CAN AI OUTPERFORM HUMAN MANAGERS IN OPERATIONAL DECISION-MAKING?

Jonathan Kuzmanko – Lucie Vrbová

Abstract

Large language models (LLMs) have emerged as a groundbreaking application of artificial intelligence, enabling more natural and sophisticated interactions between humans and AI systems. This pilot study compares the performance of human managers and LLMs in operational decision-making scenarios. We developed three scenarios, each representing common managerial challenges. Detailed solutions to these scenarios from four experienced first-line managers and three advanced LLMs (GPT-4, Claude-3 Opus, and Gemini Advanced) were evaluated by three expert raters using established frameworks for quality and creativity, resulting in 378 expert evaluations. Our findings revealed that LLMs demonstrated higher performance than human managers in both quality and creativity of decisions across all scenarios. While the study's sample size is limited, it provides valuable insights into the potential of AI in operational management. This research highlights the need for further investigation into human-AI collaboration in organizational decision-making processes. While the results suggest that LLMs have significant potential to augment and enhance managerial decision-making, they also underscore the importance of continued research to fully understand the implications and optimal integration of AI technologies in real-world organizational contexts.

Key words: AI, LLMs, Decision-Making, Operational Management

JEL Code: M10, O33, M15

Introduction

The rapid advancement of artificial intelligence (AI) has significantly transformed business operations and management. Large language models (LLMs) have emerged as a groundbreaking application of AI, enabling more natural and sophisticated interactions between humans and AI systems. LLMs like GPT-4, launched by OpenAI in 2023, demonstrate remarkable natural language capabilities with the potential to revolutionize decision-making processes (Kanbach et al., 2024). This reflects growing interest in exploring AI's impact on organizational processes and managerial roles.

Operational management relies heavily on first-line managers, who handle repetitive tasks requiring day-to-day oversight of employees and resources (Hales, 2005). AI promises to automate these tasks, optimize resource allocation, and enhance risk assessment. AI's role in operational management becomes more critical as it influences innovation, automating or augmenting decision-making (Gama & Magistretti, 2023), this could significantly improve the efficiency of operational decision-making into organizations. Moreover, by streamlining routine management activities like scheduling and resource allocation, AI frees managers to focus on higher-level strategic work, contributing to greater productivity. AI and human collaboration are crucial for strengthening organizational intelligence and adaptability (Kolbjørnsrud, 2023).

Our study focuses on understanding the practical implications of AI's use in real-world operational decision-making processes. We compare human decision-making with AI decision-making in specific scenarios, providing novel insights into AI's transformative potential within the business landscape.

This focused investigation aims to fill existing research gaps, offering theoretical contributions and practical applications. By conducting this study, we contribute to the growing body of knowledge on AI's influence on organizational processes and provide critical insights into the future of managerial roles in an AI-driven world.

1 Literature Review

AI is considered a critical technology that influences innovation capabilities, leading to augmentation or automation in decision-making. The integration of AI in operational management not only streamlines processes but also facilitates deeper analysis and better decision-making based on data-driven insights (Grover, Kar, & Dwivedi, 2022)

Several studies have demonstrated the potential advantages of AI in specific areas. AI's capacity to quickly handle large alternative sets and make decisions highlights its utility in high-velocity environments (Shrestha, Ben-Menahem, & Von Krogh, 2019). Moreover, using AI in decision-making can lead to significant replicability of decisions, ensuring consistency across similar situations, which is challenging to achieve with human decision-makers due to variability in human judgment (Shrestha et al., 2019). In the study by Eloundou, Manning, Mishkin, and Rock (2023) examining the impact of AI on the U.S. labor market, the researchers highlighted that, with the advent of LLM-powered software, between 47% - 56% of all tasks could be completed significantly faster at the same level of quality.

Research has also explored the creative potential of AI, comparing its performance to that of humans. A study by Stevenson, Smal, Baas, Grasman, and van der Maas (2022) found that while humans generally outperform AI in terms of originality and surprise, AI-generated responses were rated higher in utility, suggesting that AI can generate practical and applicable ideas, which are essential for effective decision-making in managerial roles.

AI's role in strategic decision-making includes providing insights that inform long-term planning and decision-making at the highest levels of an organization (Shrestha et al., 2019). However, the complexity of strategic decisions, often involving ambiguity and uncertainty, poses challenges for AI systems (Jarrahi, Lutz, & Newlands, 2022).

In contrast, operational management, particularly at the first-line manager level, involves a wide range of organizational activities and often faces repetitive, short-term tasks that require day-to-day management of activities and employees. A study by Hales (2005), which examined changes in first-line managers' (FLMs) roles, shows that the FLM role is characterized by narrow spans of control, vertical authority structures, and internal relationships. The authority to make decisions and participate in organizational strategy remains limited, primarily focusing on operational routines, including monitoring work performance, allocating tasks, and resolving immediate issues related to staffing or equipment. The application of generative AI in operational management promises to automate repetitive tasks, optimize resource allocation, and improve risk assessment, among other potential benefits (Barcaui & Monat, 2023).

Current research suggests that AI has the potential to outperform humans in specific tasks, particularly those that are repetitive, data-driven, and involve well-defined processes. However, there is a need for further investigation into the specific application of AI in operational managerial decision-making processes, considering the unique challenges and complexities of this domain. Several studies recommend future research directions, indicating a knowledge gap that requires in-depth examination (Shrestha et al., 2019).

Based on the reviewed literature and the identified research gap, the primary research question is: "Can AI exhibit better performance than humans in operational managerial decision-making processes?"

1.1 Research Hypotheses

Based on the literature review, we propose two hypotheses:

H1: LLMs score higher than human managers in operational decision-making scenarios in quality criteria.

H2: Human managers produce higher score responses than LLMs in operational decision-making scenarios at creativity criteria.

2 Methodology

2.1 Criteria

In order to compare the decision-making of AI and managers, two criteria were chosen to assess the decisions: creativity and quality. Creativity is crucial for generating innovative solutions to complex organizational challenges, while quality in decision-making is a critical determinant of organizational success (Keren & de Bruin, 2003; Mumford, Scott, Gaddis, & Strange, 2002). To evaluate these criteria, the study employs two frameworks. The PrOACT model (Hammond, Keeney, & Raiffa, 2015) provides a structured approach to enhance decision quality by encouraging comprehensive thinking and facilitating comparative analysis. Meanwhile, the Creative Product Analysis Matrix (CPAM) by Besemer & O'Quin (1999) evaluates creativity through novelty, resolution, and synthesis. Although initially designed to assess product creativity, the CPAM framework applies to managerial decision-making due to its adaptability and quantitative evaluation.

2.2 Scenarios

To ensure external validity, 3 scenarios were developed to address operational problems faced by first-line managers. They focus on topics of high business importance, as supported by previous studies and corroborated by second-line managers and experts. Scenario 1 presents a situation of high workload in a business team, leading to burnout and a decline in performance. Scenario 2 focuses on goal setting, where managers are asked to consider new objectives and how to set them for employees. Scenarios 3 deals with efficiency improvement, where the manager is asked to create an efficiency plan and meet the financial targets set for them.

2.3 Participants

2.3.1 Human Managers

4 experienced first-line managers working in service or operations were selected based on criteria such as managerial experience (> 3 years), tenure (> 3 years), number of employees managed (>5), and performance ratings over last three years (above-average) as reported in prior characteristics data interviews.

2.3.2 AI Models LLM's

3 advanced large language models (LLMs) were selected: GPT-4 (OpenAI), Claude-3 Opus (Anthropic), and Gemini Advanced (Google). Selection criteria included the models' popularity (>one million users), tier (highest paid), media coverage, and public availability via chatbots.

2.4 Data Collection Process

In each scenario, both human managers and LLMs were required to provide six comprehensive responses that examined their decision-making process in terms of both creativity and quality. For the quality criterion, participants were asked to demonstrate their understanding of the problem definition, clarification of objectives, generation of alternatives, understanding the implications, and selecting the best alternative according to the framework by Hammond et al. (1999). For the creativity aspect, the responses were evaluated based on the novelty they introduced, the resolution (the feasibility and effectiveness of the solution), and the answer synthesis, following the framework proposed by Besemer & O'Quin (1999). This method provided a total of 126 responses for analysis by the experts.

2.5 Sample and Evaluation Process

The primary sample for this study consists of 3 expert evaluators. We employed the expert evaluation method, selecting experts who met specific criteria: managerial experience exceeding 5 years, current positions as managers or consultants at partner or senior levels in strategic consulting firms, and substantial operational management consulting experience. This choice stems from the recognized importance of managerial experience in decision-making processes (Dane & Pratt, 2007) and the valuable insights possessed by experts with practical experience (Hoffman, Shadbolt, Burton, & Klein, 1995).

The experts assessed responses from all 7 participants (4 human managers and 3 LLMs) without knowing that some responses were generated by LLMs. The results were presented to the experts in a random order without indicating the source of each response. Using a structured assessment tool, the experts answered six key questions to evaluate decision-making quality and creativity for each response. They rated these responses on a Likert scale from 1-6, where 1 indicates the lowest evaluation and 6 the highest.

Accordingly, each expert answered 42 questions (6 questions multiply by 7 participants) for each scenario, resulting in a total of 126 responses per expert. In total, 378 expert evaluations were analyzed.

Results

2.6 Statistical measures

The research findings and data analysis were based on a comprehensive evaluation of 378 responses. Table 1 presents the number of responses collected from 3 experts, categorized by group (human, language model) and criterion (creativity, quality).

Tab. 1: Experts Responses Count

	Humans	LLMs	Total
Creativity	144	108	**252**
Quality	72	54	**126**
Total	**216**	**162**	**378**

Table 2 displays analyzed statistical data of 378 responses in total, distinguishing between the two groups according to the two criteria (experts' evaluations were provided on a Likert scale ranging from 1 to 6, where 1 indicates the lowest evaluation and six the highest).

Tab. 2: Statistical Measures for scenarios

	Humans			LLMs			Humans	LLMs
Mean	Sc1	Sc2	Sc3	Sc1	Sc2	Sc3	Total	Total
Creativity	2.98	2.94	2.33	4.36	4.33	3.92	2.75	4.20
Quality	3.21	2.92	2.58	4.72	4.28	4.22	2.90	4.41
StDev								
Creativity	1.51	1.21	1.23	1.20	1.01	1.66	1.35	1.32
Quality	1.25	1.21	1.02	0.89	1.23	1.63	1.18	1.28
Median								
Creativity	3.00	3.00	2.00	4.00	4.00	4.00	3.00	4.00
Quality	3.50	3.00	2.00	5.00	5.00	5.00	3.00	5.00

Sc=Scenario

The creativity score for LLMs (M = 4.20, SD = 1.32) was significantly higher than that of human managers (M = 2.75, SD = 1.35), t(233) = 8.51, p < .0001. In terms of quality, LLMs exhibited even more significant outperformance (M = 4.41, SD = 1.28) compared to human managers (M = 2.90, SD = 1.18), t(109) = 6.85, p < .0001. The results are consistent across all three scenarios (presented in columns CS1, CS2, CS3).

2.7 Examining agreement among experts

A calculation of Kendall's W coefficient was performed to ensure agreement among experts on the given ratings. A strong consensus among experts was observed regarding both criteria. For the quality criterion, the W value was 0.89, while for the creativity criterion, the value was 0.94. The overall W value, without separating by criteria, was 0.93, indicating substantial agreement among the experts.

3 Discussion

3.1 Interpretation of Results

While we initially assessed that LLMs would achieve better results in terms of quality but may not be as good in creativity, the findings indicate that LLMs managed to achieve superior results in both criteria. This highlights significant differences in both creativity and decision quality between LLMs and human managers. Although these results are promising, they should be interpreted cautiously due to the study's limitations. This preliminary study can serve as a foundation for further research and broader investigations to validate these initial findings on a larger scale..

3.2 Comparison with Previous Research

Our research presents a different picture from previous studies (Stevenson et al., 2021) that suggested humans have an advantage in creativity. This discrepancy might be attributed to the rapid advancements in large language models over the past three years, particularly our focus on specialized versions released in 2024. The apparent advantage of LLMs in managerial decision-making processes at lower managerial levels aligns with studies demonstrating LLM capabilities in well-defined, data-based, and repetitive tasks under conditions of certainty (Barcaui & Monat, 2023; Grover et al., 2022; Shrestha et al., 2019). As our literature review indicated, first-line managers' work often involves such tasks, which may contribute to the observed performance differences.

3.3 Limitations and Methodological Considerations

It is crucial to acknowledge the study's limitations. The small sample size of human managers, while consisting of experienced professionals, limits the generalizability of our findings. Although our scenarios were developed with input from second-line managers and experts to represent real-life scenarios, their implementation remains theoretical and untested in practice.

The study does not account for the execution phase of decisions, where factors such as intuition, emotional intelligence, and adaptability play significant roles.

To enhance the robustness of future studies, several methodological improvements could be implemented. Expanding the sample size of human managers would increase the statistical power and generalizability of the results. Incorporating a wider range of scenarios, including edge scenarios, could provide a more comprehensive evaluation of decision-making processes. Involving a larger pool of experts in the evaluation process could enhance the validity of the assessments. Additionally, longitudinal studies could track the consistency and evolution of LLM performance over time, particularly given the rapid advancement of AI technology.

3.4 Implications and Future Research Directions

These findings suggest that the potential advantage of LLMs may lie in their capacity to assist managers in making better decisions, rather than replacing human managers entirely. This research focused on specific critical decision-making processes, and further studies with larger samples and more diverse scenarios are crucial to validate these findings and explore their implications fully.

As this is a rapidly evolving field, ongoing research is necessary to keep pace with technological advancements. This study can serve as a point of comparison for future research, acknowledging its limitations while providing a foundation for more comprehensive investigations.

Conclusion

This research offers an intriguing perspective on the potential integration of large language models (LLMs) in the operational decision-making processes of first-line managers. While our findings suggest performance differences between AI and humans in terms of creativity and decision quality, we emphasize the need for cautious interpretation due to methodological limitations. The study's primary contribution lies in its potential to serve as a starting point for broader investigations, providing a foundation for future research in this rapidly evolving field. Rather than positioning LLMs as replacements for human managers, our results point towards the potential advantage of integrating AI as an assistive tool in decision-making processes. This emerging area of AI application in managerial decision-making offers fertile ground for further research and the potential improvement of management practices. Future studies should focus on addressing the limitations identified in this study, exploring a wider range of operational scenarios, and employing larger samples to provide a more comprehensive understanding of AI's potential in managerial decision-making.

Acknowledgment

During the preparation of this work, tools such as "Machine Translation" and "Grammarly" were used for translation, grammatical improvement, and writing corrections. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

Barcaui, A., & Monat, A. (2023). Who is better in project planning? Generative artificial intelligence or project managers? *Project Leadership and Society, 4(100101).*

Besemer, S. P., & O'Quin, K. (1999). Confirming the three-factor creative product analysis matrix model in an American sample. *Creativity Research Journal, 12*(4), 287-296.

Dane, E., & Pratt, M. G. (2007). Exploring intuition and its role in managerial decision making. *Academy of Management Review, 32*(1), 33-54.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Gama, F., & Magistretti, S. (2023). Artificial intelligence in innovation management: A review of innovation capabilities and a taxonomy of AI applications. *Journal of Product Innovation Management*. 1-36.

Grover, P., Kar, A. K., & Dwivedi, Y. K. (2022). Understanding artificial intelligence adoption in operations management: insights from the review of academic literature and social media discussions. *Annals of Operations Research, 308*(1), 177-213.

Hales, C. (2005). Rooted in supervision, branching into management: Continuity and change in the role of first-line manager. *Journal of Management Studies, 42*(3), 471-506.

Hammond, J. S., Keeney, R. L., & Raiffa, H. (2015). *Smart choices: A practical guide to making better decisions*. Harvard Business Review Press.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62*(2), 129-158.

Jarrahi, M. H., Lutz, C., & Newlands, G. (2022). Artificial intelligence, human intelligence, and hybrid intelligence based on mutual augmentation. *Big Data & Society 9*(2).

Kanbach, D. K., Heiduk, L., Blüher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science, 18,* 1189–1220.

Kolbjørnsrud, V. (2024). Designing the Intelligent Organization: Six Principles for Human-AI Collaboration. *California Management Review, 66*(2), 44-64.

Mumford, M. D., Scott, G. M., Gaddis, B., & Strange, J. M. (2002). Leading creative people: Orchestrating expertise and relationships. *The Leadership Quarterly, 13*(6), 705-750.

Shrestha, Y. R., Ben-Menahem, S. M., & Von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review 61*(4), 66-83.

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022). Putting GPT-3's creativity to the (alternative uses) test. *arXiv preprint arXiv:2206.08932*.

Contacts

Jonathan Kuzmanko
Prague University of Economics and Business
W. Churchill Sq. 1938/4, Prahue 3, Czech Republic
Kuzj01@vse.cz

Lucie Vrbová
Prague University of Economics and Business
W. Churchill Sq. 1938/4, Prahue 3, Czech Republic
lucie.vrbova@vse.cz