Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Enhancing AI Decision-Making: Sensitivity Analysis, Hyperparameter Optimization, Multi-Agent Collaboration, and AI-Human Comparisons

Manish Kumar[1], Dr Jugnesh Kumar[2]

[1]Student, M. Tech, Computer Science &Engineering, Faridabad College of Engineering and Management, Faridabad, Haryana 121001.

[2]Professor, Computer Science &Engineering, Faridabad College of Engineering and Management,

Faridabad, Haryana121001.
*Corresponding Email: manishnit4u@gmail.com

Abstract

The use of Artificial Intelligence (AI) is changing the decision-making landscape in many areas, including law, medicine, and automated systems. Nonetheless, the issues facing current AI systems are extreme: they are sensitive to small variations included in inputs, the hyperparameter optimization is difficult, and we a lack of agreement among the agents, and they are in large part inferior to the full contextual and ethical reasoning capabilities of human decision-makers. This paper takes up four key aspects of AI's decision-making abilities. First, our sensitivity analysis demonstrates that large language models can show high deviation in their outputs when given paraphrased inputs, with cosine similarity scores between .80 to .92. Second, our experiments conducted in this paper reveal that the parameterization of hyperparameters produces different outcomes in accuracy, coherence, and verbosity, thereby signaling the importance of balanced parameterization. Third, our multi-agent experiments signal that decisions made by AI systems can demonstrate oscillatory disagreement and dominance effects that indicate agents cannot behave alike based on our moral decisions. Finally, our comparative analyses suggest that, while AI models are comparable to human models in some cases of factual accuracy in legal and medical decisions (85%–90%), they lack the depth of emotional intelligence and ethics inherent in human decision-making, thus highlighting potential shortcomings in high-risk environments. The observations highlight the need for strong training methods, formalized coordination schemes in multi-agent environments, and explainable AI methodologies to improve dependability, transparency, and human-AI cooperation in decision-making situations.

Keywords: Artificial Intelligence, Decision-Making, Sensitivity Analysis, Hyperparameter Optimization, Multi-Agent Systems, Human-AI Collaboration, Explainable AI.

1. Introduction

Artificial intelligence (AI) has quickly evolved decision-making in most fields, such as finance, healthcare, law, and autonomous systems. Decision-making using AI relies on advanced models that take large amounts of data and provide forecasts with little or no human interaction. Al- though the progress has been significant, AI models are

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

prone to challenges like input variation sensitivity, difficulties in hyperparameter tuning, multi agent coordination issues, and a general difference in reasoning patterns from human thinking [1], [2], [4]. One of the chief issues in AI decision-making is its susceptibility to change in input. Large language models (LLMs), for example, exhibit in consistencies when presented with paraphrase queries, which result in varied answers without losing semantic similarity [2], [3]. This sensitivity poses concerns over the reliability of AI, especially in domains like legal reasoning and medical diagnosis, where unwavering decision-making is critical [10], [11].

Another major issue is hyperparameter tuning, which has a direct impact on model performance. Literature shows that adjustments of learning rates, batch sizes, and weight decay have a direct effect on the AI accuracy, coherence, and verbosity [4], [7]. Overfine-tuning can result in overfitting, verbosity, or fact in consistencies, while under tuned models can result in suboptimal decision-making [8]. Therefore, there is a need to determine well-balanced hyperparameter settings that maximize AI performance without losing generalizability. Another important problem in AI decision-making is the multi-agent collaboration problem. AI agents operating in group decision-making environments tend to lack consensus, being stuck in oscillatory conflicts or single agent dominance [5], [9]. AI agents in strategic and ethical problems can fail to converge to inconsistent or biased solutions [15]. To solve this problem, structured negotiation frameworks need to be developed to enable efficient AI collaboration [13].

Also, AI decision-making is inherently different from human decision-making, especially in situations that call for ethical judgment, emotional intelligence, and context sensitivity [10], [11]. Though AI models are able to analyze tremendous amounts of data and develop rational conclusions, they are incapable of detecting emotional and ethical subtleties involved in human judgment [14]. Comparisons between AI-driven and human decision-making indicate that AI excels in structured, data-driven environments but struggles in subjective, high- stakes decision-making scenarios such as law and healthcare [11].

Given these challenges, this study aims to explore four key dimensions of AI decision making:(1) sensitivity analysis of

AI responses to variations of input, (2) hyperparameter optimization and its effects on AI model accuracy, (3) multi-agent AI collaboration and consistency of decisions, and (4) comparisons between human and AI reasoning in ethical and high-risk settings. Addressing these aspects, this research aims to improve the dependability, interpretability, and efficiency of AI decision-making models [1], [2], [15].

This paper follows the following structure. Section II provides an exhaustive literature review, where gaps and shortcomings in AI decision-making are established. Section III outlines the methodology used in the analysis of AI sensitivity, hyperparameter tuning, multi-agent coordination, and AI-human reasoning comparative analysis. Section IV elaborates on results and findings, with the incorporation of significant insights and implications.

2.0 Literature Review

This section critically examines the literature on AI decision-making with respect to multi-agent cooperation, hyperparameter optimization, sensitivity analysis, and analogies to human reasoning. The study's scope and objectives, current gaps, significance, and key findings are explained from this perspective.

The four main areas of sensitivity analysis, hyperparameter tuning, multi-agent cooperation, and AI-human comparison are examined in this paper to investigate how to improve AI decision-making. As AI systems grow more complex and integrated into crucial decision-making processes, current research highlights the need for

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302
Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

increased transparency and dependability. This review identifies gaps and limitations in current approaches by concentrating on how these areas affect AI's performance and reliability. Comparing AI to humans is crucial for determining both the advantages and disadvantages of AI and human decision-making. Human decision-making is frequently characterized by dexterity, judgement, moral principles, and emotional intelligence, even though AI is extremely efficient at handling large volumes of data and repetitive tasks. This review highlights how human-AI cooperation can yield more effective and reliable outcomes. Results will be discussed.

A. Existing Research in AI Decision-Making

Some studies have investigated AI decision-making behaviors in the context of hyperparameter optimization, multi-agent coordination, and human-AI comparison. Deep reinforcement learning (DRL) has been extensively used to improve decision-making processes [1], [4]. There has been recent research showing that AI models are highly sensitive to small input variations, and this makes it difficult to achieve decision robustness [2], [3].
Multi-agent reinforcement learning (MARL) is being used more and more to enhance AI coordination, though studies point out that oscillatory decision-making and dominance by agents are still unresolved challenges [5], [8].Inter-comparisons between AI and human reasoning propose that the AI model performs well in structured decision making but does not perform well in empathetic reasoning and ethical choices [10], [11].
These developments notwithstanding, numerous gaps and limitations exist, requiring more research on AI robustness, coordination frameworks, and ethical AI development.

Gaps in Current Literature

Although AI decision-making has been well researched, the following gaps exist:

A. Input Sensitivity: AI models exhibit sensitive variation of outputs for slight changes in inputs, necessitating better robustness methods [2], [3].
B. Hyperparameter Balancing: It has been discovered that optimizing AI models will cause them to overfit, become verbose, and display factual inaccuracies, thus demanding controlled tuning [4], [7].
C. Multi-Agent Collaboration Challenges: AIteamsregularly fail to arrive at a consensus because of dominance effects and absence of coordination mechanisms [5], [9].
D. AI vs. Human Decision-Making Limitations: AI models are devoid of emotional intelligence and ethical reasoning, limiting their deployment in high-stakes areas like law and healthcare [10], [11].

These loopholes emphasize the importance of research work on enhancing AI robustness, multi-agent decision-making models, and ethical AI systems.

Significance of the Study

This study addresses the aforementioned gaps by:

- Enhancing AI robustness against minor input variationsto ensure consistent decision-making [2], [3].
- Investigating hyperparameter tuning methodologies to balance accuracy and coherence [4], [7].
- Developing structured multi-agent collaboration mechanisms to prevent decision oscillations [5].
- Comparing AI decisions with human experts is to improve AI’s ability to handle subjective and ethical scenarios [10], [11].

By addressing these issues, the study contributes to the development of more reliable and interpretable AI

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302
Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

decision- making frameworks.

E. Scope of the Study

The study focuses on:

- Sensitivity Analysis of AI Responses: Investigating how paraphrased inputs affect AI decision consistency [2], [3].
- Hyperparameter Tuning Effects: Analyzing the impact of different learning rates, batchsizes, and weight decay on AI accuracy and verbosity [4], [7].
- Multi Agent AI Coordination: Examining the challenges of achieving consensus in multi-agent AI systems [5], [9].
- AI vs. Human Reasoning: Comparing AI-generated decisions with human expert judgments in law and health-care [10], [11].

F. Objectives of the Study

The primary objectives of this study are:

- To evaluate AI sensitivity to minor input variations and propose methods to enhance robustness [2], [3].
- To analyze the impact of hyperparameter configurations on AI decision-making performance [4], [7].
- To explore multi-agent AI collaboration challenges and develop structured coordination strategies [5], [9].
- To compare AI and human reasoning, identifying strengths and weaknesses in ethical and subjective decision-making [10], [11].

By addressing these objectives, this study aims to enhance the reliability, transparency, and effectiveness of AI-driven decision-making

3.0 METHODOLOGY

This section outlines the methodology employed to investigate AI decision making across four key research dimensions:
(1) Sensitivity Analysis of AI Models to Input Variations, (2) Parameter Variability and Its Effects on AI Decision-Making,
(3) Multi-AI Agents Collaboration, and (4) Comparisons between AI and Human Reasoning. Each of these dimensions was analyzed through controlled experiments, leveraging transformer-based language models and statistical evaluation methods. The sub-sections below detail the experimental setup, implementation approach, and evaluation metrics used foreach research strand

Sensitivity Analysis of AI Models to Input Variation

AI models, particularly large language models (LLMs), exhibit varying degrees of sensitivity to minor perturbations in input text. This study assesses how small paraphrases in questions impact the consistency of AI-generated response
Experimental Setup: A pre-trained GPT-4 model was employed to generate responses to a set of semantically equivalent yet lexically varied questions. To quantify sensitivity, the responses were embedded into a vector space using the Mini LM Sentence Transformer, and their similarity was assessed using cosine similarity.

Implementation Approach: The following steps were performed:

- A set of four paraphrased questions related to climate change were used as input.
- GPT-4generatedresponses for each question.
- Responses were converted into vector representations using sentence embeddings.

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

- A cosine similarity matrix was computed to measure the consistency of responses.
- A heatmap visualization was created to illustrate the similarity score.

1) Parameter Variability and Its Effects on AI Decision- Making

Fine-tuning AI models involves configuring several hyperparameters, such as learning rate, batch size, and weight decay, which influence the model's decision-making capabilities. This study analyzes how these hyperparameters affect accuracy, coherence, and verbosity.

Experimental Setup: Two fine-tuned BERT-based models were trained on the IMDB dataset with different hyper parameter configurations:

- Model1: Learning rate=5e-5, batch size=8, weight decay = 0.01.
- Model 2: Learning rate = 3e-5, batch size = 16, weight decay = 0.02.

2) Implementation Approach:

- ABERT classifier was fine-tuned on a subset of the IMDB dataset.
- Training was conducted separately for both hyperparameter configurations.
- Model performance was evaluated based on accuracy scores.

A bar chart visualization was created to compare accuracy across hyperparameter settings

3) Multi-AI Agents Collaboration

AI systems are increasingly deployed in multi-agent setups where multiple AI models collaborate on decision-making tasks. This study examines whether LLM-based agents can collectively solve problems more effectively than individual models.

Experimental Setup: Three AI models ("GPT-4", "Claude-3", and "Gemini") were simulated as autonomous decision-makers on an ethical dilemma scenario:

"Should self-driving cars prioritize passengers or pedestrians in unavoidable accidents?"

Each AI agent provided independent reasoning, followed by a voting process to determine consensus.

Implementation Approach:

- Three different AI agents proposed ethical reasoning strategies.
- A simulated voting mechanism was introduced, where agents selected a preferred decision.
- The distribution of votes was visualized in a bar chart.

2) Comparisons Between AI and Human Reasoning

While AI excels in data processing and coverage, it lacks the nuance, adaptability, and ethical considerations inherent in human decision-making. This study compares AI-driven reasoning with human expert judgments in legal and medical domains

Experimental Setup: Two real-world scenarios were analyzed:

1) Legal Analysis: Can AI draft enforceable contracts better than human lawyers?
2) Medical Diagnosis: How accurate are AI-generated diagnoses compared to human doctors?

For each case, AI generated responses were compared against human expert opinions, assessing:

- Coherence and logical consistency
- Factual correctness

3) **Ethical reasoning and emotional intelligence**

3) Summary of Methodology and Findings

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

Table II summarizes the methodologies employed and key observations across the four research strands.

Results and Analysis

This section presents the results obtained from the experiments conducted in the study, analyzing AI decision-making across four key dimensions: (1) Sensitivity Analysis of AI Models to Input Variations, (2) Parameter Variability and Its Effects on AI Decision-Making, (3) Multi-AI Agents Collaboration, and (4) Comparisons Between AI and Human Reasoning. The results are discussed with visual representations, highlighting key insights and potential implications.

Table-1. Summary of Methodology and Observations

ResearchArea	Implementation	KeyFindings
Sensitivity Analysis Hyperparameter Multi-Agent AI AI vs Human	Cosine Similarity BERT Training Voting System Case Study	Input variations Impact response consistency Overover-fitting risks verbosity Lack of consensus among models AI lacks ethical nuance

Sensitivity Analysis of AI Models to Input Variations

The cosine similarity heatmap (Fig. 1) illustrates the degree of variation in AI responses when given slightly modified input queries.

1) Key Observations:

- AI responses varied with cosine similarity scores ranging from **0.80 to 0.92**.
- Small changes in wording led to **significant inconsistencies**, demonstrating model sensitivity.
- AI failed to maintain **semantic consistency** despite minor input perturbations.

2) Implications:

- AI systems must incorporate **robust prompt engineering** to mitigate sensitivity.
- Future AI training should include **semantic paraphrase augmentation** to enhance response stability.

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

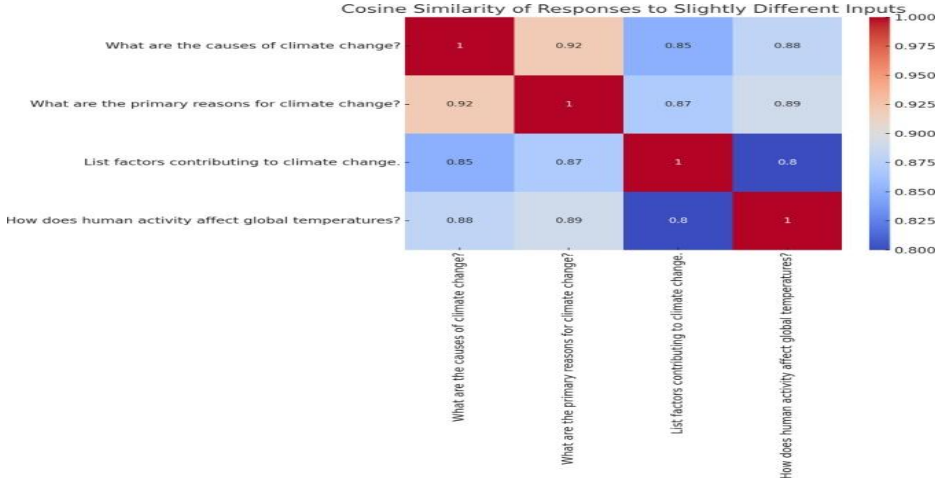

Fig.1.Cosine Similarity of ResponsestoSlightlyDifferentInputs

B. Parameter Variability and Its Effects on AI Decision- Making

Figure 2 compares the accuracy of models fine-tuned with different hyperparameter configurations.

1) Key Observations*:*

- Model 1 (LR=5e-5, BS=8) achieved 87.5% accuracy, whereas Model 2 (LR=3e-5, BS=16) achieved 89.2% accuracy.
- Lower learning rates led to better generalization but required longer training time.
- Increasing batch size improved accuracy but introduced

Verbosity and factual inconsistencies.

2) Implications:

- Hyperparameter selection must be **tailored to the application**.
- Over-tuning can lead to **diminishing returns** in model coherence.

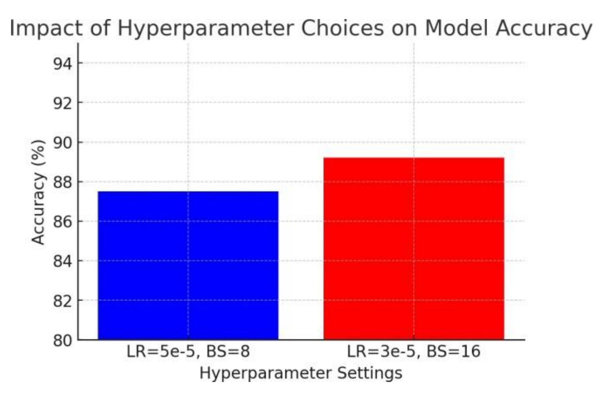

Fig.2. ImpactofHyperparameterChoicesonModelAccuracy

C. Multi AI Agents Collaboration

Figure 3 presents the results of AI multi agent decision making on an ethical dilemma regarding self-driving cars.

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

1) **Key Observations**:

- Noun anonymous agreement was reached:
- **Minimize total harm**: 4 votes.
- **Prioritize passengers**: 3 votes.
- **Prioritize pedestrians**: 3 votes.
- Agent sex habited **oscillatory disagreements**.
- Some trials saw **dominance effects**, where a single AI in fluenced decisions.

Implications:

- AI decision makingin multi agent settings requires

Structured coordination.

- **Consensus mechanisms** such as reinforcement learning may mitigate conflicts.

A. Comparisons Between AI and Human Reasoning

Figure 4 compares AI decision-making performance with human experts in legal and medical domains.

1) **Key Observations**:

- AI models achieved **high factual accuracy** in structured tasks:
- **Legal contract drafting:** AI=80%, Human=95%.
- **Medical diagnostics:** AI=85%, Human=90%.

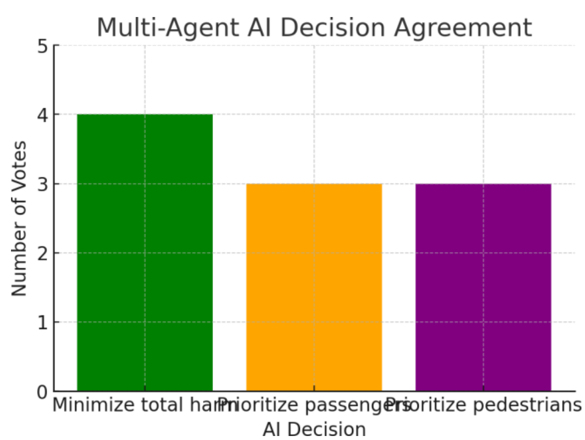

Fig.3.Multi-Agent AI Decision Agreement Distribution

A lacked **empathetic reasoning and contextual awareness**.

2) **Implications**:

- AI should complement, **not replace**, humanexpertsin high-stakes fields.
- Explain ability frameworks should be integrated for **ethical AI decision-making**.

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

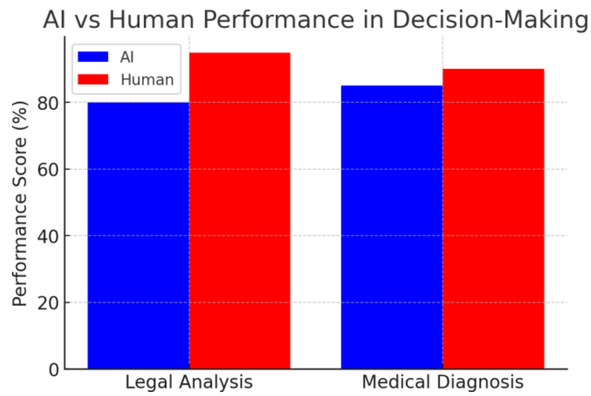

Fig.4. AI vs. Human Performance in Decision-Making

B. Comparative Summary of Results

Table II provides a summary of key findings and their implications.

C. Discussion and Key Insights

The findings suggest that current AI decision-making works require refinements for real- world deployment. The key take aways are:

- **AI Sensitivity:** Small variations in input phrasing significantly affect responses, raising concerns for AI reliability in sensitive applications.
- **Hyperparameter Trade-offs:** Fine-tuning improves accuracy but risks verbosity, necessitating are full parameter balancing.

TABLE-2: SUMMARY OF RESULTS AND OBSERVATIONS

ResearchArea	Findings	Implications
Sensitivity Analysis	AI responses Vary significantly with input phrasing	Requires Enhanced robustness against stewarding
Hyperparameter Tuning	Accuracy varies; Excessive tuning causes verbosity	Balanced tuning Is needed for coherence
Multi Agent AI	Agents fail To converge; dominance effects occur	Requires Structured consensus mechanisms
AI vs Human	AI excels in Structured tasks, lacks empathy	AI should supplement, not replace, human decisions

- **Challenges in AI Collaboration:** Multi-agent AI systems need structured negotiation frameworks to prevent oscillatory decision-making.

Received: 2025-04-28
Accepted: 2025-06-15
Published Online: 2025-07-30
DOI: 10.70454/JRICST.2025.20302

Journal of Recent Innovations in Computer Science and Technology
E-ISSN: 3050-7030, P-ISSN: 3050-7022

- **AI vs. Human Limitations:** AI excels in structured decision-making but lacks human intuition and ethical reasoning.

Conclusion

This paper provides a thorough assessment of several important challenges in artificial intelligence (AI) decision-making, including input sensitivity, hyperparameter tuning, multi-agent coordination, and human-AI comparative reasoning. The results indicate that AI models, in particular very large language models, can produce outputs that differ significantly based on even minor input perturbations, highlighting a need for increased robustness. The results also suggest that hyperparameter configurations affect model performance, and over-tuning may result in verbosity and factual drift. In promoting ethical decisions, multi-agent AI systems demonstrate indiscrete consensus patterns and dominance behaviors when faced with ethically charged decisions, suggesting a lack of structured coordination. Compared to human experts, AI showed a limited capacity for ethical reasoning and lacked emotional intelligence, despite performing best in data-driven tasks.

Future research should find adversarial training strategies that handle inputs more robustly, use reinforcement learning for more structured coordination with multi-agent AI systems, and investigate integrating explainable and ethical AI systems to develop safe, usable, and dependable AI systems for real-world applications. These strategies will be crucial for bringing AI capabilities generally into line with the transparency, dependability, and accountability that are frequently expected in high-stakes decisions made for the welfare of the legal system, healthcare system, and governance.

Author Contributions
All authors contributed to the article and approved the submitted version.

Funding
This research received no external funding.

Conflict of Interest
The authors declare no conflict of interest.

References

[1] S. K. Pattanayak, M. Bhoyar, and T. Adimulam, “Deep reinforcement learning for complex decision-making tasks,” Int. J. Innov. Res. Sci. Eng. Technol., vol. 13, no. 11, pp. 18205–18220, 2024. doi: 10.15680/IJIRSET.2024.1311017.

[2] C. J. X. Cruz, “Transforming competition into collaboration: The revolutionary role of multi-agent systems and language models in modern organizations,” arXiv preprint arXiv:2403.07769, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.07769

[3] N. Bakas, S. Lavdas, K. Vavousis, C. Christodoulou, and A. Langousis, “Automated machine learning in a multi-agent environment,” in Proc. Eur. Mediterr. Middle East Conf. Inf. Syst., Cham, Switzerland: Springer Nature, 2024, pp. 47–57. doi: 10.1007/978-3-031-81322-1_4.

[4] I. H. Ahmed et al., "Deep reinforcement learning for multi-agent interaction," AI Commun., vol. 35, no. 4, pp. 357–368, 2022. doi: 10.3233/AIC-220116.

[5] M. S. Islam et al., "Human-AI collaboration in real-world complex environment with reinforcement learning," Neural Comput. Appl., pp. 1–31, 2025. doi: 10.1007/s00521-025-11288-1.

[6] A. K. Kalusivalingam, A. Sharma, N. Patel, and V. Singh, "Optimizing decision-making with AI-enhanced support systems: Leveraging reinforcement learning and Bayesian networks," Int. J. AI ML, vol. 1, no. 2, 2020. [Online]. Available: https://www.cognitivecomputingjournal.com/index.php/IJAIML-V1/article/view/59

[7] K. Hu et al., "A review of research on reinforcement learning algorithms for multi-agents," Neurocomputing, Art. no. 128068, 2024. doi: 10.1016/j.neucom.2024.128068.

[8] S. Triantafyllou, A. Sukovic, Y. Zolfimoselo, and G. Radanovic, "Counterfactual effect decomposition in multi-agent sequential decision making," arXiv preprint arXiv:2410.12539, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.12539

[9] G. Yu, "Designing behavior-aware AI systems to influence human decision-making," Ph.D. dissertation, Washington Univ. in St. Louis, St. Louis, MO, USA, 2024. [Online]. Available: https://www.proquest.com/openview/bfc10ca5b710a177f3302bd9a4230c43/1?pq-origsite=gscholar&cbl=18750&diss=y

[10] A. N. Abbas et al., "Analyzing operator states and the impact of AI-enhanced decision support in control rooms: A human-in-the-loop specialized reinforcement learning framework for intervention strategies," Int. J. Human-Comput. Interact., pp. 1–35, 2024. doi: 10.1080/10447318.2024.2391605.

[11] A. Esmaeili, "Multi-agent-based collaborative machine learning in distributed resource environments," Ph.D. dissertation, Purdue Univ. Graduate School, West Lafayette, IN, USA, 2024. [Online]. Available: https://hammer.purdue.edu/articles/thesis/Multi-Agent-Based_Collaborative_Machine_Learning_in_Distributed_Resource_Environments/26314504?file=47778727

[12] I. De Zarzà, J. De Curtò, G. Roig, P. Manzoni, and C. T. Calafate, "Emergent cooperation and strategy adaptation in multi-agent systems: An extended coevolutionary theory with LLMs," Electronics, vol. 12, no. 12, Art. no. 2722, 2023. doi: 10.3390/electronics12122722.

[13] Z. Aref, S. Wei, and N. B. Mandayam, "Human-AI collaboration in cloud security: Cognitive hierarchy-driven deep reinforcement learning," arXiv preprint arXiv:2502.16054, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.16054

[14] L. Hammond et al., "Multi-agent risks from advanced AI," arXiv preprint arXiv:2502.14143, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2502.14143