DR-HAI: Argumentation-based Dialectical Reconciliation in Human-AI Interactions

Stylianos Loukas Vasileiou[1], **Ashwin Kumar**[1], **William Yeoh**[1], **Tran Cao Son** [2], **Francesca Toni** [3]

[1]Washington University in St. Louis
[2] New Mexico State University
[3] Imperial College London
{v.stylianos, ashwinkumar, wyeoh}@wustl.edu, stran@nmsu.edu, f.toni@imperial.ac.uk

Abstract

We introduce Dialectical Reconciliation in Human-AI Interactions (DR-HAI), a framework designed to extend Explainable AI Planning (XAIP) approaches for enhanced human-AI interaction. By adopting an argumentation-based dialogue paradigm, DR-HAI enables interactive reconciliation to address knowledge discrepancies between an AI agent and a human user. In line with the symposium's focus on adaptive AI systems, DR-HAI offers a natural and sound approach for making AI systems more understandable and effective in real-world scenarios.

Introduction

The rapid advancement and integration of AI systems into various aspects of daily life underscore the need for systems that are not only effective and adaptable but also transparent and understandable to human users. In response, the field of Explainable AI Planning (XAIP) has emerged (Fox, Long, and Magazzeni 2017), focusing on developing AI agents capable of explaining their decisions and actions in a manner comprehensible to users. At the heart of XAIP is the concept of *model reconciliation* (Chakraborti et al. 2017), a process aimed at aligning the mental models of AI agents and human users to facilitate better understanding and communication. These mental models are typically encoded in planning (Sreedharan, Chakraborti, and Kambhampati 2021) or logical formalisms (Son et al. 2021; Vasileiou et al. 2022).

However, current trends in XAIP face significant challenges, particularly in the context of adaptive AI systems. Traditional approaches often assume that the AI agent already knows the user's (mental) model. This assumption can lead to misunderstandings, as the agent might base its explanations on an inaccurate or incomplete understanding of the user's knowledge and preferences. Additionally, these methods typically rely on single-shot interactions, which may be insufficient in dynamic environments where user needs, objectives, and external conditions evolve post-deployment.

Arguing for a more interactive approach to human-AI interactions, we introduce the *Dialectical Reconciliation in Human-AI Interactions* (DR-HAI) framework. DR-HAI extends beyond traditional XAIP and model reconciliation approaches, fostering more effective human-AI interactions by enabling a deeper understanding of AI agent decisions and behavior. Specifically, DR-HAI facilitates a multi-shot, argumentation-based dialogue between an AI agent and a human user. This interaction is not based on presupposed user models; instead, it evolves dynamically as the interaction progresses, allowing for a more accurate and nuanced exchange of information. We call this kind of interaction *dialectical reconciliation* and it is aimed at enhancing the user's *understanding* of the agent's decisions. Importantly, the goal of DR-HAI is not to persuade the user to agree with the agent's decisions, but to *facilitate an understanding of these decisions from the agent's perspective—even if the user disagrees with those decisions*.

Consider an illustrative example where a user observes a robot assistant behaving unexpectedly, such as avoiding specific actions. Through dialectical reconciliation, the user can engage with the robot to delve into its decision-making process. This interaction not only allows the user to gain a clear understanding of the robot's behavior but also provides a crucial opportunity to refine or repair the robot's model. By understanding the robot's decisions from its perspective, users can more effectively identify and address gaps or inaccuracies in the robot's model, leading to improvements in the robot's functionality and behavior.

The next section outlines key aspects of the DR-HAI framework. For a deep dive into its technical details, evaluation, and related discussions, please refer to our extended version (Vasileiou et al. 2023).

DR-HAI Framework

The conceptual foundation of DR-HAI is *dialectical reconciliation*, a process resolving inconsistencies, misunderstandings, and knowledge gaps between the AI agent and the human user. This is achieved through argument exchange and dialogue moves that collaboratively construct a shared understanding of the agent's decisions.

To successfully achieve dialectical reconciliation, the agent and user follow certain dialogue protocols that guide their interaction:

- Establish a clear dialogue structure, including the use of *locutions* that define permissible speech acts and turn-taking mechanisms.
- Engage in a cooperative and collaborative manner, with the shared goal of improving the user's understanding.

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

- Employing argumentation techniques to constructively challenge each other's positions.

Key Assumptions and Goal of DR-HAI

The following key assumptions underlie DR-HAI:

- **Distinct Knowledge bases:** The AI agent is associated with a knowledge base KB_a, encapsulating its understanding of the task at hand. Conversely, the user is linked to KB_h, representing their approximation of the agent's knowledge, which can initially be empty. Importantly, the agent (resp. user) does not have explicit access to KB_a (resp. KB_h).
- **User Queries:** Initiated by the user, the dialogue starts with a query φ, where $KB_h \not\models \varphi$ (or $KB_h \models \neg\varphi$) and $KB_a \models \varphi$. The user has the flexibility to generate subsequent queries dynamically as the dialogue progresses, reflecting their evolving understanding and the need for clarification.
- **Public Commitment stores:** Both the agent and the user contribute to a *public commitment store*, akin to a "chat log", which stores their utterances throughout the dialogue. This feature allow to build complex and contextually aware arguments.

Now, the goal of a DR-HAI is formalized as follows:

> *Given an agent with knowledge base KB_a, a human user with knowledge base KB_h, and an initial query φ such that $KB_h \not\models \varphi$ (or $KB_h \models \neg\varphi$) and $KB_a \models \varphi$, the goal of a DR-HAI dialogue is to* enable $KB_h \models \varphi$ *through dialectical reconciliation.*

A crucial aspect of this formulation is successfully *enabling* $KB_h \models \varphi$. At a high level, this translates to finding a way to help the user transition from a state of *not understanding* a decision φ (i.e., $KB_h \not\models \varphi$ or $KB_h \models \neg\varphi$) to a state of *understanding* the decision (i.e., $KB_h \models \varphi$). We posit that dialectical reconciliation is ideal in achieving this goal.

DR-HAI Dialogue Structure

Inspired by Hamblin's dialectical games framework (Hamblin 1970, 1971), a DR-HAI dialogue is viewed as a game-theoretic interaction, where utterances are treated as moves governed by rules that define their applicability. In this context, moves consist of a set of *locutions*, which determine the types of permissible utterances agents can make. Specifically, we allow for the following locutions:

- **Query:** This locution is available only to the user, and it is used to request supportive arguments (e.g., explanations) on specific agent decisions.
- **Support:** This locution is only available to the agent, who uses it to provide arguments supporting its decisions, as requested by the user's query.
- **Refute:** Available to both participants, this locution permits them to provide counterarguments that refute the other's arguments. For example, the user can refute the agent's support, the agent can in turn refute the user's counterargument, and so on.
- **Agree-to-Disagree:** This locution allows both the agent and the user to acknowledge each other's perspective when no further queries (from the user) or counterarguments (from both) are possible.

These locutions are instantiated with specific formulae from the knowledge bases that make up the range of possible *dialogue moves*. To generate arguments and counterarguments from the knowledge bases, we employ logic-based argumentation techniques (Besnard and Hunter 2014). Finally, to maintain a coherent dialogue structure, the agent and the user take turns in making moves, and the dialogue terminates if and only if the user opts for the agree-to-disagree move, i.e., when the user does not have any more queries to be addressed or refutations to provide.

Discussion

Effectively bridging the gap between the AI agent's decision-making process and the human user's understanding is a key problem for current and future AI systems. With the DR-HAI framework, we have argued for an interactive, argumentation-based dialogue, namely *dialectical reconciliation*, as a solution to this problem.

While it might be tempting to view the symbolic nature of DR-HAI as a limitation compared to the capabilities of large language models (LLMs), this perspective overlooks the unique strengths and applications of each approach. LLMs, celebrated for their proficiency as few-shot learners and their skill in generating well-structured sentences, excel in processing and generating natural language (Brown et al. 2020; Lu et al. 2022). This ability significantly enhances the user experience in human-AI interactions. However, LLMs' reliance on statistical inference can lead to challenges in maintaining logical consistency and accuracy, particularly in complex reasoning and planning tasks (Rae et al. 2021; Creswell, Shanahan, and Higgins 2023; Zhang et al. 2022; Valmeekam et al. 2023)..

In contrast, the core strength of DR-HAI lies in its symbolic and logical foundations, providing a robust framework for logical reasoning and argumentation-based dialogues. The explicit representation of knowledge is not only well-suited for interpretability and explainability tasks (Evans and Grefenstette 2018; Schede, Kolb, and Teso 2019), but also allows for a deeper understanding of the knowledge it represents, its assumptions, and the reasoning processes involved (Mocanu and Belle 2023). Therefore, DR-HAI's capabilities are particularly valuable in scenarios that demand rigorous, multi-step logical reasoning and planning, areas where LLMs may fall short.

Looking ahead, the integration of DR-HAI's symbolic logic with the natural language processing strengths of LLMs presents an exciting avenue for the development of AI systems. This hybrid approach could involve translating the formal arguments and logical structures generated by DR-HAI into more intuitive, natural language expressions. Such a combination would not only enhance the accessibility of these systems for lay users but also ensure that the explanations provided are both logically coherent and easily understandable, thus leading to more advanced, trustworthy, and user-friendly AI systems.

References

Besnard, P.; and Hunter, A. 2014. Constructing argument graphs with deductive arguments: a tutorial. *Argument & Computation*, 5(1): 5–30.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chakraborti, T.; Sreedharan, S.; Zhang, Y.; and Kambhampati, S. 2017. Plan Explanations as Model Reconciliation: Moving Beyond Explanation as Soliloquy. In *IJCAI*, 156–163.

Creswell, A.; Shanahan, M.; and Higgins, I. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *ICLR*.

Evans, R.; and Grefenstette, E. 2018. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61: 1–64.

Fox, M.; Long, D.; and Magazzeni, D. 2017. Explainable Planning. *CoRR*, abs/1709.10256.

Hamblin, C. L. 1970. *Fallacies*. Methuen and Co Ltd.

Hamblin, C. L. 1971. Mathematical models of dialogue. *Theoria*, 37(2): 130–155.

Lu, K.; Grover, A.; Abbeel, P.; and Mordatch, I. 2022. Pretrained transformers as universal computation engines. In *AAAI*, 7628–7636.

Mocanu, I. G.; and Belle, V. 2023. Knowledge representation and acquisition in the era of large language models: Reflections on learning to reason via PAC-Semantics. *Natural Language Processing Journal*, 100036.

Rae, J. W.; Borgeaud, S.; Cai, T.; Millican, K.; Hoffmann, J.; Song, F.; Aslanides, J.; Henderson, S.; Ring, R.; Young, S.; et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Schede, E. A.; Kolb, S.; and Teso, S. 2019. Learning linear programs from data. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 1019–1026. IEEE.

Son, T. C.; Nguyen, V.; Vasileiou, S. L.; and Yeoh, W. 2021. Model reconciliation in logic programs. In *European Conference on Logics in Artificial Intelligence*, 393–406.

Sreedharan, S.; Chakraborti, T.; and Kambhampati, S. 2021. Foundations of explanations as model reconciliation. *Artificial Intelligence*, 301: 103558.

Valmeekam, K.; Sreedharan, S.; Marquez, M.; Olmo, A.; and Kambhampati, S. 2023. On the planning abilities of large language models (a critical investigation with a proposed benchmark). *arXiv preprint arXiv:2302.06706*.

Vasileiou, S. L.; Kumar, A.; Yeoh, W.; Son, T. C.; and Toni, F. 2023. DR-HAI: Argumentation-based Dialectical Reconciliation in Human-AI Interactions. *arXiv preprint arXiv:2306.14694*.

Vasileiou, S. L.; Yeoh, W.; Son, T. C.; Kumar, A.; Cashmore, M.; and Magazzeni, D. 2022. A Logic-based Explanation Generation Framework for Classical and Hybrid Planning Problems. *Journal of Artificial Intelligence Research*, 73: 1473–1534.

Zhang, H.; Li, L. H.; Meng, T.; Chang, K.-W.; and Broeck, G. V. d. 2022. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*.