Argumentative Human-AI Decision-Making: Toward AI Agents That Reason With Us, Not For Us

Blue Sky Ideas Track

Stylianos Loukas Vasileiou
New Mexico State University
Las Cruces, NM, USA
stelios@nmsu.edu

Antonio Rago
King's College London
London, UK
antonio.rago@kcl.ac.uk

Francesca Toni
Imperial College London
London, UK
f.toni@imperial.ac.uk

William Yeoh
Washington University in St. Louis
St. Louis, MO, USA
wyeoh@wustl.edu

ABSTRACT

Computational argumentation offers formal frameworks for transparent, verifiable reasoning but has traditionally been limited by its reliance on domain-specific information and extensive feature engineering. In contrast, LLMs excel at processing unstructured text, yet their opaque nature makes their reasoning difficult to evaluate and trust. We argue that the convergence of these fields will lay the foundation for a new paradigm: *Argumentative Human-AI Decision-Making*. We analyze how the synergy of *argumentation framework mining*, *argumentation framework synthesis*, and *argumentative reasoning* enables agents that do not just justify decisions, but engage in dialectical processes where decisions are contestable and revisable – reasoning *with* humans rather than *for* them. This convergence of computational argumentation and LLMs is essential for human-aware, trustworthy AI in high-stakes domains.

KEYWORDS

Argumentation; Human-AI Collaboration

ACM Reference Format:
Stylianos Loukas Vasileiou, Antonio Rago, Francesca Toni, and William Yeoh. 2026. Argumentative Human-AI Decision-Making: Toward AI Agents That Reason With Us, Not For Us: Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 6 pages. https://doi.org/10.65109/EKAW8904

1 INTRODUCTION

For over three decades, computational argumentation (CA) has offered formal frameworks for valid and sound reasoning, enabling transparent and verifiable decision-making through approaches such as abstract [5, 18] and structured [7] argumentation. However, its practical application has often been constrained by the need for structured inputs, such as hand-crafted knowledge bases and domain-specific rules. On the other hand, large language models (LLMs) have demonstrated an unprecedented ability to process and generate natural language [12]. Yet, their internal reasoning remains approximate and largely opaque, making their outputs difficult to verify and trust [8, 22, 37, 55].

The integration of CA and LLMs presents an opportunity to address their complementary limitations. LLMs gain structured reasoning through formal argumentation frameworks (e.g., graphs of interconnected arguments with well-defined semantics), while CA overcomes its reliance on extensive knowledge engineering and gains the ability to operate at scale on unstructured text [9]. Recent systems already demonstrate the power of this integration by using LLMs to construct formal argumentation frameworks for tasks such as explainable and contestable claim verification, which are then evaluated by a formal argumentation engine [21, 62].

We propose that the convergence of CA and LLMs enables a new paradigm: AI agents that engage in dialectical reasoning processes with humans. We organize this vision by first presenting a taxonomy of three core tasks: *argumentation mining*, *argumentation synthesis*, and *argumentative reasoning*. We then examine recent developments in each area, demonstrating how the integration of LLMs with CA is not merely enhancing existing capabilities but creating fundamentally new approaches to human-AI collaboration. These synergistic developments signal the emergence of interactive, contestable AI agents, giving systems that reason with humans rather than for them. Such agents will mark the emergence of *argumentative human-AI decision-making*.

2 TAXONOMY OF ARGUMENTATION TASKS

Traditionally, CA has centered on three core tasks related to argumentation frameworks (AFs): *argumentation mining* (extracting formal argument structures from natural text), *argumentation synthesis* (generating new arguments within formal constraints), and *argumentative reasoning* (evaluating and explaining arguments through formal argumentation). In the remainder of this section, we analyze developments across these tasks, summarized in Table 1.

2.1 Task I: Argumentation Mining

Argumentation mining is the task of automatically identifying and extracting AFs from unstructured text [31]. The application of LLMs to this task helps bridge the gap between raw natural language

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/EKAW8904

Task	Description	Representative Work
Argumentation Mining	**Definition:** Automatic identification and extraction of argumentative structures from text **Enhancement:** *LLMs* → Eliminating need for manual features and structured inputs **Subtasks**: *AF Extraction*, *Content Detection*	[10, 13, 23, 30, 32, 47, 49, 53]
Argumentation Synthesis	**Definition:** Generation of new arguments, premises, claims, and summaries **Enhancement:** *LLMs* → Enabling flexible, scalable generation without templates **Subtasks**: *AF Generation*, *AF Summarization*	[2, 16, 24, 28, 33, 40, 48]
Argumentative Reasoning	**Definition:** Application of computational argumentation for sound decision-making **Enhancement:** *Argumentation* → Providing formal semantics and transparent, contestable reasoning **Subtasks**: *Claim Verification*, *Explainable Decision-Making*	[21, 27, 34, 41, 51, 61, 62]

Table 1: A taxonomy of synergistic tasks for LLMs and CA.

and the structured representations needed for formal reasoning. This eliminates the need for many of the handcrafted features and domain-specific heuristics that limited traditional approaches.

AF Extraction: AF extraction identifies argumentative units (e.g., premises, claims) and the relations (e.g., support, attack) between them. For instance, given "The proposed method outperforms the baseline because it reduces inference time by 40%," extraction identifies "reduces inference time by 40%" as a premise supporting the claim "the proposed method outperforms the baseline." While traditional transformer baselines like RoBERTa established strong performance [49], they necessitated task-specific training. In contrast, LLMs have shifted the paradigm toward *in-context learning*, allowing general-purpose models to capture complex argumentative relations via few-shot prompting alone [23]. Beyond prompting, recent work frames extraction as a text generation task: Cabessa et al. [13] achieved state-of-the-art results by fine-tuning open-weight models, while the ArgInstruct framework [53] utilizes instruction tuning to handle unseen tasks in zero-shot settings, effectively unifying fragmented extraction pipelines.

Content Detection: Content detection focuses on identifying specific properties of argumentative text such as the author's stance (e.g., supporting/opposing a position), identifying reasoning patterns (analogical, causal), or detecting linguistic markers. For the same example above, content detection would identify the author's positive stance toward the method. Traditional pipelines typically trained single-task classifiers, often with limited cross-genre transfer. LLMs address this limitation through agentic and prompt-based formulations. Lan et al. [30] introduced COLA, which reframes stance detection not as static classification, but as a collaborative multi-agent debate, achieving state-of-the-art performance without labeled data. For cross-genre robustness, Rocha et al. [47] utilized LLMs to insert missing discourse markers, making implicit relations explicit. Furthermore, the larger LLMs have demonstrated strong few-shot capabilities in classifying complex argument schemes, a task where smaller LLMs remain inadequate [10].

2.2 Task II: Argumentation Synthesis

Argumentation synthesis involves generating new AFs, such as premises, claims, counterarguments, and summaries [15, 25, 46]. Earlier systems relied on templates or rigid generation pipelines that limited flexibility [19, 29, 50]. Like mining, synthesis is enhanced by the generative power of LLMs, which allows for the flexible creation of new AFs and summaries from natural language prompts.

AF Generation: AF generation concerns the creation of new argumentative units that are grammatically correct and logically sound. This requires three core capabilities: understanding the topic, organizing premises and claims into coherent argumentative relations, and ensuring adherence to well-defined argumentation schemes. Historically, generating arguments relied on templates and domain-specific pipelines, often resulting in repetitive or unconvincing outputs. LLMs surmount these limitations by enabling multi-step reasoning and instruction adherence. Chen et al. [16] validated this potential, showing strong performance across generation tasks. Recent architectures further improve quality through agentic interaction: AMERICANO [28] employs a generate-critique-refine loop to ensure factual grounding. Furthermore, researchers are now addressing the logical reliability of these generations. Mouchel et al. [40] introduced preference-based fine-tuning to penalize fallacious reasoning, and Gray et al. [24] utilized scheme-based prompting to enforce valid legal argumentation patterns.

AF Summarization: Summarizing argumentative discourse requires preserving central claims and key points of contention. Earlier methods struggled with coverage and salience in long or heterogeneous corpora. With LLMs and large-scale resources such as OpenDebateEvidence [48], recent systems achieve better results. Altemeyer et al. [2] integrated LLMs into AF summarization and evaluation, reporting substantial gains over traditional approaches. However, end-to-end pipelines remain challenging. Li et al. [33] introduced a multi-task dataset covering the full workflow of preparing an argumentative essay and evaluated multiple generative baselines. Their findings reveal that while LLMs perform well on individual subtasks, performance degrades when tasks are chained together, highlighting the problem of error propagation in complex pipelines.

2.3 Task III: Argumentative Reasoning

Argumentative reasoning applies CA semantics for decision-making, verification, and explanation generation [3, 4, 17]. While argumentation mining and synthesis are primarily enhanced by LLMs, argumentative reasoning inverts this relationship. In this task, LLMs handle the front-end challenge of mapping unstructured natural language onto formal AFs, while CA provides the back-end engine for transparent, robust, and verifiable reasoning. For example, while the LLM extracts the paper's claims, the formal reasoning engine determines if the paper's conclusion logically follows from the provided experimental results, or flags a contradiction between the methodology and the results. This hybrid approach overcomes the brittleness of traditional symbolic systems by enabling dynamic, explainable decision-making directly from textual data.

Claim Verification: Claim verification assesses the truthfulness of a claim (or argument) by weighing conflicting evidence. While early

rule-based approaches lacked open-domain adaptability [1, 11, 20], LLM-driven pipelines now enable verification through explicit argument construction. RAFTS [61] replaces opaque scoring with the synthesis of contrastive arguments, improving auditability, while the CHECKWHY benchmark [51] enables multi-step (causal) reasoning for verdict justification. Importantly, recent work integrates formal semantics to enforce logical consistency. Systems such as Argumentative LLMs [21] and ArgRAG [62] use LLMs to map claims/evidence into quantitative AFs, deferring the final decision to a deterministic solver. Consequently, MArgE [41] extended this neuro-symbolic approach to multi-model settings, meshing arguments from diverse LLMs to mitigate individual hallucinations.

Explainable Decision-Making: Explainability requires that decisions can be inspected and contested, with traditional approaches using handcrafted AFs (c.f. [17, 60] for overviews), as well as other forms of argumentative explanation resulting from them (c.f. [42]). LLM-based approaches now generate natural language arguments and formalize them into AFs for transparent inference. Frameworks like Argumentative LLMs [21] and ArgRAG [62] externalize reasoning into editable graphs, allowing users to explicitly add, strengthen, weaken, or remove arguments to observe outcome changes. This capability extends to specialized domains: ArgMed-Agents [27] structures clinical discussions via argumentation schemes, while recent multimodal systems utilize multi-agent debate to provide explicit justifications for image-text classifications [34].

Across these tasks, we observe the synergistic relationship between LLMs and CA. This synergy is beginning to enable capabilities neither approach could achieve independently, particularly the ability to construct, evaluate, and revise AFs through natural language interaction while maintaining formal guarantees. This is pointing toward our central thesis: the emergence of *argumentative human-AI decision-making*.

3 TOWARD ARGUMENTATIVE HUMAN-AI DECISION-MAKING

The advances we have traced across the tasks in the previous section are converging toward AI agents that do not simply assist with decisions but engage with humans as argumentative partners. The full realization of argumentative human-AI collaboration will require agents that can propose and evaluate claims, surface and weigh evidence, and adapt their reasoning as humans add context or express preferences. The goal will not be to replace the humans in the loop, but to amplify them by making complex reasoning more transparent, contestable, and adaptable to domain norms. In what follows, we sketch what such AI agents might look like, and how they may reshape practice in some high-stakes domains.

The transition from current systems to this vision requires bridging three gaps: moving from single-shot argument generation to iterative refinement, from explaining individual decisions to exposing entire reasoning processes, and from domain-agnostic models to systems aligned with professional norms. Current prototypes like Argumentative LLMs [21] and ArgRAG [62] demonstrate feasibility but operate in constrained settings. The next generation of agents must handle open-domain reasoning while maintaining formal guarantees.

Contestable Architectures: The defining property of these agents is contestability, that is, decisions must be open to inspection and revision. The architecture operates through three components. First, separation of generation and evaluation: the LLM generates candidate arguments and relations as structured outputs, while a formal argumentation solver determines acceptability. This ensures probabilistic content generation with deterministic inference. Second, end-to-end provenance, where each argument node contains pointers to source text spans, enabling auditing of hallucinatory or unsupported claims. Third, bidirectional propagation; when users modify the framework, e.g., adjusting argument strength, adding attack relations, or introducing new premises, the solver recomputes acceptability labels across the entire graph. For example, in medical diagnosis, an initial framework might conclude "prescribe medication X" based on arguments A_1 (symptoms match condition Y) and A_2 (X treats Y effectively). If a physician adds argument A_3 (patient has contraindication Z) attacking A_2, the solver propagates this change, potentially flipping the conclusion to "avoid X, consider alternative." This is not only explanation, but collaborative reasoning through formal structures.

Note that our paradigm differs from traditional XAI. Current XAI methods provide post-hoc rationalizations, i.e., explaining what an opaque model did [6, 26], or explaining why certain decisions were reached [52, 57–59]. Even interactive XAI systems that allow probing with "what-if" scenarios still treat the model's reasoning as a fixed black box. In contrast, our paradigm, which is similar in spirit to the Evaluative AI framework [38],[1] externalizes reasoning into formal AFs where inference becomes the interface. In other words, XAI explains the product (a decision), while our approach makes the process (the reasoning) the primary artifact.

Interactive Revisions: While contestable architectures make reasoning inspectable, interactive revisions make it improvable. We envision agents that engage in iterative cycles of proposal, critique, and revision, as in [44]. This is not just "chatting" with a model; it is a structured dialectical protocol where humans provide their reasons and normative constraints [56]. The agent then regenerates the AF to satisfy these constraints while maintaining logical consistency. This enables role-aware collaboration: the human sets the strategy and value axioms, while the agent handles the combinatorial complexity of connecting claims to evidence. These interactions serve a dual purpose: they solve the immediate problem and provide feedback data to align the agent with domain-specific (reasoning) norms, such as legal standards of proof or clinical risk thresholds.

Possible Applications: To understand why this paradigm shift matters, consider the following high-stakes applications in medicine and scientific peer review.

In medicine, decisions require integrating heterogeneous evidence, such as clinical trials, patient histories, treatment guidelines, under significant uncertainty. Current clinical decision support systems often function as black boxes, providing recommendations without accessible reasoning. ArgMed-Agents [27] point toward an alternative, namely AI agents that construct explicit argumentative justifications for clinical recommendations. When combined with

[1]The Evaluative AI framework aims to improve human decision-making by providing users with evidence for (or against) a decision, instead of a single recommendation.

contestable architectures, these become negotiable recommendations. An agent might argue against a treatment due to potential interactions, but may present this as an argument with explicit premises that physicians can challenge based on patient-specific factors. The physician might counter that the interaction risk is mitigated by the patient's genetic profile, leading the system to revise its recommendation. This will preserve clinical autonomy while providing sophisticated decision support. The AI agent serves as an argumentative partner that ensures all relevant evidence is considered, but the physician retains control over how that evidence is weighted given the specific patient context. This is no mean feat for an agent, however, since patients' and physicians' comprehension and trust of explanations in this context have been demonstrated to be sensitive even to the *format* in which this information is delivered [45]. This highlights the suitability of argumentation for this challenging task, however, since it has been shown to support explanations of different formats to adapt to users' preferences [43].

In scientific peer review, one of the problems is scale and consistency. Reviewers must evaluate whether conclusions follow from evidence, identify methodological flaws, and assess contribution significance, all while managing increasing submission volumes. This is a particularly urgent problem in AI, where the number of required reviewers does not scale linearly with the rapidly increasing submission numbers, meaning LLM assistance is already being trialed at top conferences.[2] Argumentative AI agents could construct detailed argument frameworks connecting a paper's claims to its evidence, explicitly modeling the inferential steps and identifying weak links. Reviewers would see not only what the paper claims but how those claims are supposedly supported. They could challenge specific inferential steps, add missing considerations, or identify unstated assumptions. Over time, such systems could help communities converge on shared standards for what constitutes sufficient argumentative support in different subfields, and could expose systematic weaknesses (e.g., recurring reliance on under-powered studies) that are hard to spot one paper at a time. Initial steps have already been made towards this goal [54], showing its potential.

Challenges: These applications share a pattern that defines the promise of argumentative human-AI decision-making. The AI agent's advantage lies in scale and structure, i.e., synthesizing vast amounts of information into coherent, revisable AFs and generating alternatives humans might overlook. The human's advantage lies in normative judgment and contextual sensitivity, i.e., setting goals, interpreting risks through value systems, and recognizing when formally valid conclusions are practically (or ethically) unacceptable. The promise is to combine these complementary strengths in ways that improve decision quality while preserving human agency. Yet realizing this vision faces several challenges.

First, current benchmarks evaluate isolated tasks (e.g., argument extraction accuracy) but not the multi-turn interactions that will characterize the human-AI systems. Specifically, traditional evaluation metrics like accuracy scores might fail to capture whether human-AI teams make better decisions than either alone. We will need new evaluation frameworks that evaluate the following questions: (i) does the system actually save time, or does inspecting and correcting arguments take longer than working independently? (ii) do collaborative decisions show fewer errors and more robust justifications than individual human or AI decisions? (iii) how easily can humans locate and correct mistakes, and do corrections properly propagate through the reasoning chains? and (iv) does the system reduce or increase the human user's mental effort required? [3]

Preventing trust erosion is also an important challenge. AI agents must communicate the uncertainty and strength of arguments without causing over-reliance or unnecessary skepticism. For instance, humans need to know when AI-generated arguments are based on strong evidence versus speculation, when multiple models disagree, and which conclusions are most sensitive to, say, premise changes. There needs to be clearly define roles and dialogue protocols, such as who initiates the dialogue and argumentative process, how disagreements are resolved, when the dialogue should terminate, all while remaining practically feasible [39]. Agents should ideally generate arguments in a timely-efficient manner, should not consume excessive computational resources that make deployment unaffordable, and should handle sensitive data (e.g., legal or medical data) with privacy guarantees. Additionally, agents must be aligned with domain norms, whether it is respecting jurisdictional legal reasoning, clinical guidelines in medicine, or methodological standards in peer review. Without proper governance mechanisms, AI agents may produce arguments that appear stylistically correct but violate substantive domain requirements.

Even with these challenges, we view the trajectory outlined above as encouraging. As evidenced by recent works, we see that the technical ingredients are beginning to align. What remains is to treat argumentation not merely as a representational formalism or a set of NLP tasks, but as a design paradigm for human–AI collaboration, one in which the output is a decision, and the product is the argumentative process that justifies, qualifies, and, when necessary, changes that decision under scrutiny.

4 CONCLUSION

We have discussed how advances in argumentation framework mining and synthesis and argumentative reasoning are not isolated improvements but interconnected capabilities that enable a new paradigm: *argumentative human-AI decision-making*.

The path forward requires contributions from multiple communities. AI researchers must develop architectures that support contestable reasoning and interactive revisions. HCI researchers must design interfaces that make complex argumentation accessible. Domain experts (e.g., scientists, doctors) must help define what constitutes good argumentative practice in their fields. Ethicists and policymakers must establish governance frameworks that ensure these systems respect human values and professional norms.

To reemphasize, the goal is not to automate human judgment but to augment it, to create AI agents that amplify human expertise rather than replace it. If successful, argumentative human-AI decision-making could transform how we approach complex problems, combining the processing power and consistency of AI agents with the values and contextual understanding that humans possess. The result would be decisions that are not only more accurate but more transparent, more justifiable, and more aligned with humans.

[2] https://tinyurl.com/AAAI-LLM-Press-Release

[3] Note that while there is some work on evaluating human-AI interactions [14, 35, 36], adapting these to contestable argumentation systems is a big open challenge.

ACKNOWLEDGMENTS

This research is partially supported by the National Science Foundation (award #2232055) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement #101020934). The views and conclusions expressed in this paper are those of the authors and do not necessarily reflect the official policies or positions of the sponsoring organizations, agencies, or governments.

REFERENCES

[1] Naser Ahmadi, Joohyung Lee, Paolo Papotti, and Mohammed Saeed. 2019. Explainable Fact Checking with Probabilistic Answer Set Programming. In *Proceedings of the 2019 Truth and Trust Online Conference (TTO)*.

[2] Moritz Altemeyer, Steffen Eger, Johannes Daxenberger, Yanran Chen, Tim Altendorf, Philipp Cimiano, and Benjamin Schiller. 2025. Argument Summarization and its Evaluation in the Era of Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, (EMNLP)*. 35490–35511.

[3] Leila Amgoud. 2009. Argumentation for decision making. In *Argumentation in Artificial Intelligence*. Springer, 301–320.

[4] Katie Atkinson, Pietro Baroni, Massimiliano Giacomin, Anthony Hunter, Henry Prakken, Chris Reed, Guillermo Simari, Matthias Thimm, and Serena Villata. 2017. Towards artificial argumentation. *AI Magazine* 38, 3 (2017), 25–36.

[5] Pietro Baroni, Antonio Rago, and Francesca Toni. 2018. How Many Properties Do We Need for Gradual Argumentation?. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 1736–1743.

[6] Vaishak Belle and Ioannis Papantonis. 2021. Principles and practice of explainable machine learning. *Frontiers in Big Data* 4 (2021), 688969.

[7] Philippe Besnard and Anthony Hunter. 2014. Constructing Argument Graphs with Deductive Arguments: A Tutorial. *Argument & Computation* 5, 1 (2014), 5–30.

[8] Henrike Beyer and Chris Reed. 2025. Lexical Recall or Logical Reasoning: Probing the Limits of Reasoning Abilities in Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 13532–13557.

[9] Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2024. Towards Dialogues for Joint Human-AI Reasoning and Value Alignment. *arXiv preprint arXiv:2405.18073* (2024).

[10] Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can Large Language Models Understand Argument Schemes?. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 13666–13681.

[11] Nouf Bindris, N. Cristianini, and J. Lawry. 2020. Claim Consistency Checking Using Soft Logic. *Machine Learning and Knowledge Extraction* 2 (2020), 147–171.

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. 1877–1901.

[13] Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument Mining with Fine-Tuned Large Language Models. In *Proceedings of the International Conference on Computational Linguistics (COLING)*. 6624–6635.

[14] Federico Cabitza, Caterina Fregosi, and Lucia Vicente. 2025. Too Sure for Trust. The Paradoxical Effect of Calibrated Confidence in Case of Uncalibrated Trust in Hybrid Decision Making. In *Proceedings of the World Conference on Explainable Artificial Intelligence (XAI)*. 233–254.

[15] Giuseppe Carenini and Johanna D Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence* 170, 11 (2006), 925–952.

[16] Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the Potential of Large Language Models in Computational Argumentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 2309–2330.

[17] Kristijonas Cyras, Antonio Rago, Emanuele Albini, Pietro Baroni, and Francesca Toni. 2021. Argumentative XAI: A Survey. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 4392–4399.

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

[19] Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the International Conference on Natural Language Generation (INLG)*. 54–64.

[20] Alex Estes, Nikhita Vedula, Marcus D. Collins, Matt Cecil, and Oleg Rokhlenko. 2022. Fact Checking Machine Generated Text with Dependency Trees. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[21] Gabriel Freedman, Adam Dejl, Deniz Gorur, Xiang Yin, Antonio Rago, and Francesca Toni. 2025. Argumentative large language models for explainable and contestable decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[22] Debela Gemechu, Ramon Ruiz-Dolz, Henrike Beyer, and Chris Reed. 2025. Natural Language Reasoning in Large Language Models: Analysis and Evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 3717–3741.

[23] Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can Large Language Models Perform Relation-based Argument Mining?. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[24] Morgan Gray, Li Zhang, and Kevin D Ashley. 2025. Generating case-based legal arguments with LLMs. In *Proceedings of the Symposium on Computer Science and Law*. 160–168.

[25] Nancy Green, Rachael Dwight, Kanyamas Navoraphan, and Brian Stadler. 2011. Natural language generation of biomedical argumentation for lay audiences. *Argument & Computation* 2, 1 (2011), 23–50.

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[27] Shengxin Hong, Liang Xiao, Xin Zhang, and Jianxia Chen. 2024. ArgMed-Agents: explainable clinical decision reasoning with LLM disscusion via argumentation schemes. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 5486–5493.

[28] Zhe Hu, Hou Pong Chan, and Yu Yin. 2024. AMERICANO: Argument Generation with Discourse-driven Decomposition and Agent Interaction. In *Proceedings of the International Natural Language Generation Conference (INLG)*.

[29] Xinyu Hua, Zhe Hu, and Lu Wang. [n.d.]. Argument Generation with Retrieval, Planning, and Realization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pages=2661–2672, year=2019*.

[30] Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*. 891–903.

[31] John Lawrence and Chris Reed. 2020. Argument mining: A survey. *Computational Linguistics* 45, 4 (2020), 765–818.

[32] Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025. Large Language Models in Argument Mining: A Survey. *arXiv preprint arXiv:2506.16383* (2025).

[33] Hao Li, Yuping Wu, Viktor Schlegel, R. Batista-Navarro, Tharindu Madusanka, Iqra Zahid, Jiayan Zeng, Xiaochi Wang, Xinran He, Yizhi Li, and Goran Nenadic. 2024. Which Side Are You On? A Multi-task Dataset for End-to-End Argument Summarisation and Evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[34] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference (WWW)*. 2359–2370.

[35] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards human-ai deliberation: Design and evaluation of llm-empowered deliberative ai for ai-assisted decision-making. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–23.

[36] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are you really sure?" Understanding the effects of human self-confidence calibration in AI-assisted decision making. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–20.

[37] R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences* 121, 41 (2024), e2322420121.

[38] Tim Miller. 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. 333–342.

[39] Sanjay Modgil. 2017. Dialogical Scaffolding for Human and Artificial Agent Reasoning. In *Proceedings of the 5th International Workshop on Artificial Intelligence and Cognition, (AIC)*. 58–71.

[40] Luca Mouchel, Debjit Paul, Shaobo Cui, Robert West, Antoine Bosselut, and Boi Faltings. 2025. A Logical Fallacy-Informed Framework for Argument Generation. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*. 7296–7314.

[41] Ming Pok Ng, Junqi Jiang, Gabriel Freedman, Antonio Rago, and Francesca Toni. 2025. MArgE: Meshing Argumentative Evidence from Multiple Large Language Models for Justifiable Claim Verification. *arXiv preprint arXiv:2508.02584* (2025).

[42] Antonio Rago. 2024. A Little of That Human Touch: Achieving Human-Centric Explainable AI via Argumentation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 8565–8570.

[43] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David A. Lagnado, and Francesca Toni. 2021. Argumentative explanations for interactive recommendations. *Artificial Intelligence* 296 (2021), 103506.

[44] Antonio Rago, Hengzhi Li, and Francesca Toni. 2023. Interactive Explanations by Conflict Resolution via Argumentative Exchanges. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (KR)*. 582–592.

[45] Antonio Rago, Bence Pálfi, Purin Sukpanichnant, Kavyesh Vivek, Hannibal Nabli, Olga Kostopoulou, James Kinross, and Francesca Toni. 2025. Exploring the Effect of Explanation Content and Format on User Comprehension and Trust in Healthcare. In *28th European Conference on Artificial Intelligence (ECAI) - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS)*. 5400–5407.

[46] Chris Reed and Derek Long. 1998. Generating the structure of argument. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 1091–1097.

[47] Gil Rocha, Henrique Lopes Cardoso, Jonas Belouadi, and Steffen Eger. 2024. Cross-genre argument mining: Can language models automatically fill in missing discourse markers? *Argument & Computation* Preprint (2024), 1–41.

[48] Allen Roush, Yusuf Shabazz, Arvind Balaji, Peter Zhang, Stefano Mezza, Markus Zhang, Sanjay Basu, Sriram Vishwanath, and Ravid Shwartz-Ziv. 2024. Opendebateevidence: A massive-scale argument mining and summarization dataset. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. 34270–34293.

[49] Ramon Ruiz-Dolz, Jose Alemany, Stella M Heras Barberá, and Ana García-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems* 36, 6 (2021), 62–70.

[50] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) - System Demonstrations*. 109–114.

[51] Jiasheng Si, Yibo Zhao, Yingjie Zhu, Haiyang Zhu, Wenpeng Lu, and Deyu Zhou. 2024. CHECKWHY: Causal Fact Verification via Argument Structure. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 15636–15659.

[52] Sarath Sreedharan, Tathagata Chakraborti, and Subbarao Kambhampati. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence* 301 (2021), 103558.

[53] Maja Stahl, Timon Ziegenbein, Joonsuk Park, and Henning Wachsmuth. 2025. ArgInstruct: Specialized Instruction Fine-Tuning for Computational Argumentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[54] Purin Sukpanichnant, Anna Rapberger, and Francesca Toni. 2024. PeerArg: Argumentative Peer Review with LLMs. *arXiv preprint arXiv:2409.16813* (2024).

[55] Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models-a critical investigation. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*. 75993–76005.

[56] Stylianos Loukas Vasileiou, Ashwin Kumar, William Yeoh, Tran Cao Son, and Francesca Toni. 2024. Dialectical Reconciliation via Structured Argumentative Dialogues. In *International Conference on Principles of Knowledge Representation and Reasoning (KR)*. 777–787.

[57] Stylianos Loukas Vasileiou, Alessandro Previti, and William Yeoh. 2021. On exploiting hitting sets for model reconciliation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 6514–6521.

[58] Stylianos Loukas Vasileiou, William Yeoh, Alessandro Previti, and Tran Cao Son. 2025. On generating monolithic and model reconciling explanations in probabilistic scenarios. *Journal of Artificial Intelligence Research* 84 (2025).

[59] Stylianos Loukas Vasileiou, William Yeoh, Tran Cao Son, Ashwin Kumar, Michael Cashmore, and Daniele Magazzeni. 2022. A Logic-Based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research* 73 (2022), 1473–1534.

[60] Alexandros Vassiliades, Nick Bassiliades, and Theodore Patkos. 2021. Argumentation and Explainable Artificial Intelligence: A Survey. *Knowledge Engineering Review* 36 (2021), e5.

[61] Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. 10331–10343.

[62] Yuqicheng Zhu, Nico Potyka, Daniel Hernández, Yuan He, Zifeng Ding, Bo Xiong, Dongzhuoran Zhou, Evgeny Kharlamov, and Steffen Staab. 2025. ArgRAG: Explainable retrieval augmented generation using quantitative bipolar argumentation. In *Proceedings of the Internation Conference on Neurosymbolic Learning and Reasoning (NeSy)*.