In AAAI 2025 Spring Symposium on Current and Future Varieties of Human-AI Collaboration (HAI 2025)
San Francisco, CA, USA

Leveraging LLMs for Collaborative Human-AI Decision Making

Anthony Favier, Pulkit Verma, Ngoc La, and **Julie A. Shah**
MIT CSAIL
Cambridge, USA
{antfav24, pulkitv, ntmla}@mit.edu, julie_a_shah@csail.mit.edu

1 Introduction

Human-AI collaboration is a rapidly evolving field that seeks to leverage the complementary strengths of humans and artificial intelligence (AI) to solve complex problems. An area where such collaboration holds significant promise is in decision-making tasks, particularly in automated planning [2]. Classical automated planners have been widely used to address such problems, but they often struggle with large, complex, and realistic scenarios. Due to their PSPACE complexity [1], these systems typically cannot generate optimal solutions for complex problems. Instead, they rely on simplification techniques like heuristics and relaxation methods to manage computational demands [5]. While effective in constrained settings, these approaches often fall short when applied to real-world problems that require adaptability and scalability.

In recent years, Large Language Models (LLMs) have emerged as a potential solution to these challenges. LLMs, with their ability to process and generate natural language, offer a unique combination of abstraction and generalization. However, despite their promise, LLMs alone are not sufficient for solving complex planning problems [3]. Their limitations in reasoning, consistency, and reliability highlight the need for hybrid approaches that combine the strengths of symbolic reasoning with the flexibility of LLMs. An overview of such hybrid approach is depicted in figure 1. In this extended abstract, we use a disaster recovery scenario as a running example to discuss how to leverage LLMs in a Human-AI collaboration.

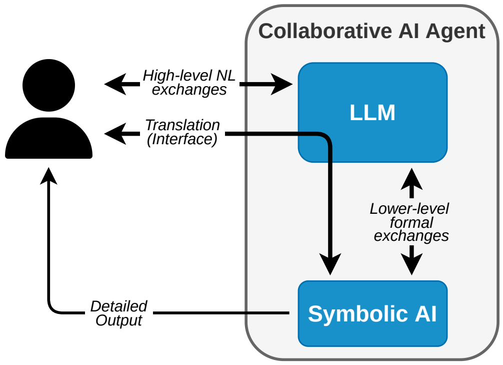

Figure 1: Overview of a generic Hybrid Collaborative AI Agent

2 Human-AI Collaboration Roles in Decision-Making

In this context, different roles can be identified. Classical symbolic AI should be in charge of **low-level, complex and long-horizon reasoning**. They ensure sound and detailed solutions. On the other hand, humans should act as **high-level reasoners and critics**, leveraging their inherent common sense and intuition. In this context, LLMs can play several roles beneficial to symbolic systems, humans, and the collaboration. First, LLMs can act as an **interface translator** converting human natural language inputs into formal, structured language (e.g. PDDL [4]) for symbolic systems. This would avoid the need for specialized technical knowledge, which can impede rapid response and effective coordination. For instance, an LLM could translate a human operator's verbal instruction to "prioritize children and injured individuals" into a formal constraint for the planner. LLMs can **guide the search process** of symbolic planners by informing heuristic functions or suggesting high-level strategies. For instance, in a disaster scenario, an LLM might propose prioritizing rescue routes with lower risks or higher practicability based on weather predictions and fuel constraints. LLMs can **generate ideas** to explore novel alternative solutions that may not be immediately apparent to human planners. For example, an LLM might suggest using drones to scout inaccessible areas or rerouting ground vehicles to avoid flooded paths. LLMs can also **highlight information** to emphasize critical information within a problem domain, such as areas with high survival probabilities or locations with rapidly deteriorating weather conditions.

Challenges and Considerations LLMs have limitations, notably the challenge of verifying their outputs beyond syntax checks. They can hallucinate, misinterpret, or omit information, leading to errors. For instance, an LLM might misread weather data and suggest an impassable route. Thus, systems should integrate LLMs as part of a broader decision-making process, e.g. a vote in a committee, to avoid reliance on a single point of failure.

3 Incorporating Transparency, Trust, and Dependability

Effective Human-AI collaboration requires transparency, trust, and dependability. Transparency helps operators understand AI decisions, such as why a particular route was prioritized or why a rescue team was assigned to a specific location. Trust is built through consistent and reliable performance, especially in high-stakes situations where errors can have severe consequences. Dependability requires robust mechanisms to handle errors and uncertainties, especially when LLMs are involved. Uncertainties or ambiguous outputs can be flagged for human review, ensuring that critical decisions are not based on unreliable information. For example, LLM-generated routes can be cross-checked with weather data and human input.

4 Measuring Performance and Addressing Tradeoffs

Evaluating human-AI teams in decision-making requires metrics that assess overall effectiveness and individual contributions. These should consider task completion time, solution quality, and adaptability to challenges. In rescue operations, for instance, key metrics include the number of people rescued, response time, and survival estimate accuracy. Human-AI collaboration also involves benefit-risk tradeoffs: while LLMs can enhance decision-making by providing insights and reducing cognitive load, they may introduce biases. For example, prioritizing survivors with higher probabilities could raise ethical concerns. Balancing these tradeoffs demands careful system design and continuous evaluation.

5 Conclusion

The integration of LLMs into collaborative decision-making systems represents a promising avenue for addressing complex automated planning problems. By combining the strengths of symbolic reasoning and LLMs, hybrid approaches can overcome the limitations of traditional methods while enabling more accessible and effective human-AI collaboration. However, realizing this potential requires addressing challenges related to verification, transparency, and dependability, as well as developing robust performance metrics and strategies for managing benefit-risk tradeoffs. As research in this area continues to advance, it has the potential to transform how humans and AI systems work together to solve some of the most challenging problems across diverse application domains.

Acknowledgment

This work was supported by the ONR under grant N000142312883.

References

[1] T. Bylander. Complexity results for planning. In *Proc. IJCAI*, 1991.

[2] M. Ghallab, D. Nau, and P. Traverso. *Automated Planning and Acting*. Cambridge University Press, 2016.

[3] S. Kambhampati, K. Valmeekam, L. Guan, M. Verma, K. Stechly, S. Bhambri, L. P. Saldyt, and A. B. Murthy. LLMs can't plan, but can help planning in LLM-modulo frameworks. In *Proc. ICML*, 2024.

[4] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. S. Weld, and D. Wilkins. PDDL – The planning domain definition language. Technical report, Yale Center for Computational Vision and Control, 1998.

[5] H. Müller-Merbach. Heuristics and their design: A survey. *Journal of EJOR*, 1981.