Literature Review on Agentic AI

Seminar paper

Ehmann, Alexa-Sophie, FH Wedel, Wedel, Germany, stud106926@fh-wedel.de

Abstract

Agentic AI: autonomous systems that plan, decide, and act independently, represents a paradigmatic shift in Artificial intelligence, yet the field lacks conceptual clarity. Following Webster and Watson (2002), a systematic literature review across four academic databases yielded 42 relevant sources. Findings identify four defining characteristics: autonomy, tool use, persistent memory, and goal-directedness. The ReAct framework, hierarchical planning approaches, and multi-agent systems are currently the dominant research areas. While hallucination, prompt injection, and coordination errors remain key challenges. For IT management, three action areas emerge: defining autonomy boundaries, building agent-aware security architectures, and prioritizing use cases systematically.

Keywords: Agentic AI, AI Agents, Autonomy, IT Management

Table of Contents

1 Introduction

Artificial intelligence (AI) has undergone remarkable developments in recent years. Earlier systems were essentially answering machines, you asked a question, you got an answer. With the development of agentic AI, this is changing fundamentally. These systems no longer merely react, but actively pursue goals, make decisions independently, and carry out actions. Technically speaking, this represents a shift away from stateless, prompt-driven models toward systems that can perceive, plan, act, and adapt (Wang et al., 2024).

This transformation goes far beyond technology. It raises entirely new questions for organizations and their IT management. Recent figures show that the topic has long since become a reality: KPMG has already dubbed 2025 the "Year of Agentic AI." A survey from the fourth quarter of 2024 shows that over half of the organizations surveyed (51%) are actively exploring the use of AI agents, 37% are already testing the technology in concrete pilot projects (KPMG, 2024). At the same time, a significant gap exists between technological progress and organizational maturity. According to a 2025 McKinsey study, while nearly nine in ten companies regularly use AI, only about one-third have truly moved beyond the pilot and experimentation phase (McKinsey & Company, 2025).

Despite this growing significance, the scientific picture remains inconsistent. The field is evolving rapidly that even basic terms are not used consistently, modern neural systems and older symbolic models are often lumped together. Abou Ali et al. (2026) refer to this as "conceptual retrofitting." To bring clarity to this issue and organize the state of research, a systematic review of the existing literature is needed.

That is precisely the aim of this seminar paper. It examines the conceptual foundations of Agentic AI, its core capabilities and levels of autonomy, as well as technical approaches and practical applications. Based on this, conclusions are drawn for IT management. The central research question is:

"What is the current state of research on Agentic AI?"

The paper follows a clear structure: Chapter 2 introduces the conceptual foundations and distinguishes Agentic AI from related concepts. Chapter 3 outlines the methodology of the literature review, while Chapter 4 presents the key findings from the literature. These results are discussed in Chapter 5 drawing implications for IT Management and identifying open research questions/identifying research gaps. The paper concludes with a summary in Chapter 6.

2 Background

2.1 From Artificial Intelligence to Generative AI

The term "artificial intelligence" was established by John McCarthy in 1956, with the goal of building machines that could mimic or even surpass human thought and action. Russell and Norvig (2021) describe AI as the science of rational agents: systems that can act autonomously, perceive their environment, adapt over time, and pursue their own goals.

Within this broad field, generative AI has developed a dynamic all its own since around 2022. Systems like ChatGPT have demonstrated on a large scale for the first time what is possible: language models that generate coherent text, code, or other content based on simple inputs. Agentic AI takes a decisive step further. Whereas Generative AI relies on prompt-driven interaction and comparatively shallow reasoning, agentic systems fundamentally. It restructures these capabilities, they interact with environments and tools, plan several steps ahead, and can reflect on their own reasoning (Plaat, A. et al., 2025).

The difference is not a gradual one, but a fundamental one. Generative models follow the prompt, while agent-based models follow a goal. While the Generative models are based on large neural networks trained on static datasets, agentic systems integrate additional components, such as reinforcement learning, search algorithms, multi-agent frameworks, and control loops, that enable learning through interaction and feedback (Plaat, A. et al., 2025).

2.2 The concept of agency: What distinguishes an agent from a LLM?

The question of what defines an agent has long been a focus of AI research. Russell and Norvig (1995) defined an agent simply as anything that can perceive its environment and interact with it. Inputting data via sensors and outputting actions via actuators. This deliberately open-ended definition encompasses everything from a simple thermostat to highly complex autonomous systems. Agency is therefore not a yes-or-no proposition, but a spectrum.

In the context of today's AI, this distinction becomes more concrete and relevant. A large language model such as GPT-4 or Claude is, in its basic form, primarily a statistical transformation model: input in, output out, without its own goals, without tools, without independent action. An AI agent, on the other hand, goes far beyond this. It independently designs workflows, makes decisions, solves problems, and interacts with external systems and environments (IBM, 2025).

Two characteristics are particularly clear here: Autonomy and tool use. Recent literature adds further features that distinguish agentic AI systems from individual AI agents: multi-agent collaboration, dynamic task decomposition, persistent memory, and coordinated autonomy (Sapkota et al., 2025).

In summary, the distinction can be illustrated on three levels:

Dimension	**LLM(Basic)**	**AI-Agent**	**Agentic AI**
Control	Prompt-driven	Goal-driven	Self-directed
Tool Use	None	Limited (APIs)	Dynamic, multi-tool
Memory	Context window	Session-based	Persistent
Autonomy	Reactive	Semi-autonomous	Highly autonomous
Coordination	Single model	Single agent	Multi-agent systems

Table 1: Distinction between LLM – AI Agent – Agentic AI (own representation based on Abou Ali et al., 2026; Wang et al., 2024)

This three-way distinction serves as a common thread throughout the entire literature review and forms the conceptual lens through which the following chapters examine the research literature.

3 Literature Review

3.1 Methodological Approach

Methodologically, this literature review follows the framework of Webster and Watson (2002), widely used in business informatics for concept-centered literature reviews. Webster and Watson describe what such a review is intended to achieve systematically synthesizing knowledge, stimulating theoretical development, providing an overview of a research field, and highlighting blind spots. Unlike quantitative reviews, such as those common in medicine, this approach is particularly well-suited for fields in which conceptual, qualitative, and empirical work coexist and together form a coherent picture (Webster and Watson, 2002).

For this work, this is the right choice for one reason. Agentic AI is a field that is still in the process of being fully established. The literature is heterogeneous, the terminology is sometimes inconsistent, and a rigid quantitative approach would not do justice to this reality.

3.2 Search Strategy and Databases

For literature-search I scanned four databases that are particularly relevant to business informatics and related fields: Google Scholar, AISeL (AIS eLibrary), Web of Science, and arXiv for recent preprints in AI and machine learning. Since Agentic AI only really gained momentum as an independent field of research starting around 2022, my search was limited to publications between 2019 and 2025, with a clear focus on the years 2023 to 2025.

The search terms I defined were specifically tailored to AI- and agent-specific terminology and used both individually and in combination:

- “Agentic AI”
- “AI agent” / “autonomous agent”
- “LLM-based agent” / “LLM agent”
- “Artificial Intelligence agent”
- “multi-agent system”
- “AI autonomy”

The search primarily targeted the titles, abstracts, and keywords of the respective publications, to ensure that the results are actually relevant to the topic and not just tangentially related to it.

3.3 Inclusion and exclusion criteria

To ensure that the source pool is robust, clear inclusion and exclusion criteria were established.

I included sources that meet the following conditions:

- Peer-reviewed journal articles, conference papers (particularly A- and B-ranked outlets according to VHB-JOURQUAL), and highly cited preprints
- Published between 2019 and 2025
- Relevant in content to the definition, architecture, capabilities, or application areas of AI agents and agentic AI
- Publications in English

The following were excluded:

- Purely hardware-oriented papers with no reference to AI agents

- Gray literature without recognizable scientific quality assurance
- Works in which AI agents appear only marginally

3.4 Overview of the Selected Literature

During the database search I found several hundred initial hits. The high volume itself is revealing search terms like “agentic AI” or “LLM-based agent” return not only scientific contributions but also white papers, blog posts, and vendor documentation, materials that do not meet the quality standards set in Section 3.3. Systematic titles and abstract screening were therefore essential, not merely procedural. After applying the inclusion and exclusion criteria, 42 sources were retained for the final review. Notably, the vast majority of this date from 2023 onward. Earlier publications on autonomous systems or symbolic AI were included only where they provided foundational theoretical grounding, for instance, Russell and Norvig (2021) on rational agency or Webster and Watson (2002) on literature review methodology.

The resulting corpus is deliberately heterogeneous in structure. Its methodological and theoretical core consists of peer-reviewed articles from academic journals and established review papers, particularly from journals such as *Artificial Intelligence Review, Information Fusion,* and *Neurocomputing*, as well as foundational works from conference proceedings (e.g., ICLR) and standard texts such as Russell and Norvig (2021). These peer-reviewed sources constitute the largest portion of the literature base (approx. 50–55%).

In addition, recent conference papers and high-quality preprints from the arXiv community were included, particularly in cases where, due to the highly dynamic nature of the research field, no final, peer-reviewed publications are yet available (approx. 25–30%).

Furthermore, so-called gray literature in the form of industry and consulting reports (including Gartner, McKinsey, KPMG, and IBM) was integrated, provided that these provide empirical insights into the acceptance, market penetration, and practical implementation of agentic AI systems, which have so far been only limitedly available in the scientific literature (approx. 15–20%).

This composition reflects a deliberate methodological choice: excluding preprints and practitioner reports entirely would have left a significant blind spot, since much of the technically substantive work on agentic architecture has not yet completed peer review. At the same time, this concession to recency comes at a cost to verifiability, a limitation addressed in Section 5.4.

Table 2 maps 16 of these sources against the five analytical dimensions used throughout this review. Peripheral aspects such as general LLM benchmarking or AI ethics without direct relevance to agentic system design were excluded from the mapping, even if retained in the broader literature corpus. The resulting pattern is instructive: coverage of conceptual foundations and autonomy levels is relatively dense, whereas systematic empirical work on real-world deployments and security risks remains sparse. This asymmetry directly informs the research gaps identified in Section 5.3.

Author	Conceptual Foundations	Levels of Autonomy	Technical Approaches	Real-world use cases	Challenges & Risks
(Abou Ali et al., 2026)	•			•	•
Feng et al. (2025)		•			
Gartner (2024)				•	•
IBM (2025)	•				
Kasirzadeh & Gabriel (2025)		•			
Kim et al. (2026)					•
KPMG (2024)				•	
McKinsey & Company (2025)				•	
Plaat et al. (2025)	•				
(Raza et al., 2026)					•
Russell & Norvig (2021)	•				
(Sapkota et al., 2025)	•				
(Wang et al., 2024)	•		•		
Webster & Watson (2002)					
Xu et al. (2025)	•		•		
Yao et al. (2023)		•	•		

Table 2: Concept Matrix

4 Findings

4.1 Conceptual Foundations of Agentic AI

In the literature, agentic AI is recognized as being a reorientation rather that a further development of previous AI paradigms, not as a further development, but as a reorientation. Modern agentic systems plan proactively, possess contextual memory, use tools in sophisticated ways, and adapt their behavior based on environmental feedback. They thus act less as on-demand problem solvers and more as collaborative partners (Abou Ali et al., 2026).

When examining the architecture, two dominant lines can be distinguished in the literature: symbolic-classical and neural-generative approaches. Symbolic systems utilize algorithmic planning and persistent state, whereas neural-generative systems operate based on stochastic generation and prompt-driven orchestration. If both lines are uncritically mixed, which is not uncommon in the literature, this is referred to as "conceptual retrofitting" (Abou Ali et al., 2026).

The Belief-Desire-Intention (BDI) model has established itself as a common theoretical foundation. It provides a useful framework. Beliefs represent the agent's perception of the state of the world and its memory, desires represent its goals and constraints, and intentions represent the specific plans and tool invocations it pursues. Reactive architectures map perceptions directly to actions; deliberative architectures, on the other hand, maintain explicit world models and select actions through search and planning (Wang et al., 2024).

Significant progress has also been made at the perception and memory levels. Agents today no longer process only text, but also visual, multimodal, and auditory inputs, which enables, for example, the interpretation of complex graphical interfaces. Memory has evolved from short-lived context windows to persistent storage based on vector databases and Retrieval Augmented Generation (RAG), thereby supporting continuity over long periods of time (Xu et al., 2025).

4.2 Levels of autonomy

One of the central questions in the literature is: At what point is a system autonomous enough to be considered an agent? A common answer to this question is a taxonomy with five levels of autonomy. Ranging from simple, heavily controlled systems (Level 1) to fully autonomous ones (Level 5). The distinguishing criteria here are the breadth of functions and the degree of human control (Yao et al., 2023).

A slightly different approach describes autonomy not in terms of system capabilities, but in terms of the role that humans still play in the interaction process, ranging from operator to collaborator and advisor to approver and mere observer. This model makes it clear that autonomy is not an either/or proposition, but a fluid spectrum on which systems are positioned differently depending on the use case (Feng et al., 2025).

Kasirzadeh and Gabriel (2025) go one step further and propose a four-dimensional framework that describes AI agents along the dimensions of autonomy, effectiveness, goal complexity, and generality. These so-called "agent profiles" help to make the governance challenges of different system classes tangible. In practice, according to the consensus in the literature, most current systems fall within the middle range of this spectrum, roughly levels 3 to 4, where humans retain strategic control and act as a validation authority (Kasirzadeh & Gabriel, 2025).

4.3 Technical Approaches

On a technical level, several architectural patterns have emerged in the literature that dominate the discourse. Arguably the most influential of these is the ReAct framework. Developed by. It combines chain-of-thought reasoning with the use of external tools. A ReAct agent operates within a simple yet effective cycle: think, act, observe, and repeat this until the task is completed (Yao et al., 2023).

In addition, the literature describes hierarchical planning architectures and multi-agent systems as increasingly important patterns. A common approach breaks down LLM-based agents into six modular dimensions. Core components (perception, memory, action, profile), cognitive architecture (planning and reflection), learning, multi-agent systems, environments, and evaluation. At the same time, a shift is emerging, away from rigid API calls and toward open standards such as the Model Context Protocol (MCP) and native computer use (Xu et al., 2025).

MCP has established itself as the de facto standard for the integration layer. It solves a practical problem referred to in the literature as the N×M integration problem without a common standard, each of the N AI applications would need its own connectors for each of the M data sources, an exponentially growing effort. MCP reduces this complexity to N×M implementations, making the infrastructure significantly leaner and more maintainable (Xu et al., 2025).

4.4 Real-world use cases

The literature uses concrete examples to demonstrate where agent-based systems are already making a demonstrable difference today.

In software engineering, studies show that agents are increasingly capable of independently handling complex, multi-step tasks, such as troubleshooting, code reviews, and test automation. In healthcare, agentic AI enables the comprehensive analysis of patient data, clinical notes, and operational systems. Multi-agent frameworks take on the role of specialized teams: different agents cross-check results, orchestrate workflows, and ground their outputs in verified medical literature using RAG, which significantly improves accuracy (Abou Ali et al., 2026). In the finance and compliance sectors, agents support continuous risk assessment, underwriting, and rapid response to market changes.

Current forecasts indicate that the topic is on the verge of entering the mainstream. Gartner expects that by 2028, one-third of all enterprise software will already have integrated agentic AI capabilities. Compared to less than one percent in 2024. At the same time, Gartner warns that over 40 percent of agent-based AI projects could be discontinued by 2027 because the business value remains unclear (Gartner, 2024).

A common thread runs through the literature on what successful use cases have in common, clearly defined workflows, measurable goals, and well-designed human-machine interfaces. Where tasks are vague and data is scarce, however, even agent-based systems quickly reach their limits.

4.5 Challenges and Risks

As promising as the potential applications are, the literature also makes it clear that agent-based systems face serious challenges. These can be broadly divided into two categories: technical and security-related risks.

On the technical level, hallucination is the most frequently discussed problem. LLM-based agents produce outputs that sound plausible but are factually incorrect, and when running autonomously, this can lead to a cascade of subsequent errors. This is particularly critical for the

reasoning component. Even small changes to the input can lead to misalignment or incorrect conclusions. Added to this is the problem of coordination errors in multi-agent systems, where the interaction of multiple agents can lead to emergent, hard-to-trace behaviors, in the worst case, cascading errors that ripple through the entire system (Raza et al., 2026).

From a security perspective, prompt injections are the most dangerous attack vector. In multi-agent configurations, one agent's output becomes the input for the next. This opens up attack vectors that do not exist in traditional single-system architectures. Malicious content can spread from agent to agent like an infection and specifically manipulate communication channels. Added to this are so-called trust-authorization mismatches, agents trust other agents or services too blindly and perform actions that go far beyond their intended scope (Kim et al., 2026).

A real-world example demonstrates that these are not purely theoretical scenarios. The Echo-Leak exploit against Microsoft Copilot made it clear how manipulated email messages could cause an agent to automatically exfiltrate sensitive data, without any user interaction whatsoever. This underscores an important practical implication. Security architecture for agent-based systems must not be limited to the model level. They must equally incorporate memory, tools, communication channels, and human oversight (Kim et al., 2026).

5 Discussion

5.1 Similarities, Differences, and Contradictions in the Literature

Overall, the reviewed literature presents a largely coherent picture, although several notable contradictions remain.

There is broad agreement on one point. Agentic AI is not a gradual improvement on existing systems, but a genuine paradigm shift. Autonomy, tool use, persistent memory, and goal-directedness are cited as constitutive features in nearly all sources reviewed (Abou Ali et al., 2026; Sapkota et al., 2025; Wang et al., 2024).

Researchers are less in agreement regarding terminology. Sapkota et al. (2025) make a clear distinction between the AI agent as an individual system and Agentic AI as a higher-level, multi-agent paradigm. Other sources, however, use both terms synonymously. This is no coincidence, but rather symptomatic of a fundamental problem in the field. A consolidated, interdisciplinary terminology does not yet exist. Abou Ali et al. (2026) summarize this when they speak of "conceptual retrofitting", older concepts are retroactively assigned new terms without clear dividing lines being drawn.

Opinions also differ regarding the degree of autonomy. Yao et al. (2023) and Feng et al. (2025) describe autonomy as a spectrum and each develop their own taxonomies, which, however, do not fully converge. Kasirzadeh and Gabriel (2025) bring to the forefront a question that has so far received little technical attention. At what point does autonomy actually become irresponsible? The fact that this normative dimension has so far received limited attention in the technical literature represents one of the central gaps identified in this review.

5.2 Implications for IT management

Strategic Level: Agentic AI as an Organizational Issue

The literature is consistent in emphasizing one key point. The successful deployment of agent-based systems is not purely a technical matter. Before organizations consider implementation, they must answer a more fundamental question, namely, which tasks and decisions should be delegated to autonomous systems, and which should not.

To define the limits of autonomy, specifically to determine at what level of autonomy spectrum system operate lies within the specific responsibility of IT Management. Kasirzadeh and Gabriel (2025) make it clear that this is not a one-time decision, but an ongoing governance task, because as system capabilities grow, these boundaries must be regularly renegotiated.

The figures demonstrate that the topic has long since entered practical application. According to KPMG (2024), 37% of the companies surveyed are already testing agent-based AI in pilot projects. However, the McKinsey study (2025) reveals the flip side, less than a third have truly moved beyond the experimental phase. It is precisely this gap between technological availability and organizational maturity that represents the central challenge IT management must address in the coming years.

Operational Level: Governance, Tools, and Integration

At the operational level, the literature identifies three areas of action that are directly relevant to IT management.

First: Integration. With the growing relevance of agent-based systems, a well-thought-out integration strategy is becoming essential. As introduced in Chapter 4.3, the Model Context Protocol (MCP) addresses the N×M integration problem and significantly reduces the effort required to connect data sources and tools (Xu et al., 2025). For IT departments, this implies that: Meaning that those IT departments relying on open standards and infrastructure planning today lay the foundation for the future scalability of agent-based systems while those who do not create a debt that will prove costly.

Second: Security. The attack surface as outlined in Chapter 4.5. of agent-based systems differs from that of traditional software, resulting in traditional security approaches falling short here. Prompt injection, trust-authorization mismatches, and cascading errors in multi-agent systems cannot be adequately addressed with conventional concepts (Kim et al., 2026). IT management must develop security architectures that treat the model, memory, tools, and communication channels as a cohesive attack surface, not as separate layers.

Third: Use-case prioritization. The literature consistently shows that agent-based systems are most effective where workflows are clearly structured and sufficient data is available (Abou Ali et al., 2026). The decision on which cases to use should therefore not depend solely on technical feasibility, but rather on process clarity, data quality, and regulatory requirements. Gartner (2024) warns that over 40% of agent-based AI projects could be discontinued by 2027, because the business value remains unclear. Structured prioritization is thus not an optional best practice, but simply an economic necessity.

Executive Level: Human-in-the-Loop and Change Management

The levels of autonomy described in Chapter 4.2. make it obvious that: humans are not disappearing from the decision-making process, instaed their role is changing fundamentally. Feng et al. (2025) aptly describe this shift as a transition from operator to observer. Those who used

to perform tasks will increasingly monitor, validate, and intervene in the future, while directly performing tasks less frequently.

This has concrete implications for IT management. Teams must be prepared for the deployment of agent-based systems not only technically, but also organizationally and culturally. Employees need to understand how collaboration with autonomous systems works: when they need to intervene, how to assess the quality of outputs, and where the limits of system autonomy lie. This extends beyond technological considerations and becomes a question of organizational leadership.

5.3 Gaps in research and future research needs

Despite the field's rapid growth, the literature still contains some significant gaps. Most notably, there is a lack of empirical studies on organizational implementation.

The vast majority of the sources reviewed are conceptual or technical in nature. How companies actually implement agentic AI, which governance structures prove effective in practice, and what facilitates the transition from the pilot phase to full-scale operation have hardly been investigated to date.

Secondly, there is a lack of robust metrics for measuring degrees of autonomy in real-world application scenarios. The taxonomies developed (Yao et al., 2023; Feng et al., 2025; Kasirzadeh & Gabriel, 2025) are conceptually valuable, but remain limited in their practical operationalization. IT managers need standardized assessment frameworks that can actually be used to classify specific systems.

Furthermore, the question of liability in multi-agent systems remains largely unsettled. If an agent act based on another agent and causes an error in the process, who bears the responsibility? This question is unresolved both legally and organizationally and highlights the need for interdisciplinary research at the intersection of computer science, law, and organizational theory.

5.4 Limitations of the study

Like any literature review, this study has limitations that should be considered when interpreting the results.

The focus on English-language publications from the period 2019–2025 implies that earlier foundational work on agent systems was included only selectively. Adding to this, the nature of the fuel itself is a limitation as the field of agentic AI is evolving at a rapid speed which results in some of the literature reviewed, particularly pre-prints, being already outdated by the time this paper is published due to more recent findings.

Finally, the selection of sources reflects the databases used and the search terms chosen; relevant contributions in other outlets cannot be ruled out.

6 Conclusion

This literature review has examined the current state of research on agentic AI on three levels.

Conceptually, the literature makes it clear that agentic AI does not represent a gradual progression, but rather a genuine break with previous AI approaches. Four characteristics distinguish agentic systems from reactive language models: autonomy, tool use, persistent memory, and goal-directedness. The three-part classification developed in Chapter 2, LLM, AI agent, and Agentic AI, has proven to be a helpful framework for navigating the conceptual confusion often found in the literature.

At the technical level, three approaches dominate the research discourse: the ReAct framework, hierarchical planning architectures, and multi-agent systems. Initiatives such as the Model Context Protocol (MCP) suggest that the underlying infrastructure is maturing. Nevertheless, key problems remain unresolved, hallucination, prompt injection, and coordination errors in multi-agent systems are still real obstacles that hinder widespread productive deployment.

When examining specific application areas, a clear picture emerges agent-based systems demonstrate their strengths in areas where workflows are well-structured and data is abundant, for example, in software engineering, healthcare, or the financial sector. Domains characterized by high uncertainty, sparse data, or strict regulatory requirements, on the other hand, remain challenging.

Agentic AI is at a turning point. The technological foundations are in place, but organizations and regulations are struggling to keep up. For IT management, this means that the key skill in the coming years will not be implementing agentic systems. The focus will be on managing them responsibly. That is, knowing how much autonomy makes sense in which situations, how to establish governance structures, and how to lead teams where humans and machines work together.

Research still has some catching up to do in this area. Technical progress alone is not enough, we need robust empirical studies on how organizations actually implement and use these systems. Three areas in particular are likely to become more important in the coming years: hybrid neuro-symbolic architectures, governance models that adapt to different contexts, and metrics that allow us to meaningfully measure degrees of autonomy in the first place.

Ultimately, agentic AI is not just a technical issue. It is a socio-technical challenge, and thus one of the central future topics in business informatics.

References

Abou Ali, M., Dornaika, F. and Charafeddine, J. (2026) Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions. Artificial Intelligence Review, 59(11). https://doi.org/10.1007/s10462-025-11422-4 (Accessed: 4 June 2026).

Feng, K.J. et al. (2025) Levels of Autonomy for AI Agents. arXiv preprint arXiv:2506.12469. https://arxiv.org/abs/2506.12469 (Accessed: 4 June 2026).

Gartner (2024) Predicts 2025: Agentic AI Is the New GenAI. Stamford, CT: Gartner Research.

IBM (2025) What are AI agents? https://www.ibm.com/think/topics/ai-agents (Accessed: 4 June 2026).

Kasirzadeh, A. and Gabriel, I. (2023) In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. Philosophy & Technology, 36, Article 27. https://doi.org/10.1007/s13347-023-00606-x (Accessed: 4 June 2026).

Kim, J. et al. (2026) The Attack and Defense Landscape of Agentic AI: A Comprehensive Survey. Neurocomputing (Elsevier). https://doi.org/10.1016/j.neucom.2026.134049 (Accessed: 4 June 2026).

KPMG (2024) AI Pulse Survey Q4 2024. https://kpmg.com/ (Accessed: 4 June 2026).

McKinsey & Company (2025) The State of AI in 2025. https://www.mckinsey.com/ (Accessed: 4 June 2026).

Plaat, A. et al. (2025): Generative to Agentic AI: Survey, Conceptualization, and Challenges. arXiv:2504.18875.

Raza, S., Sapkota, R., Karkee, M. and Emmanouilidis, C. (2026) TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. AI Open, 7, pp. 71–95. https://doi.org/10.1016/j.aiopen.2026.02.006 (Accessed: 4 June 2026).

Russell, S. and Norvig, P. (2021) Artificial Intelligence: A Modern Approach. 4th edn. Harlow: Pearson.

Sapkota, R., Roumeliotis, K.I. and Karkee, M. (2025) AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. Information Fusion, 126, 103599. https://doi.org/10.1016/j.inffus.2025.103599 (Accessed: 4 June 2026).

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z. and Wen, J. (2024) A Survey on Large Language Model Based Autonomous Agents. Frontiers of Computer Science, 18, 186345. https://doi.org/10.1007/s11704-024-40231-1 (Accessed: 4 June 2026).

Webster, J. and Watson, R.T. (2002) Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26(2), pp. xiii–xxiii.

Xu, B., Gangadharan, G.R. and Buyya, R. (2026) Agentic Artificial Intelligence (AI): Architectures, Taxonomies, and Evaluation of Large Language Model Agents. arXiv preprint arXiv:2601.12560. https://arxiv.org/abs/2601.12560 (Accessed: 4 June 2026).

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2023) ReAct: Synergizing Reasoning and Acting in Language Models. In: Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). Kigali, Rwanda. https://arxiv.org/abs/2210.03629 Accessed: 4 June 2026).

Declaration of Authorship

I hereby declare under oath that I have completed this thesis independently and without using any resources other than those cited; any ideas taken directly or indirectly from external sources, as well as any content created or edited by artificial intelligence such as ChatGPT, are clearly identified as such.

This work has not previously been submitted to any other examination committee in the same or a similar form, nor has it been published.

Hamburg, 12.06.2026

Alexa-Sophie Ehmann