algorithmic foundations. We formalise EAI as a *ranking-based* problem over hypotheses, given their pro and con evidence, as suggested by [4]. For example, in healthcare, an EAI system may rank possible diagnoses given a patient's symptoms, not as a final decision but as the system's evidence-based "preference" over hypotheses. We propose computational argumentation, particularly weighted Quantitative Bipolar Argumentation Frameworks (wQBAFs) [11, 12, 13], as a formal and computational foundation for ranking-based EAI. Compared with the Weight of Evidence method originally proposed by [14], wQBAFs better support structured reasoning over hypotheses and evidence, argumentative explainability [15, 16], human contestability of evidence, hypotheses, and rankings [17], and distributed EAI in which diverse agents deliberate using different sources and reasoning styles.

Motivating Illustration. Figure 1 shows the skeleton of an argumentative solution to a ranking-based EAI problem in a healthcare scenario[1]. It consists of arguments organised hierarchically across

The 26th International Workshop on Computational Models of Natural Argument (CMNA'26)
*Corresponding author.
xy620@ic.ac.uk (X. Yin); timothy.miller@uq.edu.au (T. Miller); potykan@cardiff.ac.uk (N. Potyka); antonio.rago@kcl.ac.uk (A. Rago); ft@ic.ac.uk (F. Toni)

[1]This is a simplified, non-clinically validated example. The three-layer organisation in Figure 1 is illustrative rather than required; in other applications, the argumentative structure and the conceptual interpretation of intermediate arguments may be domain-dependent.

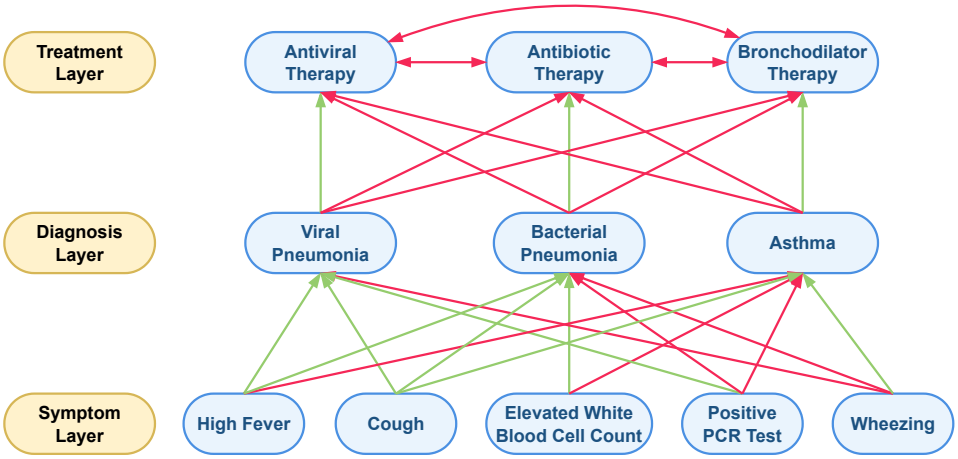

Figure 1: Skeleton of an argumentative solution to a ranking-based EAI problem in healthcare. Blue nodes denote arguments and green/red edges, resp., support/attack relations. If, in the corresponding wQBAF, all arguments have an initial weight of 0.5 (the neutral value in [0,1]) and the edges all have the weight of 1 (the top value in [0,1]), then QE [18] yields the ranking `Anti-viral` $\succ$ `Antibiotics` $\succ$ `Bronchodilator` with corresponding strengths 0.3564, 0.2368, 0.1479.

three layers via support and attack relations. The symptom layer represents observable evidence, the treatment layer represents possible clinical decisions, and the intermediate diagnosis layer plays a dual role: as hypotheses with respect to symptoms and as evidence for treatments.[2] A wQBAF augments this skeleton with initial weights for arguments, reflecting importance or confidence, and relation weights for edges, capturing their influence. Quantitative evaluation methods then compute the final strength of each argument recursively from its initial weight and the strengths of its supporters and attackers, enabling conflict resolution [12, 13, 15, 19, 20]. The resulting strengths induce a ranking over treatment hypotheses. The wQBAF and its strengths are interpretable and can support explanations of rankings, hypotheses, or evidence in established argumentative XAI formats [15, 16, 21, 22, 23, 24, 25, 26, 27]. Moreover, users can contest components of the QBAF or the resulting ranking [28, 29], for example by challenging initial weights or adding pro/con evidence or hypotheses. This supports human-in-the-loop decision-making, a central goal of EAI. Finally, the wQBAF may represent one agent's opinion and can be combined with others through argumentative exchanges or QBAF aggregation [30, 31].

Contributions. This paper makes two initial contributions. First, EAI can be formally understood as a ranking-based problem (Section 3). Second, wQBAFs from the field of (computational) argumentation provide a principled paradigm for realising ranking-based EAI in an explainable and contestable manner suitable for human-centred EAI systems (Section 4).

2. Related Work

Evaluative AI (EAI). EAI is a recent paradigm [4], and its practical realisation remains at an early stage. Existing approaches are still limited. For example, [14] use Weight of Evidence (WoE) [32] to quantify the direction and magnitude of the impact of evidence on candidate hypotheses, while [33] use LLMs to generate pro and con evidence conversationally. In contrast, our ranking-based argumentative approach supports richer dependencies between hypotheses and evidence, including cases where hypotheses influence one another. It also separates evidential stance from strength, determines hypothesis ratings and rankings using established wQBAF evaluation methods and their formal and practical guarantees [34, 28], derives faithful explanations from these evaluations, and helps mitigate hallucination risks when LLMs are used to generate evidence and hypotheses, by exploiting the contestable nature of argumentative solutions [35].

[2] wQBAFs can in general accommodate any structure.

Argumentative XAI. Argumentation is an influential paradigm for XAI, supporting structured, transparent, and human-aligned reasoning [15, 16]. Our approach falls under *intrinsic* argumentative explanations [15], where the underlying model is itself argumentative, namely a wQBAF. This allows us to explain rankings between hypotheses using established attribution-based formats, e.g. by identifying the evidence most influential for or against ranked hypotheses [21, 22, 24, 25, 23, 26, 27]. For example, the top-ranked hypothesis `Anti-viral` in Figure 1 has strength 0.40, with `High Fever`, `Cough`, and `Positive PCR test` identified as the most influential symptoms following [22]. Faithfulness to the underlying model is an important desideratum for XAI [36]. Since explanations from intrinsically argumentative solutions are faithful by design [35, 37], they provide a strong basis for trustworthy EAI.

3. The Ranking-based EAI Problem

Let $\mathscr{H}$ be a finite, non-empty set of *hypotheses*, and $\mathscr{E}$ be a finite set of (pieces of) *evidence*. Hypotheses and evidence can be associated with an *initial weight* that capture how plausible they are before considering any dependencies amongst them.

Definition 1. *The function* $\tau : (\mathscr{E} \cup \mathscr{H}) \to [0, 1]$ *maps each* $x \in (\mathscr{E} \cup \mathscr{H})$ *to its* initial weight $\tau(x)$.

In the motivating illustration, the initial weight of a symptom (e.g., `cough`) or of a medical condition (e.g., `asthma`) may capture its perceived severity for a given patient. For instance, a `higher fever` corresponds to a higher initial weight, reflecting the greater severity of the symptom.

(Positive and negative) dependencies within hypotheses and evidence can be modelled in terms of (pro and con, resp.) relations.

Definition 2. *The* pro *and* con relations *are disjoint binary relations* $\mathsf{pro}, \mathsf{con} \subseteq (\mathscr{E} \times (\mathscr{E} \cup \mathscr{H})) \cup (\mathscr{H} \times \mathscr{H})$ *(resp.). For all* $x, y \in (\mathscr{E} \cup \mathscr{H})$*,* x *is called* pro/con evidence or hypothesis *for* y *iff* $(x, y) \in \mathsf{pro}$/$(x, y) \in \mathsf{con}$*, resp.*

Note that here we consider, besides direct relations from evidence to hypotheses ($\mathscr{E} \times \mathscr{H}$) as in [4], relations among evidence ($\mathscr{E} \times \mathscr{E}$) and among hypotheses ($\mathscr{H} \times \mathscr{H}$). In this way, evidence may also attack or support other evidence. This allows multi-step reasoning and evaluation.

To quantify how strongly a piece of evidence affects the plausibility of another element (hypothesis or evidence), we use the notion of *relation weight*.

Definition 3. *The function* $w : (\mathsf{pro} \cup \mathsf{con}) \to [0, 1]$ *maps each* $r \in (\mathsf{pro} \cup \mathsf{con})$ *to its* relation weight $w(r)$.

In the motivating illustration, the relation weight for an evidence-hypothesis pair (e.g., `(cough,asthma)`) represents the strength of the influence from the symptom (e.g., `cough`) to the diagnosis (e.g., `asthma`). A higher weight indicates a stronger influence.

We next formally define EAI as a *ranking-based problem*.

Definition 4. *For* $E = \langle \mathscr{E}, \mathscr{H}, \mathsf{pro}, \mathsf{con}, \tau, w \rangle$*, the* ranking-based EAI problem *is to determine a total preorder* $\succcurlyeq_E$ *on* $\mathscr{H}$*.* [3]

Intuitively, $\succcurlyeq_E$ represents the relative *plausibility* of hypotheses given the weighted pro/con relations and the initial weights of evidence/hypotheses. For $h_1, h_2 \in \mathscr{H}$, $h_1 \succcurlyeq_E h_2$ means that h_2 is not more plausible than h_1 given E.

Formalising EAI as a ranking-based problem offers two key advantages. First, ranking naturally accommodates multiple hypotheses rather than forcing a single recommendation, which aligns directly with the core philosophy of EAI. Second, ranking offers a concrete computational framework, both for generating rankings (e.g., [38, 39]) and for evaluating them (e.g., Kendall's τ [40]), making EAI amenable to formal analysis and algorithmic implementation. We discuss next how to use argumentation to identify solutions to the ranking-based EAI problem.

[3] A total preorder is a reflexive, transitive and total relation.

4. An Argumentative Solution

In this section, we show how argumentation can lead to the identification of solutions to the ranking-based EAI problem with desirable properties. Specifically, we use *weighted Quantitative Bipolar Argumentation Frameworks (wQBAFs)* [11, 12, 13], which are quintuples $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^-, \mathcal{R}^+, \tau, w \rangle$ where $\mathcal{A}$ is a finite set of *arguments*, $\mathcal{R}^-, \mathcal{R}^+ \subseteq \mathcal{A} \times \mathcal{A}$ are disjoint binary relations called *attack* and *support*, $\tau : \mathcal{A} \to [0, 1]$ is a *base score function*, and $w : \mathcal{R}^- \cup \mathcal{R}^+ \to [0, 1]$ is an *edge weight function*. These wQBAFs can be used to represent EAI problems.

Definition 5. *The* wQBAF representation of $E = \langle \mathcal{E}, \mathcal{H}, \text{pro}, \text{con}, \tau, w \rangle$ *is the wQBAF* $\mathcal{Q} = \langle \mathcal{A}, \mathcal{R}^+, \mathcal{R}^-, \tau, w \rangle$ *such that* $\mathcal{A} = \mathcal{E} \cup \mathcal{H}$*;* $\mathcal{R}^- = \text{con}$*; and* $\mathcal{R}^+ = \text{pro}$.

We define $\mathcal{A}$ so that debate can take place about all evidence and hypotheses. As for the other components, we use those of the EAI problem directly.

Once we map the EAI problem into a wQBAF, identifying solutions to the ranking-based EAI problem requires two steps. First, we compute the strengths of arguments (both evidence and hypotheses) using quantitative evaluation methods, which update the initial weights of arguments by taking support and attack relations into account (see, e.g., [12, 13]).[4] Second, we derive a ranking over hypotheses by comparing their strengths, as illustrated in the Introduction.

The wQBAF solution to the ranking-based EAI problem is naturally explainable: qualitatively, the graphical structure shows the relationship between different arguments and users can see the reasoning path from a piece of evidence to a hypothesis; quantitatively, the final strength of each hypothesis is explainable via (adaptations of) many possible explanation methods for QBAFs, such as attribution method [22, 24] and counterfactual explanations [42, 28], affording also contestability by humans who can modify the weights of arguments or relations to change the strength of a hypothesis to a desired one. For example, for Figure 1, an attribution explanation may identify PCR test as the most influential evidence on the strength of Anti-viral Therapy. A clinician who questions the reliability of the PCR test could then contest its initial weight; recomputing the argument strengths would reveal whether and how the treatment ranking changes. The proposed argumentative solution also motivates a number of desirable principles for EAI, introduced next. We present these principles informally as desiderata rather than as a formal axiomatization; they serve both as design guidelines for selecting or developing ranking methods and as criteria for evaluating them.

Principle 1 (Monotonicity). Monotonicity is a fundamental guarantee to ensure that rankings remain aligned with rational intuition and expectation. Monotonicity requires that the relative ranking of an individual hypothesis should change in a way consistent with the direction of change in the underlying factors, such as the introduction of new evidence or updates to the weights of existing evidence. For instance, adding additional pro evidence for a hypothesis should never cause its rank to fall; and increasing the initial weight of its piece of pro evidence should not lower its position relative to others. Many widely used evaluation methods for wQBAFs (e.g., O-QuAD [12], MLP-based semantics [13]) are known to satisfy Principle 1 [28].

Principle 2 (Balance). Balance ensures that a ranking remains unchanged when the pro and con evidence are balanced. In particular, any modification of the evidence that preserves the equality between the overall strength of pro and con evidence should not affect the relative ranking of hypotheses. This includes, for instance, redistributing weights among existing pro and con evidence or introducing additional evidence, provided that the total pro and total con influence remain equal and therefore cancel each other out. The aforementioned semantics also satisfy Principle 2.

Principle 3 (Equivalence). Equivalence states that two hypotheses must be ranked equally if they share the same set of pro and con evidence with the same initial weight. This principle is fundamental for ranking fairness, ensuring symmetric treatment for evidentially symmetric cases. Beyond fairness, it is also a prerequisite for rational and intuitive ranking method behaviour. The aforementioned semantics also satisfy Principle 3.

[4] Existing ranking-based semantics [41] are defined for abstract argumentation frameworks and cannot be directly adapted to our quantitative setting.

Principle 4 (Dominance). The Dominance principle establishes a preference in the ranking under controlled conditions. For example, a hypothesis must be ranked no lower than another if, given an equivalent set of con evidence against both, either (1) its set of pro evidence is a superset of the other's, meaning that more pro evidence ranking higher; or (2) the aggregated weight of its pro evidence is greater. This principle is fundamental because it ensures the ranking objectively rewards hypotheses with superior pro evidence (either in number or in strengths), providing a clear and fair rationale for comparative assessment. The aforementioned semantics also satisfy Principle 4.

Principle 5 (Robustness). Robustness requires that the overall ranking exhibits stability when minor perturbations to the underlying evidence or its associated weights occur. Specifically, small variations in the evidence should not induce large-scale or counter-intuitive reversals in the resulting ranking order, unless such variations correspond to application-specific changes for which a substantial change in the ranking is warranted. This principle is critical for ensuring the reliability and practical dependability of the EAI rankings, as users must be able to trust that its evaluations are not brittle or overly sensitive to noise.

It is unclear whether existing evaluation methods for wQBAFs ensure robustness at the ranking level; if not, new robustness-oriented ranking semantics may be needed.

Principle 6 (Explainability). Explainability requires that the ranking is generated through a mechanism whose logic is transparent and whose outcomes can be clearly explained. This entails providing both local explanations (e.g., why one hypothesis is ranked above another) and a global rationale for the overall ordering. This principle is fundamental to transforming the ranking from an opaque output to an understandable result, fostering user comprehension and trust. As we already argued, a large array of argumentative explanations drawn from wQBAFs can support this principle.

Principle 7 (Contestability). Contestability ensures that the EAI ranking can be challenged by users [17]. This includes the ability to question not only the veracity of individual evidence, or their initial weights, but also the final ranking itself. By explicitly supporting such challenges, this principle reaffirms the role of EAI as a deliberative partner in a human-AI team, empowering users to engage with and refine the system's reasoning process critically. Again, wQBAFs show promise to support this principle.

5. Discussion

This work takes an initial step towards making EAI formally tractable by establishing a ranking- and argumentation-based foundation. Furthermore, it inherently paves the way towards explainability, as the argumentative structure makes the entire evaluation process explicit and auditable. Finally, this explicit reasoning process empowers contestability, transforming the AI from a static recommender into a dynamic deliberative aider that users can meaningfully challenge and refine, thereby fostering a new paradigm of collaborative and trustworthy human-in-the-loop decision-making.

Future work. Our framework also opens a natural path towards multi-agent EAI. In realistic evaluative settings, different agents may construct complementary wQBAFs from different evidence sources, expertise, or reasoning semantics. These wQBAFs could then be exchanged, aligned, or aggregated to support group-level evaluation. When such exchanges take the form of debates, computational argumentation may further provide principled mechanisms for post-hoc debate judgement [43]. Such a perspective may help reduce individual bias and accommodate pluralistic viewpoints, while preserving the explainability and contestability of the evaluative process. Finally, while we have focused on quantitative argumentation through wQBAFs, future work could investigate whether qualitative argumentation formalisms, such as assumption-based argumentation or priority-based non-monotonic reasoning, can also provide suitable foundations for EAI. Such approaches may offer complementary ways of representing qualitative preferences and supporting explainability.

Limitations. Several limitations remain for future development. In practical applications, instantiating our framework requires mapping domain information into hypotheses and evidence, identifying their pro/con dependencies, and assigning appropriate initial and relation weights. This mapping is application-dependent and may require a combination of domain expertise and automated extraction. The interpretation and provenance of these weights are likewise application-dependent: initial weights may encode quantities such as plausibility, confidence, or severity, while relation weights represent the strength of influence between arguments. They may be elicited from domain experts or estimated from data, rather than having a universal probabilistic interpretation. Our framework currently assumes that the resulting wQBAFs are correctly constructed, whereas in practice the reliable extraction of argumentative structures remains a core bottleneck; advances in LLM-based argument mining may help alleviate this challenge [44, 45]. Even when a wQBAF is successfully constructed, the openness required for contestability may introduce a second limitation: users may strategically or unintentionally provide misleading or low-quality evidence, creating risks of manipulation, analogous to shilling attacks in recommender systems [46]; mechanisms such as permissioned contribution and trust-aware weighting may offer partial safeguards. Finally, in the multi-agent setting discussed above, agents may naturally exhibit deep and sometimes irreducible disagreement [47]. Rather than enforcing consensus, this limitation suggests the need for managing such pluralism, supporting structured disagreement [48], and helping users navigate conflicting evaluations, something that argumentation is well-positioned to address.

Acknowledgments

Yin and Toni were partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101020934, ADIX). Toni was also funded by EPSRC (grant UKRI3928, NeSyDebates). We thank the anonymous reviewers for their careful reading and detailed, constructive comments, which helped us improve the clarity and presentation of this work. Any views or opinions expressed herein are solely those of the authors.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT 5.5 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[2] C. Tong, X. Yin, J. Li, T. Zhu, R. Lv, L. Sun, J. J. Rodrigues, An innovative deep architecture for aircraft hard landing prediction based on time-series sensor data, Applied Soft Computing 73 (2018) 344–349.

[3] C. Tong, X. Yin, S. Wang, Z. Zheng, A novel deep learning method for aircraft landing speed prediction based on cloud-based sensor data, Future Generation Computer Systems 88 (2018) 552–558.

[4] T. Miller, Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai, in: Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, 2023, pp. 333–342.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[6] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.
[7] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.
[8] Z. Buçinca, M. B. Malaya, K. Z. Gajos, To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making, Proceedings of the ACM on Human-computer Interaction 5 (2021) 1–21.
[9] K. Z. Gajos, L. Mamykina, Do people engage cognitively with ai? impact of ai assistance on incidental learning, in: Proceedings of the 27th International Conference on Intelligent User Interfaces, 2022, pp. 794–806.
[10] V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn, A. Perer, Ignore, trust, or negotiate: understanding clinician acceptance of ai-based treatment recommendations in health care, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–18.
[11] T. Mossakowski, F. Neuhaus, Modular semantics and characteristics for bipolar weighted argumentation graphs, CoRR abs/1807.06685 (2018). URL: http://arxiv.org/abs/1807.06685. arXiv:1807.06685.
[12] H. Chi, Y. Lu, B. Liao, L. Xu, Y. Liu, An optimized quantitative argumentation debate model for fraud detection in e-commerce transactions, IEEE Intelligent Systems 36 (2021) 52–63.
[13] N. Potyka, Interpreting neural networks as quantitative argumentation frameworks, in: AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 6463–6470.
[14] T. Le, T. Miller, L. Sonenberg, R. Singh, Towards the new xai: A hypothesis-driven approach to decision support using evidence, in: ECAI 2024, IOS Press, 2024, pp. 850–857.
[15] K. Čyras, A. Rago, E. Albini, P. Baroni, F. Toni, Argumentative xai: a survey, arXiv preprint arXiv:2105.11266 (2021).
[16] A. Vassiliades, N. Bassiliades, T. Patkos, Argumentation and explainable artificial intelligence: a survey, The Knowledge Engineering Review 36 (2021) e5.
[17] F. Leofante, H. Ayoobi, A. Dejl, G. Freedman, D. Gorur, J. Jiang, G. Paulino-Passos, A. Rago, A. Rapberger, F. Russo, et al., Contestable ai needs computational argumentation, in: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, volume 21, 2024, pp. 888–896.
[18] N. Potyka, Continuous dynamical systems for weighted bipolar argumentation, in: 16th International Conference on Principles of Knowledge Representation and Reasoning (KR), 2018.
[19] H. Ayoobi, N. Potyka, F. Toni, Sparx: Sparse argumentative explanations for neural networks, in: European Conference on Artificial Intelligence (ECAI), volume 372, 2023, pp. 149–156.
[20] N. Potyka, X. Yin, F. Toni, Explaining random forests using bipolar argumentation and markov networks, in: AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 9453–9460.
[21] T. Kampik, K. Čyras, Explaining change in quantitative bipolar argumentation 1, in: Computational Models of Argument, IOS Press, 2022, pp. 188–199.
[22] X. Yin, N. Potyka, F. Toni, Argument attribution explanations in quantitative bipolar argumentation frameworks, in: European Conference on Artificial Intelligence (ECAI), volume 372, 2023, pp. 2898–2905.
[23] L. Amgoud, J. Ben-Naim, S. Vesic, Measuring the intensity of attacks in argumentation graphs with shapley value, in: International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 63–69.
[24] X. Yin, N. Potyka, F. Toni, Explaining arguments' strength: Unveiling the role of attacks and supports, in: International Joint Conference on Artificial Intelligence (IJCAI), 2024.
[25] T. Kampik, N. Potyka, X. Yin, K. Cyras, F. Toni, Contribution functions for quantitative bipolar argumentation graphs: A principle-based analysis, Int. J. Approx. Reason. 173 (2024) 109255. URL: https://doi.org/10.1016/j.ijar.2024.109255. doi:10.1016/J.IJAR.2024.109255.
[26] C. Al Anaissy, J. Delobelle, S. Vesic, B. Yun, Impact measures for gradual argumentation semantics, in: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent

Systems, 2025, pp. 69–77.
[27] F. Naudot, A. Brännström, V. Torra, T. Kampik, Set contribution functions for quantitative bipolar argumentation and their principles, arXiv preprint arXiv:2509.14963 (2025).
[28] X. Yin, N. Potyka, A. Rago, T. Kampik, F. Toni, Contestability in Edge-Weighted Quantitative Bipolar Argumentation Frameworks, in: Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning, 2026, pp. 676–687. URL: https://doi.org/10.24963/kr.2026/64. doi:10.24963/kr.2026/64.
[29] T. Kampik, X. Yin, N. Potyka, F. Toni, Strength change explanations in quantitative argumentation, arXiv preprint arXiv:2603.00008 (2026).
[30] A. Rago, H. Li, F. Toni, Interactive explanations by conflict resolution via argumentative exchanges, in: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning, volume 19, 2023, pp. 582–592.
[31] C. Cayrol, M. Lagasquie-Schiex, Weighted argumentation systems: A tool for merging argumentation systems, in: IEEE 23rd International Conference on Tools with Artificial Intelligence, ICTAI 2011, Boca Raton, FL, USA, November 7-9, 2011, IEEE Computer Society, 2011, pp. 629–632. URL: https://doi.org/10.1109/ICTAI.2011.99. doi:10.1109/ICTAI.2011.99.
[32] D. A. Melis, H. Kaur, H. Daumé III, H. Wallach, J. W. Vaughan, From human explanation to model interpretability: A framework based on weight of evidence, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, volume 9, 2021, pp. 35–47.
[33] A. Ermellino, L. Malandri, F. Mercorio, N. Nobani, A. Serino, et al., An approach to evaluative ai through large language models, in: European Conference on Artificial Intelligence, volume 3803, 2024, pp. 1–15.
[34] P. Baroni, A. Rago, F. Toni, How many properties do we need for gradual argumentation?, in: 32nd AAAI Conference on Artificial Intelligence, 2018.
[35] G. Freedman, A. Dejl, D. Gorur, X. Yin, A. Rago, F. Toni, Argumentative large language models for explainable and contestable claim verification, in: T. Walsh, J. Shah, Z. Kolter (Eds.), AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, AAAI Press, 2025, pp. 14930–14939. URL: https://doi.org/10.1609/aaai.v39i14.33637. doi:10.1609/AAAI.V39I14.33637.
[36] N. Potyka, X. Yin, F. Toni, Towards a theory of faithfulness: Faithful explanations of differentiable classifiers over continuous data, arXiv preprint arXiv:2205.09620 (2022).
[37] L. Chen, X. Yin, F. Toni, Latent debate: A surrogate framework for interpreting llm thinking, arXiv preprint arXiv:2512.01909 (2025).
[38] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression (2000).
[39] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: Proceedings of the 22nd international conference on Machine learning, 2005, pp. 89–96.
[40] M. G. Kendall, A new measure of rank correlation, Biometrika 30 (1938) 81–93.
[41] L. Amgoud, J. Ben-Naim, Ranking-based semantics for argumentation frameworks, in: International Conference on Scalable Uncertainty Management, Springer, 2013, pp. 134–147.
[42] X. Yin, N. Potyka, F. Toni, CE-QArg: counterfactual explanations for quantitative bipolar argumentation frameworks, in: Proceedings of the 21st International Conference on Principles of Knowledge Representation and Reasoning, 2024, pp. 697–707.
[43] X. Yin, A. Dejl, A. Rago, L. Chen, F. Toni, A theory of post-hoc debate judgement, arXiv preprint arXiv:2608.19002 (2026).
[44] D. Gorur, A. Rago, F. Toni, Can large language models perform relation-based argument mining?, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 8518–8534.
[45] D. Gorur, A. Rago, F. Toni, Retrieval and argumentation enhanced multi-agent llms for judgmental forecasting, arXiv preprint arXiv:2510.24303 (2025).
[46] C. Tong, X. Yin, J. Li, T. Zhu, R. Lv, L. Sun, J. J. Rodrigues, A shilling attack detector based on convolutional neural network for collaborative recommender system in social aware network,

The Computer Journal 61 (2018) 949–958.
[47] Z. Wu, T. Ito, The hidden strength of disagreement: Unraveling the consensus-diversity tradeoff in adaptive multi-agent systems, arXiv preprint arXiv:2502.16565 (2025).
[48] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, Z. Tu, Encouraging divergent thinking in large language models through multi-agent debate, in: Proceedings of the 2024 conference on empirical methods in natural language processing, 2024, pp. 17889–17904.