

Cognitive AI agents: From Reasoning and Planning to Autonomous Action

Nasini PA*, Ms.Riya K Prakash**, Dr.L C Manikandan***

*B.Tech Student, Department of CSE, Universal Engineering College, Thrissur, Kerala, India.

Email: nasininajeeb61@gmail.com

** Assistant Professor, Department of CSE, Universal Engineering College, Thrissur, Kerala, India.

Email: riyajothish13@gmail.com

***Professor, Department of CSE, Universal Engineering College, Thrissur, Kerala, India.

Email: lmanikandan@gmail.com

Abstract:

Cognitive AI agents represent an emerging paradigm in artificial intelligence that combines large language models (LLMs) with reasoning, planning, memory, tool use, perception, and feedback to accomplish complex goals with increasing levels of autonomy. Unlike conventional task-specific AI systems, modern agentic systems can interpret high-level objectives, decompose them into subtasks, reason over available information, construct and revise plans, interact with external tools and environments, and adapt their actions according to observed outcomes. Recent advances in LLM-based reasoning, retrieval-augmented generation, tool-augmented agents, reflective reasoning, multimodal interaction, and multi-agent collaboration have significantly expanded the capabilities of autonomous AI systems. This survey presents a structured analysis of cognitive AI agents with particular emphasis on the transition from reasoning and planning to autonomous action. It examines the major architectural components and cognitive capabilities of modern agents, including goal interpretation, reasoning, task decomposition, adaptive planning, memory, tool selection, environmental interaction, reflection, and feedback. The review further explores the development of single-agent and multiagent systems and their applications in autonomous systems, healthcare, recommender systems, robotics, energy, and other complex decision-making environments. The analysis shows that reliable autonomy cannot be achieved through reasoning or planning alone. Instead, effective agents require a continuous interaction between reasoning, planning, memory, tool use, action, observation, and replanning. Despite rapid progress, important challenges remain in long-horizon planning, memory reliability, tool safety, multi-agent coordination, evaluation, security, explainability, and human oversight.

Keywords — Cognitive AI Agents, Agentic AI, Large Language Models, Autonomous Agents, Reasoning, Planning, Memory, Tool Use, Multi-Agent Systems, Autonomous Action.

I. INTRODUCTION

Artificial Intelligence (AI) has developed considerably from older systems that were built to carry out fixed and task focused activities to more advanced systems that can think, make choices, adjust themselves, and engage with complicated environments [1], [3]. Previously, AI agents typically depended on set rules, static knowledge bases, or particular algorithms. Although these methods performed well in structured settings, their lack of flexibility reduced their effectiveness when

applied to unpredictable, changing, and open-ended real-world scenarios. The development of Large Language Models (LLMs) has created a new approach for building more adaptable autonomous systems, offering significant abilities in understanding natural language, processing knowledge, reasoning, and supporting decision-making [13], [14].

LLM-based autonomous agents enhance the abilities of language models by incorporating features like planning, memory, retrieval, tool usage, and interaction with the environment [13], [15]. These

agents are capable of understanding a broad goal, breaking it down into smaller tasks, considering different ways to solve the problem, creating a plan of action, using additional information and tools, carrying out actions, evaluating the results, and adjusting their approach as needed [8], [13], [15]. The shift from generating language to performing goal oriented actions has led to the quick expansion of research in agentic AI. Recent surveys have looked at these systems from various angles, such as how agents are built, how multiple agents work together, their real-world uses, ways to assess their performance, security issues, and potential future problems [1], [3], [4], [12], [13].

Reasoning and planning play a key role in this development as autonomous agents need to convert high-level objectives into practical steps for action [3], [13], [14]. Various methods of reasoning, such as Chain-of-Thought, Tree-of Thoughts, self-consistency, and reasoning-acting frameworks, have been studied to enhance the capability of large language models in tackling complex tasks [8], [14], [15]. Memory and retrieval mechanisms help support extended interactions by keeping track of the task context, past experiences, and important external information [8], [13], [15]. Tool integration enables agents to communicate with databases, APIs, software systems, simulators, and other external environments, which enhances their abilities beyond just generating text [7], [8], [15].

The growth of multi-agent systems has broadened the range of capabilities for autonomous AI. Rather than placing all the duties on one agent, several agents with specific skills can talk to each other, work together, and combine their efforts to handle difficult tasks [3], [12]. Surveys on systems that use large language models as part of multi-agent setups emphasize key elements like agent profiles, how agents perceive their environment, their ability to take actions, interactions between agents, and the process of system evolution [12]. These systems have been studied in various fields such as software development, scientific research, simulations, recommendations, healthcare, autonomous systems, and energy management [5], [6], [7], [10], [11], [12].

Despite these advancements, achieving dependable autonomous operation continues to be difficult. Current studies highlight several challenges in the field, including difficulties with long-term planning, instances of generating false information, inconsistencies in memory retention, improper use of tools, concerns regarding security and privacy, coordination between multiple agents, methods for assessing performance, the need for transparency in decision-making, and the importance of human supervision [1], [3], [4], [9], [13], [15]. Security research specifically indicates that as autonomy increases, it brings about risks throughout the entire agent process, such as reasoning, memory, tool usage, and communication [4], [9]. Therefore, creating autonomous agents needs more than just improved reasoning and planning abilities; it also demands systems that ensure verification, safety, reliability, and safe interaction with outside environments [4], [5], [9], [10].

Driven by these advancements, this survey centers on Cognitive AI Agents, exploring their capabilities in reasoning, planning, and executing autonomous actions. It offers a detailed examination of the key elements and functions of cognitive AI agents, covering areas such as their design structure, ways of reasoning, planning processes, memory and information retrieval, integration with tools, ability to act independently, teamwork among multiple agents, practical uses, security features, reliability aspects, and ongoing research issues [1], [3], [7], [12]–[15]. The goal is to explore how these components function together to change LLM-based systems from simple answer providers into flexible, purpose-driven agents that can take on more independent actions in changing environments.

II. METHODOLOGY

The methodology of this survey focuses on the key components that enable a cognitive AI agent to transform a broad goal into a sequence of actions that can be performed autonomously. Rather than viewing an intelligent agent as only a language model, the approach considers it as a complete system that includes environment perception, reasoning, planning, memory, information retrieval, tool interaction, action execution, feedback, and

verification. These components work together to achieve goals in dynamic environments.

The process begins when an agent receives a goal or task from a user or an external source. The goal is interpreted according to the surrounding context, constraints, and desired outcome. The agent then identifies the information required to complete the task and applies suitable reasoning mechanisms to analyse the available information. By decomposing complex objectives into smaller subtasks, the agent can construct an appropriate plan. Memory and retrieval mechanisms provide relevant information from the current task and previous interactions, while external tools enable the agent to obtain additional information or perform actions.

Once a plan is selected, the agent executes the required actions using suitable tools or interfaces. The resulting observations are then provided back to the agent as feedback. The agent evaluates whether the obtained result satisfies the original objective. If the outcome is unsatisfactory or the environment changes, the agent can revise the current plan and continue the task. This creates a continuous cycle of reasoning, planning, action, and feedback, which supports adaptive and autonomous behaviour in cognitive AI agents [1], [3], [12], [13].

A. Cognitive AI agent Architecture

The Modular Architecture of a cognitive AI Agent shown in Fig. 1.

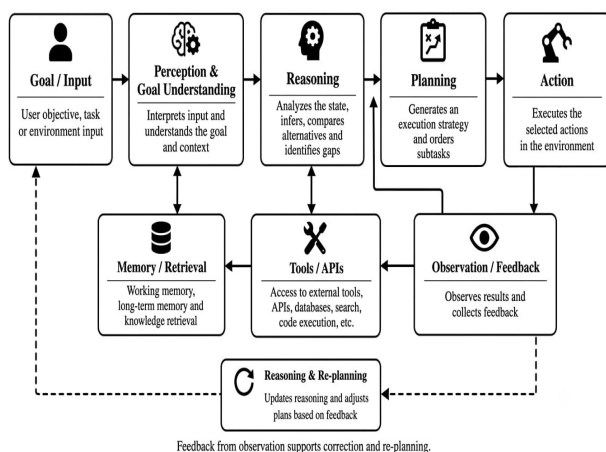


Fig. 1 Modular Architecture of a cognitive AI agent showing the flow from goal understanding to action and feedback-supported replanning.

A cognitive AI agent can be represented as a modular architecture in which different components cooperate to achieve a common objective. The architecture normally begins with an input or goal layer and progresses through reasoning and planning before reaching the action layer. Memory, retrieval and external tools support the decision-making process, while feedback allows the agent to continuously evaluate its performance [1], [3], [8], [12]–[15]. The overall architecture of LLM-based autonomous agents commonly integrates planning, memory, action, and interaction mechanisms to support goal-directed behaviour.

The perception and goal-understanding layer converts the incoming task into a representation that can be processed by the agent. The input may be natural language, structured data, images, sensor information, documents, or a combination of different modalities. Understanding the objective correctly is important because an incorrectly interpreted goal can affect every subsequent stage [7], [12], [13]. LLM-based multi-agent research highlights perception as an important stage for acquiring environmental information and supporting subsequent decision-making.

The reasoning layer analyses the current state and determines possible approaches to the task. It can compare alternatives, infer relationships, identify missing information, and decide whether additional information should be retrieved. Modern LLM-based agents can perform this reasoning using prompting strategies, external knowledge, or iterative reasoning mechanisms [3], [12]–[15].

The planning layer converts the reasoning outcome into an executable strategy. A complex task can be divided into smaller subtasks, dependencies can be identified, and an ordered sequence of actions can be generated. For dynamic tasks, the plan should not be considered permanent; it may need to be revised when the agent receives new observations [3], [8], [12]–[15]. The reviewed autonomous-agent literature identifies planning as a core capability for transforming goals into actionable steps.

The memory and knowledge-retrieval layer provides information that may not be available in the immediate context. Working memory can maintain the current task state, while long-term memory can preserve useful information from previous

interactions. Retrieval mechanisms can also provide relevant external documents or knowledge during decision making [8], [13], [15]. Research on autonomous agents specifically identifies memory as an important component for maintaining information and supporting continued task execution.

The tool and API layer connects the agent with external capabilities. Depending on the application, these capabilities may include search systems, databases, software applications, code execution environments, APIs, simulators, or physical devices. Tool integration allows the agent to move beyond generating information and perform operations in an external environment [7], [8], [15]. Tool use is also recognized as an important mechanism for extending LLM agents beyond their internal knowledge and enabling interaction with external environments.

Finally, the action and feedback layer executes the selected operation and collects the resulting observation. The agent evaluates the observation against the expected outcome. If the task has not been successfully completed, the information can be sent back to the reasoning and planning stages for correction or replanning [1], [3], [7], [8], [12]–[15]. This iterative interaction between action, observation, and replanning is central to the development of autonomous agent systems.

B. Reasoning Mechanisms

Reasoning is an important skill that cognitive AI agents have. It allows an agent to move from simply handling basic information to making decisions in a more structured way. When handling tough tasks, an agent has to consider many things at once, understand how different pieces of information are related, and choose the action that seems most logical [3], [13]–[15].

People have considered various approaches to how agents using large language models think. Chain-of-thought reasoning helps the model think through each step of a problem, especially when the problem has multiple parts that need to be solved one at a time. Self-consistency checks various ways of thinking to find a more accurate answer. Tree-of-Thoughts takes this idea further by allowing you to explore and examine various paths of thinking [13]–[15].

Another important approach is ReAct, which combines thinking and action. The agent doesn't

pause to finish thinking before acting; it moves back and forth between thinking and taking actions in the environment. The result of what you do can influence the next part of your thinking. Reflection-based methods allow an agent to review previous decisions and identify potential errors [13]–[15].

For cognitive agents, reasoning is like a process that takes place one step at a time, instead of being a fast, single choice. The way a person thinks clearly influences how they decide and plan their next actions. Just thinking about the steps isn't enough to know if something is correct, so important decisions should be supported by real evidence, tool results, rules, or checks to make sure everything is right [1], [3], [4], [8], [13], [15].

C. Planning Mechanisms

Planning is especially important when an agent is required to complete tasks that involve multiple dependent steps. In such situations, the agent must determine not only what action should be performed, but also the appropriate order in which the actions should be executed. Dependencies between subtasks are considered during plan generation so that an action is performed only when its required conditions are satisfied [3], [12]–[15].

Another important aspect of planning is context-aware decision making. The same goal may require different action sequences depending on the available resources, environmental conditions, user requirements, and previous actions. Cognitive agents can therefore use the current state of the environment together with information retrieved from memory and external sources to generate a contextually appropriate plan [8], [12]–[15].

Planning may also be performed iteratively. Instead of generating a complete plan for a long task at the beginning, an agent can generate an initial plan, execute a portion of it, observe the resulting state, and then determine the next set of actions. This approach is useful in dynamic environments where conditions can change during execution [3], [7], [8], [13], [15].

The planning process can further incorporate constraints and priorities. Safety requirements, time limitations, resource availability, task deadlines, and user preferences can influence the selection of a

particular plan. A suitable plan should therefore satisfy the required constraints while achieving the intended objective efficiently [3], [4], [5], [7].

In autonomous systems, planning is closely connected with reasoning and execution. Reasoning helps the agent understand the problem and compare possible strategies, while planning organizes the selected strategy into executable steps. Execution then provides real-world feedback that can be used to evaluate and modify the plan. This interaction makes planning an adaptive component rather than a fixed sequence of instructions [3], [7], [12], [13], [15].

Thus, effective planning enables cognitive AI agents to handle complex objectives through task decomposition, dependency management, candidate-plan generation, plan evaluation, constraint handling, execution monitoring, and dynamic replanning. These capabilities allow an agent to adapt its behaviour when unexpected events occur and progressively move toward successful task completion [1], [3], [12]–[15].

D. Memory and Knowledge Retrieval

Memory helps keep track of what you're doing when you're working on something that hasn't finished yet. During a multistep process, the agent must keep track of which tasks have already been done, which actions are left to do, and what information was gathered from earlier steps. Keeping the task state helps avoid doing the same work again and lets the agent keep going with what it's currently doing [3], [8], [13], [15].

Memory can also help keep track of the context from one interaction to another. In chats or apps that are used over a long time, details from a previous conversation can be helpful when doing a new task later. A good memory system can keep important information and bring it back when needed. Keeping track of every past conversation isn't always a good idea because too much or unnecessary information can make the context too big and might make it harder to make good decisions. So, choosing the right memories and filtering out the less important ones are really important [12], [13], [15].

Knowledge retrieval helps by getting information from outside sources, which works along with memory. An agent can get the needed information,

like documents or records, right when they need to do a task, instead of only using what's already stored in the language model. This is especially helpful when the job requires knowledge that's specific to a field, changes often, or is kept outside the main system [1], [7], [8], [13], [15].

The usual way to get information is by first understanding what you need, then creating a search or retrieval question, gathering possible related information, choosing the most relevant parts, and sending that information to the next step for further processing. The reasoning process can then use this information to help with planning or decide what to do next [1], [3], [13], [15].

Memory can also be linked to the feedback loop of an autonomous agent. The outcome from doing something might give you new knowledge that you can save to use later. In this way, the agent's memory serves as both a place to store past information and a way to keep learning and updating itself by interacting with the environment [8], [12], [13].

Another important consideration is memory management. A smart thinker needs to tell the difference between facts that only last a little while and facts that are useful for a long time. Once a task is done, the temporary details about it can be deleted. However, any useful knowledge or key experiences from the task should be kept. This kind of management helps keep the memory size in check and makes the information that is retrieved more relevant [12], [13], [15].

The trustworthiness of the information we get is just as important. Information from outside sources might not be complete, up-to-date, consistent, or accurate. So, before using any information for big decisions, agents need to check where it came from, how important it is, and how up-to-date it is. Verification methods can help lower the chance that wrong information affects later decisions and actions [1], [4], [13], [15].

Memory and knowledge retrieval give the basic information needed for good cognitive behavior. Memory helps keep things consistent over time, retrieval brings in new information from outside, and reasoning helps decide how to use the information we have. Their integration lets cognitive AI agents manage longer and more complicated tasks, and it also cuts down on the need to rebuild information

over and over again at each step of the process [1], [3], [8], [13], [15].

E. Tool Use and External Interaction

Tool integration also enables cognitive agents to operate beyond the limitations of their internal knowledge. When a task requires real-time information, numerical computation, data retrieval, or interaction with an external service, the agent can identify the appropriate tool and incorporate its output into the ongoing task. This makes the agent more capable of handling tasks that cannot be completed through language generation alone [1], [7], [8], [13], [15].

Tools may include web search systems, databases, APIs, code execution environments, software applications, simulators, and other external services. By selecting and using these tools according to the task requirements, an agent can obtain additional information or perform operations in an external environment [7], [8], [15].

An important part of tool use is tool selection. The agent must determine which available tool is appropriate for the current task and provide the required inputs or parameters correctly. The output produced by the tool must then be interpreted and integrated into the reasoning process so that it can support subsequent planning or action [3], [7], [13], [15].

Tool interaction can also be iterative. An agent may use one tool, examine the result, and then decide whether another tool or another action is required. This allows the agent to combine information from different sources and progressively complete a complex task [1], [7], [8].

However, external interaction also introduces reliability and security concerns. Incorrect tool outputs, inappropriate tool selection, invalid parameters, or unrestricted access to external services can lead to incorrect or harmful actions. Therefore, tool permissions, input validation, output verification, and appropriate monitoring are important for safe agent operation [4], [9], [15].

Overall, tool use provides the connection between an agent's reasoning capabilities and the external world. By combining reasoning with appropriate tools and external interaction, cognitive AI agents can retrieve information, perform computations,

manipulate software systems, and execute actions in dynamic environments [1], [3], [7], [8], [13], [15].

F. Autonomous Action and Feedback

The continuous reasoning-planning-action-feedback loop supporting adaptive autonomous behaviour is shown in Fig. 2.

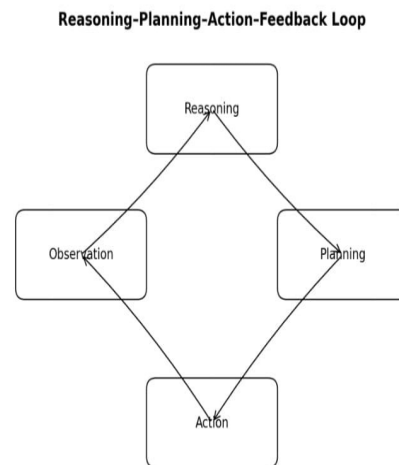


Fig. 2 Continuous reasoning-planning-action-feedback loop supporting adaptive autonomous behaviour.

Autonomous action represents the transition from decision making to the actual execution of tasks. Once the agent has thoroughly thought through and chosen the most effective approach, it must convert that plan into clear, actionable steps that can be implemented. These operations may involve making a request to an external service, controlling a software application, or interacting with a physical environment [1], [7], [8], [13], [15].

For reliable autonomous behavior, every action needs to be related to the current environmental situation. Before performing the action, the agent must verify whether the required conditions for the action are still valid. This is particularly important in situations where conditions change, and the information available at the time of planning might not match the information available when the action is actually taken [3], [7], [12], [13].

Once an action is performed, the agent receives an observation or a result. This observation is compared with the expected outcome. If the anticipated condition is satisfied, the agent can proceed to the next step. If the outcome is not what was anticipated, the agent might return to the thinking or planning

stage to explore an alternative approach [3], [8], [12], [13], [15].

This process creates a closed-loop control system in which actions and observations continuously influence each other. Unlike a fixed automation workflow, the agent does not consistently follow the same sequence of steps, regardless of the conditions in the environment. Instead, it is capable of adjusting its actions based on the information it collects during the execution of the task [7], [8], [13], [15].

Feedback can also assist in recognizing and addressing errors. If an operation fails, the agent can identify whether the failure occurred because of incorrect information, an unavailable tool, an incorrect parameter, a change in the environment, or a plan that did not function as expected. Based on this evaluation, the agent can decide to try the operation again, select a different tool, modify the plan, or request assistance from a human [3], [4], [7], [15]. Another important feature is progress monitoring. During long-running operations, the agent is capable of monitoring both completed and currently active subtasks. This allows us to determine whether the primary objective is getting closer to completion and prevents the system from repeating the same task continuously [3], [8], [13].

In situations where safety is of utmost importance or the consequences are major, autonomous operations must be carried out under appropriate human oversight. The agent is capable of handling routine and low-risk tasks independently, but it requires approval from a human before performing actions that are irreversible, significant, or could result in serious consequences. This provides a good mix of automated operations and attentive control [4], [5], [9], [10].

Therefore, autonomous action should not be viewed as simply following a plan that was established beforehand. It becomes more evident as a recurring process that includes planning, performing actions, observing the results, evaluating the outcomes, and making required changes. This capability is essential for transitioning from systems that can reason in an intelligent way to fully autonomous agents that are focused on achieving specific goals [1], [3], [7], [12]–[15].

G. Multi-agent Collaboration

For difficult tasks, several thinking agents can collaborate instead of depending on just one all-purpose agent. Each agent can have a particular job, like getting information, making plans, carrying them out, checking if they're done, or keeping an eye on things. This setup lets the whole system break down a complicated problem into smaller, easier tasks [3], [9], [12].

A coordinator can give subtasks to different agents who are experts in specific areas. Each agent carries out their specific tasks and shares the results they get along the way. A verification agent can check the results before the final decision is made. Depending on the task, agents might work one after another, with the result from one agent used as the starting point for the next, or they might work at the same time, each handling different parts of the task on their own [3], [9], [12].

Communication is a key part of how multiple agents work together. Agents should share task details, progress updates, what they see, and their choices in a clear and organized way. Good communication helps avoid doing the same tasks over and over, and lets experts use their skills to help reach the main goal [9], [12].

Another advantage of multi-agent collaboration is specialization. Instead of having one agent do all the tasks, different agents can each handle specific skills. For example, one agent can get the needed information, another can create a plan, another can carry out the chosen actions, and a fourth agent can check if the result is correct. Focusing on a specific area can help make complicated tasks easier to manage [3], [9], [12].

Having more agents doesn't always mean better results. As the system gets more complicated, there might be more communication between parts, outputs that don't match up, decisions that clash, and delays in getting things done together. So, when you have multiple agents working together, you need to make sure they have the right way to divide tasks, communicate with each other, settle any disagreements, and check if their results are correct [3], [9], [12].

Multi-agent systems can also help teams work together to solve problems. When agents come up with different ways to solve the same problem, their

results can be looked at and checked before choosing the best one. This adds another step of thinking and checking, especially when there are several ways to solve a problem [3], [12].

Another important consideration is error propagation. If one agent makes a wrong result and another agent takes it without checking, the mistake can keep going through the next steps. So, having separate checks and knowing who is responsible for what helps keep the system working well [4], [9], [12].

Using multiple agents working together helps tackle complicated tasks more effectively by bringing together different expert skills. How well these systems work depends on more than just how smart each agent is. It also relies on how well they work together, share information, assign tasks, check each other's work, and make decisions as a group [3], [9], [12].

H. Applications of Cognitive AI Agents

Cognitive AI agents are being explored across a wide range of application domains because their reasoning, planning, memory, retrieval, and tool-use capabilities can be adapted to different environments. Unlike conventional systems that are designed for a single task, cognitive agents can interpret objectives, perform multiple operations, interact with external resources, and adapt their behaviour according to changing conditions [1], [3], [7], [13]–[15].

In autonomous systems and robotics, cognitive agents can interpret mission objectives, analyse sensor information, generate action plans, use external tools, and modify their behaviour according to environmental changes. LLM-based agents have been investigated for autonomous driving, unmanned aerial vehicles, navigation, planning, and other embodied tasks [7], [8], [13], [15].

In healthcare and medicine, cognitive agents can assist with information retrieval, clinical documentation, workflow coordination, and decision-support activities. They can retrieve relevant information from large medical knowledge sources and support healthcare professionals in organizing complex information. However, human supervision and validation remain important when agent outputs may influence critical decisions [6].

In recommender systems, agentic approaches allow systems to reason about user preferences, retrieve

information, interact with databases, and generate more adaptive recommendations. Compared with static recommendation methods, agentic recommender systems can dynamically use multiple agents and external tools to support recommendation workflows [2].

In software engineering, cognitive agents can support requirements analysis, planning, coding, debugging, testing, and verification. Multiple specialized agents can divide software development activities among different roles and collaborate to complete complex development tasks [3], [12].

In energy and industrial systems, cognitive agents can monitor operational conditions, analyse system data, assist with forecasting, maintenance, optimization, and coordinated decision making. In wind-energy systems, multi-agent architectures have been proposed at turbine, farm, and grid levels to support forecasting, control, maintenance, and resilience [5].

Cognitive AI agents are also being studied in simulation and organizational environments. LLM-based agents can model interactions among individuals, support complex decision making, and simulate behaviours in cyber, physical, social, and hybrid environments [10], [12]. Neuro symbolic multi-agent approaches have additionally been explored to provide more explainable and traceable decision-making in organizational settings [10].

Overall, the applications of cognitive AI agents demonstrate their ability to combine understanding, reasoning, planning, memory, tool interaction, and autonomous execution within a unified workflow. However, the level of autonomy required differs across applications, and high-impact domains require stronger verification, security mechanisms, and human oversight [4], [6], [7], [9].

I. Security, Reliability and Trust

As intelligent entities develop the capability to retrieve information from outside sources and take action, ensuring security becomes a fundamental part of the approach. A harmful input might affect how someone reasons, recalls information, makes plans, or chooses tools. Therefore, security needs to be taken into account not just at the model level but throughout the entire agent architecture [4], [9], [13].

Prompt injection is a significant issue that must be addressed when working with large language model-based agents. An attacker could give directions meant to trick the agent into disregarding its original goal or carrying out a different action than intended. The risk increases when the agent can use external tools or access confidential information [4], [9].

Data and memory security are also important. If incorrect or harmful information gets stored in the agent's memory, it could affect the decisions the agent makes later on. Persistent memory can therefore become a possible point of vulnerability for attacks. Systems should implement suitable validation and filtering processes prior to storing critical information for later use [4], [9], [13].

Tool access introduces another security challenge. An autonomous agent might be allowed to access databases, APIs, files, or other services. Excessive permissions can lead to a greater risk if an action is done incorrectly or with harmful intent. Least privilege access, proper authentication, authorization processes, and restricted tool interfaces can help lower these risks [4], [9].

Reliability can also be impacted by false information, wrong reasoning, mistakes in planning, and unexpected results from tools. A system that generates an output that seems realistic might still have errors in facts or reasoning. Therefore, key results should be validated through reliable sources, external confirmation, consistency assessments, or manual evaluation [1], [3], [4], [13], [15].

Monitoring and logging can provide additional protection. Agent actions, the use of tools, the information gathered, and key choices can all be documented for future review. These records can assist in finding problems, looking into unusual actions, and making the system better as time goes on [4], [9].

Human-in-the-loop mechanisms are especially helpful when dealing with actions that have significant consequences. The system can be designed to ask for human approval before carrying out actions that involve sensitive information, financial impacts, permanent changes, or decisions that affect safety [4], [5], [6], [9].

Trust in cognitive AI agents relies on more than just their capacity to finish tasks. Users should also be assured that the system operates in a secure,

predictable, transparent manner, and remains within established boundaries [1], [4], [9], [10].

j. Research Challenges and Future Directions

Despite rapid progress, cognitive AI agents still face limitations in long-horizon reasoning, reliable planning, persistent memory, tool safety, computational efficiency, and real-world evaluation. An agent may produce a reasonable plan but fail during execution because of an incorrect assumption or an unexpected environmental condition [1], [3], [4], [7], [13], [15].

One major challenge is long-horizon task execution. As the number of steps increases, small errors made during an early stage can influence later decisions. Maintaining the correct task state and continuously verifying intermediate results are therefore important for reliable long-duration operation [3], [7], [13], [15].

Another challenge is reliable memory. Agents need to retain useful information without accumulating irrelevant, outdated, or incorrect information. Future memory systems may require better mechanisms for information selection, updating, forgetting, and verification [1], [8], [13], [15].

Planning reliability is another important research direction. Current systems can generate plausible action sequences, but plausibility does not always mean that the plan will succeed in a real environment. Future approaches may combine language model reasoning with search algorithms, symbolic constraints, simulators, and verification mechanisms [3], [7], [8], [13], [15].

The integration of multimodal information is also expected to become increasingly important. Future cognitive agents may need to combine text, images, audio, video, sensor data, and structured information to understand complex environments. This can improve their ability to operate in physical and digital environments [6], [7], [13].

Another research direction is explainable autonomous decision making. When an agent performs an important action, users may need to understand why a particular plan was selected. Providing understandable explanations and maintaining traceable decision records can improve transparency and user confidence [4], [9], [10], [13].

Scalability is also a concern in multi-agent systems. Increasing the number of collaborating agents can improve specialization but may also increase communication and computational costs. Future research should therefore investigate efficient coordination strategies that maintain performance without unnecessary communication [3], [9], [12].

Standardized evaluation frameworks are needed to compare cognitive agents across different tasks. Evaluation should consider not only final task accuracy but also planning quality, tool-use reliability, adaptability, safety, efficiency, and the ability to recover from failures [1], [4], [6], [7], [13], [15]. Future systems are expected to focus on stronger verification, structured memory, efficient planning, multimodal perception, secure tool interfaces, explainability, and standardized evaluation. Hybrid architectures combining neural models with symbolic reasoning, search, constraints, or external verification may provide better reliability for safety-critical applications [3], [4], [7], [9], [10], [15].

III. RESULT ANALYSIS

The review of current research on Cognitive AI Agents shows a clear shift from systems that mainly produce responses to systems that can reason, plan, interact with external tools, and carry out actions aimed at achieving specific goals [1], [3], [7], [13]–[15]. The evaluated methods show that merging various cognitive abilities can enhance the capacity of AI agents to manage complicated and multi-step tasks [1], [3], [12], [13].

A. Analysis of Reasoning Capabilities

Reasoning plays a crucial role in cognitive AI agents as it enables them to handle complex tasks that cannot be effectively addressed by simply generating direct responses. Techniques like structured reasoning, iterative reasoning, reflection, and reasoning along with action give agents ways to examine intermediate states and choose appropriate solutions [3], [8], [13]–[15].

The analysis indicates that reasoning is more effective when it is linked with outside information and responses from the environment [7], [8], [13], [15]. However, the quality of reasoning can still

differ based on the complexity of the task, the amount of available context, and how reliable the underlying language model is [1], [3], [13], [15].

B. Analysis of Planning Performance

Planning mechanisms allow agents to break down broad goals into smaller, manageable tasks that can be carried out effectively. Task decomposition and hierarchical planning are especially helpful when dealing with complex goals that require several interdependent tasks [3], [12]–[15].

The analysis shows that the ability to replan dynamically is crucial for autonomous systems. A predetermined plan can lose its effectiveness if the surrounding conditions change or if a step in between does not work as intended. Agents that are able to see the outcomes of their actions and adjust their plans accordingly are more effective in environments that change over time [3], [7], [8], [13], [15].

C. Analysis of Memory and Retrieval

Memory and knowledge retrieval help maintain consistency throughout extended tasks. Working memory helps maintain the current task, whereas long-term memory stores important information that can be used over time. RAG-based methods also enable agents to gather information from outside knowledge sources [1], [8], [13], [15].

However, the analysis also points out that memory reliability is a significant issue. Incorrect, outdated, or irrelevant information that is stored in memory can affect future decisions. Therefore, managing memory and verifying information continue to be essential requirements [4], [8], [13], [15].

D. Analysis of Tool Integration

Tool integration greatly enhances the real-world abilities of cognitive AI agents. External search systems, databases, APIs, code execution environments, and software tools enable agents to carry out tasks that cannot be accomplished through language generation alone [1], [7], [8], [13], [15]. The analysis indicates that using tools effectively requires choosing the right tools, creating accurate parameters, understanding the results properly, and confirming their reliability. Unrestricted access to tools can also lead to security and reliability issues,

especially when agents are permitted to carry out sensitive tasks [4], [7], [9], [15].

E. Analysis of Autonomous Action

The transition from planning to autonomous action represents one of the most important developments in agent based AI. An autonomous agent can execute a selected plan, observe the resulting state, evaluate its progress, and modify subsequent actions when necessary [3], [7], [8], [13], [15].

The reasoning–planning–action–feedback loop provides greater adaptability than fixed automation workflows. However, autonomous action also increases the consequences of errors. For this reason, verification and human supervision remain important for high-impact applications [4], [5], [6], [9], [10].

F. Comparative Analysis of Cognitive Agent Capabilities

The high-level capability comparison based on the reviewed survey literature is shown in Fig. 3.

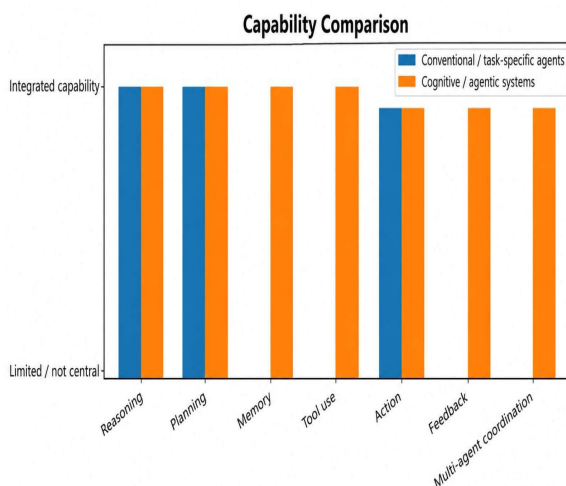


Fig. 3 High-level capability comparison based on the reviewed survey literature; the figure emphasizes integrated capabilities rather than a numerical benchmark.

Cognitive AI agents offer better adaptability than traditional systems designed for specific tasks. Conventional agents typically rely on set rules and fixed tasks, while cognitive agents can use various skills and adjust their actions based on the needs of the task and information from the environment. The use of memory, external tools, and feedback systems enables cognitive agents to keep track of context and

adjust their responses as conditions change during the performance of tasks [1], [3], [7], [13], [15].

IV. CONCLUSIONS

Cognitive AI agents mark a significant advancement in artificial intelligence, as they integrate large language models with capabilities such as reasoning, planning, memory storage, information retrieval, tool utilization, interaction with the environment, and the ability to learn from feedback. The reviewed research indicates that these components function together to enable AI systems to move beyond basic response generation and develop towards goal-directed and more autonomous behaviour [1], [3], [12]–[15]. Reasoning enables agents to comprehend issues and assess potential solutions, whereas planning converts these choices into a series of actions that can be carried out. Memory and retrieval allow agents to access necessary information, while the integration of tools helps them interact with outside systems and environments [7], [8], [13], [15]. Multi-agent systems build upon these capabilities by enabling different agents to interact, work together, and perform intricate tasks through communication and coordination [3], [9], [12]. The survey also indicates that cognitive AI agents are being studied in various fields such as autonomous systems, healthcare, recommender systems, robotics, software engineering, energy systems, and simulation [2], [5]–[7], [10]–[12]. However, the growing degree of autonomy brings significant challenges concerning long-term planning, hallucinations, reliability of memory, improper use of tools, security, privacy, coordination among multiple agents, evaluation processes, explainability, and the need for human supervision [1], [4], [9], [13], [15]. Overall, reliable independence cannot be achieved simply through reasoning or planning alone. Effective cognitive AI agents need a constant interaction among reasoning, planning, memory, using tools, taking actions, observing results, receiving feedback, and adjusting plans accordingly. Future studies should therefore aim to create agents that are not only effective and able to adjust, but also safe, reliable, transparent, verifiable, and able to function consistently in unpredictable real-world situations [1], [3], [4], [9], [12]–[15].

ACKNOWLEDGMENT

The author acknowledges the guidance and support of Prof. Riya K Prakash and Prof.(Dr).L.C.Manikandan during the preparation of this study.

REFERENCES

- [1] A. K. Pati, "Agentic AI: A comprehensive survey of technologies, applications, and societal implications," *IEEE Access*, vol. 13, 2025.
- [2] I. da Silva Portugal, P. Alencar, and D. Cowan, "Agentic recommender systems: A systematic literature review," *IEEE Transactions on Software Engineering*, vol. 52, no. 4, Apr. 2026, doi: 10.1109/TSE.2026.3657900.
- [3] D. B. Acharya, K. Kuppan, and B. Divya, "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, 2025.
- [4] A. Chhabra, S. Datta, S. K. Nahin, and P. Mohapatra, "Agentic AI security: Threats, defenses, evaluation, and open challenges," *IEEE Access*, vol. 14, 2026, doi: 10.1109/ACCESS.2026.3675554.
- [5] A. Hasan, E. Kandemir, D. Mordasov, and D. T. Nguyen, "Agentic AI in wind energy systems: Multi-agent architectures for optimization and resilience," *IEEE Access*, vol. 14, 2026, doi: 10.1109/ACCESS.2026.3654529.
- [6] S. Vatsal, H. Dubey, and A. Singh, "Agentic AI in healthcare and medicine: A seven-dimensional taxonomy for empirical evaluation of LLM-based agents," *IEEE Access*, vol. 14, 2026, doi: 10.1109/ACCESS.2026.3651218.
- [7] M. A. Ferrag, A. Lakas, N. Tihanyi, and M. Debbah, "LLM and AI agents for autonomous systems: A survey of applications, datasets, and security challenges," *IEEE Open Journal of the Industrial Electronics Society*, vol. 7, 2026, doi: 10.1109/OJITS.2026.3665677.
- [8] A. Khamis, "Agentic AI systems: Architecture and evaluation using a frictionless parking scenario," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3590264.
- [9] Y. Wang, Y. Pan, S. Guo, and Z. Su, "Security of Internet of Agents: Attacks and countermeasures," *IEEE Open Journal of the Communications Society*, vol. 6, 2025.
- [10] A. Mertzani, P. G. Giannopoulos, V. Malamas, P. T. Chountalas, and T. K. Dasaklis, "A neurosymbolic multiagent system for explainable and traceable decision-making in organizational contexts," *IEEE Access*, vol. 14, 2026.
- [11] C. Gao, X. Lan, N. Li, Y. Yuan, J. Ding, Z. Zhou, F. Xu, and Y. Li, "Large language models empowered agentbased modeling and simulation: A survey and perspectives," *Humanities and Social Sciences Communications*, vol. 11, art. 1259, 2024, doi: 10.1057/s41599-024-03611-3.
- [12] X. Li, S. Wang, S. Zeng, Y. Wu, and Y. Yang, "A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges," *Vicinagearth*, vol. 1, art. 9, 2024, doi: 10.1007/s44336-024-00009-2.
- [13] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, art. 186345, 2024, doi: 10.1007/s11704-024-40231-1.
- [14] L. Dinc, kal, "Large language model-based autonomous agents: Trends and directions," *AIPA's International Journal on Artificial Intelligence: Bridging Technology, Society and Policy*, vol. 1, no. 1, pp. 13–24, 2024, doi: 10.5281/zenodo.14293261.
- [15] S. S. Chowda, R. Alvi, S. S. Rahman, M. A. Rahman, M. A. K. Raiaan, M. R. Islam, M. Hussain, and S. Azam, "From language to action: A review of large language models as autonomous agents and tool users," *Artificial Intelligence Review*, 2026, doi: 10.1007/s10462-025-11471-9.