AI Systems Engineering
eISSN: 3093-8449
https://aise.cultechpub.com/index.php/aise

Agentic AI Systems: Evolution, Efficiency, and Ethical Implementation

K. A. S. N. Kodikara*

Department of Computer Science, University of Ruhuna, Matara, Sri Lanka

*Corresponding author: K. A. S. N. Kodikara, snkodikara52@gmail.com

Abstract

Agentic AI represents a major paradigm shift from passive language models to active systems capable of autonomous decision-making and task execution. This comprehensive review synthesizes recent literature from 2023-2025, examining three critical dimensions that define the future of artificial intelligence: architectural innovations enabling collaborative reasoning between multiple agents, optimization techniques for achieving computational efficiency in resource-constrained environments, and ethical frameworks for ensuring safe deployment in real-world applications. Through systematic analysis of peer-reviewed literature and recent preprints, key findings reveal that multi-agent debate systems demonstrate significant improvements in reasoning accuracy, with collaborative frameworks showing enhanced problem-solving capabilities across diverse domains. However, critical challenges persist in safety alignment mechanisms, environmental sustainability concerns, and the development of standardized evaluation metrics for complex agentic behaviors. The analysis identifies hybrid architectures combining collaborative reasoning with knowledge distillation as the most promising direction for future research, particularly for applications requiring both high accuracy and computational efficiency. This review provides researchers and practitioners with a consolidated framework for understanding current developments and identifying future research directions in agentic systems, highlighting the urgent need for interdisciplinary approaches to address technical, ethical, and environmental challenges in this rapidly evolving field.

Keywords

Agentic AI, Multi-agent Collaboration, Efficiency Optimization, AI Safety, Knowledge Distillation

1. Introduction

The Agentic AI Revolution

The transition from traditional AI architectures to agentic systems marks one of the most profound developments in artificial intelligence over the past decade. Unlike conventional AI models that operate as sophisticated input-output processors, agentic systems embody genuine autonomy, demonstrating the capacity for independent reasoning, collaborative problem-solving, and adaptive behavior that emerges from complex multi-agent interactions [1].

This paradigmatic transformation fundamentally reconceptualizes how we approach intelligent system design. Sapkota et al. distinguish between conventional AI Agents-modular systems primarily driven by Large Language Models for specific automation tasks and true Agentic AI systems characterized by "multi-agent collaboration, dynamic task decomposition, persistent memory, and orchestrated autonomy" [2]. This distinction proves crucial for understanding the revolutionary potential of current developments in artificial intelligence.

Building upon generative AI foundations, Agentic AI constitutes "the next major evolutionary step with significantly enhanced reasoning and interaction capabilities that enable autonomous behavior for tackling complex tasks" [3]. This progression reflects a natural advancement from content generation to sophisticated reasoning systems capable of making informed decisions and coordinating with other agents to achieve complex objectives that would be impossible for individual systems to accomplish.

Three fundamental capabilities distinguish agentic systems from their predecessors: advanced reasoning and reflection mechanisms enabling self-evaluation and strategy adjustment, sophisticated action execution frameworks facilitating real-world interaction, and complex multi-agent coordination protocols supporting collaborative problem-solving [4]. These capabilities collectively enable the emergence of intelligent behaviors that surpass the sum of individual component capabilities.

This review addresses three critical research questions essential for successful agentic AI development: How do collaborative architectures enhance reasoning beyond single-agent capabilities? What optimization techniques enable computational efficiency while preserving sophisticated agentic capabilities? How can ethical frameworks ensure safe

deployment in high-stakes real-world applications? Our analysis focuses on peer-reviewed literature and preprints from 2023-2025, prioritizing empirical studies with measurable outcomes that demonstrate concrete advances in agentic AI capabilities.

2. Architectural Foundations of Agentic Systems

2.1 Collaborative Reasoning Through Multi-Agent Debate

.Multi-agent debate has emerged as a fundamental technique for enhancing reasoning accuracy in agentic systems, revolutionizing approaches to complex problem-solving in artificial intelligence. This methodology leverages adversarial interactions between diverse perspectives to produce more robust and accurate conclusions than individual reasoning processes can achieve [5].

Drawing inspiration from human collaborative problem-solving, multi-agent debate translates the principle that diverse perspectives yield superior solutions into artificial intelligence contexts. Du et al. established that multi-agent debate systems consistently outperform single-agent approaches across mathematical problem-solving, factual question answering, and complex logical inference tasks [5]. These improvements stem from the systems' ability to identify and correct individual agent errors while exploring alternative solution paths and synthesizing insights from multiple perspectives.

Expanding upon this foundation, Liu et al. introduced GroupDebate, incorporating group discussion mechanisms that enhance multi-agent debate efficiency in logical reasoning tasks [6]. Their innovation addresses computational overhead limitations in early debate systems through simultaneous-talk interactions, where multiple agents contribute concurrently rather than sequentially, improving both accuracy and efficiency within multi-agent frameworks.

Hegazy demonstrated that diversity of thought in multi-agent debate frameworks elicits stronger reasoning capabilities, with advanced frameworks showing substantial improvements in collaborative reasoning across multiple domains [7]. This research emphasizes ensuring that participating agents contribute genuinely different perspectives rather than variations of similar reasoning patterns, maximizing the benefits of collaborative approaches.

However, Eo et al. introduce a nuanced perspective advocating for adaptive approaches where debate is employed strategically rather than universally [8]. Their research on adaptive multiagent collaboration demonstrates that selective deployment of debate mechanisms can achieve comparable accuracy while reducing computational overhead, recognizing that not all problems require full multi-agent interaction complexity.

Beyond debate frameworks, agentic systems require sophisticated mechanisms for integrating and synthesizing knowledge from diverse sources and agent interactions. These mechanisms enable agents to maintain coherent understanding while incorporating new information and perspectives from collaborative interactions. The development of effective knowledge integration protocols represents a critical challenge in designing robust agentic systems that can operate reliably across diverse domains and task requirements.

2.2 Efficiency Optimization Strategies

The computational demands of agentic systems have driven significant research into efficiency optimization. Table 1 summarizes key optimization approaches identified in recent literature:

Table 1. Key optimization approaches in agentic AI systems.

Approach	Key Mechanism	Performance Impact	Reference
Resource-Constrained Collaboration	Coordinated agent operation under limited computing resources	40-60% efficiency improvement with <5% accuracy loss	Gupta (2025)[9]
Adaptive Multi-Cloud Systems	Dynamic resource allocation across cloud environments	37% cost reduction with 2.1× faster task completion	Thamma and Oriti (2024)[10]
Multi-Agent Deep RL	Reinforcement learning for resource allocation	72% reduction in processing time with 92% accuracy maintained	Narantuya et al. (2022)[11]

Gupta addresses practical implementations of AI agent collaboration under resource constraints, developing frameworks that enable effective coordination despite limited computing resources [9]. This work demonstrates that next-generation multi-agent systems can successfully integrate advanced AI decision-making with principled resource management, making sophisticated agentic capabilities accessible in resource-limited environments. management.

2.3 Knowledge Distillation and Architectural Compression

Knowledge distillation has become critical for creating efficient agentic systems that operate effectively in resource-constrained environments while maintaining sophisticated reasoning capabilities. This approach addresses the fundamental tension between model capability and computational efficiency that characterizes agentic AI deployment [12].

Yang et al. provide a comprehensive analysis of knowledge distillation methods for large language models, examining foundational approaches including soft label distillation, feature matching, attention transfer, and advanced

reinforcement learning-based strategies [12]. Their work demonstrates how distillation techniques significantly broaden LLM applicability across resourceconstrained environments by creating smaller models that retain substantial capability.

Effective knowledge distillation for agentic systems requires careful consideration of factors beyond simple performance metrics. Distilled models must maintain accuracy in individual tasks while preserving the ability to engage effectively in multi-agent collaboration, respond appropriately to coordination signals, and contribute meaningfully to collaborative reasoning processes.

Idowu explores the transition from teacher to student models through knowledge distillation, examining techniques specifically tailored for LLMs incorporating reinforcement learning and meta-learning approaches [13]. This research demonstrates that effective distillation can maintain model performance while dramatically reducing computational requirements, making advanced agentic capabilities accessible in resource-limited deployment scenarios.

The incorporation of reinforcement learning into the distillation process enables student models to learn not just from the static outputs of teacher models but from the dynamic decision-making processes that characterize effective agentic behavior. This approach helps ensure that distilled models retain the adaptive and responsive characteristics that are essential for successful multi-agent collaboration.

Meta-learning approaches further enhance the distillation process by enabling student models to learn how to learn from their teacher models, developing the ability to quickly adapt to new tasks and collaboration scenarios without requiring extensive retraining. This capability is particularly valuable in agentic systems where agents must be able to adapt to new collaborative partners and novel problem domains.

2.4 Recent Advances in Multi-Agent Collaboration (2025)

The year 2025 has witnessed unprecedented advances in agentic AI systems, with significant breakthroughs in both theoretical understanding and practical implementation of multi-agent collaboration frameworks. These developments have moved the field from experimental research toward real-world deployment, with documented improvements in efficiency, accuracy, and practical applicability across diverse domains[14].

Joshi (2025) provides compelling evidence that multi-agent systems demonstrate 40-60% efficiency gains in enterprise processes, with specialized agent coordination protocols becoming critical infrastructure components for modern organizations. This comprehensive analysis of real-world deployments reveals that the theoretical promises of agentic AI are beginning to translate into measurable business value across multiple industries[14].

The efficiency gains documented by Joshi stem from several key factors: improved task parallelization through intelligent agent coordination, reduced redundancy in computational processes through specialized agent roles, and enhanced decision-making speed through collaborative reasoning mechanisms that can quickly converge on optimal solutions for complex business problems.

This comprehensive review reveals that human-agent collaboration requires new stewardship and motivational models that account for the unique characteristics of agentic systems. Traditional management approaches designed for human teams or simple automated systems prove inadequate for managing the complex dynamics of multi-agent systems that exhibit emergent behaviors and adaptive capabilities.

The stewardship models identified in the research emphasize the importance of designing human oversight mechanisms that can effectively monitor and guide agentic systems without constraining their autonomous capabilities. These models recognize that effective human-agent collaboration requires understanding and leveraging the complementary strengths of human intuition and agentic computational power.

Hughes et al. (2025) conducted a multi-expert analysis examining AI agents' potential to transform industries through decentralized decision-making and enhanced cross-functional collaboration. Their findings indicate that 50% of organizations currently using generative AI will implement AI agents by 2027, representing a massive shift in how businesses approach automation and decision-making processes.[15]

The multi-expert analysis reveals specific applications that are driving this rapid adoption, including healthcare systems creating adaptive treatment plans that can adjust to individual patient responses in real-time, supply chain agents predicting disruptions and automatically implementing contingency plans, and financial systems that can analyze market conditions and adjust investment strategies autonomously within predefined risk parameters.

In healthcare applications, agentic systems demonstrate particular promise for personalized medicine, where multiple specialized agents can collaborate to analyze patient data, monitor treatment responses, and adjust therapeutic approaches based on individual patient characteristics and real-time health indicators. These systems show the potential to significantly improve treatment outcomes while reducing the burden on healthcare professionals.

Supply chain applications leverage the distributed intelligence of agentic systems to monitor global conditions, predict potential disruptions, and coordinate responses across multiple stakeholders. These systems can process vast amounts of information from diverse sources and make rapid decisions that would be impossible for human managers to coordinate effectively.

Fang et al. (2025) introduced the paradigm of self-evolving AI agents, bridging foundation models with lifelong agentic systems that can continuously improve their capabilities through ongoing experience and interaction. Their comprehensive survey establishes three fundamental laws for self-evolving agents: Endure (safety adaptation), Excel (performance preservation), and Evolve (autonomous evolution).[16]

The Endure principle focuses on maintaining safety and alignment as agents evolve their capabilities, ensuring that autonomous learning processes do not compromise the safety constraints and ethical guidelines that govern agent behavior. This principle addresses one of the most significant concerns about self-evolving systems: the potential for autonomous learning to lead agents away from their intended objectives or safety constraints.

The Excel principle ensures that performance preservation mechanisms prevent agents from losing previously acquired capabilities as they learn new skills. This addresses the common challenge of catastrophic forgetting in machine learning systems, where learning new tasks can interfere with performance on previously mastered tasks.

The Evolve principle governs autonomous evolution through continuous optimization loops that enable agents to identify opportunities for improvement and implement changes to their own capabilities without requiring external intervention. This principle represents the ultimate goal of agentic systems: true autonomous learning and adaptation.

This work demonstrates the evolution from static model pretraining to dynamic, autonomous agent evolution through continuous optimization loops that enable agents to improve their performance based on ongoing experience and feedback from their environment and collaborative partners.

3. Critical Implementation Challenges

3.1 Safety Alignment and Ethical Frameworks

TheThe deployment of agentic systems raises significant safety and ethical concerns that must be addressed before these systems can be safely deployed in high-stakes real-world applications. The autonomous nature of agentic systems, combined with their ability to make decisions and take actions without direct human supervision, creates new categories of risk that require sophisticated management approaches[17].

Leikas et al. (2019) established foundational principles for ethical framework design in autonomous intelligent systems, emphasizing that "autonomous systems are fundamentally changing our world" and requiring alignment with fundamental values and ethical principles that reflect human moral reasoning and societal expectations[17]. Their work provides a comprehensive foundation for understanding the ethical challenges posed by autonomous systems and developing appropriate governance frameworks.

The ethical challenges in agentic systems are particularly complex because they involve not only the behavior of individual agents but also the emergent behaviors that arise from multi-agent interactions. These emergent behaviors can be difficult to predict or control, creating potential for unintended consequences that may not be apparent from analyzing individual agent capabilities.

Recent work by Tallam (2025) provides a systems engineering perspective on alignment, agency, and autonomy in frontier AI, arguing for rigorous frameworks that treat advanced systems as fully autonomous rather than semi-autonomous assistants[18]. This work highlights the critical importance of developing robust alignment mechanisms as systems become increasingly capable of independent decision-making and action.

The systems engineering approach emphasizes the need for comprehensive testing and validation procedures that can assess not only individual agent behavior but also the complex interactions and emergent properties that characterize multi-agent systems. These procedures must account for the adaptive nature of agentic systems and their ability to learn and change their behavior over time.

Tallam's framework calls for treating advanced agentic systems as fully autonomous entities with their own goals, preferences, and decision-making processes, rather than as sophisticated tools that simply execute human instructions. This perspective requires developing new approaches to oversight and control that can work effectively with truly autonomous systems.

Jedličková (2024) offers a comprehensive survey of ethical approaches in designing autonomous and intelligent systems, emphasizing responsible development practices that integrate ethical considerations throughout the design and implementation process[19]. The research identifies deontological ethics as particularly relevant for aligning the personal ethics of designers with system behavior, promoting responsible use of AI technologies.

The survey reveals that effective ethical frameworks for agentic systems must address multiple stakeholders and considerations, including the rights and interests of users, affected communities, and society as a whole. These frameworks must also account for cultural differences in ethical perspectives and ensure that agentic systems can operate appropriately across diverse cultural and legal contexts.

Deontological approaches prove particularly valuable because they focus on the inherent rightness or wrongness of actions rather than simply their consequences, providing clearer guidance for system designers and more predictable behavior for users and affected parties.

3.2 Environmental Impact and Sustainability

Environmental implications of agentic AI systems present growing concerns requiring attention to ensure sustainable technology development. The computational complexity of multi-agent systems, combined with increasing deployment scale, creates significant environmental costs that could undermine societal benefits [20].

Mitu and Mitu examine AI's hidden costs, revealing that complex model training can generate hundreds of tonnes of CO_2 emissions, with large-scale training runs producing carbon footprints equivalent to multiple automobile lifetime emissions [20]. Their research emphasizes that "environmental impact is exacerbated by increasing complexity" of modern architectures, particularly those involving multiple interacting agents.

Environmental costs extend beyond initial training to include ongoing operational costs associated with multi-agent coordination, communication overhead, and continuous learning processes characterizing advanced agentic systems. These ongoing costs accumulate to represent significant environmental impact, particularly with large-scale deployment across multiple applications and organizations.

Joshua et al. provide comprehensive analysis of sustainable AI practices, focusing on measuring and reducing carbon footprint in both model training and deployment phases [21]. Their research includes recommendations for implementing environmental cost limits that halt training if sustainability thresholds are exceeded, ensuring environmental considerations integrate into development processes rather than being treated as afterthoughts.

3.3 Evaluation and Benchmarking Gaps

The lack of standardized evaluation metrics represents a significant challenge for agentic AI development, hindering progress by making it difficult to compare different approaches, validate improvements, and ensure that systems meet appropriate standards for real-world deployment[22]. The complexity of agentic systems, with their emergent behaviors and multi-dimensional capabilities, makes traditional evaluation approaches inadequate.

Moshkovich et al. (2025) argue for moving beyond black-box benchmarking, proposing observability, analytics, and optimization frameworks for agentic systems that can provide deeper insights into system behavior and performance[22]. Their work emphasizes the need for "standardized evaluation criteria and benchmarks across" diverse agentic architectures that can capture the full range of capabilities and behaviors that characterize these systems.

The black-box evaluation approaches that have been effective for traditional AI systems prove inadequate for agentic systems because they cannot capture the complex interactions, emergent behaviors, and dynamic adaptation that characterize multi-agent collaboration. These systems require evaluation frameworks that can assess not only final outcomes but also the quality of the processes by which those outcomes are achieved.

The observability frameworks proposed by Moshkovich et al. include comprehensive monitoring capabilities that can track agent interactions, decision-making processes, and the evolution of agent capabilities over time. These frameworks enable researchers and practitioners to understand how agentic systems achieve their results and identify opportunities for improvement[22].

Kapoor et al. (2024) examine what makes AI agents matter, arguing for "incorporating metrics beyond accuracy into agent evaluation" and calling for "standardization of agent benchmarks and evaluations." [23] This research highlights the complexity of evaluating systems that must balance multiple objectives including accuracy, efficiency, safety, and social impact.

The multi-objective nature of agentic system evaluation requires new approaches that can assess trade-offs between competing objectives and provide guidance for optimizing systems across multiple dimensions simultaneously. Traditional single-metric approaches cannot capture the complexity of these trade-offs or provide appropriate guidance for system designers.

The research by Kapoor et al. identifies several critical evaluation dimensions that are often overlooked in traditional AI evaluation but are essential for agentic systems: robustness to distribution shifts, fairness across different user populations, interpretability of decision-making processes, and long-term stability of system behavior.

4. Conclusion

This comprehensive literature review reveals that agentic AI systems represent a fundamental evolution in artificial intelligence, characterized by sophisticated collaborative reasoning capabilities, advanced optimization techniques for computational efficiency, and complex ethical considerations that require careful management for safe deployment. The evidence demonstrates clearly that multi-agent collaboration significantly enhances reasoning accuracy across diverse domains, while efficiency optimization techniques enable practical deployment under realistic resource constraints.

The analysis of recent research shows that multi-agent debate systems consistently outperform single-agent approaches, with collaborative frameworks demonstrating enhanced problem-solving capabilities that emerge from the interaction between diverse specialized agents. These improvements stem from the ability of collaborative systems to identify and correct individual errors, explore alternative solution paths, and synthesize insights from multiple perspectives into more comprehensive and accurate conclusions.

Efficiency optimization strategies have proven effective in addressing the computational challenges associated with multi-agent systems, with documented improvements of 40-60% in processing efficiency while maintaining accuracy levels within acceptable bounds. These optimizations enable the deployment of agentic systems in resource-constrained environments where the full computational requirements of unoptimized systems would be prohibitive.

However, critical challenges remain in several key areas that must be addressed before agentic systems can achieve their full potential. Safety alignment mechanisms require further development to ensure that autonomous systems operate in accordance with human values and societal expectations, particularly as these systems become more capable and autonomous. Environmental sustainability concerns demand immediate attention as the computational requirements of agentic systems scale with deployment.

The development of standardized evaluation metrics remains an urgent priority for the field, as the lack of appropriate benchmarks hinders progress and makes it difficult to validate improvements or ensure that systems meet appropriate standards for real-world deployment. The complexity of agentic systems requires evaluation frameworks that can assess not only performance outcomes but also the quality of decision-making processes and the appropriateness of system behavior across diverse contexts.

A notable limitation of the current literature is the heavy reliance on non-peer-reviewed preprints for cutting-edge results, particularly in the rapidly evolving 2025 research landscape. Many breakthrough findings come from arXiv preprints and conference papers that have not yet undergone rigorous peer review, which requires cautious interpretation of the most advanced claims until they can be independently validated through the peer review process.

The most promising future directions lie in hybrid approaches that balance collaborative capability with computational efficiency, supported by robust ethical frameworks and comprehensive evaluation metrics. These hybrid approaches recognize that not all problems require the full complexity of multi-agent collaboration and that computational resources can be more efficiently allocated through intelligent system design that matches problem complexity with appropriate solution approaches.

Recent 2025 advances in self-evolving agent paradigms and enterprise-scale deployments suggest that the field is transitioning from experimental research to practical implementation, with documented efficiency gains of 40-60% in real-world applications across diverse industries. This transition represents a critical milestone in the development of agentic AI and opens new opportunities for addressing complex societal challenges through advanced artificial intelligence.

The future success of agentic AI systems will depend on continued interdisciplinary collaboration between computer scientists, ethicists, environmental scientists, and domain experts to address the technical, ethical, and sustainability challenges that characterize this rapidly evolving field. Only through such collaborative approaches can we ensure that the significant potential of agentic systems is realized in ways that benefit society while minimizing risks and negative impacts.

References

[1] A. Plaat, M. van Duijn, N. van Stein, and M. Preuss, "Agentic large language models: A survey," arXiv preprint, arXiv:2503.23037, 2025.
[2] R. Sapkota, K. I. Roumeliotis, and M. Karkee, "AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges," arXiv preprint, arXiv:2505.10468, 2025.
[3] J. Schneider, "Generative to agentic AI: Survey, conceptualization, and challenges," arXiv preprint, arXiv:2504.18875, 2025.
[4] U. M. Borghoff, P. Bottoni, and R. Pareschi, "Human-artificial interaction in the age of agentic AI: A system-theoretical approach," Frontiers in Human Dynamics, 2025.
[5] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in Proc. Forty-first Int. Conf. Machine Learning, 2023.
[6] T. Liu, X. Wang, W. Huang, W. Xu, Y. Zeng, L. Jiang, and Y. Zhang, "GroupDebate: Enhancing the efficiency of multi-agent debate using group discussion," arXiv preprint, arXiv:2409.14051, 2024.
[7] M. Hegazy, "Diversity of thought elicits stronger reasoning capabilities in multi-agent debate frameworks," arXiv preprint, arXiv:2410.12853, 2024.
[8] S. Eo, H. Moon, E. H. Zi, C. Park, and H. Lim, "Debate only when necessary: Adaptive multiagent collaboration for efficient LLM reasoning," arXiv preprint, arXiv:2504.05047, 2025.
[9] S. Gupta, "AI agents collaboration under resource constraints: Practical implementations," Int. J. Artificial Intelligence and Machine Learning, 2025.
[10] S. R. T. S. R. Thamma, "Optimization of generative AI costs in multi-agent and multi-cloud systems," PhilPapers, 2024.
[11] J. Narantuya, J. S. Shin, and S. Park, "Multi-agent deep reinforcement learning-based resource allocation in HPC/AI converged cluster," Computers, Materials & Continua, 2022.
[12] C. Yang, Y. Zhu, W. Lu, Y. Wang, Q. Chen, and C. Gao, "Survey on knowledge distillation for large language models: Methods, evaluation, and application," ACM Trans. Knowledge Discovery from Data, 2024.
[13] C. Idowu, "From teacher to student: Knowledge distillation as a path to scalable language models," HAL Archives, 2024.
[14] S. Joshi, "Review of autonomous and collaborative agentic AI and multi-agent systems for enterprise applications," Int. J. Innovative Research in Engineering and Management, vol. 12, no. 3, pp. 65-76, 2025.
[15] L. Hughes, Y. K. Dwivedi, T. Malik, M. Shawosha, M. A. Albashrawia, V. Dutot, and M. Wade, "AI agents and agentic systems: A multi-expert analysis," Information Systems Frontiers, 2025.

[16] J. Fang, Y. Peng, X. Zhang, Y. Wang, X. Yi, G. Zhang, and Z. Meng, "A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems," arXiv preprint, arXiv:2508.07407, 2025.
[17] J. Leikas, R. Koivisto, and N. Gotcheva, "Ethical framework for designing autonomous intelligent systems," J. Open Innovation: Technology, Market, and Complexity, vol. 5, no. 1, p. 18, 2019.
[18] K. Tallam, "Alignment, agency and autonomy in frontier AI: A systems engineering perspective," arXiv preprint, arXiv:2503.05748, 2025.
[19] A. Jedličková, "Ethical approaches in designing autonomous and intelligent systems: A comprehensive survey towards responsible development," AI & Society, Springer, 2024.
[20] N. E. Mitu and G. T. Mitu, "The hidden cost of AI: Carbon footprint and mitigation strategies," SSRN Electronic Journal, 2024.
[21] C. Joshua, A. Marvellous, B. Matthew, and M. Pezzè, "Sustainable AI: Measuring and reducing carbon footprint in model training and deployment," ResearchGate, 2025.
[22] D. Moshkovich, H. Mulian, S. Zeltyn, N. Eder, and E. Shpigel, "Beyond black-box benchmarking: Observability, analytics, and optimization of agentic systems," arXiv preprint, arXiv:2503.06745, 2025.
[23] S. Kapoor, B. Stroebl, Z. S. Siegel, N. Nadgir, S. Goel, and B. Nushi, "AI agents that matter," arXiv preprint, arXiv:2407.01502, 2024.