Expert Systems With Applications 291 (2025) 128404

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0

Francesco Piccialli *, Diletta Chiaro , Sundas Sarwar , Donato Cerciello , Pian Qi , Valeria Mele

University of Naples Federico II , Department of Mathematics and Applications "R. Caccioppoli", Mathematical mOdelling and Data AnaLysis (M.O.D.A.L.) Research Group, Via Cintia, Naples, 80126, Italy

ARTICLE INFO

Keywords:
AgentAI
AgenticAI
Industry 4.0
Distributed artificial intelligence
Autonomous decision-making

ABSTRACT

AgentAI represents a transformative approach within distributed Artificial Intelligence (AI) in which autonomous agents work either individually or collaboratively in decentralized environments to address challenging problems. AgentAI enhances scalability, robustness, and flexibility by utilizing advanced communication, learning, and decision-making capabilities, making it integral to diverse applications in Industry 4.0. The ability of AI systems to interpret sensory data in open-world environments has seen significant advancements in recent years. This progress emphasizes the need to move beyond reductionist approaches and embrace more embodied and cohesive systems, which integrate foundational models into agent-driven actions. Existing surveys often focus on isolated domains or specific autonomy levels, lacking a cohesive analysis that spans the full spectrum of AgentAI development in Industry 4.0. This survey explicitly fills this gap by introducing a multi-domain taxonomy and by systematically analyzing both non-autonomous and fully autonomous AgentAI systems, offering a comprehensive synthesis not previously available in the literature. Additionally, the paper extends the discussion to Industry 5.0 and 6.0, exploring the evolution of AgentAI from automation to collaboration and, ultimately, to fully autonomous systems. This comprehensive analysis highlights the potential of AgentAI in driving industries toward a more efficient, sustainable, and adaptable future.

1. Introduction

The recent developments in AI, especially with Visual Language Models (VLMs) and Large Language Models (LLMs), denote a significant transformation toward the development of agent-oriented systems (Wang et al., 2024b) that integrates context memory, language proficiency, intuitive reasoning, visual cognition, and adaptability. Influenced by the advancements in AI systems discussed at the Dartmouth Conference in 1956, and subsequent studies such as Minsky's "Copy Demo" (Steels & Brooks, 2019), these systems are now positioned to go beyond traditional, specialized AI functions. They are expected to take on dynamic, autonomous roles in complex environments (McCarthy, Minsky, Rochester, & Shannon, 2006).

In this context, Aristotelian holism provides a valuable philosophical framework for understanding how these AI systems function. Holism emphasizes the importance of understanding a system as a whole, rather than analyzing components in isolation. This approach suggests that the full effectiveness of a system is realized when its parts – such as memory, reasoning, and perception – are developed and integrated in relation to the entire system. In AI, this holistic perspective is essential for creating systems that not only function as individual units, but work together cohesively to achieve the AI's ultimate "final cause" – its intended purpose or goal (Youvan, 2024). This holistic design paradigm is central to the development of agents capable of adapting, functioning, and reasoning in complex environments. By applying these principles, AI agents can plan, reason, and execute actions across a range of domains, including healthcare, robotics, and gaming, thereby demonstrating enhanced capabilities in adaptive behavior and multi-step decision-making (FAIR et al., 2022). This development has the potential to significantly transform industries, reshaping human-AI interactions and raising critical ethical considerations as AI systems take on agentic and autonomous roles within social frameworks (Ouyang et al., 2022).

Furthermore, the development of AgentAI architectures has been significantly influenced by the integration of Large Foundation Models (LFMs), LLMs, and VLMs, which enhance their capabilities beyond basic language processing (Bubeck et al., 2023). In particular, LFMs

* Corresponding author.
E-mail addresses: francesco.piccialli@unina.it (F. Piccialli), diletta.chiaro@unina.it (D. Chiaro), sundas.sarwar@unina.it (S. Sarwar), donato.cerciello@unina.it (D. Cerciello), pian.qi@unina.it (P. Qi), valeria.mele@unina.it (V. Mele).

https://doi.org/10.1016/j.eswa.2025.128404
Received 13 December 2024; Received in revised form 21 May 2025; Accepted 29 May 2025
Available online 2 June 2025

decompose complex tasks into smaller subtasks, leveraging pre-trained knowledge to plan and execute sophisticated actions across various domains (Huang, Abbeel, Pathak, & Mordatch, 2022). When combined with LLMs, these models enable advanced decision-making and task execution by processing complex inputs and providing detailed outputs in dynamic environments. This synergy is further enhanced by Embodied AI (Sharma et al., 2019), which integrates these cognitive capabilities with a physical form by enabling the AI to interact with the real world through sensory inputs and motor outputs. By incorporating LLMs into such systems, task planning and action execution in physical environments are optimized, allowing agents to perform tasks with greater precision and adaptability. These embodied agents leverage zero-shot proficiency and incorporate environmental feedback to continuously refine their actions, improving performance over time (Ahn et al., 2022).

In addition, interactive learning models contribute to this process by allowing AI systems to learn from user interactions, adjusting their outputs based on real-time feedback, and analyzing trends as they emerge. This iterative learning process enables the system to become more efficient, improving its ability to interact socially and execute tasks effectively. As a result, these advanced AI systems illustrate how AgentAI can automate complex real-world tasks, advancing the potential for intelligent, autonomous systems across diverse domains (Zha et al., 2024).

1.1. Research questions

To ensure a structured and purposeful contribution, this survey is guided by the following primary and specific research questions:

- **RQ1:** How are autonomous and non-autonomous AgentAIs currently integrated into Industry 4.0? What is their potential?
- **RQ2:** Which core technologies underpin the evolution of AgentAI? How do they enhance autonomy and adaptability?
- **RQ3:** What are the real-worlds domains where AgentAI is being applied?
- **RQ4:** What are the key challenges that interfere with the adoption of AgentAI in industrial environment?
- **RQ5:** How will the role of AgentAI evolve in the evolution of intelligent systems beyond the core principles of Industry 4.0?

1.2. Contribution

AgentAI is increasingly becoming a cornerstone of Industry 4.0 due to its ability to support autonomous decision-making, foster adaptive systems, and enable collaborative industrial applications. To the best of our knowledge, existing literature only examines AgentAI in fragmented or application-specific contexts, with limited focus on agent autonomy and architectural frameworks, and lacks a comprehensive, cross-sectoral perspective. A key innovation of this work, as shown in Table 1, lies in its multi-domain taxonomy, which consolidates AgentAI research across nine sectors, a scope not found in existing surveys. Furthermore, the integration of both non-autonomous and fully autonomous AgentAI within a unified analytical framework marks a significant advancement in how autonomy is studied in the context of Industry 4.0 and beyond.

The main contributions of this survey can be summarized as follows:

1. **Comprehensive overview on AgentAI in Industry 4.0:** This paper provides a thorough examination of how AgentAI contributes to core Industry 4.0 objectives.
2. **Comparison with existing literature:** It identifies and compares the most relevant surveys on AgentAI.
3. **Taxonomy and cross-domain classification:** A structured taxonomy of AgentAI research is presented, covering its applications across multiple domains.
4. **Analysis of ethical implications:** The broader consequences of adopting AgentAI in industrial settings are discussed, including ethical concerns, and the impact on human-machine collaboration.
5. **Identification of challenges and future directions:** This survey outlines open challenges related to real-time adaptability, context awareness, and agent autonomy, and highlights future directions towards human-centric systems in Industry 5.0 and fully autonomous cognitive frameworks in Industry 6.0.

1.3. Paper organization

The survey is organized as follows: Section 2 provides a detailed account of the literature collection process, including search terms, eligibility criteria, and evaluation methodology, ensuring the study's rigor

Table 1
Existing surveys on AgentAI and our contributions.

Paper	Fields	Key contribution	Ethics	Non autonomous Agents	Fully autonomous Agents	Taxonomy	Architecture
Gridach, Nanavati, Abidine, Mendes, and Mack (2025)	Chemistry, Biology	Provided a comprehensive overview of AgentAI in scientific discovery.	✓	✓	✓	✓	✗
Acharya, Kuppan, and Divya (2025)	Healthcare, Finance	Proposed a framework for AgentAI 's safe integration into society.	✓	✗	✓	✗	✓
Zhang et al. (2025)	Networking	Introduced a framework for telecom-specific planning.	✗	✗	✓	✗	✗
Durante et al. (2024)	Embodied agents, Virtual Reality	Explored next-embodied action prediction, incorporated external knowledge, and envisioned virtual reality for agent interactions.	✗	✗	✓	✗	✓
Masterman, Besen, Sawtell, and Chao (2024)	Gaming	Analyzed AgentAI 's current capabilities and limitations, offered guidance for future designs, including architectural choices.	✗	✗	✓	✓	✗
Kshetri (2025a)	Marketing	Analyzed how AgentAI can address inefficiencies in global healthcare, leading to reduced costs and higher quality care.	✗	✗	✓	✗	✓
Ours	Transportation, Energy, Healthcare, Networking, Defence, Gaming, Governance, Marketing, E-Learning	Provided a detailed multi-domain taxonomy, discussing the AgentAI advancements and challenges, within Industry 4.0 and beyond.	✓	✓	✓	✓	✓

and transparency. Section 3 delves into the foundational aspects of AgentAI, covering its principles, components, fundamental concepts and algorithms. Section 4 explores the enabling technologies for Industry 4.0. Section 5 introduces a detailed taxonomy of AgentAI applications in Industry 4.0. Section 6 provides the challenges faced by AgentAI systems and potential future research directions to address these challenges. Section 7 concludes with a brief summary.

2. Literature collection

This survey follows the PRISMA guidelines (Page et al., 2021) to ensure scientific rigor and suitability for a comprehensive and systematic review.

The literature search was conducted using four key databases: *Scopus*, *Web of Science*, *IEEE Xplore*, and *ScienceDirect* given their comprehensive coverage of relevant papers. Each database was meticulously chosen for its distinctive strengths: *Scopus* and *Web of Science* provide multidisciplinary articles with robust citation indexing, *IEEE Xplore* serves as a premier source for cutting-edge research in AI and ML, and *ScienceDirect* provides access to high-impact journals and publications in computer science and machine learning (ML). The primary search term "AgentAI" OR "AgenticAI" OR "Agentic AI" was employed to search through the title, abstract, and keywords of the selected publications in order to ensure the comprehensive coverage of relevant literature. The search in all databases employed a specific query: TITLE-ABS-KEY: ("AgentAI" OR "AgenticAI" OR "Agentic AI"). Fig. 1 provides an overview of the literature on the retrieval process.

The initial search across the selected databases using key terms yielded a total of 255 articles, among these 130 articles were sourced from *Scopus*, 64 articles from *Web of Science*, 25 articles from *ScienceDirect*, and 36 articles from *IEEE Xplore*. After excluding non-english articles and the duplicates, a total of 144 articles were obtained. Next, the title, abstract, and keywords of the selected articles were analyzed to eliminate those that did not directly contribute to the field of AgentAI or those that only mention "agent", but did not align their research on the principles and applications of AgentAI.

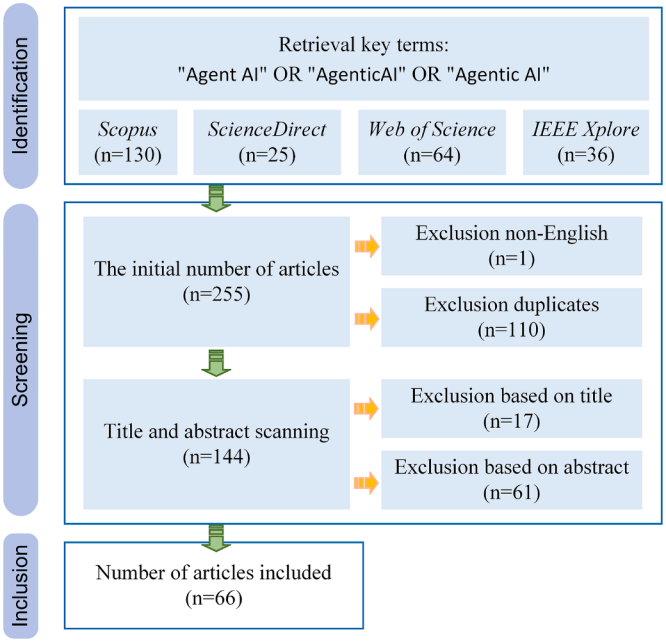

Fig. 1. The literature retrieval strategy encompassed the following three stages: identification, screening, and inclusion. In the identification stage, the appropriate keywords were established to effectively retrieve research articles pertinent to the topic being investigated. Next, in the screening stage, specific criteria were established to screen and exclude literature that did not meet the research objectives and needs. Finally, in the inclusion stage, literature that met the established criteria and requirements of the study was identified and integrated into this paper.

It is essential to note that the term "agent" constitutes a broad concept employed in various domains, such as biological agents or cells executing particular functions, which differ significantly from the computational and decision-making orientation of AgentAI within the realm of AI. As the focus was specifically on experimental articles, review articles and surveys were excluded from the analysis. In addition, we also excluded very old articles and only focused on articles within the past decade. Following this screening process, 78 articles were eliminated, resulting in a final set of 66 articles that were deemed suitable for inclusion in the survey, which will be systematically reported in Section 5. To enhance transparency and ensure reproducibility in the literature collection process, we provide a link that contains all collected articles, as well as those that remained after each step of the screening process. The link can be accessed: https://github.com/MODAL-UNINA/AgenAI-literature-search.

3. AgentAI

This section presents a concise overview of AgentAI, including its foundational principles, components, core techniques, and essential algorithms. This preamble provides the scientific basis necessary for understanding the mechanisms and methodologies underlying AgentAI, setting the stage for the rest of the paper.

In the broader domain of AgentAI, a distinction is made between non-autonomous systems and fully autonomous ones. Non-autonomous agents are typically used in simpler or more controlled environments where tasks are predefined and require minimal adaptive decision-making. These systems rely on human supervision or predefined rules and are not designed to adapt autonomously over time(Chan et al., 2023).

In contrast, autonomous AgentAI systems operate independently, responding to dynamic and complex environments with minimal human intervention. These agents function independently and are goal-driven, based on adaptive frameworks that execute actions, sensory data, and acquired knowledge in complex, dynamic environments (Nourani, 1999). They respond to changing environmental signals with strategy adjustments, without requiring continuous human supervision. Through multimodal inputs, such as text, images, and audio, AgentAI interprets complex contexts, and refines its responses in real-time using feedback and predictive modeling. In multi-agent environments, AgentAI employs reinforcement learning (RL) and evolutionary algorithms to facilitate adaptive decision-making, where agents refine their actions and engage in effective collaboration to fulfill common objectives. The fundamental principle of autonomy enables AgentAI to autonomously execute decisions, adjust through various learning paradigms, and proficiently manage real-time tasks, thereby reducing the need for continuous oversight in complex environments (Shankar, 2024).

The AgentAI systems continuously gather multisensory environmental feedback, which they process to refine actions and enhance contextual awareness for intuitive, context-specific interactions (Zhang, Vinyals, Munos, & Bengio, 2018). The knowledge agents within AgentAI integrate implicit understanding derived from pretrained models with explicit data sourced from structured databases, thereby constructing robust inference frameworks that guarantee accurate and contextually pertinent decision-making. The combination of LLMs and VLMs allows AgentAI to adeptly manage multimodal interactions, process linguistic, visual, and contextual information to yield versatile and integrated responses in tasks necessitating a confluence of visual and linguistic comprehension (Radford et al., 2021).

The integration of memory and continuous learning functionalities considerably enhances the decision-making performance of AgentAI, thereby facilitating its adaptation to novel data and user interactions over time with minimal retraining prerequisites (Parisi, Kemker, Part,

Kanan, & Wermter, 2019). However, when significant changes or improvements are needed–such as adapting to new domains, handling edge cases more effectively, or improving accuracy–retraining becomes mandatory. In practice, there are three primary strategies (Liu, Singh, Liu, Payani, & Zheng, 2025) used to refine AgentAI's performance:

- *Zero-shot learning* allows the agent to perform tasks without prior examples, relying on its pre-existing knowledge from initial training. It's efficient for tasks with no explicit data, yet still needing useful responses;
- *Few-shot learning* provides the agent with a small number of examples to guide task-specific behavior, allowing for quick adaptation without extensive training;
- *Fine-tuning* adjusts the agent's model using new task-specific data, helping it become more specialized, improve accuracy and handle edge cases, by retraining with additional examples.

Fine-tuning allows the AgentAI to handle specialized tasks more effectively and can significantly improve the agent's ability to perform in dynamic environments. Once fine-tuned, the agent can execute tasks more efficiently, requiring fewer examples for each subsequent interaction, optimizing computational costs, and reducing latency (Zooey Nguyen et al., 2024).

Ethical considerations are prioritized within the architecture of AgentAI to promote fairness, inclusivity, and accountability (Floridi & Cowls, 2022). Implementing mechanisms aimed at alleviating biases and ensuring respectful interactions among diverse demographic cohorts is of utmost importance, particularly for agents functioning within sensitive domains such as healthcare and finance (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). Moreover, AgentAI is designed with interpretability as a core attribute, ensuring that its decision-making processes are both transparent and comprehensible to human users, this is an essential element in fields that require accountability and trust, such as medical diagnostics and legal analysis (Doshi-Velez & Kim, 2017a). Fig. 2 illustrates the workflow of an AgentAI system, showing how it interacts with users, databases, and ML models to deliver actionable outputs. The user provides input to the AI Agent, which retrieves information from a database or a vector database (storing structured and semantic data, respectively). The LLM processes the retrieved data for tasks like context understanding, reasoning, and response generation, then executes actions according to the output of the LLM. The data flywheel captures feedback from these actions to continuously refine the system, while model customization ensures the AI adapts to specific tasks and environments over time, creating a dynamic and efficient operational loop.

3.1. Components of agentAI

The foundational elements of AgentAI systems enable for intelligent and autonomous operations in dynamic environments through making decisions, processing information, and executing actions. This section presents the core components of AgentAI, which provide a critical basis for understanding and responding to complex real-world scenarios in Industry 4.0.

The main components of AgentAI, as shown in Fig. 3, could be clustered in *perception module*, *cognition module* and *action module*.

- In the *perception module*, agents use VLMs and LLMs to interpret environmental data, integrating multimodal inputs such as images and textual instructions. This enables a fundamental comprehension of the environment through the integration of visual stimuli with verbal directives to augment cross-modal understanding (Kim, Yu, & Lee, 2017).
- In *cognition module*, AgentAI employs RL and interactive feedback loops to enhance decision-making processes by, integrating complex inputs and learned knowledge. Agents use explicit and implicit knowledge bases to engage in logical reasoning and acquire contextual understanding, thereby demonstrating a capacity for effective adaptation across diverse tasks (Valmeekam, Sreedharan, Marquez, Olmo, & Kambhampati, 2023).
- *Action module* translates an agent's cognitive decisions into real-life or virtual actions, enabling dynamic interactions with the environment. The action strategies employed by the agent are enhanced by RL through rewards and penalties, which allows for optimal responses and task execution with minimal human intervention (Kim et al., 2017). While traditional agents use a predefined set of hard-coded actions, modern AgentAI systems–powered by LLMs and modular tools–can dynamically expand their action space at runtime, enabling flexible and adaptive behaviors in open-ended environments (Hou, Zhao, & Wang, 2025).

3.2. AgentAI techniques and algorithms

This section provides an overview of the fundamental algorithms and techniques that underpin AgentAI systems. These algorithms are designed to facilitate autonomy, learning, and decision-making in multi-agent environments.

3.2.1. Reinforcement learning

RL is a ML paradigm in which agent learns by interacting with its environment, and refining their behavior through rewards and penalties. It plays a critical role in enabling agents to develop intelligent behaviors by optimizing state-action associations.

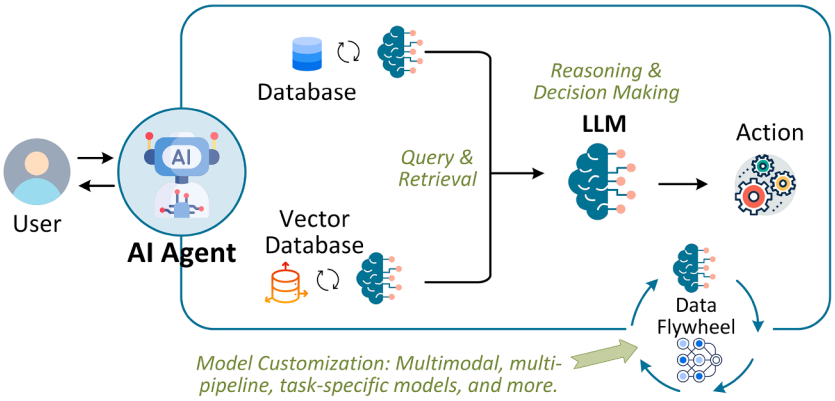

Fig. 2. The workflow of an AgentAI system, showing interactions with users, databases, and an LLM. The AI Agent receives user input, retrieves data from structured or semantic databases, processes it via an LLM for context understanding and reasoning, executes actions based on LLM outputs, and incorporates feedback through a data flywheel to continuously refine the system.

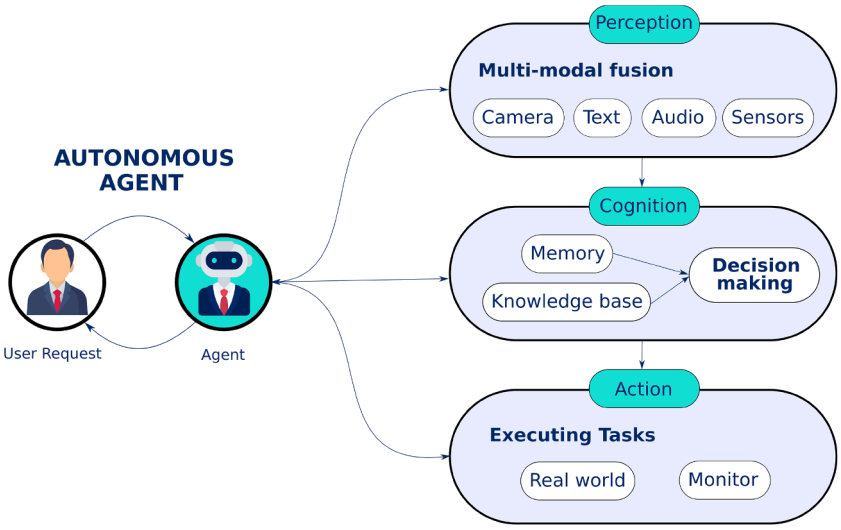

Fig. 3. Overview of the main components of AgentAI: perception, cognition, and action modules, working together for autonomous and intelligent operations.

1. **Deep reinforcement learning:** Deep Reinforcement Learning (DRL) integrates RL with deep neural networks, allowing agents to handle high-dimensional inputs and make efficient decisions in complex environments. DRL operates directly through interactions with the environment, thereby enabling autonomous learning without the need for system modeling (Valladares et al., 2019; Xie, Ajagekar, & You, 2023).
2. **Multi-agent reinforcement learning (MARL):** Multi-Agent Reinforcement Learning (MARL) involves multiple agents in a shared environment, optimizing both individual and collective goals to enhance system performance. It addresses challenges like non-stationarity and coordination (Foerster et al., 2017), ensuring that agents balance local and global objectives. Algorithms like QMIX enable decentralized actions through centralized training (Rashid et al., 2020), whereas G2ANet employs attention mechanisms to prioritize observations and actions, thereby fostering collaboration and stability in complex scenarios (Nam, Heo, Kim, & Yoo, 2023).

3.2.2. *Transformers*

In modern ML, the architecture of transformer is a foundational framework, specifically effective for processing sequential data such as text or time series. At its core, it uses a self-attention mechanism that allows the model to focus dynamically on different parts of the input sequence (Vaswani, 2017). This mechanism assigns attention scores to pairs of elements, or tokens in the sequence, enabling the model to identify relationships and dependencies, even across long distances.

The model employs multi-head attention to enhance its ability to recognize distinct patterns, which creates numerous parallel attention mechanisms that simultaneously analyze different aspects of the input (Vaswani, 2017). Since the transformer does not process inputs sequentially, positional encoding is added to the token embeddings to incorporate information about the order of elements in the sequence, while feedforward layers independently process each token for deeper representations. To ensure efficient gradient flow and stable training, the architecture additionally includes residual connections (which prevent certain layers) and layer normalization. Transformers consist of stacked encoder and decoder layers: encoders focus on understanding the input sequence (Kenton & Toutanova, 2019), while decoders generate outputs, making the architecture highly suitable for tasks such as machine translation and text generation (Radford, 2018). A key advantage of Transformers lies in their ability to leverage transfer learning, they are frequently pre-trained on extensive large datasets and fine-tuned for specific tasks to achieve high performance with limited labeled data.

3.2.3. *Large language models*

Based on the transformer architecture, LLMs represent a significant advancement in Natural Language Processing (NLP). The self-attention mechanism inherent in LLMs calculates the dependencies between tokens, thereby allowing for contextually relevant focus (Triem & Ding, 2024). Parallel processing enables faster training on large datasets. LLMs are pre-trained on vast unlabeled text datasets to learn general language patterns and are fine-tuned for specific tasks to enhance performance. Transformer variants like BERT and GPT specialize in tasks: BERT uses masked language modeling for bidirectional context understanding (Kenton & Toutanova, 2019), whereas GPT employs autoregressive language modeling for text generation (Radford, 2018). LLMs excel in text completion, question answering, and conversational AI.

The following highlight key advancements and models that have contributed significantly to the advancement of the capabilities of transformer-based LLMs.

- **OpenAI and chatGPT:** OpenAI has been at the forefront of advancing the transformer model with the GPT series, which includes GPT-3 (Benjamin Mann, 2023) and GPT-4 (OpenAI, 2024). These models learn language patterns and information by pre-training on vast internet data, followed by fine-tuning on more specific datasets to improve performance on conversational or targeted tasks, such as conversational assistance and text generation. ChatGPT leverages supervised learning and RL, specifically RL from Human feedback, for training. The training process involves three steps: supervised fine-tuning, a reward model, and maximum policy optimization. This advancement enhances ChatGPT's ability to generate human-like responses for conversation and assistant tasks.
- **Llama:** An open-source foundational model similar to OpenAI and GPT another language model Llama was conceptualized (Touvron et al., 2023). It is part of Meta's initiative to provide a powerful language model that is suitable for text generation and supporting various applications. Llama models have unique configurations, with their training datasets drawn from a mix of text sources, emphasizing efficient and large-scale natural language understanding.
- **Gemini:** Gemini is a multimodal large language model developed by DeepMind, designed to integrate and process text, images, audio, and code. It builds upon Google's PaLM architecture and combines neural networks with symbolic reasoning capabilities (Team, 2024). Gemini models are trained to perform complex tasks involving reasoning, retrieval, and multimodal input, pushing the boundaries of general-purpose AI systems.

Table 2
Comparative analysis of agentAI paradigms.

Paradigm	Interpretability	Scalability	Adaptability	Suitable Tasks
Symbolic Reasoning Agents	●●●	●○○	●◐○	Explainable planning, regulatory systems
Graph-Based Policies	●●○	●●○	●●○	Logistics, semantic QA, multi-step planning
Behaviour Tree Agents	●●○	●●●	●●○	Real-time control, robotics, interactive simulation
Probabilistic and Bayesian Agents	●●○	●●○	●●◐	Human-aware robotics, uncertain navigation, decision support
Transformer-based LLMs	●◐○	●●●	●●●	Language processing, reasoning, tool-augmented agents

- **DeepSeek:** DeepSeek is a series of open-source language models, developed by High-Flyer company, and focused on providing scalable and high-performing models for both code and natural language tasks (DeepSeek-AI et al., 2025). Notably, DeepSeek-V2 introduced dense hybrid architectures that improve efficiency and performance across multilingual benchmarks. The models are trained on diverse high-quality datasets and optimized for real-world application needs.

3.2.4. Alternative agent paradigms

While transformer-based Large Language Models (LLMs) currently dominate the design of AgentAI systems, alternative agentic paradigms offer complementary strengths and are particularly relevant in domains where interpretability, structured reasoning, or adaptive coordination are required.

The following highlights some key alternative agentic paradigms.

- **Symbolic reasoning agents:** Symbolic agents operate on explicit representations of knowledge–typically logic rules, ontologies, or formal grammars–to perform inference and planning. These agents excel when transparency and verifiability are critical, for instance in automated verification or regulatory compliance (Bougzime, Jabbar, Cruz, & Demoly, 2025). Modern neurosymbolic hybrids increase this approach by coupling deep perception modules with symbolic planners, as in the Neuro-Symbolic Concept Learner (Mao, Gan, Kohli, Tenenbaum, & Wu, 2019), or by implementing structured recurrent reasoning via compositional attention (Hudson & Manning, 2018).
- **Graph-based policies:** Graph-based policies represent the agent's decision space as a structured graph, where nodes typically correspond to entities and edges to their relations. A common form of such a structure is the Knowledge Graph (KG), which encodes semantic relationships between entities in a way that supports reasoning and generalization (Hogan et al., 2021). Within this framework, agents can apply RL to learn how traverse the graph efficiently (Xiong, Hoang, & Wang, 2018) or use Graph Neural Network (GNN) to propagate and aggregate relational information (Schlichtkrull et al., 2017). These approaches are particularly effective in domains like logistics, sematic query answering, and multi-step planning, where the relational structure of the environment is the key in order to solve the task.
- **Behaviour tree agents:** Behaviour Trees (BTs) model agent performances as a modular, hierarchical structure. Each tree node defines either an atomic action (e.g., move, check, pick) (Colledanchise & Ögren, 2018) or control operation (e.g., sequence, fallback, repeat) (Iovino, Scukins, Styrud, Ögren, & Smith, 2022). Their clean separation of logic and execution flow allows BTs to be easily extended, reused or interrupted, making them suitable for robotics, games, and simulations. BTs are often preferred in real-time or reactive contexts because they ensure predictable and safe execution, even under dynamic environmental changes.
- **Probabilistic and Bayesian agents:** These agents explicitly represent and reason about uncertainty. By modeling probability distributions over observations, hidden states, and future outcomes, they can make informed decisions in environments with incomplete or noisy information. Key algorithm tools include Markov Decision Processes (MDPs) (Triantafyllou, 2023), Bayesian belief updates, and probabilistic graphical models. In practical terms, such agents are effective for risk-sensitive planning, robot navigation under uncertainty, and human-agent interaction, where ambiguity must be actively managed (Hubmann, Becker, Althoff, Lenz, & Stiller, 2017).

To highlight the diversity of strengths and limitations across these approaches, Table 2 provides a comparative overview of the paradigms discussed above along four criteria: interpretability, scalability, adaptability, and their most suitable tasks.

3.3. AgentAI life cycle

In an AgentAI Life Cycle, the birth of an autonomous agent begins with the collaboration of three primary parties: the model developer, that creates the AI model that powers the agent, defining its core capabilities and behaviors, the system developer, that builds the larger system around this model, integrating it into tools and interfaces, and the user, that interacts with the agent, setting goals and providing the system with task-specific instructions (Shavit et al., 2023). Together, these parties ensure the agent's creation, customization, and alignment with the desired objectives.

Once the initial creation of the AgentAI is completed, its deployment requires sophisticated infrastructures capable of supporting distributed, low-latency, and scalable execution. These infrastructures typically rely on a combination of cloud-native platforms, containerized environments (e.g., Docker), and orchestration tools, which enable dynamic scheduling and elasticity (Muzumdar, Bhosale, Basyal, & Kurian, 2024; Senjab, Abbas, Ahmed, & Khan, 2023). These systems rely on advanced communication protocols (e.g. TCP/IP, QUIC, gRPC, MQTT) and cloud networking technologies, such as Software Defined Network (SDN) and Network Function Virtualization (NFV) to dynamically allocate resources, manage connections, and ensure secure communications (Mayer, 2024).

Following deployment, migration becomes very important when adapting agents to evolving environments or transferring them between platforms. Migration may involve model fine-tuning or re-training with new domain data. Smooth migration ensures that agents remain efficient and accurate in new contexts, avoiding performance degradation or ethical pitfalls due to domain shift. To support this adaptability, AgentAI systems must be capable of interpreting and responding to changes in network structure, through service discovery mechanisms, which allow agents to dynamically locate and connect to the appropriate services based on availability, proximity, or system load (Angelis & Kousiouris, 2025).

This marks the transition from technical development to applied use, where AgentAI systems begin to generate tangible value in real-world industrial ecosystems. Several major companies–such as Microsoft, Salesforce, and Accenture–have integrated AgentAI models into their platforms to improve operational efficiency and decision-making. In particular, Microsoft Dynamics 365 uses agents to automate customer service processes, Salesforce Agentforce supports sales and Customer Relationship Management (CRM) automation, and Accenture AI Refinery enhance workflow management in industrial environments (Kshetri, 2025b).

Despite growing autonomy, human-in-the-loop control remains essential in high-risk domains for output verification, ethical alignment, and exception handling (Manzini et al., 2024; Roesler, Rieger, & Langer, 2025). Interruptibility, or the ability to safely shut down an autonomous AgentAI, is a crucial part of its lifecycle to prevent harm, particularly in the event of malfunctions, when the agent is no longer needed, or

Table 3
Domain-specific differences in AgentAI lifecycle attributes.

Domain	Fine-Tuning	Migration	Interruptibility
Transportation	Enhance coordination and routing; improve logistics; support driver training	Edge-to-cloud dynamic deployment	Prevent traffic disruptions and ensure safety
Energy	Enable accurate load forecasting; adapt to demand patterns; improve grid efficiency	Grid-level updates; highly distributed	Preserve energy reliability
Healthcare	Personalize diagnosis and care pathways; ensure fairness and clinical accuracy	Patient specific deployment; ensure compliance.	Allow shutdowns with fallback and safety guarantees
Networking	Dynamically adjust to traffic and security needs; optimize network throughput	Service discovery and topology-driven relocation	Maintain uptime during reconfiguration
Defence	Enhance strategic decision-making; simulate threats	Tactical migration across systems; supports mission variability	Minimal interruption; essential for safety and scenario integrity
Gaming	Improve NPC behavior and challenge level; optimize performance during gameplay	Seamless across platforms	Interrupt for responsiveness to player actions or system events
Governance	Ensure transparency, accountability, and value alignment; prevent misuse or bias	Deployment constrained by jurisdiction	Mandatory interruptibility for ethical and legal compliance
Marketing	Personalize messaging; adapt strategies based on behavior	Cross-platform migration via APIs systems	Non-critical but user-sensitive
E-Learning	Tailor instruction to individual learners; support engagement and content personalization	Support multi-device learning; update content dynamically	Enhance adaptive feedback; encourage experimentation

in high-risk environments (Hu, NA, Yellamati, & Goktas, 2025). The shutdown process must be designed to stop operations without causing disruption, which may involve halting specific actions or fully terminating the agent's function (Shavit et al., 2023). To ensure safety, agents should be able to accept shutdown requests, even during complex tasks, and have fallback procedures in place, such as notifying users if a task is incomplete. It's essential that agents allow users to shut them down without interference, even in emergency situations (Rashid et al., 2025). An overview of the entire AgentAI Life Cycle is summarized in Fig. 4.

While the presented AgentAI lifecycle model outlines a domain-independent sequence of stages, its implementation in real-world scenarios is far from uniform. Each phase encounters different operational, regulatory, and technical constraints depending on the industrial context. For instance, the frequency and complexity of agent updates, the criticality of system interruption, and the level of customization required can vary widely across domains. Factors such as legal compliance, real-time responsiveness, integration with physical systems, and user-specific adaptation all shape the lifecycle trajectory of AgentAI systems (Doshi-Velez & Kim, 2017b; Shneiderman, 2020). These divergences are not merely technical, but reflect the differing priorities of each sector, such as safety in healthcare, explainability in governance, or scalability in marketing automation. To illustrate these sector-specific variations, Table 3 provides a comparative overview of key lifecycle attributes across the domains outlined in the taxonomy.

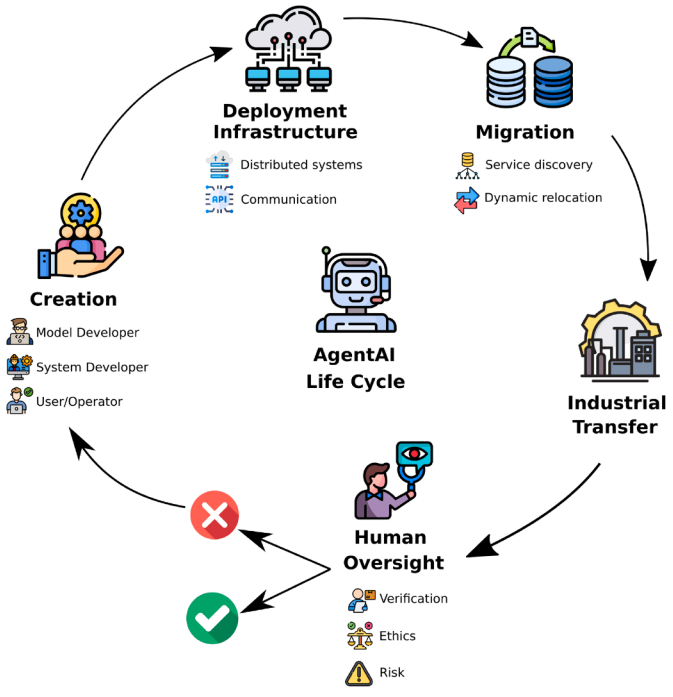

Fig. 4. The diagram shows the sequential phases in the life cycle of an autonomous agent: from its initial creation by developers and users, through deployment in distributed infrastructures, adaptive migration across network nodes, and integration into industrial environments. It also highlights the essential roles of human-in-the-loop supervision and controlled shutdown procedures, which are critical for ensuring ethical compliance, safe decision-making, and graceful interruption of agent operations in high-risk or failure scenarios.

4. Industry 4.0 framework

Industry 4.0, also known as the fourth industrial revolution, emphasizes the convergence of digital technologies to create intelligent, interconnected, and autonomous systems within industrial environments (Kagermann, Wahlster, & Helbig, 2013). Widely adopted across manufacturing sectors, it focuses on modernization, interoperability, and system-level integration, where seamless information exchange and cross-layer collaboration serve as fundamental enablers (Lu, 2017).

At the core of Industry 4.0 are Cyber-Physical Systems (CPS), which integrate computational and physical processes by collecting and analyzing data across various stages of production. Through technologies such as digital twins and advanced analytics, CPS enable real-time monitoring, predictive maintenance, and data-driven decision-making, thereby improving manufacturing efficiency, adaptability, and resilience (Oks et al., 2022). In recent years, the evolution of CPS has increasingly involved the integration of AI agents. Chae, Lee, Jang, Hong, and Park (2023) proposed an AI-enhanced CPS architecture comprising five interrelated layers: physical systems, information systems, users (AgentAI), networks, and data. Within this framework, intelligent agents play a central role in interpreting sensor data, monitoring system states, and making autonomous decisions. These agents operate without human intervention by leveraging AI techniques such as machine learning, enabling CPS to manage complex and dynamic industrial conditions more effectively. Applications range from resource optimization to intelligent scheduling and real-time production coordination. Together with the Internet of Things (IoT), CPS facilitates the creation of fully connected industrial ecosystems, enabling seamless communication between machines, humans, and products. This lays the foundation for autonomous decision-making, self-optimization, and predictive capabilities in smart factories (Qin, Liu, & Grosvenor, 2016).

Building upon the capabilities of AI-driven CPS, cloud computing emerges as a crucial enabler that supports the scalable deployment, training, and coordination of intelligent agents. Cloud platforms offer shared access to models and data, allowing AI agents deployed across multiple sites to exchange insights and collaboratively improve performance. For example, Rehman et al. (2021) presents a cloud-integrated multi-agent production framework combining JADE-based agent systems, machine learning, and a simulated Factory I/O environment. In this system, cloud-based classifiers assess product quality, while distributed agents make routing decisions. The framework demonstrates flexibility and efficiency in handling both qualified and defective parts, illustrating the potential of cloud-supported agent architectures in real-world manufacturing scenarios.

As physical execution remains a critical component of CPS, robotics plays a foundational role in bridging digital intelligence with physical action. Modern industrial robots, embedded with sensors and AI capabilities, can autonomously adapt to environmental changes, perform complex tasks alongside humans, and reconfigure themselves without manual reprogramming. AI agents frequently interface with these robotic systems, translating high-level decisions into physical operations. Additionally, advanced techniques such as computer vision and anomaly detection empower robots to inspect product quality and make real-time adjustments, reducing waste and downtime (Bahrin, Othman, Azli, & Talib, 2016). Beyond controlling physical systems, intelligent agents also serve as intermediaries between machines and human operators. Augmented Reality (AR) facilitates this human-agent interaction by providing intuitive, context-aware visualizations of system status and decision logic. AR overlays AI-generated insights onto the physical workspace, offering real-time alerts, guided instructions, and system feedback to support operations such as maintenance, inspection, and quality control (De Pace, Manuri, & Sanna, 2018). A notable example is presented by Li, Zheng, Li, Pang, and Lee (2022), who developed a collaborative manufacturing system driven by AR-assisted digital twins and reinforcement learning-based agents. In this system, the digital twin of a robot is rendered as a holographic projection via AR glasses, allowing operators to monitor its real-time status and environment. The AI agent autonomously plans motion paths based on user-defined goals received through the AR interface, effectively integrating human oversight with autonomous control. Here, AR serves as the operator's 'eyes and hands', while the AI agent functions as the cognitive core of the system.

5. AgentAI within industry 4.0

Through the enhancement of complex decision-making mechanisms, the provision of real-time adaptability, and the encouragement of improved collaborative strategies, AgentAI powers intelligent automation and optimized resource management across industrial processes. These models provide powerful capabilities for adaptation, learning from experience, and generating context-aware responses, making them particularly valuable for sectors like smart manufacturing, predictive maintenance, and intelligent automation, where efficiency, personalization, and resilience are paramount.

In this survey, we examined how AgentAI, through these foundational models, is shaping the landscape of Industry 4.0 by driving efficiency, personalization, and resilience in industrial processes. For example, as shown in Fig. 5, the entire workflow of a predictive maintenance example is shown, which supports early fault detection and optimized maintenance planning within the Industry 4.0 framework. Specifically, from querying historical sensor and maintenance data to inputting it into a machine learning model (e.g., built using TensorFlow/Keras) for fault prediction and RUL estimation. The results are then interpreted and presented through a generative model (e.g., GenAI/RAG), providing intuitive decision support for engineers. The flow illustrates seamless collaboration among these elements, highlighting the synergy between AI agents and Industry 4.0 technologies.

As our investigation centers on the application of AgentAI within the context of Industry 4.0, we systematically classified the reviewed literature into distinct sub-fields, resulting in a comprehensive taxonomy that offers a clear and organized framework for analysis. The 66 articles included in our review were grouped into nine specific sub-fields. Fig. 6 illustrates the distribution of publications across these categories, along with their respective publication years, while Table 4 summarizes the

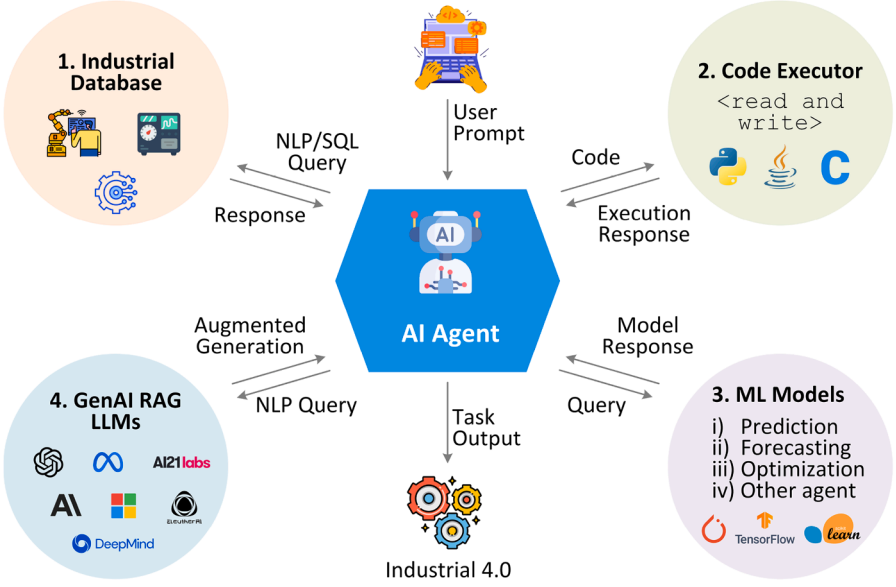

Fig. 5. The integration of Industry 4.0 with AgentAI. (1) The agents send NLP or Structured query language (SQL) queries to industrial databases and receive structured responses. (2) The agents dispatch code (e.g., Python scripts) to execution environments, enabling read/write operations and processing the returned results. (3) The agents query machine learning models, which provide predictive outputs. (4) The agents interact with generative AI models by sending text-based prompts and receiving synthesized or enhanced content.

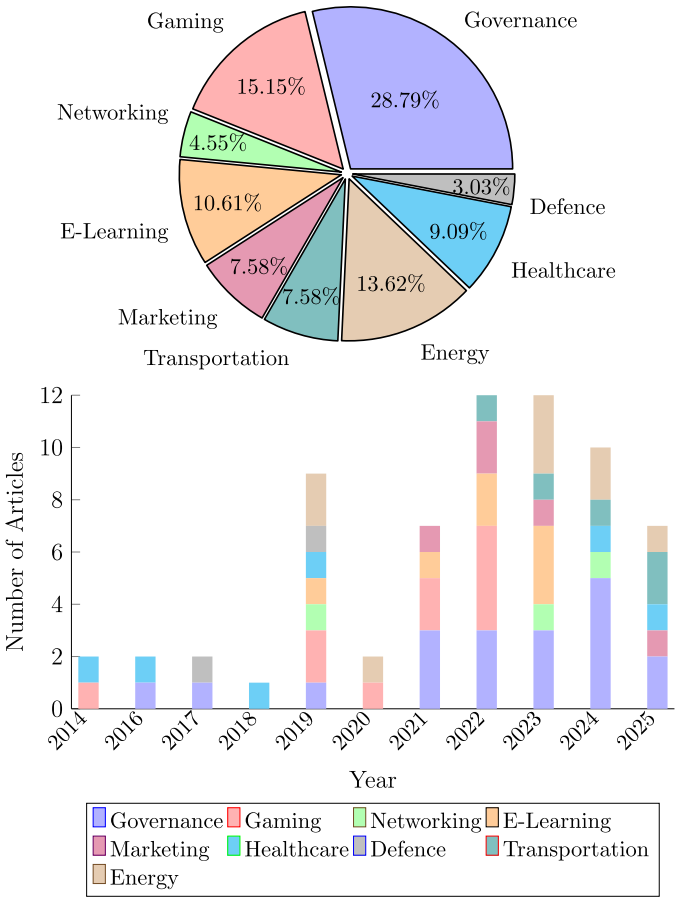

Fig. 6. Distribution and temporal evolution of AgentAI research articles in Industry 4.0 across domains. (Top: Domain-wise distribution; Bottom: Year-wise trend).

main AgentAI applications, contributions, and techniques identified in each presented domain.

5.1. Transportation

In the context of Industry 4.0, smart transportation systems increasingly integrate AI-driven training environments to enhance human-machine interaction and operational safety (Delanovic et al., 2023). One notable application combines supervised learning with virtual reality (VR) technologies–specifically, a random forest algorithm within an Oculus VR and Unity 3D environment–to simulate complex traffic scenarios for training purposes (Sangeetha, Vamsidharan, Saran, & Shrikesh, 2022). Originally developed for educating children on traffic safety (including pedestrian behavior, traffic signals, and vehicle rules), such immersive frameworks can be adapted to industrial contexts such as autonomous vehicle operator training, logistics fleet coordination, and smart city traffic management.

A recent AI framework developed for the 'University City of the Future' underscores the role of Generative AI agents in promoting more sustainable and intelligent urban environments (De Silva et al., 2025). This system follows a five-step process–Acquisition, Preparation, Orchestration, Dissemination, and Retrospection–to manage urban data and optimize operations. Through integration of advanced AI capabilities, it supports adaptive learning and data-driven decision-making. Particularly in domains such as human mobility, the framework has demonstrated promise for improving urban traffic planning, safety, and environmental outcomes. Additionally, research on LLM-based multi-agent systems shows that integrating LLMs with existing urban information infrastructures significantly enhances smart city management, especially in transportation coordination and infrastructure optimization. These systems improve query routing accuracy (94 %-99 %) and response relevance, effectively shortening decision-making cycles from days to hours (Kalyuzhnaya et al., 2025). Their ability to synthesize multimodal data for traffic facility design and mobility demand analysis further highlights AI's transformative role in urban mobility, while also addressing challenges such as computational scalability and real-time data integration.

By enabling realistic and interactive simulations, these systems help users better understand AI behavior in dynamic transportation settings and adapt their responses accordingly (Lähdeaho & Hilmola, 2024). This aligns with the broader Industry 4.0 objective of building cyber-physical systems that not only function autonomously but also interact smoothly with humans in complex real-world environments. However, despite their promise, AI-VR training systems still face several limitations. The transferability of skills learned in virtual environments to real-world contexts remains insufficiently validated, raising concerns about generalization and long-term effectiveness. Moreover, the computational intensity and hardware costs of immersive VR setups may limit scalability, especially in resource-constrained industrial applications. A further challenge lies in the lack of standardization regarding how human feedback is incorporated into AI learning loops, potentially reducing adaptability in rapidly changing environments. These issues point to the need for more rigorous evaluation frameworks and improved integration between virtual simulations and real-world deployment.

5.2. Energy

In the context of Industry 4.0, AgentAI enhances sustainability by leveraging intelligent technologies across multiple sectors. In particular, it facilitates greener and more energy-efficient industrial practices through improved resource utilization and coordination (Rahman, Chawla, Yaqot, & Menezes, 2025; Tortorelli, Sabina, & Marchetti, 2024). One prominent area of application is smart energy systems, where AgentAI increasingly enables dynamic adaptation to renewable energy sources (Renjith et al., 2024), thereby accelerating the transition toward cleaner and more resilient industrial infrastructures.

Within smart grids, multi-agent systems enable dynamic energy consumption adjustment (Nam et al., 2023; Xie et al., 2023), peak load mitigation (Zhang et al., 2023), and seamless incorporation of renewable sources such as solar and wind (Renjith et al., 2024). By facilitating intelligent energy scheduling and load balancing, these systems help maximize the use of renewable energy. For instance, in decentralized energy management, AgentAI architectures significantly enhance the operation of modular nanogrids (Renjith et al., 2024), promoting local energy sharing and improving grid reliability. In such systems, the smart grid is divided into autonomous controllers, each managing a microgrid integrated with photovoltaic and wind units. A fuzzy logic controller (FLC), powered by AI, enables real-time decision-making by monitoring electricity prices and power flow to select optimal energy suppliers and improve energy quality.

Beyond power distribution, AgentAI also advances environmental sustainability through applications in pollution treatment and process optimization. In wastewater treatment, for example, multi-agent reinforcement learning (MARL) algorithms such as QMIX (Rashid et al., 2020) and G2ANet (Nam et al., 2023) optimize operational parameters under variable influent conditions. These methods have achieved up to a 25 % reduction in aeration energy consumption and a 7 % improvement in effluent quality. Moreover, platforms like CRYSTAL (Gomes et al., 2019) leverage AI agents to accelerate material discovery for energy-efficient chemical processes, illustrating AgentAI's transformative role in both resource optimization and environmental protection.

Additionally, AgentAI contributes to energy conservation in buildings through intelligent control of heating, ventilation, and air conditioning (HVAC) systems. By training on ten years of simulated data, these systems can adaptively manage thermal comfort, air quality,

Table 4
AgentAI applications across various domains.

Domain	Application	Contribution	Techniques
Transportation	City coordination	Improve coordination, routing, and driver training via AI-VR frameworks Sangeetha et al. (2022); simulate complex traffic environments (De Silva et al., 2025); support human-machine interaction (Delanovic et al., 2023; Lähdeaho & Hilmola, 2024)	Random Forest, VR simulation, DRL
Energy	Environmental Control	Optimize energy efficiency (Zhang et al., 2023), load balancing, HVAC control (Valladares et al., 2019), pollution treatment (Nam et al., 2023; Rashid et al., 2020), and renewable integration (Gomes et al., 2019; Renjith et al., 2024; Valladares et al., 2019)	MARL, Fuzzy Logic Controllers
Healthcare	Psychology	Analyze psychological states (Hong, 2018) through memory-based models Ellwart and Kluge (2019) for adaptive learning and social deduction techniques (Katagami et al., 2014).	ML, Schema Theory
	Decision making	Enhance AI-driven data analysis to improve diagnostics (Crowder, 2016) and provide personalized treatments using cognitive augmentation instruments (Yager, 2024).	NLP, VLM
Networking	Traffic Prediction	Optimize traffic and resource allocation in SDN and 5G networks to enhance efficiency and user satisfaction Cao, Wang, Chen, and Barnawi (2019).	SDN, ML
	Network Resource Optimization	Allocate resources in 6G networks (Zhang & Zhu, 2023), achieving QoS requirements such as low latency and high throughput (Wang et al., 2024a).	Massive MIMO, NFV, LLM
Defence	Tactical Decision Making	Optimize advanced simulations, optimal escape routes for evacuation systems (Yan, Jia, Hu, Guo, & Zhu, 2019), and leverage tactical maneuvers for military operations (Tiao-Ping, Jing, & Jian-Hui, 2017).	CXBR, CAS, RFES
Gaming	Strategy	Improve decision-making flexibility (Franklin & Markley, 2014; Pan, Xue, & Ge, 2022; Sanjaya, Wang, & Yang, 2022) and modularity of complex real-world tasks (Barriga, Stanescu, Besoain, & Buro, 2019).	Markov Models, Nash Clustering, Monte Carlo Tree Search
Governance	Ethics	Critique (Boyles, 2024; De Vreede, Raghavan, & De Vreede, 2021; Gratch & Fast, 2022; Shams, Beynier, Bouveret, & Maudet, 2022) or promote ethical decision-making (Manzini et al., 2024; Srivastava, Lilly, & Feigh, 2024) through predictive simulations or frameworks and collaborative reasoning (Cabitza, Campagner, & Simone, 2021; Choung, Seberger, & David, 2023; Gervais, 2023) to address AI bias (Chan et al., 2024; Chandar et al., 2017).	GOGAR, BDMA, KA, MDPs
Marketing	Customer Services	Improve AI chatbots for business management (Sepanosian, Milosevic, & Blai, 2025) and customer services (Baabdullah, Alalwan, Algharabat, Metri, & Rana, 2022; Jeon, 2022) and managerial roles Chong, Yu, Keeling, and de Ruyter (2021); He et al. (2023).	LLM, RL
E-Learning	Teaching	Enhance student engagement, provide real-time feedback (Khabarov & Volegzhanina, 2019; Lee et al., 2022), and support hands-on art (Deshpande & Magerko, 2021; Kang et al., 2023), and education (Kumar, Tian, Celepkolu, Israel, & Boyer, 2022; Lee, Lim, & Nagarajan, 2023; Tian et al., 2023).	AI-FML, RF, SketchRNN

and energy use. Compared to traditional control methods, AI-enhanced HVAC systems not only reduce energy consumption but also lower CO_2 levels and enhance indoor comfort, demonstrating the potential of AgentAI in managing complex built environments while supporting sustainable industrial operations (Valladares et al., 2019).

Although AgentAI demonstrates significant promise across energy systems–from smart grids to HVAC control–several challenges remain. Many multi-agent frameworks are built on idealized assumptions of communication and coordination, which often break down in real-world energy networks characterized by latency, noise, and incomplete data. Additionally, the scalability of AI-based control strategies in large-scale, heterogeneous environments–especially under uncertain renewable energy inputs–remains a technical bottleneck. Computational constraints and the fast-changing nature of energy markets further hinder real-time responsiveness. Moreover, integrating advanced AI agents with traditional industrial or municipal infrastructure introduces technical and operational complexities, including compatibility and system integration issues. These challenges underscore the need for robust and adaptable AgentAI architectures capable of maintaining performance amid uncertainty and system variability.

5.3. Healthcare

While Industry 4.0 traditionally focuses on manufacturing systems, cross-domain applications of AgentAI–particularly in healthcare–provide valuable insights for the future of industrial cognitive systems and human-agent collaboration (Good & Horn, 2025; Hong, 2018). In healthcare, AgentAI models enhance clinical decision-making and cognitive augmentation by providing data-driven guidance and automating complex workflows (Crowder, 2016). These capabilities closely parallel the goals of intelligent decision support systems in industrial environments. For example, wearable sensors and digital health platforms continuously collect physiological and behavioral data, enabling adaptive, personalized interventions–a model that may inform future human-centric industrial systems. Cognitive augmentation systems such as the "science exocortex" (Yager, 2024) also illustrate how interconnected AI agents can assist researchers by automating data analysis, knowledge integration, and experiment management. Similar architectures can be envisioned in smart factories, where AgentAI supports engineers by managing system diagnostics, visualizing production data, and guiding maintenance. Furthermore, AI agents that process multimodal cues, such as facial expressions and gestures, have been explored in social deduction games (Katagami et al., 2014). These studies demonstrate how social signal processing and behavioral analysis can enhance human-machine interaction, offering potential applications in industrial safety monitoring, worker intent recognition, and collaborative robotics.

These cross-domain examples demonstrate how innovations in AgentAI for healthcare and cognitive science can inspire future developments in industrial intelligence, especially in areas requiring human-aware autonomy, adaptive decision-making, and seamless human-agent collaboration (Ellwart & Kluge, 2019). While cross-domain insights from healthcare and cognitive science offer compelling directions for industrial AgentAI systems, significant challenges remain in transferring these models across contexts. Human-agent collaboration in healthcare typically involves high-quality, individualized data in controlled environments, whereas industrial settings often face noisy, sparse, or fragmented data streams. Furthermore, the ethical, privacy, and interpretability standards in healthcare may not directly translate to industrial use cases, requiring careful adaptation. Multimodal AI, especially systems interpreting social signals like gestures or facial expressions, faces additional barriers in terms of robustness, cultural variability, and real-time responsiveness in complex, high-speed industrial environments. These discrepancies underscore the need for domain-aware adaptation strategies and flexible AgentAI architectures that can bridge

the gap between cognitive augmentation in healthcare and real-world industrial applications.

5.4. Networking

In the networking sector, including AgentAI models into SDN optimizes traffic prediction (Cao et al., 2019) and resource allocation, thereby enhancing the efficiency of network. Service prediction and resource scheduling across the user, controller layer, and base station are the main objectives of the proposed AI agent framework for 5G SDNs (Cao et al., 2019). These agents augment conventional approaches to communication network optimization and service quality enhancement by using ML to anticipate user needs and dynamically allocate resources, addressing issues such as network congestion, retransmissions, and interference. This adaptive, multi-layered approach promotes faster, more efficient networks, with a focus on minimizing retransmissions and reducing the impact of interference and improving user satisfaction.

In the context of 6G networks, AI-enabled architecture integrates Massive Multiple-Input-Multiple-Output (Massive MIMO), NFV, and SDN technologies to dynamically allocate network resources according to statistical Quality of Services (QoS) requirements (Zhang & Zhu, 2023). This system employs edge AI frameworks and federated learning for real-time decision-making and facilitates support for multiple services by visualizing the network into slices, each tailored to specific QoS needs, such as low latency or high throughput. Extensive simulations demonstrate that the proposed approach improves network efficiency and consistently meets QoS demands, thereby establishing it as a robust solution for contemporary 6G applications.

In addition, advanced LLM-driven agents are becoming essential to optimize performance and automate processes for optical network management (Wang et al., 2024a). These agents are positioned within the network's control layer using a suggested framework, which allows for intelligent coordination between the application and physical layers. By using retrieval augmentation and engineering techniques, they improve the accuracy of data handling and maximize high-speed data transmission.

Collectively, these techniques enable more responsive, efficient, and precise network operations, ensuring optimal performance and seamless communication across increasingly complex systems. By effectively addressing the high demands of Industry 4.0 environments, they contribute to the development of networks that can dynamically adapt to changing requirements, support diverse applications, and maintain high service quality.

While AI-driven network management significantly advances the flexibility and intelligence of modern communication systems, critical challenges remain. Many proposed frameworks, particularly in 5G and 6G scenarios, depend on accurate, large-scale data for training and real-time decision-making, yet such data may be scarce or fragmented in emerging networks. Moreover, integrating federated learning and edge AI raises concerns about communication overhead, model convergence under heterogeneous conditions, and data privacy. Lastly, as AI assumes greater control over network infrastructure, the lack of standardized benchmarks for performance, reliability, and robustness becomes a barrier to broader deployment. These challenges underscore the importance of developing resilient, interpretable AgentAI systems tailored to the unique demands of next-generation networks.

5.5. Defence

In the defence sector, AgentAI supports advanced simulations and tactical decision-making, particularly in anti-submarine warfare (ASW). A context-aware combat helicopter computer-generated force (ACGF) model, based on an AgentAI framework integrating context-based reasoning (CXBR) and neural networks, autonomously or semi-autonomously executes tasks like target tracking, weapon engagement, and tactical decision-making, dynamically adapting to environmental feedback and leveraging tactical knowledge databases for realistic and effective ASW training. Using complex adaptive system (CAS) theory and the DI-Guy AI multi-agent simulation platform, marine unit offensive operations are modeled to simulate realistic military scenarios where autonomous agents, such as soldiers or vehicles, make decisions on pathfinding, attacking, and adapting to environmental factors, offering insights into complex systems warfare through simplified air operations that account for dual initiative and terrain-influenced maneuvers (Tiao-Ping et al., 2017).

In the same way that AgentAI frameworks support complex tactical decision-making, the RFES-ACO algorithm optimizes real-time fire evacuation efforts, adapting to dynamic fire scenarios and obstacles such as smoke (Yan et al., 2019). Mobile devices are optimized to enhance evacuation efficiency by using lightweight three-dimensional (3D) models and multi-threading, making the tool effective for conducting virtual fire drills. For example, Fig. 7 illustrates a simulation of a metro fire with varying numbers of agents, as presented in research work (Yan et al., 2019).

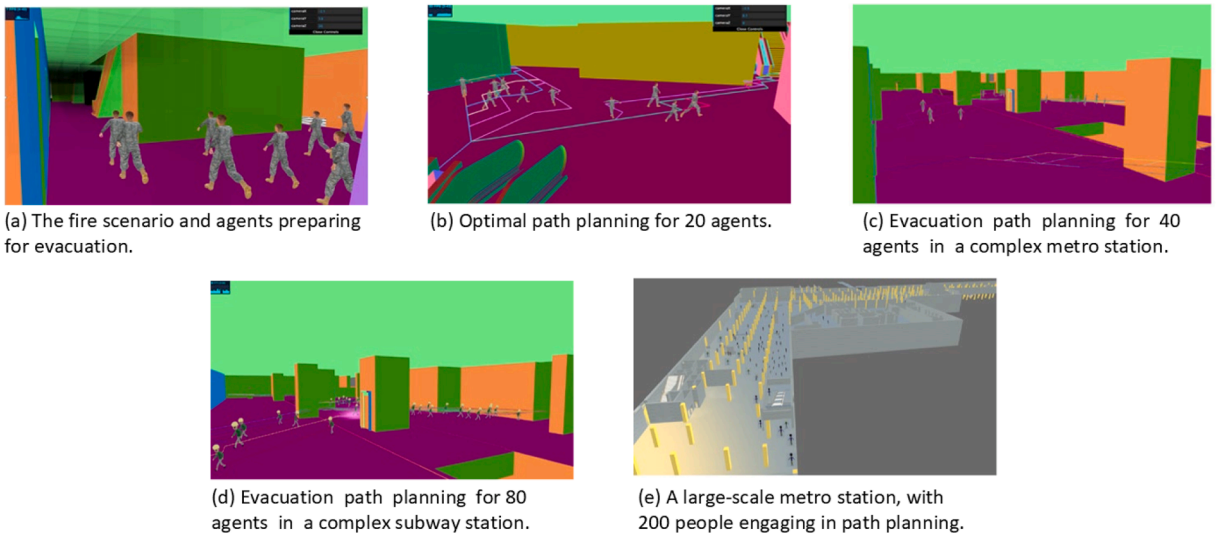

Fig. 7. The simulation of a metro fire with varying numbers of agents, as presented in Yan et al. (2019).

In summary, AgentAI applications showcase transformative capabilities in defence by enhancing tactical decision-making, enabling realistic military simulations, and improving emergency response systems. By leveraging advanced frameworks such as CAS theory, CXBR, and lightweight 3D modeling, these systems dynamically adapt to complex scenarios, providing more effective and responsive solutions in both military and emergency contexts. However, despite these advancements, current AgentAI-based simulations still face limitations in their real-world applicability. For instance, emergency response simulations, though optimized through lightweight 3D modeling and multithreading, may fall short in capturing the unpredictability and complexity of real-life disasters, thereby affecting their operational reliability. Additionally, the adaptive performance of such systems remains heavily dependent on the quality of initial modeling and training data, making them potentially vulnerable to inaccuracies when deployed in environments that diverge from the simulated conditions.

5.6. Gaming

In gaming, AgentAI is transforming the industry by enabling adaptive, realistic, and context-aware behaviors in non-player characters (NPCs) and multi-agent systems through advanced techniques.

For example, in strategic board games, AI agents exploit non-transitivity and uncertainty to develop adaptive strategies. For example, an exploitative approach allows AI to outperform Markov model-based agents in blackjack (Franklin & Markley, 2014; Pan et al., 2022), while chess-based research identifies circular strategy patterns and employs methods like Nash Clustering and fictitious play to enhance multi-agent performance through strategic flexibility (Sanjaya et al., 2022). AgentAI also integrates supervised learning to improve decision-making and modularity in real-time strategy games, enabling complex tasks to be executed with greater flexibility (Barriga et al., 2019).

Several studies have demonstrated how RL and Q-learning enable multi-agent AI systems to self-organize efficiently to allocate resources, with agents adapting collectively in dynamic environments (Zhang, Zhang, Chen, & Liu, 2020). For instance, Zhang et al. (2019) on cooperative behaviors in evolutionary games revealed that RL-driven agents exhibit cooperation and unique oscillatory patterns, as observed in game types like the snowdrift and rock-paper-scissors. Additionally, AI systems that use behavior models and 3D map structures can automatically evaluate spatial designs by simulating path finding and decision-making (Bai et al., 2021). A sound-based AI developed utilizing Proximal Policy Optimization (PPO) exhibits innovative applications by reacting to audio signals alone in a fighting game, thereby expanding possibilities for both sound design and AI-driven gaming (Bakhtin et al., 2022; Van Nguyen, Dai, Khan, Thawonmas, & Pham, 2022). Within the modular gaming context using methodologies such as Minimax, Monte Carlo Tree Search (MCTS), and evolutionary algorithms, the ColorShapeLinks framework employs AgentAI to construct AI agents (Fachada, 2021). This platform serves for teaching AI concepts, fostering creativity, and bridging education, research, and industry.

The examples presented reveal how AgentAI in gaming has advanced from static rule-based systems to highly adaptive, multi-agent frameworks capable of strategic reasoning, environmental sensing, and dynamic interaction. A common thread in these studies is the emphasis on learning, whether through reinforcement learning, supervised approaches, or evolutionary paradigms, as a driver of realism and emergent behavior within the game. Such forward strides also come with technical challenges. Generalizability to game genres or architectures remains low, as most systems are tailored to specific mechanics or datasets. Data scarsity is a problem for procedurally generated games, where agents must learn from limited or synthetic interaction patterns. Finally, modularity and audio input add expressiveness to agents but also make coordination and evaluation harder. As gaming increasingly converges with education, simulation, and research, the demand grows for AgentAI systems that are explainable, interoperable, scalable and that can adapt without sacrificing transparency or creative control.

5.7. Governance

Governance ensures ethics, transparency, accountability, and trust in AgentAI systems through policies, standards, and explainable AI systems that conform with legal and ethical norms (Chan et al., 2024; Srivastava et al., 2024). In governance, AgentAI models address the critical challenges of accountability (Chan et al., 2024), transparency (Srivastava et al., 2024), ethical decision-making, and user trust (Manzini et al., 2024; Roesler et al., 2025), which are of paramount importance in the context of Industry 4.0.

Research in this area focuses on the development of AI frameworks and governance mechanisms to ensure ethical AI deployment, incorporating predictive simulations and dynamic evaluations to ensure moral decision-making and trustworthiness in autonomous systems as demonstrated by multi-agent simulations like the Belief-Desire Model of Agency and Game of Giving and Asking for Reasons (GOGAR), which enable agents to justify claims, navigate discourses, and manage conflicting assertions through structured social interactions (Evans, 2016). Furthermore, the study (Manheim, 2019) delineates the failure modes characteristic of multi-agent systems, including coordination failures, adversarial conduct, and input spoofing, highlighting the risks in competitive environments. In addition, another study (Jebari & Lundborg, 2021) argued that systems like AlphaZero, regardless of their extensive specialization, lack the productive desires vital for general agency, thereby contesting the assertion that sophisticated intelligence can achieve artificial general intelligence or super intelligence solely through self-improvement.

Various studies (Cabitza et al., 2021; Choung et al., 2023; Gervais, 2023) emphasizes shifting the AI perspective as autonomous agents to Knowledge Artifacts (KA) that support human decision-making, emphasizing collaborative reasoning to mitigate challenges like automation bias and human prejudice, thereby enhancing decision-making quality in critical fields such as healthcare and finance. To ensure ethical behavior, research has focused on visibility, predictive simulation, and activity logging to ensure ethical behavior in autonomous AI, especially in high-stakes environments (Chan et al., 2024; Chandar et al., 2017). Study (Cioroaica, Buhnova, & Tomur, 2022) introduced a framework for developing artificial altruistic behavior, enabling AI agents to foresee and prevent harm through real-time consequence simulation.

However, AI systems–especially AI assistants–can create psychological distance that leads to unethical behavior or reduced social accountability (De Vreede et al., 2021; Gratch & Fast, 2022; Shams et al., 2022). This is particularly relevant in high-stakes environments, where trust and transparency are paramount. For instance, research into trust dynamics in AI decision support found that participants displayed a greater level of trust in human support compared to AI, particularly when the information provided was less quantifiable, underscoring the importance of clear and transparent AI systems (Roesler et al., 2025). To address this, the research (Manzini et al., 2024) proposes a socio-technical framework for evaluating AI trustworthiness at three levels–design, organizational practices, and third-party oversight–ensuring that AI systems act competently and align with user values.

Ethical guidance in autonomous AI also includes enhancing shared situation awareness and monitoring visibility for accountability. Studies (Triantafyllou, 2023; von Rütte, Anagnostidis, Bachmann, & Hofmann, 2024) have explored methods for guiding LLMs via latent space control, refining abstract concepts, and have proposed ethical decision-making methods for attribution of responsibility in multi-agent AI employing causality, gaming, and MDPs. Another study (Boyles, 2024) critiqued bottom-up artificial moral agents (AMAs), arguing that they lack true moral grounding because ethical judgments cannot emerge solely from factual data. Furthermore, in recommender systems enhancing shared awareness between humans and AI improves decision-making, aligning

human judgment with AI insights thus fostering balanced trust and performance (Srivastava et al., 2024).

In the context of AI governance, the risks of deception and manipulation posed by advanced generative AI, particularly large language models (LLMs), have been emphasized (Tarsney, 2025). These AI systems, with their near-human linguistic abilities, present increased dangers, especially regarding deception. AI-generated content should be held to stricter transparency standards than human content, with deceptive AI content defined as that which misleads individuals away from their ideal beliefs. To mitigate these risks, two strategies are proposed: implementing 'extreme transparency' for AI-generated content, ensuring clear disclosure of its origins and context, and establishing 'defensive systems' to provide users with contextual information to counter misleading statements.

The studies reviewed in this section converge on the necessity of AgentAI to design ethical, transparent, and accountable systems, particularly in high-stakes, decision-intensive environments. Across the various frameworks, simulations, and socio-technical models discussed, a single trend emerges: while AgentAI systems offer novel possibilities for monitoring, explanation, and coordination, they also expose persistent limitations in moral grounding, social accountability, and user trust. Contradictions arise especially in the conception of agency, between framing AI as autonomous decision-makers and knowledge artifacts in support of human judgment. Additionally, the need for visibility, predictive transparency, and value alignment is shared across applications, but practical implementation of these principles remains uneven. They need a deeper integration of technical design and normative frameworks to operationalize ethical decision-making, prevent abuse, and balance human-AI interaction in the frame of emerging governance architectures.

5.8. Marketing

In marketing, AgentAI is transforming the field by the introduction of advanced capabilities for simulating complex interactions, enhancing customer service Chong et al. (2021), and driving user engagement through intelligent systems (Baabdullah et al., 2022).

AI-powered chatbots are revolutionizing how businesses manage customer service, essentially serving as a jack of all trades in the retail sector while allowing for improved interactions and driving sales through better customer experiences. This study develops a framework that explores the effect of chatbot design on service delivery and customer perceptions, based on roles such as assistant, coach, and coworker. Through an analysis of agency: self, proxy, and collective, these roles can be employed to further enhance the customer experience in a meaningful and effective manner (Chong et al., 2021; He et al., 2023). AI-powered chatbots with managerial titles, such as AI manager enhance the customer satisfaction, purchase intentions, and brand perception. A study demonstrated that transitioning customers from AI agents to AI managers in service encounters positively impacts customer perceptions leading to boost trust, engagement, and marketing outcomes (Jeon, 2022). Another study (Baabdullah et al., 2022) demonstrated that interactive features in virtual agents including, personalization, responsiveness, transparency, and readability are essential for fostering a seamless user engagement and satisfaction, particularly in service-oriented sectors like shipping and courier, offering valuable insights for designing agents that enable meaningful continuous interactions.

A recent study explores the development of AgentAI for retirement planning scenarios, introducing the crewAI framework (Sepanosian et al., 2025). This system consists of four specialized agents–Policy, Industry, Advisory, and QA Experts–that sequentially collaborate to generate a comprehensive report. For testing, a fictional client, was used, with the system providing a detailed comparison of his performance against industry averages, assessing investments, and offering policy recommendations.

These applications demonstrate how AgentAI is revolutionizing marketing through more interactive, adaptive, and role-aware customer experience. From the transformation of chatbots into AI managers to agentic infrastructures that support complex financial planning, these systems exhibit a personalized, simulation-based form of engagement. But with greater autonomy comes the question: preserving consistency of user trust, managing data-driven personalization without compromising on transparency, and integrating multiple agent roles without watering down the customer journey. These emergent trends require greater emphasis on design ethics, explainability, and harmonization across systems to achieve the full potential of AgentAI in marketing usage.

5.9. E-Learning

E-Learning is intended to further advance educational experiences using personalized and interactive tools through AgentAI models (Lee et al., 2022). These systems offer dynamic, learner-centered experiences, using technologies like virtual reality, intelligent tutoring, and adaptive content delivery to support diverse educational needs.

An example is an ontology-based intelligent agent that updates its contents using fuzzy logic, neural networks, and evolutionary computation for high school students on the AI-Fuzzy Markup Language (FML) Metaverse platform to promote greater participation in computational intelligence (Khabarov & Volegzhanina, 2019; Lee et al., 2022). Therefore, it uses these things, which are called "learning thermometers" to measure student performance and adjust guidance on the fly based on real-time interaction in an experiential learning environment. The platform combines human intellect and ML to provide a cooperative learning experience, enabling students to apply computational intelligence principles to real-world practical problems.

AMBY is a web-based tool for middle school students to learn about AI literacy and to build and test their own conversational agents (Kumar et al., 2022). AMBY provides a clear and simple-to-use platform, with visual dialog management panel, a testing panel, and speech input around the system, that is easily usable for students regardless of individual level of computer literacy. The proposed model provides the ability to generate training datasets, customize avatars, and visualize a dialog flow in real-time, thereby enabling hands-on experience without prior programming knowledge. With iterative improvement based on the inputs of students and experts, the platform was successfully applied in a two-week summer camp where learners created conversational AI projects and hands-on experience learning about AI (Lee et al., 2023; Tian et al., 2023).

Another study (Kang et al., 2023) proposed AI-driven online dance education system that integrates a 2D pose estimation model with human tutor advice to enhance remote instructional experiences. A two-week case study showed its effectiveness in analyzing student movements in comparison to those of the teacher, enabling improved dance learning and collaborative refinement in remote environments. Fig. 8 shows the process of online learning conducted by AI agent and human teacher, proposed by this study (Kang et al., 2023).

Creativity-focused education has also benefited from AgentAI. A multi-agent AI system called Drawcto enables collaborative non-representational art generation with humans (Deshpande & Magerko, 2021). Inspired by Gestalt principles and learning-based techniques like SketchRNN, the system utilizes rule-based logic and enhances creativity and explainability in human-AI interaction (Deshpande & Magerko, 2021).

These diverse applications demonstrate the growing potential of AgentAI to enhance educational experiences across multiple dimensions. What unites these initiatives is a common goal of enabling more adaptive, engaging, and personalized learning environments. However, these systems also reveal a shared set of challenges that cut across technological and pedagogical lines. Many rely on real-time responsiveness and user-generated data, which introduce concerns related to data sparsity, system latency, and scalability across contexts with different access to

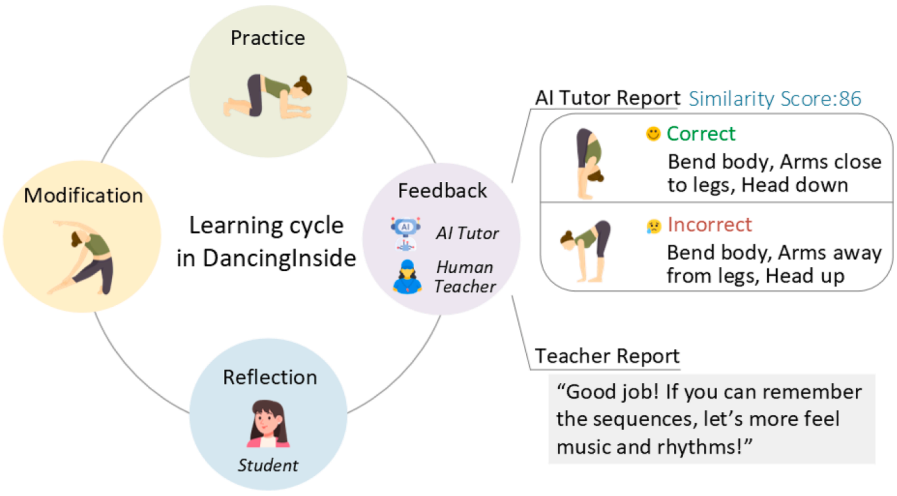

Fig. 8. The process of online learning conducted by AI agent and human teacher, as presented in Kang et al. (2023).

hardware or connectivity. Moreover, the increasing role of agents in instructional processes raises important questions about transparency, alignment with educational goals, and the division of responsibility between AI and human facilitators. Together, these aspects underscore the need for further cross-disciplinary research that not only advances the technical capabilities of AgentAI in E-Learning, but also addresses its integration into diverse educational ecosystems.

6. Discussion

6.1. Challenges and potential issues

Despite the notable progress of AgentAI technologies within Industry 4.0, such as improved automation, adaptive decision-making, and dynamic optimization–significant challenges remain, especially in dynamic environments characterized by high-modality observations. Agents often struggle with interpreting evolving goals, adapting to changing contexts, and processing visual inputs that influence both high-level intentions and low-level behaviors.

Empathetic reasoning in agents requires the ability to process heterogeneous object types and leverage common-sense knowledge for context-aware decision-making–capabilities critical for domains such as healthcare, governance, and human-facing applications like e-learning and personalized marketing. Seamless collaboration in such contexts demands that AgentAI systems comprehend nuanced goals and constraints, requiring not only advanced natural language processing but also interdisciplinary cooperation to address complex, unpredictable environments. A key tension in AgentAI systems is balancing autonomy with human oversight to ensure ethical, transparent, and efficient operation. This becomes increasingly important as agents take on roles that influence safety-critical and socially sensitive outcomes.

To address these limitations, future AgentAI frameworks must integrate real-time adaptability, context-aware reasoning, and privacy-preserving mechanisms at their core. As Industry 5.0 emerges, the emphasis is shifting from pure automation toward human-centric AI systems, where optimal collaboration between humans and agents is essential (Martini, Bellisario, & Coletti, 2024).

Ethical and societal considerations have long accompanied the evolution of AI. Traditionally seen as passive tools, AI agents are now becoming collaborative partners that influence and participate in decision-making processes (Cañas, 2022). Approaching AI ethics from this collaborative standpoint requires more than just aligning system outputs with predefined goals–it calls for careful design of shared responsibilities, mechanisms for mutual supervision, and alignment of joint outcomes with ethical norms. Ultimately, building a trustworthy human-AI partnership requires agents to support and enhance human capabilities while operating under shared ethical principles.

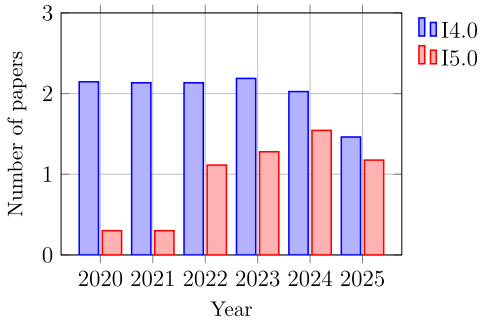

Fig. 9. Trends in the number of publications indexed in Scopus from 2020 to 2025 for Industry 4.0 and Industry 5.0 in AgentAI. The graph is displayed on a logarithmic scale to better highlight the relative changes over time. It can be observed that Industry 4.0 remains steady over the years while Industry 5.0 shows significant growth, reflecting increasing research interest in this new field.

6.2. Extending the agentAI paradigm: Toward responsible autonomy

As AgentAI transitions from its foundational role in Industry 4.0 to the more complex landscapes of Industry 5.0 and 6.0, its scope of influence is expanding from automation and optimization toward deeper autonomy, contextual understanding, and ethical alignment. As reflected in Fig. 9, recent publication trends underscore this shift, with growing interest in Industry 5.0 applications of AgentAI. This evolution brings both promising opportunities and significant challenges across technical, organizational, and societal dimensions.

In Industry 4.0, AgentAI has already demonstrated its value in domains such as energy management, traffic optimization, and intelligent manufacturing through multimodal data analysis and real-time responsiveness. However, these implementations have also surfaced limitations, including fragmented data silos, limited interoperability across systems, and poor generalizability of simulation-trained models. As these agents are increasingly deployed in high-stakes sectors such as healthcare and governance, the quality, fairness, and representativeness of their training data become critical–biased or incomplete data may lead to flawed, opaque, and even unsafe decisions.

The emergence of Industry 5.0 introduces a shift toward human-centric collaborative intelligence, emphasizing shared control, situational awareness, and empathetic AI reasoning (Adel & Alani, 2024).

AgentAI must therefore evolve from tools of automation to partners in decision-making–capable of ethical judgment, explainable reasoning, and privacy-aware action. Embedding governance frameworks into the system architecture is essential to ensure transparency, accountability, and respect for user autonomy. Concepts such as algorithmic explainability, bias mitigation, and the "right to be forgotten" must move from theoretical ideals to operational design principles.

Looking ahead, Industry 6.0 envisions AgentAI as a cognitively capable and self-regulating entity, able to learn and act autonomously in dynamic, uncertain environments (Lykov et al., 2024). These agents will harness the capabilities of quantum computing to overcome computational bottlenecks, integrate neural-symbolic reasoning to unify logic and learning, and leverage edge intelligence for secure, low-latency, decentralized coordination. Each agent will function as part of a distributed ecosystem–individually autonomous, collectively resilient, and ethically grounded.

These developments set the stage for a new question: What should AgentAI look like in the era of Industry 6.0? The following presents a strategic blueprint for designing intelligent, resilient, and ethically aligned agents in next-generation industrial systems.

6.3. Toward a blueprint for agentAI in industry 6.0

As Industry 6.0 advances toward highly autonomous, adaptive, and self-regulating systems, AgentAI is poised to evolve from task-specific utilities into intelligent, autonomous entities capable of high-level reasoning, social coordination, and decentralized infrastructure control. This transformation will be shaped by several key innovation domains:

A) *Emerging Multi-Agent Coordination*. Future industrial ecosystems will comprise vast networks of autonomous agents operating at the edge, across organizational layers and distributed physical environments. These agents must collaborate in real time through decentralized negotiation, collaborative planning, and distributed consensus protocols. For instance, in a global supply chain, agents representing suppliers, manufacturers, and logistics providers could dynamically negotiate contracts, reschedule production, and reroute deliveries in response to disruptions, enhancing both agility and resilience.

B) *Intention Inference and Contextual Reasoning*. Beyond coordination, AgentAI must be capable of inferring the latent intentions of humans and machines within complex, multimodal environments. This requires integrating behavioral signals, linguistic input, and sensor data to understand context and predict goals. In a smart factory, for example, an agent detecting a human operator's attempt to override a safety mechanism could proactively intervene or provide context-aware support, thereby improving safety and reducing cognitive burden.

C) *Decentralized Self-Management of Infrastructure*. Industry 6.0 envisions a paradigm shift from centralized control systems to distributed, self-governing infrastructure. Embedded within machines, sensors, and production units, AgentAI will autonomously manage local resources, monitor system health, and execute recovery actions. Through edge intelligence, blockchain-enabled trust mechanisms, and federated learning, these agents will form a cyber-resilient, energy-efficient, and self-healing industrial ecosystem.

Consider a smart manufacturing environment comprising thousands of autonomous production cells, each governed by local AgentAI. These agents continuously learn from localized feedback, share optimized workflows with peers, and autonomously coordinate task scheduling and predictive maintenance. Upon detecting an imminent mechanical failure, an agent could initiate repair protocols, notify neighboring agents to rebalance workloads, and update global production forecasts–all without centralized oversight.

Realizing this vision will demand interdisciplinary research spanning graph-based agent communication models, explainable multi-agent reinforcement learning, intent-aware planning algorithms, and ethical frameworks for trust-centric autonomy. These innovations are essential for establishing the technical, organizational, and regulatory foundations of the next-generation AgentAI ecosystem.

7. Conclusion

In this paper, we present a state-of-the-art literature review on the concept of AgentAI and its applications to Industry 4.0. In distributed AI, AgentAI applies a transformative paradigm that enables collaborative and autonomous decision-making and adapts to dynamic industrial environments. In the emerging context of Industry 4.0, the vital capabilities for improved efficiency and scalability are collaborative coordination, autonomous decision-making, and real-time adaptability. In this paper, our motivation arises from the lack of a comprehensive review and taxonomy of AgentAI methodologies in the Industry 4.0 context.

We first introduced our research methodology and provided an overview of the existing literature on AgentAI techniques and technologies of Industry 4.0. We also explored the benefits of implementing AgentAI in practical applications, emphasizing its role in improving efficiency, productivity, and safety within industrial settings. We have highlighted how AgentAI is transforming industries and solving several complex challenges through an analysis of diverse applications across various fields.

Additionally, we discuss the current challenges and propose potential future research directions in the emerging domains of Industry 5.0 and 6.0, where human-centricity, resilience, and intelligent automation are expected to play a central role. This survey is the first to provide a taxonomy of Industry 4.0 in AgentAI, serving as a foundation for further exploration and development in this area. Through our extensive discussions, we foster research interest in the field of AgentAI.

CRediT authorship contribution statement

Francesco Piccialli: Supervision, Validation, Project administration, Writing – review & editing; **Diletta Chiaro:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision; **Sundas Sarwar:** Investigation, Resources, Writing – original draft; **Donato Cerciello:** Formal analysis, Investigation, Writing – original draft, Supervision; **Pian Qi:** Visualization, Resources, Writing – review & editing, Supervision; **Valeria Mele:** Resources, Writing – review & editing, Supervision.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by PNRR Centro Nazionale HPC, Big Data e Quantum Computing, (CN 00000013) (CUP: E63C22000980007). The authors thank the IBiSco project (Infrastructure for BIg data and Scientific COmputing), PON R&I 2014–2020 under Call 424–2018 - Action II.1, for the support and use of the HPC Cluster. The authors extend special thanks to Dr. Luisa Carracciuolo and Eng. Davide Bottalico for their constant and continuous support in utilizing the IBiSco HPC Cluster.

References

Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous intelligence for complex goals–a comprehensive survey. *in IEEE Access*, *13*, 18912–18936. https://doi.org/10.1109/ACCESS.2025.3532853

Adel, A., & Alani, N. H. S. (2024). Human-centric collaboration and industry 5.0 framework in smart cities and communities: Fostering sustainable development goals 3, 4, 9, and 11 in society 5.0. *Smart Cities*, *7*(4), 1723.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, J., Jeffrey, K., … Zeng, A. (2022). Do as i can, not as i say: Grounding language in robotic affordances. arXiv preprint arXiv:2204.01691

Angelis, A., & Kousiouris, G. (2025). A survey on the landscape of self-adaptive cloud design and operations patterns: Goals, strategies, tooling, evaluation and dataset perspectives. 2503.06705.

Baabdullah, A. M., Alalwan, A. A., Algharabat, R. S., Metri, B., & Rana, N. P. (2022). Virtual agents and flow experience: An empirical examination of AI-powered chatbots. *Technological Forecasting and Social Change*, *181*, 121772.

Bahrin, M. A. K., Othman, M. F., Azli, N. H. N., & Talib, M. F. (2016). Industry 4.0: A review on industrial automation and robotic. *Jurnal teknologi*, *78*(6–13).

Bai, Y., Xue, Z., Zhang, Y., Wang, M., Ren, Y., & Tan, J. (2021). A method for testing the feasibility of 3d map design based on multi-agent AI-driven. In *2021 2nd international conference on big data economy and information management (BDEIM)* (pp. 429–433). IEEE.

Bakhtin, A., Wu, D. J., Lerer, A., Gray, J., Jacob, A. P., Farina, G., Miller, A. H., & Brown, N. (2022). Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. arXiv preprint arXiv:2210.05492

Barriga, N. A., Stanescu, M., Besoain, F., & Buro, M. (2019). Improving RTS game AI by supervised policy learning, tactical search, and deep reinforcement learning. *IEEE Computational Intelligence Magazine*, *14*(3), 8–18.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Proceedings of the 34th international conference on neural information processing systems* Red Hook, NY, USA: Curran Associates Inc..

Bougzime, O., Jabbar, S., Cruz, C., & Demoly, F. (2025). Unlocking the potential of generative AI through neuro-symbolic architectures: Benefits and limitations. 2502.11269.

Boyles, R. J. M. (2024). Can'T bottom-up artificial moral agents make moral judgements? *Filosofija. Sociologija*, *35*(1), 14–22.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712

Cabitza, F., Campagner, A., & Simone, C. (2021). The need to move away from agential-AI: Empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies*, *155*, 102696.

Cañas, J. J. (2022). Ai and ethics when human beings collaborate with ai agents. *Frontiers in Psychology*, *13*, 836650.

Cao, Y., Wang, R., Chen, M., & Barnawi, A. (2019). Ai agent in software-defined network: Agent-based network service prediction and wireless resource scheduling optimization. *IEEE Internet of Things Journal*, *7*(7), 5816–5826.

Chae, J., Lee, S., Jang, J., Hong, S., & Park, K.-J. (2023). A survey and perspective on industrial cyber-physical systems (ICPS): From ICPS to AI-augmented ICPS. *IEEE Transactions on Industrial Cyber-Physical Systems*, *1*, 257–272.

Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N. et al. (2024). Visibility into AI agents. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 958–973).

Chan, A. et al. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (p. 651–666). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3593013.3594033

Chandar, P., Khazaeni, Y., Davis, M., Muller, M., Crasso, M., Liao, Q. V., Shami, N. S., & Geyer, W. (2017). Leveraging conversational systems to assists new hires during onboarding. In *Human-computer interaction-INTERACT 2017: 16th IFIP TC 13 international conference, Mumbai, India, september 25–29, 2017, proceedings, part II 16* (pp. 381–391). Springer.

Chong, T., Yu, T., Keeling, D. I., & de Ruyter, K. (2021). Ai-chatbots on the services frontline addressing the challenges and opportunities of agency. *Journal of Retailing and Consumer Services*, *63*, 102735.

Choung, H., Seberger, J. S., & David, P. (2023). When AI is perceived to be fairer than a human: understanding perceptions of algorithmic decisions in a job application context. *International Journal of Human–Computer Interaction*, (pp. 1–18).

Cioroaica, E., Buhnova, B., & Tomur, E. (2022). Towards trusting the ethical evolution of autonomous dynamic ecosystems. In *Proceedings of the 1st workshop on software engineering for responsible AI* (pp. 13–16).

Colledanchise, M., & Ögren, P. (2018). Behavior trees in robotics and AI. https://doi.org/10.1201/9780429489105

Crowder, J. A. (2016). Ai inferences utilizing occam abduction. In *2016 annual conference of the north american fuzzy information processing society (NAFIPS)* (pp. 1–6). IEEE.

De Pace, F., Manuri, F., & Sanna, A. (2018). Augmented reality in industry 4.0. *American Journal of Computer Science and Technology*, *6*(1), 17.

De Silva, D., Mills, N., Moraliyage, H., Rathnayaka, P., Wishart, S., & Jennings, A. (2025). Responsible artificial intelligence hyper-automation with generative AI agents for sustainable cities of the future. *Smart Cities*, *8*(1), 34.

De Vreede, T., Raghavan, M., & De Vreede, G.-J. (2021). Design foundations for AI assisted decision making: a self determination theory approach.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2025). Deepseek-v3 technical report. 2412.19437.

Delanovic, A., Chiu, C., Kolen, J., Gnanasekaran, A., Surana, A., Srivastava, K., Zhu, H. A., Lin, Y., Bukingolts, N., Willis, D. et al. (2023). Results of the airlift challenge: A multi-agent AI planning competition. In *Artificial intelligence and machine learning for multi-domain operations applications v* (pp. 374–393). SPIE (*12538*).

Deshpande, M., & Magerko, B. (2021). Drawcto: A multi-agent co-creative AI for collaborative non-representational art. In *Aiide workshops*.

Doshi-Velez, F., & Kim, B. (2017a). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608

Doshi-Velez, F., & Kim, B. (2017b). Towards a rigorous science of interpretable machine learning. 1702.08608.

Durante, Z., Huang, Q., Wake, N., Gong, R., Park, J. S., Sarkar, B., Taori, R., Noda, Y., Terzopoulos, D., Choi, Y., Ikeuchi, K., Vo, H., Fei-Fei, L., & Gao, J. (2024). Agent AI: Surveying the horizons of multimodal interaction. arXiv preprint arXiv:2401.03568

Ellwart, T., & Kluge, A. (2019). Psychological perspectives on intentional forgetting: an overview of concepts and literature. *KI-Künstliche Intelligenz*, *33*(1), 79–84.

Evans, R. P. (2016). Computer models of constitutive social practice. *Fundamental issues of artificial intelligence*, Springer (pp. 391–411).

Fachada, N. (2021). Colorshapelinks: A board game AI competition for educators and students. *Computers and Education: Artificial Intelligence*, *2*, 100014.

Meta Fundamental AI Research Diplomacy Team (FAIR)† et al. (2022). Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, *378*(6624), 1067–1074. https://doi.org/10.1126/science.ade9097

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, (pp. 535–545).

Foerster, J., Nardelli, N., Farquhar, G., Afouras, T., Torr, P. H. S., Kohli, P., & Whiteson, S. (2017). Stabilising experience replay for deep multi-agent reinforcement learning. In *International conference on machine learning* (pp. 1146–1155). PMLR.

Franklin, D. M., & Markley, K. L. (2014). Multi-agent artificial intelligence in pursuit strategies: Breaking through the stalemate. In *The twenty-seventh international flairs conference*.

Gervais, D. J. (2023). Towards an effective transnational regulation of AI. *AI & Society*, *38*(1), 391–410.

Gomes, C. P., Bai, J., Xue, Y., Björck, J., Rappazzo, B., Ament, S., Bernstein, R., Kong, S., Suram, S. K., van Dover, R. B. et al. (2019). Crystal: A multi-agent ai system for automated mapping of materials' crystal structures. *MRS Communications*, *9*(2), 600–608.

Good, A. P., & Horn, E. (2025). Unlocking autism's complexity: The mov e initiative's path to comprehensive motor function analysis. *Frontiers in Integrative Neuroscience*, *18*, 1496165.

Gratch, J., & Fast, N. J. (2022). The power to harm: AI assistants pave the way to unethical behavior. *Current Opinion in Psychology*, *47*, 101382.

Gridach, M., Nanavati, J., Abidine, K. Z. E., Mendes, L., & Mack, C. (2025). Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. arXiv preprint arXiv:2503.08979

He, J., Piorkowski, D., Muller, M. J., Brimijoin, K., Houde, S., & Weisz, J. D. (2023). Understanding how task dimensions impact automation preferences with a conversational task assistant. In *AutomationXP@ CHI*.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G. D., Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S. et al. (2021). Knowledge graphs. *ACM Computing Surveys*, *54*(4), 1–37.

Hong, J.-W. (2018). Bias in perception of art produced by artificial intelligence. In *Human-computer interaction. interaction in context: 20th international conference, HCI international 2018, Las Vegas, NV, USA, july 15–20, 2018, proceedings, part II 20* (pp. 290–303). Springer.

Hou, X., Zhao, Y., & Wang, H. (2025). The next frontier of LLM applications: Open ecosystems and hardware synergy. 2503.04596.

Hu, Y., NA, A. N., Yellamati, D. D., & Goktas, Y. (2025). Leveraging generative AI tools for proactive risk mitigation in design. In *2025 annual reliability and maintainability symposium (RAMS)* (pp. 1–6). IEEE.

Huang, W., Abbeel, P., Pathak, D., & Mordatch, I. (2022). Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning* (pp. 9118–9147). PMLR.

Hubmann, C., Becker, M., Althoff, D., Lenz, D., & Stiller, C. (2017). Decision making for autonomous driving considering interaction and uncertain prediction of surrounding vehicles. In *2017 IEEE intelligent vehicles symposium (IV)* (pp. 1671–1678). https://doi.org/10.1109/IVS.2017.7995949

Hudson, D. A., & Manning, C. D. (2018). Compositional attention networks for machine reasoning. arXiv:1803.03067.

Iovino, M., Scukins, E., Styrud, J., Ögren, P., & Smith, C. (2022). A survey of behavior trees in robotics and AI. *Robotics and Autonomous Systems*, *154*, 104096. https://doi.org/10.1016/j.robot.2022.104096

Jebari, K., & Lundborg, J. (2021). Artificial superintelligence and its limits: Why alphazero cannot become a general agent. *AI & SOCIETY*, *36*(3), 807–815.

Jeon, Y. A. (2022). Let me transfer you to our AI-based manager: Impact of manager-level job titles assigned to AI-based agents on marketing outcomes. *Journal of Business Research*, *145*, 892–904.

Kagermann, H., Wahlster, W., & Helbig, J. (2013). Recommendations for implementing the strategic initiative industrie 4.0-final report of the industry 4.0 working group; communication promoters group of the industry-science research alliance, acatech: Frankfurt am main, Germany, Germany. https://www.acatech.de/Publikation/recommendations-for-implementing-the-strategicinitiativeindustrie, (pp. 4–0).

Kalyuzhnaya, A., Mityagin, S., Lutsenko, E., Getmanov, A., Aksenkin, Y., Fatkhiev, K., Fedorin, K., Nikitin, N. O., Chichkova, N., Vorona, V. et al. (2025). Llm agents for smart city management: Enhancing decision support through multi-agent ai systems. *Smart Cities (2624–6511)*, *8*(1).

Kang, J., Kang, C., Yoon, J., Ji, H., Li, T., Moon, H., Ko, M., & Han, J. (2023). Dancing on the inside: A qualitative study on online dance learning with teacher-AI cooperation. *Education and Information Technologies*, *28*(9), 12111–12141.

Katagami, D., Takaku, S., Inaba, M., Osawa, H., Shinoda, K., Nishino, J., & Toriumi, F. (2014). Investigation of the effects of nonverbal information on werewolf. In *2014 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 982–987). IEEE.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-HLT* (p. 2). Minneapolis, Minnesota (*vol. 1*).

Khabarov, V., & Volegzhanina, I. (2019). Knowledge management system of an industry-specific research and education complex. In *Iop conference series: earth and environmental science* (p. 012197). IOP Publishing (*vol. 403*).

Kim, S., Yu, Z., & Lee, M. (2017). Understanding human intention by connecting perception and action learning in artificial agents. *Neural Networks*, *92*, 29–38.

Kshetri, N. (2025a). Economics of agentic AI in the health-care industry. *IT Professional*, *27*(1), 14–19.

Kshetri, N. (2025b). From predictive and generative to agentic AI: shaping the future of marketing operations and strategies. *Computer*, *58*(4), 121–129.

Kumar, A., Tian, X., Celepkolu, M., Israel, M., & Boyer, K. E. (2022). Early design of a conversational ai development platform for middle schoolers. In *2022 IEEE symposium on visual languages and human-centric computing (VL/HCC)* (pp. 1–3). IEEE.

Lähdeaho, O., & Hilmola, O.-P. (2024). An exploration of quantitative models and algorithms for vehicle routing optimization and traveling salesman problems. *Supply Chain Analytics*, *5*, 100056.

Lee, C.-S., Wang, M.-H., Huang, S.-H., Yang, F.-J., Tsai, C.-H., & Wang, L.-Q. (2022). Fuzzy ontology-based intelligent agent for high-school student learning in AI-FML metaverse. In *2022 IEEE international conference on fuzzy systems (FUZZ-IEEE)* (pp. 1–8). IEEE.

Lee, K. A., Lim, S.-B., & Nagarajan, S. N. (2023). A study of the effectiveness of english speaking of teachable agent using AI chatbot. In *Icaart (1)* (pp. 308–314).

Li, C., Zheng, P., Li, S., Pang, Y., & Lee, C. K. M. (2022). Ar-assisted digital twin-enabled robot collaborative manufacturing system with human-in-the-loop. *Robotics and Computer-Integrated Manufacturing*, *76*, 102321.

Liu, Y., Singh, J., Liu, G., Payani, A., & Zheng, L. (2025). Towards hierarchical multi-agent workflows for zero-shot prompt optimization. arXiv preprint arXiv:2405.20252

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10.

Lykov, A., Cabrera, M. A., Konenkov, M., Serpiva, V., Gbagbe, K. F., Alabbas, A., Fedoseev, A., Moreno, L., Khan, M. H., Guo, Z. & Tsetserukou, D. (2024). Industry 6.0: New generation of industry driven by generative AI and swarm of heterogeneous robots. arXiv preprint arXiv:2409.10106

Manheim, D. (2019). Multiparty dynamics and failure modes for machine learning and artificial intelligence. *Big Data and Cognitive Computing*, *3*(2), 21.

Manzini, A., Keeling, G., Marchal, N., McKee, K. R., Rieser, V., & Gabriel, I. (2024). Should users trust advanced AI assistants? justified trust as a function of competence and alignment. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 1174–1186).

Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. 1904.12584.

Martini, B., Bellisario, D., & Coletti, P. (2024). Human-centered and sustainable artificial intelligence in industry 5.0: Challenges and perspectives. *Sustainability*, *16*(13), 5448.

Masterman, T., Besen, S., Sawtell, M., & Chao, A. (2024). The landscape of emerging AI agent architectures for reasoning, planning, and tool calling: A survey. arXiv preprint arXiv:2404.11584

Mayer, T. (2024). Future directions in cloud networking for AI and LLM applications. *Advances in Computer Sciences*, *7*(1).

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, *27*(4), 12.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Muzumdar, P., Bhosale, A., Basyal, G. P., & Kurian, G. (2024). Navigating the docker ecosystem: A comprehensive taxonomy and survey. *Asian Journal of Research in Computer Science*, *17*(1), 42–61. https://doi.org/10.9734/ajrcos/2024/v17i1411

Nam, K., Heo, S., Kim, S., & Yoo, C. (2023). A multi-agent AI reinforcement-based digital multi-solution for optimal operation of a full-scale wastewater treatment plant under various influent conditions. *Journal of Water Process Engineering*, *52*, 103533.

Nourani, C. F. (1999). Multiagent AI implementations an emerging software engineering trend. *Engineering Applications of Artificial Intelligence*, *12*(1), 37–42.

Oks, S. J., Jalowski, M., Lechner, M., et al. (2022). Cyber-physical systems in the context of industry 4.0: A review, categorization and outlook. *Information Systems Frontiers*, *26*, 1731–1772. https://doi.org/10.1007/s10796-022-10252-x

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). GPT-4 technical report. arXiv:2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E. et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ (Clinical Research edition)*, *372*.

Pan, Z., Xue, J., & Ge, T. (2022). Intuitive searching: An approach to search the decision policy of a blackjack agent. In *Proceedings of sixth international congress on information and communication technology: ICICT 2021, London, volume 2* (pp. 869–887). Springer.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, *113*, 54–71.

Qin, J., Liu, Y., & Grosvenor, R. (2016). A categorical framework of manufacturing for industry 4.0 and beyond. *Procedia CRIP*, *52*, 173–178.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763). PMLR.

Rahman, S. A., Chawla, S., Yaqot, M., & Menezes, B. (2025). Leveraging large language models for supply chain management optimization: A case study. In M. Dassisti, K. Madani, & H. Panetto (Eds.), *Innovative intelligent industrial production and logistics* (pp. 175–197). Cham: Springer Nature Switzerland.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., & Whiteson, S. (2020). Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, *21*(178), 1–51.

Rashid], S. M. Z. U., Montasir], I., Haq], A., & Ahmmed], M. (2025). Securing agentic AI: Threats, risks and mitigation. https://doi.org/10.13140/RG.2.2.18420.67206

Rehman, H. U., Pulikottil, T., Estrada-Jimenez, L. A., Mo, F., Chaplin, J. C., Barata, J., & Ratchev, S. (2021). Cloud based decision making for multi-agent production systems. In *Progress in artificial intelligence: 20th EPIA conference on artificial intelligence, EPIA 2021, virtual event, september 7–9, 2021, proceedings 20* (pp. 673–686). Springer.

Renjith, P. N., Alfurhood, B. S., Prashanth, K. V., Patil, V. S., Sharma, N., & Chaturvedi, A. (2024). Coordination of modular nano grid energy management using multi-agent AI architecture. *Computers and Electrical Engineering*, *115*, 109112.

Roesler, E., Rieger, T., & Langer, M. (2025). Numeric vs. verbal information: The influence of information quantifiability in human-AI vs. human-human decision support. *Computers in Human Behavior: Artificial Humans*, *3*, 100116. https://doi.org/10.1016/j.chbah.2024.100116

Sangeetha, V., Vamsidharan, V., Saran, R., & Shrikesh, S. P. (2022). AI interfaced learning module for road safety using virtual reality. In *2022 international conference on applied artificial intelligence and computing (ICAAIC)* (pp. 60–65). IEEE.

Sanjaya, R., Wang, J., & Yang, Y. (2022). Measuring the non-transitivity in chess. *Algorithms*, *15*(5), 152.

Schlichtkrull, M., Kipf, T. N., Bloem, P., van den Berg, R., Titov, I., & Welling, M. (2017). Modeling relational data with graph convolutional networks. 1703.06103.

Senjab, K., Abbas, S., Ahmed, N., & Khan, A. u. R. (2023). A survey of kubernetes scheduling algorithms. *Journal of Cloud Computing*, *12*(1), 87.

Sepanosian, T., Milosevic, Z., & Blai, A. (2025). Scaling AI adoption in finance: Modelling framework and implementation study. In *Enterprise design, operations, and computing. EDOC 2024 workshops* (pp. 221–236). Springer Nature Switzerland.

Shams, P., Beynier, A., Bouveret, S., & Maudet, N. (2022). Fair in the eyes of others. *Journal of Artificial Intelligence Research*, *75*, 913–951.

Shankar, V. (2024). Managing the twin faces of AI: A commentary on "is AI changing the world for better or worse?". *Journal of Macromarketing*, (p. 02761467241286483).

Sharma, S., Devreaux, P., Sree, S., Scribner, D., Grynovicki, J., & Grazaitis, P. (2019). Artificial intelligence agents for crowd simulation in an immersive environment for emergency response. *Electronic Imaging*, *31*, 1–8.

Shavit, Y. et al. (2023). Practices for governing agentic AI systems. *Research Paper, OpenAI*. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. 2002.04087.

Srivastava, D., Lilly, J. M., & Feigh, K. M. (2024). Exploring the role of judgement and shared situation awareness when working with AI recommender systems. *Cognition, Technology & Work*, (pp. 1–18). https://doi.org/10.1007/s10111-024-00771-9

Brooks, R.A. (1995). Intelligence without reason. *Proceedings of the 12th international joint conference on artificial intelligence*, *1*, (pp. 569–595). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc..

Tarsney, C. (2025). Deception and manipulation in generative AI. *Philosophical Studies*. https://doi.org/10.1007/s11098-024-02259-8

Team, G., et al. (2024). Gemini: A family of highly capable multimodal models. 2312.11805.

Tian, X., Kumar, A., Solomon, C. E., Calder, K. D., Katuka, G. A., Song, Y., Celepkolu, M., Pezzullo, L., Barrett, J., Boyer, K. E. et al. (2023). Amby: A development environment for youth to create conversational agents. *International Journal of Child-Computer Interaction*, *38*, 100618.

Tiao-Ping, F., Jing, Q., & Jian-Hui, Z. (2017). Marine unit offensive operation simulation based on multi-agent. In *2017 IEEE international conference on unmanned systems (ICUS)* (pp. 424–428). IEEE.

Tortorelli, A., Sabina, G., & Marchetti, B. (2024). A cooperative multi-agent q-learning control framework for real-time energy management in energy communities. *Energies (19961073)*, *17*(20).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F. et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971

Triantafyllou, S. (2023). Forward-looking and backward-looking responsibility attribution in multi-agent sequential decision making. In *Proceedings of the 2023 international conference on autonomous agents and multiagent systems* (pp. 2952–2954).

Triem, H., & Ding, Y. (2024). "Tipping the balance": Human intervention in large language model multi-agent debate. *Proceedings of the Association for Information Science and Technology*, *61*(1), 361–373.

Valladares, W., Galindo, M., Gutiérrez, J., Wu, W.-C., Liao, K.-K., Liao, J.-C., Lu, K.-C., & Wang, C.-C. (2019). Energy optimization associated with thermal comfort and indoor air control via a deep reinforcement learning algorithm. *Building and Environment*, *155*, 105–117.

Valmeekam, K., Marquez, M., Sreedharan, S., & Kambhampati, S. (2023). On the planning abilities of large language models: a critical investigation. *Proceedings of the 37th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc..

Van Nguyen, T., Dai, X., Khan, I., Thawonmas, R., & Pham, H. V. (2022). A deep reinforcement learning blind AI in darefightingICE. In *2022 IEEE conference on games (cog)* (pp. 632–637). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., (2017). Attention is all you need. *Proceedings of the 31st international conference on neural information processing systems*, (p. 6000–6010). Red Hook, NY, USA: Curran Associates Inc..

von Rütte, D., Anagnostidis, S., Bachmann, G., & Hofmann, T. (2024). A language model's guide through latent space. *Proceedings of the 41st international conference on machine learning* (pp. 33). Vienna, Austria: JMLR.org.

von Rütte, D., Anagnostidis, S., Bachmann, G., & Hofmann, T. (2024). A language model's guide through latent space. arXiv preprint arXiv:2402.14433

Wang, D., Wang, Y., Jiang, X., Zhang, Y., Pang, Y., & Zhang, M. (2024a). When large language models meet optical networks: paving the way for automation. *Electronics*, *13*(13), 2529.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y. et al. (2024b). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, *18*(6), 186345.

Xie, J., Ajagekar, A., & You, F. (2023). Multi-agent attention-based deep reinforcement learning for demand response in grid-responsive buildings. *Applied Energy*, *342*, 121162.

Xiong, W., Hoang, T., & Wang, W. Y. (2018). Deeppath: A reinforcement learning method for knowledge graph reasoning. 1707.06690.

Yager, K. G. (2024). Towards a science exocortex. *Digital Discovery*, *3*(10), 1933–1957.

Yan, F., Jia, J., Hu, Y., Guo, Q., & Zhu, H. (2019). Smart fire evacuation service based on internet of things computing for web3d. *Journal of Internet Technology*, *20*(2), 521–532.

Artime, O., De Domenico, M., (2022). From the origin of life to pandemics: emergent phenomena in complex systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *380*(2227). https://doi.org/10.1098/rsta.2020.0410

Youvan, D. C. (2024). From aristotle to AI: A philosophical journey of holism and reductionism in the context of emergent phenomena.

Zha, L., Cui, Y., Lin, L.-H., Kwon, M., Arenas, M. G., Zeng, A., Xia, F., & Sadigh, D. (2024). Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. In *2024 IEEE international conference on robotics and automation (ICRA)* (pp. 15172–15179). IEEE.

Zhang, C., Vinyals, O., Munos, R., & Bengio, S. (2018). A study on overfitting in deep reinforcement learning. arXiv preprint arXiv:1804.06893

Zhang, R., Tang, S., Liu, Y., Niyato, D., Xiong, Z., Sun, S., Mao, S., & Han, Z. (2025). Toward agentic AI: Generative information retrieval inspired intelligent communications and networking. arXiv preprint arXiv:2502.16866

Zhang, S.-P., Zhang, J.-Q., Chen, L., & Liu, X.-D. (2020). Oscillatory evolution of collective behavior in evolutionary games played with reinforcement learning. *Nonlinear Dynamics*, *99*(4), 3301–3312.

Zhang, S.-P., Zhang, J.-Q., Huang, Z.-G., Guo, B.-H., Wu, Z.-X., & Wang, J. (2019). Collective behavior of artificial intelligence population: Transition from optimization to game. *Nonlinear Dynamics*, *95*, 1627–1637.

Zhang, X., & Zhu, Q. (2023). Ai-enabled network-functions virtualization and software-defined architectures for customized statistical qos over 6g massive mimo mobile wireless networks. *IEEE Network*, *37*(2), 30–37.

Zhang, Z., Kircher, K. J., Cai, Y., Brearley, J. G., Birge, D. P., & Norford, L. K. (2023). Mitigating peak load and heat stress under heatwaves by optimizing adjustments of fan speed and thermostat setpoint. *Journal of Building Performance Simulation*, *16*(4), 493–506.

Nguyen, Z., Annunziata, A., Luong, V., Dinh, S., Le, Q., Ha, A.H., Le, C., Phan, H.A., Raghavan, S., Nguyen, C., (2024). Enhancing Q&A with domain-specific fine-tuning and iterative reasoning: a comparative study. arXiv:2404.11792