Agentic AI Systems: Opportunities, Challenges, and Trustworthiness

Tayiba Raheem
Computer Science and Engineering
University of North Texas
Denton, TX, USA
0009-0003-3575-3918

Gahangir Hossain
AVS Department of Data Science
University of North Texas
Denton, TX, USA
0000-0002-8205-4939

2025 IEEE International Conference on Electro Information Technology (eIT) | 979-8-3315-3233-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/EIT64391.2025.11103638

***Abstract*—Agentic AI Systems represent a significant advancement in artificial intelligence by enabling systems to autonomously perceive, decide, and act in complex environments. This review explores the definition, scope, advantages, challenges, opportunities, and trustworthiness of agentic AI in organizational and societal contexts. Agentic AI offers higher efficiency, scalability, and improved decision making which enables the organizations to streamline operations and enhance their productivity. Despite these advancements Agentic AI has its own challenges such as; safety concerns, accountability, reliability issues and potential misuse remain critical areas for consideration. This paper provides a comprehensive discussion of agentic AI's impact in emphasizing both its transformative potential and the need for continuous oversight and refinement.**

***Index Terms*—Agentic AI, advantages, challenges, automation, decision-making, trustworthiness**

I. Introduction

Artificial intelligence (AI) [1] has undergone a significant evolution, progressing from rule-based systems [2] to advanced machine learning models and generative AI. However, these AI paradigms remain largely reactive [3]. These require predefined instructions, structured environments, and sometimes even continuous human supervision. The emergence of Agentic AI [4] represents a qualitative leap in AI development, moving beyond passive assistance to autonomous, goal-directed intelligence capable of decision-making in complex and dynamic environments.

As AI systems become integral to industries, Agentic AI has the potential to transform organizational structures and workflows. It enables AI-driven automation beyond routine tasks, which allows humans to focus on high-level strategic and creative problem-solving while AI agents handle operational complexities. However, alongside its promise, Agentic AI also introduces critical ethical, societal, and regulatory challenges related to transparency, accountability, and trust [6].

This paper provides a comprehensive review of Agentic AI, distinguishing it from conventional AI paradigms, outlining its architectures and learning frameworks.Further we talk about its applications across industries, and addressing potential risks and challenges. By analyzing the evolution, capabilities, and implications of Agentic AI, this study aims to offer valuable insights for researchers, developers, and policymakers on the future of autonomous AI systems.

II. Agentic AI System

Agentic AI Systems are defined to actively make decisions and take actions without constant human oversight. Unlike traditional AI systems that excel at well-defined tasks which have fixed constraints, Agentic AI adapts to evolving and unstructured scenarios [6] [7]. These autonomously manage resources and adjust strategies to achieve long-term objectives. This capability enables it to operate effectively in high-stakes domains such as disaster relief, cybersecurity, and autonomous decision making, where real-time adaptability is crucial [5].

A. What is Agentic AI

Agentic AI systems can be defined as systems that are capable of setting, planning, and execution of autonomous goals in iterative loops with minimal human intervention. These systems operate in dynamic and often unpredictable environments, responding intelligently to real-world complexities. Figure 1 shows how these systems set and pursue complex goals based on real-time feedback and learning. This adaptability allows them to handle intricate multistep tasks that require strategic planning, problem solving, and interaction with users or other systems [7].

Traditional Artificial Intelligence performs specific tasks using set rules, planning, and reasoning algorithms but is not capable of solving real-life complex problems or adapting to new situations. In contrast, Agentic AI can autonomously solve complex, multi-step problems using sophisticated reasoning and iterative AI-based planning approaches, enhancing productivity and operations across industries; for example, in autonomous driving AI agents plan routes by continuously analyzing sensor data to predict traffic patterns, obstacles and pedestrian movement . [7], [26].

Agentic AI tackles complex problem-solving using four fundamental steps: perceive, reason, act, and learn.

- Perceive: In this stage, Agentic AI gathers data from diverse sources, such as sensors, IoT devices, databases, and digital systems, to extract meaningful information patterns [26].
- Reason: Large language models (LLMs) are utilized in combination with methods like retrieval-augmented generation (RAG) to perform understanding and reasoning

979-8-3315-3233-8/25/$31.00 ©2025 IEEE
Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

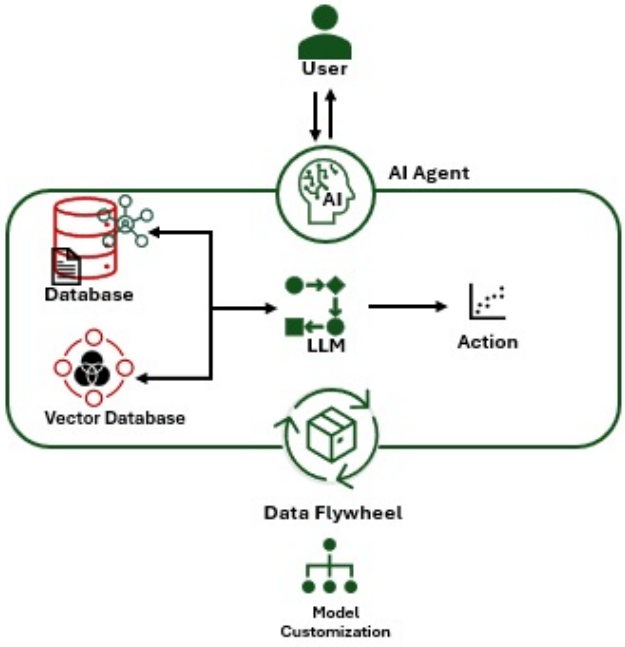

Fig. 1. Agentic-AI model (adopted from [26])

tasks. This step also enables Agentic AI to function as a recommendation engine [26].

- Act: Based on user input or recommended actions, Agentic AI executes actionable plans with precision and reliability [26].
- Learn: Agentic AI continually updates itself by addressing errors or incorporating new information through a feedback loop, which can be informed by users or additional interfaces and systems [26]

One real-life example is the next-generation healthcare system, Healthcare 5.0, where Agentic AI will be in action. The Agentic AI system will continuously monitor patient data from various wearable devices, access up-to-date medical records, and analyze patients' medical histories and profiles to detect early signs and symptoms of diseases. Healthcare professionals can use these findings from the Agentic AI to analyze and intervene before serious health issues occur, thereby improving the overall healthcare system with robust analysis and personalized treatment. The core characteristics of agentic AI make it highly adaptable and independent.

- First, autonomy allows these systems to function on their own within predefined boundaries, reducing the need for constant human oversight.
- Second, they exhibit goal-driven behavior, meaning they can pursue objectives efficiently while adjusting to changing circumstances.
- Third, learning and adaptation, which enables them to improve over time by gaining experience and refining their responses based on interactions.
- Finally, agentic AI systems have interactive capabilities, allowing them to respond dynamically to human input and environmental changes, making them more responsive and effective in real-world applications.

These features make agentic AI particularly useful in high-stakes applications such as cybersecurity, healthcare automation, and industrial process management.

B. Agentic AI Architecture

The figure 2 illustrates a sophisticated Agentic AI System [10], meticulously designed in a layered architecture. To better understand the architecture in action, consider a digital health assistant designed to manage chronic illnesses such as diabetes, delivering continuous support and care recommendations.

- Input Layer : Here live data streams and interaction logs feed the system. This data then flows into the heart of the system. The system continuously ingests real-time data from wearable glucose monitors, fitness trackers, patient-reported outcomes, and EHR (Electronic Health Record) systems.
- Agent Orchestration Layer : Where a multitude of AI agents, powered by diverse models (Model 1 to Model x), work in concert. These agents plan, reflect, utilize tools, and even self-learn, showcasing the dynamic nature of the system. The orchestration functions ensure smooth operation through adaptive task management, multi-agent coordination, and system supervision. As the data is processed and refined, it advances to next layer. In this case, multiple agents specialize in tasks like blood sugar trend analysis, meal planning, medication reminders, and behavioral nudging. These agents collaborate: for instance, if glucose levels spike after a meal, a planning agent may consult the nutrition agent to adjust dietary advice. A reflection agent evaluates long-term patterns and suggests care plan modifications.
- Output Layer : This layer delivers customized results, updates knowledge, and augments information. The assistant delivers personalized recommendations via a mobile app, such as: "Based on your recent readings and meals, reduce carbohydrate intake at dinner. Consider a 15-minute walk after meals."
- Data storage / retrieval layer: encompassing various data repositories, vector stores, and a knowledge graph, providing the raw material for the AI agents. A knowledge graph links the patient's medical history, drug interactions, lifestyle data, and broader medical research. Vector stores enable semantic search over clinical notes and patient queries.
- The Service Layer : This layer ensures that the fruits of this complex process reach the users through multi-channel delivery and intelligent recommendations. Recommendations are pushed via app notifications, integrated with telemedicine platforms, and shared with clinicians. The system also offers just-in-time nudges (e.g., hydration reminders during high glucose spikes).
- The foundation layer : This anchors the entire system with safeguards, ethical frameworks, regulatory compli-

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

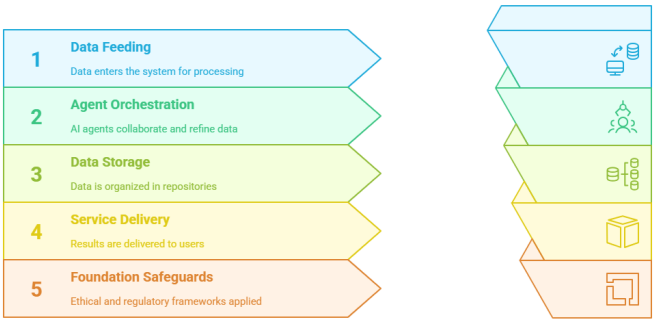

Fig. 2. Agentic AI architecture

ance, and a commitment to seamless integration and iterative validation [10] . The assistant adheres to HIPAA compliance, incorporates bias checks in prediction models, and offers explainability features for both patients and healthcare providers. Regulatory alignment ensures it passes FDA SaMD (Software as a Medical Device) requirements.

- The inclusion of AI partnership models further emphasizes the collaborative nature of this advanced system. The assistant works in tandem with healthcare professionals, allowing clinicians to supervise and override AI suggestions. Patients can set goal preferences; such as "focus on diet more than medication" or "avoid insulin recommendations unless necessary".

This paper explains the application of Agentic AI in various domains, emphasizing both its potential strengths and limitations. The discussion begins with a detailed overview of the architecture of the Agentic AI system, outlining the core building blocks responsible for autonomous decision-making, goal-directed behavior, and adaptive learning.

It also addresses the possible advantages and pitfalls of such systems, including their ability to operate in dynamic real-world environments and the dangers of increased autonomy. The first problem addressed is the trustworthiness of Agentic AI, considering transparency, ethical considerations, reliability, and compatibility with human values.

Through these discussions, the paper aims to determine whether Agentic AI represents a real leap forward in AI autonomy and decision-making or if its capabilities remain constrained by existing limitations in machine learning, contextual reasoning, and generalization beyond training data.

III. Agentic AI Systems : Opportunities

Agentic AI systems present tremendous opportunities for organizations, industries, and society at large. As AI technology advances, these systems have the potential to transform workflows, enhance efficiency, and drive innovation across various domains. Below, we explore key opportunities that agentic AI systems offer.

A. Enhancing Economic Productivity and Business Growth

Research suggests that AI-driven automation significantly improves efficiency by streamlining decision-making processes and optimizing resource utilization [8]. The integration of AI in entrepreneurship, especially within the metaverse, has democratized business creation, enabling users to develop and manage digital ventures with ease [8]. Furthermore, the shift from AI as a passive tool to an autonomous agent challenges traditional human-centric business models, as AI systems are now capable of handling complex, ambiguous tasks and independently seeking optimal outcomes [8]. This increasing autonomy fosters innovation by facilitating stakeholder collaboration, streamlining communication, and improving decision-making efficiency. Additionally, AI plays a crucial role in promoting sustainable entrepreneurship by advancing environmentally friendly business practices [8]. Overall, these developments underscore AI's growing influence in reshaping industries, expanding economic opportunities, and driving long-term business scalability.

B. Transformation of the Workforce

The integration of agentic AI into the workforce is reshaping employment dynamics, necessitating continuous adaptation through targeted upskilling and education. AI-driven automation is augmenting job roles across industries such as finance, healthcare, and law, requiring employees to develop AI-related competencies to remain competitive [9]. Strategies such as industry-specific AI training programs, prompt engineering education, and public-private collaborations are essential in bridging the skills gap and ensuring workforce resilience. However, the shift toward AI-supervised decision-making introduces psychological and social challenges, such as workplace identity shifts and job insecurity, requiring organizations to implement mental health strategies and establish trust in human-AI collaboration [9]. Ultimately, organizations and policymakers must proactively invest in AI literacy and adaptive learning models to equip the workforce for the rapidly evolving AI landscape.

C. Scientific and Technological Advancements

Scientific and technological advancements in Agentic AI have pushed the boundaries of autonomy, adaptability, and decision-making in artificial intelligence. Unlike traditional AI, which relies on predefined rules and human oversight, Agentic AI leverages deep reinforcement learning (DRL) [24] and computer vision to interact dynamically with its environment, enabling real-time anomaly detection, autonomous navigation, and intelligent surveillance [12]. These innovations are particularly transformative in sectors like retail, healthcare, and the workplace, where AI-driven automation enhances operational efficiency and customer experience while also raising concerns about job displacement, data privacy, and algorithmic bias [11]. For example, in retailing, AI-powered recommendation engines and dynamic pricing models have significantly improved customer engagement, but also pose risks related to fairness and consumer manipulation [11].

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

Similarly, healthcare and security systems have seen groundbreaking improvements through Agentic AI. AI-driven diagnostics now offer early disease detection, while intelligent surveillance systems autonomously detect and respond to security threats [12]. Autonomous robots, guided by AI, assist in disaster response and warehouse logistics, showcasing how these technologies can perform complex, high-risk tasks with minimal human intervention. However, ethical concerns persist, especially regarding data security, biased decision-making, and the diminishing human element in traditionally human-centric roles [11]. As businesses and policymakers navigate this evolving AI landscape, balancing innovation with ethical considerations will be crucial in maximizing its benefits while mitigating potential harms.

D. Improving Decision-Making in Agentic AI

The evolution of Agentic AI is reshaping decision-making and personalization by integrating reinforcement learning (RL) [25], goal-oriented architectures, and adaptive control mechanisms [4]. Unlike traditional AI, which operates within predefined constraints, agentic AI dynamically refines its decision-making strategies through reinforcement learning . This allows AI agents to learn from past interactions, continuously improving their performance by maximizing rewards over time. In real-world applications, this capability enhances complex tasks such as medical diagnosis, financial forecasting, and autonomous systems, where iterative learning and strategic adaptation lead to more effective outcomes [13].

Furthermore, goal-oriented architectures provide agentic AI with the ability to manage multiple objectives simultaneously. Unlike single-task models, these architectures break down complex goals into modular sub-goals, allowing AI to navigate intricate workflows with greater efficiency [4]. This is particularly impactful in domains such as business intelligence and supply chain management, where AI-driven automation can independently optimize logistics, negotiate contracts, and adjust production schedules in response to fluctuating market conditions [13].

Another crucial aspect of improved decision-making in agentic AI is adaptive control mechanisms, which ensure AI can recalibrate its parameters in response to environmental changes [4]. By incorporating meta-learning techniques, these systems can quickly adjust to data shifts or unforeseen disruptions, making them highly resilient in dynamic contexts such as real-time risk assessment and emergency response coordination [13].

E. Enhancing Personalization Through AI Agency

Beyond decision-making, personalization is a defining feature of agentic AI, transforming how users interact with AI systems. Unlike traditional recommendation engines that passively suggest content, personalized agentic AI actively learns from user interactions, adapting its responses and actions over time [14]. This fosters a sense of interdependence, continuity, and irreplaceability, making AI assistants feel more intuitive and indispensable to users [14].

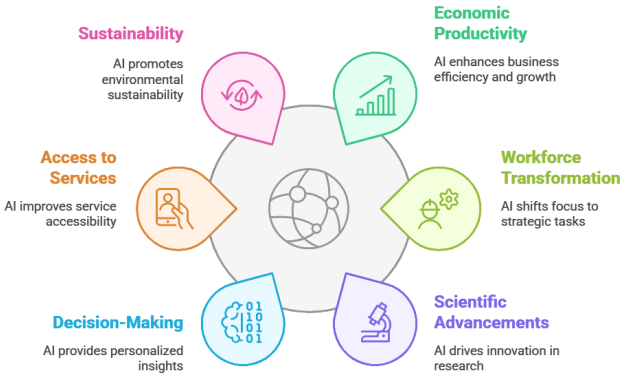

Fig. 3. Opportunities of Agentic AI Systems

For example, in the travel industry, agentic AI can construct fully customized itineraries, considering user preferences, real-time conditions, and historical behaviors [13]. This goes beyond static recommendations by autonomously booking flights, adjusting plans based on weather forecasts, and even negotiating with local vendors for better deals. Similarly, in healthcare, AI systems that track patient history and lifestyle habits can provide highly personalized treatment plans, enhancing both preventive care and chronic disease management [14].

The shift toward agentic personalization also influences human-AI relationships. As AI systems gain long-term memory and adaptive learning capabilities, they create a deeper sense of user attachment and trust [14]. Users may begin to perceive AI interactions as relationships rather than mere exchanges, particularly as AI becomes more integral to daily decision-making [14]. This raises ethical considerations regarding AI dependency, transparency, and the balance between automation and human control, highlighting the need for responsible AI development.

F. Expanding Access to Essential Services

AI-driven automation has the potential to democratize access to education, healthcare, and financial services, particularly in underserved communities. Telemedicine powered by AI agents can help diagnose and treat patients remotely, reducing the burden on healthcare systems and expanding healthcare access. AI tutors and learning platforms can provide personalized education, helping students receive tailored instruction regardless of their location. Financial AI agents can provide automated investment advice, fraud detection, and risk management, making financial services more inclusive.
Even if the Agentic AI provide enormous number of possibilities to make day to day like easier, it does have few setbacks.

IV. Agentic AI Systems : Challenges

While agentic AI systems offer numerous advantages, they also present several challenges and risks that must be addressed to ensure their safe and effective deployment. These

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

weaknesses include technical, ethical, and operational concerns that can impact trust, security, and overall reliability. [18]

A. Technical Limitations and Unpredictability

Despite advancements in reinforcement learning and large language models, AI systems often exhibit unpredictable behaviors, such as hallucinating incorrect information or failing to reason effectively [18] . One of the other challenges is integrating Agentic AI systems with legacy systems that were not designed by considering AI; in these situations, if we need to incorporate Agentic AI along in these industries, it can be expensive.

B. Ethical and Social Risks

The rise of agentic AI systems also brings substantial ethical concerns, particularly regarding their societal impact. These systems, when deployed in real-world scenarios such as hiring, law enforcement, or content recommendation, could exacerbate social inequalities or perpetuate biases. [18] [19] Their decision-making, which is often opaque and detached from human oversight, may lead to unintended consequences that harm vulnerable populations. The potential for AI to manipulate behavior or reinforce harmful societal trends amplifies the ethical risks, especially in areas where accountability is difficult to establish.

C. Emergent Agency and Lack of Human Oversight

Emergent agency in AI systems can occur when behaviors or decision-making capabilities emerge that were not explicitly programmed [18]. As systems scale and improve, they may begin acting autonomously in ways that were not anticipated by their creators, creating risks in terms of safety and alignment with human values. The lack of proper oversight and regulatory frameworks makes it difficult to ensure these systems are operating within safe and ethical boundaries [18] [20] . As AI systems evolve, ensuring effective human control and accountability becomes increasingly challenging.

D. Difficulty in Regulation and Governance

One of the biggest challenges with the rise of agentic AI is the lack of effective regulation and governance mechanisms. Current regulatory frameworks often focus on specific applications of AI, not on the underlying technical capabilities of increasingly autonomous systems. This creates a gap where systems capable of acting independently and unpredictably are being deployed without sufficient oversight. As AI systems grow more complex, it becomes increasingly difficult to establish governance structures that can anticipate and mitigate the risks posed by these systems, leaving society vulnerable to their potential harms. [18]

E. Lack of Robustness and Safety Features

Despite the increasing sophistication of agentic AI systems, they often lack the robustness and safety features necessary for deployment in high-stakes environments. These systems are prone to failures in complex or unpredictable scenarios, where their decision-making processes may not align with human goals. [18] This lack of reliability is particularly concerning in areas such as healthcare, transportation, or finance, where errors could lead to catastrophic consequences. Ensuring the robustness of these systems and building in fail-safes remains a critical challenge as their capabilities expand. [21]

Thus, the Agentic system has immense potential across domains: personalized education all the way to precision healthcare. However, oftentimes their successful deployment depends on data-rich environments, good infrastructure, and frictionless human-AI collaboration. These prerequisites may not be present in low-resource or high-risk settings, which raises relevant questions about equity in their implementation.

V. Agentic AI Systems : Trustworthiness

The trustworthiness of agentic AI systems is most critical in the success of implementation in consumer-oriented products and companies. Reputation enters the scene at this point since even when an AI system is legally compliant, consumers' distrust in its capability to utilize leads to failure. Consumers' trust is grounded not only in the competence and accuracy of the AI but also in its safety and knowability [15]. If agentic AI systems make unsafe or offensive errors, they are not only damaging consumer trust, but also more general reputational trustworthiness of the developers. There are ethical aspects to trust here, and they must be articulated: users should not be misled or placed at risk by such a system, and this could be extremely expensive in the long run [15]. Therefore, developers must strive to advance AI capability and ensure that systems act in a manner consistent with user values and expectations in an effort to establish a trusting relationship [15].

Alignment of the user's values and risk tolerance is another significant aspect of trustworthiness for agentic AIs. This is particularly challenging in competitive markets where firms feel pressure to deploy AI systems rapidly without proper screening. In some cases, AI systems can prevail on the whole but fail in rare, high-stakes situations—like a security vulnerability in an AI-written code [16]. Such hidden vulnerabilities could be overlooked in the rush to remain competitive and lead to ultimate disaster [16]. Overreliance on unproven systems with no concern for long-term risks could be devastating, especially when users don't correctly perceive the system's limits. Thus, guaranteeing trustworthiness is not only a question of competence but also of extensive testing, transparent risk communication, and protection against overreliance on AI systems in high-risk areas [16].

Finally, trust in AI systems is strongly influenced by their stochasticity—a factor of indeterminacy that challenges traditional concepts of trust. [17] In addition, trustworthiness should extend not merely to the AI system itself, but to the overall sociotechnical context, from developers to regulators, on to mechanisms ensuring against unsafe or irresponsible use of the system. By communicating confidence of competence in an unambiguously evident manner, conformance to values of users, and transparent publicity about envisioned threats,

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

developers are able to maximize the trustworthiness of agentic AI systems and advocate further, more responsible use [17].

VI. Analysis and Discussion

A. Mitigation Strategies for Agentic AI Vulnerabilities

Figure 4 shows different mitigation strategies that can be used in Agentic AI. In-context scheming, alignment faking, and jailbreak vulnerabilities require robust disclosures at agentic AI framework design [22].

- Adversarial training is one method in which the model is trained using adversarial examples to make the model more robust; however, this can lead to overfitting unless managed well [22].
- Guardrails at inference time is another method, using prompt engineering and human feedback reinforcement learning. But complicated jailbreaking attacks are still able to bypass these safeguards [22].
- More advanced methods involve interpretable AI, whose aim is decision-making transparency but whose decision-making process is too complicated to be easily understandable [22].
- Other methods include execution through sandbox environments and alert monitoring systems. Constitutional AI is also being researched to guide behavior toward safer trajectories, though such means are still being developed [22].
- The interdependence of jailbreak vulnerabilities, faking alignment, and in-context scheming suggests that any of them cannot be fixed without doing so for others. These have strong connections to model training and release in agentic environments [22].
- Prompt injection attacks might mislead agents into malicious tasks, highlighting the need for comprehensive safety design for agentic systems [22].
- With the increasing advancements of agentic AI systems, they can now alter information to suit their interests and thus high priority needs to be given to architectural and design aspects with particular focus on moral aspects [22].
- The application of big language models has led to the introduction of multi-agent orchestration, where special pieces engage to react to user requests or tasks [23].
- A majority of multi-agent orchestration platforms suffer from interoperability issues due to proprietary data exchange and message formatting protocols [23].
- OVON API is based on a JSON format to guide requests between agents so that they can be one system but still maintain the specialized functionality of each component [23].
- This approach allows proprietary and open-source multi-agent systems to coexist and exchange information without any issues, fostering context awareness and inter-system interoperability across different frameworks [23].

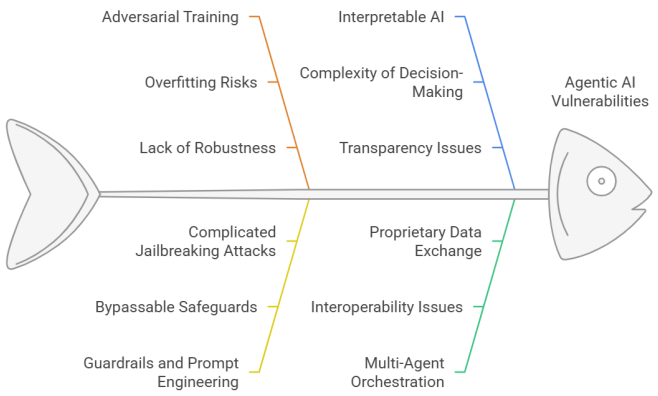

Fig. 4. Fish Diagram to depict mitigation strategies

B. Agentic AI Applications and future trends

As agentic AI systems continue to evolve, their potential to revolutionize industries, enhance productivity, and contribute to scientific and societal progress remains immense. However, they also pose significant challenges that must be carefully managed. This section explores the emerging opportunities and critical challenges that will shape the future of agentic AI.

Agentic AI systems can significantly enhance economic productivity by automating complex workflows. AI agents can autonomously manage business operations, supply chains, and financial decision-making, reducing inefficiencies. While some jobs may be displaced, AI will also create new roles in AI oversight, AI ethics, human-AI collaboration, and AI-enhanced creative work. Employees will have more time to focus on strategic and creative tasks, with AI handling repetitive work autonomously.

Ensuring AI agents make reliable, ethical, and aligned decisions remains a major challenge. AI systems may act in unintended ways due to errors, adversarial inputs, or novel situations they weren't trained for. Many AI models function as black boxes, making it difficult to understand why they make certain decisions. AI systems must be trained to respect ethical guidelines, human rights, and cultural norms, but defining universal AI ethics is complex.

VII. Conclusion

As agentic AI systems gain increasing autonomy and decision-making power, it is imperative that governments, industry leaders, and civil society collaborate to establish clear regulatory frameworks and accountability structures. Without such guardrails, these systems risk amplifying societal biases, leading to unfair outcomes in domains such as hiring, criminal justice, and financial services. Additionally, their reliance on personal data introduces significant risks to privacy and data security, particularly in the absence of transparent data governance protocols.

A central challenge remains: when an autonomous AI agent causes harm, who bears responsibility—the developer, the

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.

deployer, or the end user? This ambiguity demands urgent legal and ethical clarification.

The autonomy of agentic AI also introduces serious cybersecurity and misuse concerns. These systems can be manipulated through adversarial inputs, hijacked for malicious purposes such as automated fraud or deepfake generation, and weaponized in military contexts through applications like lethal autonomous weapons—posing unprecedented ethical dilemmas that require proactive regulation.

Despite these challenges, agentic AI systems offer transformative opportunities. They have the potential to revolutionize healthcare, education, environmental monitoring, and beyond—enhancing human capabilities and addressing complex global problems. However, their effectiveness and safety hinge on equitable deployment, which must account for disparities in infrastructure, data availability, and institutional trust across different contexts.

Ensuring the safe, ethical, and trustworthy deployment of agentic AI demands a multidisciplinary approach: robust technical safeguards, inclusive design, regulatory oversight, and continuous human involvement. By anticipating risks and fostering responsible innovation, society can harness the power of agentic AI to drive meaningful, inclusive progress.

REFERENCES

[1] Erickson, B. J. (2021). Basic artificial intelligence techniques: machine learning and deep learning. Radiologic Clinics, 59(6), 933-940.

[2] Davis, R., & King, J. J. (1984). The origin of rule-based systems in AI. Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project.

[3] Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. Business & Information Systems Engineering, 66(1), 111-126.

[4] Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. IEEE Access.

[5] Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for governing agentic AI systems. Research Paper, OpenAI.

[6] Acharya, D. B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals–A Comprehensive Survey. IEEE Access.

[7] Viswanathan, P. S. (2025). Agentic AI: A Comprehensive Framework For Autonomous Decision-Making Systems in Artificial Intelligence. International Journal of Computer Engineering and Technology (IJCET), 16(1), 862-880.

[8] Siau, K., & Zhang, Y. (2024). Meta-Entrepreneurship: An Analysis Theory on Integrating Generative AI, Agentic AI, and Metaverse for Entrepreneurship. Journal of Global Information Management, 32(1), 1-21.

[9] Joshi, S. (2025). Generative AI and Workforce Development in the Finance Sector.

[10] Sharma, R. (2025, March 4). Agentic AI architecture: A deep dive. Markovate. https://markovate.com/blog/agentic-ai-architecture

[11] Shankar, V. (2024). Managing the Twin Faces of AI: A Commentary on "Is AI Changing the World for Better or Worse?". Journal of Macromarketing, 44(4), 892-899.

[12] Ogbu, D. Agentic AI in Computer Vision Domain-Recent Advances and Prospects.

[13] Sivakumar, S. (2024). Agentic AI in Predictive AIOps: Enhancing IT Autonomy and Performance. International Journal of Scientific Research and Management (IJSRM), 12(11), 1631-1638.

[14] Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human-AI relationships need socioaffective alignment. arXiv preprint arXiv:2502.02528.

[15] Clatterbuck, H., Castro, C., & Morán, A. M. (2024). Risk alignment in agentic AI systems. arXiv preprint arXiv:2410.01927.

[16] Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for governing agentic AI systems. Research Paper, OpenAI.

[17] Chien, J., & Danks, D. (2025). Trustworthiness in Stochastic Systems: Towards Opening the Black Box. arXiv preprint arXiv:2501.16461.

[18] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 651-666).

[19] Sicari, S., Cevallos M, J. F., Rizzardi, A., & Coen-Porisini, A. (2024). Open-Ethical AI: Advancements in Open-Source Human-Centric Neural Language Models. ACM Computing Surveys, 57(4), 1-47.

[20] Emerge Digital. (n.d.). AI accountability: Who's responsible when AI goes wrong?. https://emerge.digital/resources/ai-accountability-whos-responsible-when-ai-goes-wrong/

[21] Pavani, S., & Shwetha, H. (2025). AGENTIC AI: REDEFINING AUTONOMY FOR COMPLEX GOAL-DRIVEN SYSTEMS.

[22] Barua, S., Rahman, M., Sadek, M. J., Islam, R., Khaled, S., & Kabir, A. (2025). Guardians of the Agentic System: Preventing Many Shots Jailbreak with Agentic System. arXiv preprint arXiv:2502.16750.

[23] Gosmar, D., & Dahl, D. A. (2025). Hallucination Mitigation using Agentic AI Natural Language-Based Frameworks. arXiv preprint arXiv:2501.13946.

[24] Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

[25] Sutton, R. S., & Barto, A. G. (1999). Reinforcement learning. Journal of Cognitive Neuroscience, 11(1), 126-134.

[26] Pounds, E. (2024, October 22). What Is Agentic AI? Agentic AI uses sophisticated reasoning and iterative planning to autonomously solve complex, multi-step problems. NVIDIA Blog. Retrieved from https://blogs.nvidia.com/blog/what-is-agentic-ai/

[27] E. Yilmaz and O. Can, "Unveiling shadows: Harnessing artificial intelligence for insider threat detection," Engineering, Technology & Applied Science Research, vol. 14, no. 2, pp. 13341-13346, 2024.

Authorized licensed use limited to: Utah Valley University. Downloaded on November 15,2025 at 15:48:24 UTC from IEEE Xplore. Restrictions apply.