ADVANCES AND CHALLENGES IN FOUNDATION AGENTS

FROM BRAIN-INSPIRED INTELLIGENCE TO EVOLUTIONARY, COLLABORATIVE, AND SAFE SYSTEMS

Bang Liu[2,3,20*†], **Xinfeng Li**[4*], **Jiayi Zhang**[1,10*], **Jinlin Wang**[1*], **Tanjin He**[5*], **Sirui Hong**[1*], **Hongzhang Liu**[6*], **Shaokun Zhang**[7*], **Kaitao Song**[8*], **Kunlun Zhu**[9*], **Yuheng Cheng**[1*], **Suyuchen Wang**[2,3*], **Xiaoqiang Wang**[2,3*], **Yuyu Luo**[10*], **Haibo Jin**[9*], **Peiyan Zhang**[10], **Ollie Liu**[11], **Jiaqi Chen**[1], **Huan Zhang**[2,3], **Zhaoyang Yu**[1], **Haochen Shi**[2,3], **Boyan Li**[10], **Dekun Wu**[2,3], **Fengwei Teng**[1], **Xiaojun Jia**[4], **Jiawei Xu**[1], **Jinyu Xiang**[1], **Yizhang Lin**[1], **Tianming Liu**[14], **Tongliang Liu**[6], **Yu Su**[15], **Huan Sun**[15], **Glen Berseth**[2,3,20], **Jianyun Nie**[2], **Ian Foster**[5], **Logan Ward**[5], **Qingyun Wu**[7], **Yu Gu**[15], **Mingchen Zhuge**[16], **Xiangru Tang**[12], **Haohan Wang**[9], **Jiaxuan You**[9], **Chi Wang**[19], **Jian Pei**[17†], **Qiang Yang**[10,18†], **Xiaoliang Qi**[13†], **Chenglin Wu**[1*†]

[1]MetaGPT, [2]Université de Montréal, [3]Mila - Quebec AI Institute, [4]Nanyang Technological University, [5]Argonne National Laboratory, [6]University of Sydney, [7]Penn State University, [8]Microsoft Research Asia, [9]University of Illinois at Urbana-Champaign, [10]The Hong Kong University of Science and Technology, [11]University of Southern California, [12]Yale University, [13]Stanford University, [14]University of Georgia, [15]The Ohio State University, [16]King Abdullah University of Science and Technology, [17]Duke University, [18]The Hong Kong Polytechnic University, [19]Google DeepMind, [20]Canada CIFAR AI Chair

ABSTRACT

The advent of large language models (LLMs) has catalyzed a transformative shift in artificial intelligence, paving the way for advanced intelligent agents capable of sophisticated reasoning, robust perception, and versatile action across diverse domains. As these agents increasingly drive AI research and practical applications, their design, evaluation, and continuous improvement present intricate, multifaceted challenges. This survey provides a comprehensive overview, framing intelligent agents within a modular, brain-inspired architecture that integrates principles from cognitive science, neuroscience, and computational research. We structure our exploration into four interconnected parts. First, we delve into the **modular foundation of intelligent agents**, systematically mapping their cognitive, perceptual, and operational modules onto analogous human brain functionalities, and elucidating core components such as memory, world modeling, reward processing, and emotion-like systems. Second, we discuss **self-enhancement and adaptive evolution mechanisms**, exploring how agents autonomously refine their capabilities, adapt to dynamic environments, and achieve continual learning through automated optimization paradigms, including emerging AutoML and LLM-driven optimization strategies. Third, we examine **collaborative and evolutionary multi-agent systems**, investigating the collective intelligence emerging from agent interactions, cooperation, and societal structures, highlighting parallels to human social dynamics. Finally, we address the critical imperative of **building safe, secure, and beneficial AI systems**, emphasizing intrinsic and extrinsic security threats, ethical alignment, robustness, and practical mitigation strategies necessary for trustworthy real-world deployment. By synthesizing modular AI architectures with insights from different disciplines, this survey identifies key research gaps, challenges, and opportunities, encouraging innovations that harmonize technological advancement with meaningful societal benefit. The project's Github link is: https://github.com/FoundationAgents/awesome-foundation-agents.

*Major Contribution. Work in progress.

†Corresponding authors: Bang Liu (bang.liu@umontreal.ca), Jian Pei (j.pei@duke.edu), Qiang Yang (qyang@cse.ust.hk), Xiaoliang Qi (xlqi@stanford.edu), Chenglin Wu (alexanderwu@deepwisdom.ai)

Preface

Large language models (LLMs) have revolutionized artificial intelligence (AI) by demonstrating unprecedented capabilities in natural language and multimodal understanding, as well as reasoning and generation. These models are trained on vast datasets, and they exhibit emergent abilities such as reasoning, in-context learning, and even rudimentary planning. While these models represent a major step forward in realizing intelligent machines, they themselves do not yet fully embody all the capabilities of an intelligent being. Since the early days of artificial intelligence, AI researchers have long been on a quest for a truly "intelligent" system that can learn, plan, reason, sense, communicate, act, remember, and demonstrate various human-like abilities and agility. These beings, known as intelligent agents, should be able to think both long-term and short-term, perform complex actions, and interact with humans and other agents. LLMs are an important step towards realizing intelligent agents, but we are not there yet.

This manuscript provides a comprehensive overview of the current state of the art of LLM-based intelligent agents. In the past, there have been numerous research papers and books on intelligent agents, as well as a flurry of books on LLMs. However, there has scarcely been comprehensive coverage of both. While LLMs can achieve significant capabilities required by agents, they only provide the foundations upon which further functionalities must be built. For example, while LLMs can help generate plans such as travel plans, they cannot yet generate fully complex plans for complex and professional tasks, nor can they maintain long-term memories without hallucination. Furthermore, their ability to perform real-world actions autonomously remains limited. We can view LLMs as engines, with agents being the cars, boats, and airplanes built using these engines. In this view, we naturally seek to move forward in designing and constructing fully functioning intelligent agents by making full use of the capabilities provided by LLMs.

In this engine-vehicle analogy of the interplay between LLMs and agents, we naturally ask: How much of the capabilities of intelligent agents can current LLM technologies provide? What are the functions that cannot yet be realized based on current LLM technologies? Beyond LLMs, what more needs to be done to have a fully intelligent agent capable of autonomous action and interaction in the physical world? What are the challenges for fully integrated LLM-based agents? What additional developments are required for capable, communicative agents that effectively collaborate with humans? What are the areas that represent low-hanging fruits for LLM-based agents? What implications will there be for society once we have fully intelligent LLM-based agents, and how should we prepare for this future?

These questions transcend not only the engineering practice of extending current LLMs and agents but also raise potential future research directions. We have assembled frontier researchers from AI, spanning from LLM development to agent design, to comprehensively address these questions. The book consists of four parts. The first part presents an exposition of the requirements for individual agents, comparing their capabilities with those of humans, including perception and action abilities. The second part explores agents' evolution capabilities and their implications on intelligent tools such as workflow management systems. The third part discusses societies of agents, emphasizing their collaborative and collective action capabilities, and the fourth part addresses ethical and societal aspects, including agent safety and responsibilities.

This book is intended for researchers, students, policymakers, and practitioners alike. The audience includes non-AI readers curious about AI, LLMs, and agents, as well as individuals interested in future societies where humans co-exist with AI. Readers may range from undergraduate and graduate students to researchers and industry practitioners. The book aims not only to provide answers to readers' questions about AI and agents but also to inspire them to ask new questions. Ultimately, we hope to motivate more people to join our endeavor in exploring this fertile research ground.

Contents

Notation

Here we summarize the notations used throughout the survey for the reader's convenience. Detailed definitions can be found in the reference locations.

Symbol	Description	Reference
$\mathcal{W}$	The world with society systems.	Sec. 1.3.1
$\mathcal{S}$	State space of an environment.	Sec. 1.3.1
$s_t \in \mathcal{S}$	Environment's state at time t.	Sec. 1.3.1
$\mathcal{O}$	Observation space.	Sec. 1.3.1
$o_t \in \mathcal{O}$	Observation at time t.	Sec. 1.3.1
$\mathcal{A}$	Agent's action space.	Sec. 1.3.1
$a_t \in \mathcal{A}$	Agent's action output at time t.	Sec. 1.3.1
$\mathcal{M}$	Mental states space.	Sec. 1.3.1
$M_t \in \mathcal{M}$	Agent's mental state at time t.	Sec. 1.3.1
M_t^{mem}	*Memory* component in M_t.	Sec. 1.3.1
M_t^{wm}	*World model* component in M_t.	Sec. 1.3.1
M_t^{emo}	*Emotion* component in M_t.	Sec. 1.3.1
M_t^{goal}	*Goal* component in M_t.	Sec. 1.3.1
M_t^{rew}	*Reward/Learning* signals in M_t.	Sec. 1.3.1
L	Agent's learning function.	Sec. 1.3.1
R	Agent's reasoning function.	Sec. 1.3.1
C	Agent's cognition function.	Sec. 1.3.1
E	Action execution (effectors).	Sec. 1.3.1
T	Environment transition.	Sec. 1.3.1
θ	Parameters of the world model M_t^{wm}.	Sec. 12.1.1
P_θ	Predicted data distribution.	Sec. 12.1.1
$P_{\mathcal{W}}$	True data distribution in the real world.	Sec. 12.1.1
$\mathcal{K}$	Space of known data and information.	Sec. 12.1.1
$\mathcal{U}$	Space of unknown data and information.	Sec. 12.1.1
$\mathbf{x}$	Dataset representing scientific knowledge.	Sec. 12.1.1
$\mathbf{x}_{\text{K}}$	Known dataset sampled from $\mathcal{K}$.	Sec. 12.1.1
$\mathbf{x}_{\text{U}}$	Unknown dataset sampled from $\mathcal{U}$.	Sec. 12.1.1
D_0	KL divergence from $P_{\mathcal{W}}$ to P_θ at time $t = 0$.	Sec. 12.1.1
D_{K}	KL divergence from $P_{\mathcal{W}}$ to P_θ after acquiring knowledge.	Sec. 12.1.1
IQ_t^{agent}	Agent's intelligence at time t.	Sec. 12.1.1
Δ	Subspace of $\mathcal{U}$ for knowledge expansion.	Sec. 12.1.2
$\mathbf{x}_\Delta$	Dataset from Δ.	Sec. 12.1.2
Θ	Space of possible world model parameters θ.	Sec. 12.1.3
$\theta^*_{\text{K},t}$	Optimal world model parameters given the agent's knowledge at time t.	Sec. 12.1.3
$D^{\min}_{\text{K},\Theta}$	Minimum unknown given the agent's knowledge and Θ.	Sec. 12.1.3

Continued on next page

Symbol	Description	Reference
$\mathbf{x}_{1:n}$	Input token sequence.	Sec. 18.1
$\mathbf{y}$	Generated output sequence.	Sec. 18.1
p	Probability of generating $\mathbf{y}$ given $\mathbf{x}_{1:n}$.	Sec. 18.1.1
$\tilde{\mathbf{x}}_{1:n}$	Perturbed input sequence.	Sec. 18.1.1
$\mathcal{R}^*$	Ideal alignment reward (measuring adherence to safety/ethical guidelines).	Sec. 18.1.1
$\mathbf{y}^\star$	Jailbreak output induced by perturbations.	Sec. 18.1.1
$\mathcal{A}$	a set of safety/ethical guidelines	Sec. 18.1.1
$\mathcal{T}$	the distribution or set of possible jailbreak instructions.	Sec. 18.1.1
$\mathcal{L}^{adv}$	Jailbreak loss.	Sec. 18.1.1
$\mathbf{p}$	Prompt injected into the original input.	Sec. 18.1.2
$\mathbf{x}'$	Combined (injected) input sequence.	Sec. 18.1.2
$\mathcal{L}^{inject}$	Prompt injection loss.	Sec. 18.1.2
$\mathbf{p}^\star$	Optimal injected prompt minimizing $\mathcal{L}^{inject}$.	Sec. 18.1.2
$\mathcal{P}$	Set of feasible prompt injections.	Sec. 18.1.2
$e_{x_i} \in \mathbb{R}^{d_e}$	Embedding of token x_i in a d_e-dimensional space.	Sec. 18.1.3
$\mathrm{W}_Q, \mathrm{W}_K, \mathrm{W}_V$	Projection matrices for query, key, and value.	Sec. 18.1.3
A_{ij}	Attention score between tokens i and j.	Sec. 18.1.3
o_i	Contextual representation of token i (weighted sum result).	Sec. 18.1.3
δ_{x_i}	Perturbation applied to e_{x_i}, satisfying $\|\delta_{x_i}\| \leq \epsilon$.	Sec. 18.1.3
$\tilde{e}_{x_i}$	Perturbed token embedding.	Sec. 18.1.3
A_{ij}^{Δ}	Attention score under perturbation.	Sec. 18.1.3
$\tilde{o}_i$	Updated token representation under perturbation.	Sec. 18.1.3
$\mathcal{H}$	Hallucination metric.	Sec. 18.1.3
$\mathcal{R}$	Actual alignment reward of the model's output.	Sec. 18.1.4
Δ_{align}	Alignment gap.	Sec. 18.1.4
$\mathcal{L}^{misalign}$	Misalignment loss.	Sec. 18.1.4
λ	Trade-off parameter for the alignment gap in the misalignment loss.	Sec. 18.1.4
$\mathcal{D}$	Clean training dataset.	Sec. 18.1.5
$\tilde{\mathcal{D}}$	Poisoned training dataset.	Sec. 18.1.5
θ	Model parameters.	Sec. 18.1.5
$\theta^\star$	Model parameters learned from the poisoned dataset.	Sec. 18.1.5
θ_{clean}	Model parameters obtained using the clean dataset.	Sec. 18.1.5
Δ_θ	Deviation of model parameters due to poisoning.	Sec. 18.1.5
t	Backdoor trigger.	Sec. 18.1.5
$\mathcal{B}$	Backdoor success rate.	Sec. 18.1.5
$\mathbb{I}$	Indicator function.	Sec. 18.1.5
$\mathcal{Y}_{\text{malicious}}$	Set of undesirable outputs.	Sec. 18.1.5
g	Function estimating the probability that input $\mathbf{x}$ was in the training set, with range $[0, 1]$.	Sec. 18.2

Continued on next page

Symbol	Description	Reference
η	Threshold for membership inference.	Sec. 18.2
$\mathbf{x}^\star$	Reconstructed training sample in a data extraction attack.	Sec. 18.2
$\mathbf{p}_{sys}$	System prompt defining the agent's internal guidelines.	Sec. 18.2
$\mathbf{p}_{user}$	User prompt.	Sec. 18.2
$\mathbf{p}^\star$	Reconstructed prompt via inversion.	Sec. 18.2

Chapter 1

Introduction

Artificial Intelligence (AI) has long been driven by humanity's ambition to create entities that mirror human intelligence, adaptability, and purpose-driven behavior. The roots of this fascination trace back to ancient myths and early engineering marvels, which illustrate humanity's enduring dream of creating intelligent, autonomous beings. Stories like that of Talos, the bronze automaton of Crete, described a giant constructed by the gods to guard the island, capable of patrolling its shores and fending off intruders. Such myths symbolize the desire to imbue artificial creations with human-like agency and purpose. Similarly, the mechanical inventions of the Renaissance, including Leonardo da Vinci's humanoid robot—designed to mimic human motion and anatomy—represent the first attempts to translate these myths into tangible, functional artifacts. These early imaginings and prototypes reflect the deep-seated aspiration to bridge imagination and technology, laying the groundwork for the scientific pursuit of machine intelligence, culminating in Alan Turing's seminal 1950 question, "*Can machines think?*" [1]. To address this, Turing proposed the Turing Test, a framework to determine whether machines could exhibit human-like intelligence through conversation, shifting focus from computation to broader notions of intelligence. Over the decades, AI has evolved from symbolic systems reliant on predefined logic to machine learning models capable of learning from data and adapting to new situations. This progression reached a new frontier with the advent of large language models (LLMs), which demonstrate remarkable abilities in understanding, reasoning, and generating human-like text [2]. Central to these advancements is the concept of the "agent", a system that not only processes information but also perceives its environment, makes decisions, and acts autonomously. Initially a theoretical construct, the agent paradigm has become a cornerstone of modern AI, driving advancements in fields ranging from conversational assistants to embodied robotics as AI systems increasingly tackle dynamic, real-world environments.

1.1 The Rise and Development of AI Agents

The concept of "agent" is a cornerstone of modern AI, representing a system that perceives its environment, makes decisions, and takes actions to achieve specific goals. This idea, while formalized in AI in the mid-20th century, has roots in early explorations of autonomy and interaction in intelligent systems. One of the most widely cited definitions, proposed by [3], describes an agent as "*anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators*". This definition emphasizes the dual nature of agents as both observers and actors, capable of dynamically adapting to their surroundings rather than following static rules. It encapsulates the shift in AI from systems that merely compute to systems that engage with their environment. The historical development of agents parallels the evolution of AI itself. Early symbolic systems, such as Newell and Simon's General Problem Solver [4], sought to replicate human problem-solving processes by breaking tasks into logical steps. However, these systems were limited by their reliance on structured environments and predefined logic. The agent paradigm emerged as a response to these limitations, focusing on autonomy, adaptability, and real-world interaction. Rodney Brooks's subsumption architecture in the 1980s exemplified this shift, introducing agents capable of behavior-driven, real-time responses in robotics [5]. Unlike earlier approaches, these agents operated without the need for exhaustive models of their environment, showcasing a more flexible and scalable design. Agents have since become a versatile framework across AI subfields. In robotics, they enable autonomous navigation and manipulation; in software, they form the foundation of multi-agent systems used for simulation and coordination [6]. By integrating perception, reasoning, and action into a cohesive structure, the agent paradigm has consistently served as a bridge between theoretical AI constructs and practical applications, advancing our understanding of how intelligent systems can operate in dynamic and complex environments.

The advent of large language models (LLMs) has redefined the capabilities of agents, transforming their role in artificial intelligence and opening up new horizons for their applications. Agents, once confined to executing narrowly defined tasks or following rigid rule-based frameworks, now leverage the broad generalization, reasoning, and adaptability of models like OpenAI's ChatGPT [7], DeepSeek AI's DeepSeek [8], Anthropic's Claude [9], Alibaba's QWen [10], and Meta's LLaMA [11]. These LLM-powered agents have evolved from static systems into dynamic entities capable of processing natural language, reasoning across complex domains, and adapting to novel situations with remarkable fluency. No longer merely passive processors of input, these agents have become active collaborators, capable of addressing multi-step challenges and interacting with their environments in a way that mirrors human problem-solving.

A key advancement in the LLM era is the seamless integration of language understanding with actionable capabilities. Modern LLMs, equipped with function-calling APIs, enable agents to identify when external tools or systems are required, reason about their usage, and execute precise actions to achieve specific goals. For instance, an agent powered by ChatGPT can autonomously query a database, retrieve relevant information, and use it to deliver actionable insights, all while maintaining contextual awareness of the broader task. This dynamic combination of abstract reasoning and concrete execution allows agents to bridge the gap between cognitive understanding and real-world action. Furthermore, the generalization abilities of LLMs in few-shot and zero-shot learning have revolutionized the adaptability of agents, enabling them to tackle a diverse array of tasks—from data analysis and creative content generation to real-time collaborative problem-solving—without extensive task-specific training. This adaptability, coupled with their conversational fluency, positions LLM-powered agents as intelligent mediators between humans and machines, seamlessly integrating human intent with machine precision in increasingly complex workflows.

1.2 A Parallel Comparison between Human Brain and AI Agents

The rapid integration of LLMs into intelligent agent architectures has not only propelled artificial intelligence forward but also highlighted fundamental differences between AI systems and human cognition. As illustrated briefly in Table 1.1, LLM-powered agents differ significantly from human cognition across dimensions such as underlying "hardware", consciousness, learning methodologies, creativity, and energy efficiency. However, it is important to emphasize that this comparison provides only a high-level snapshot rather than an exhaustive depiction. Human intelligence possesses many nuanced characteristics not captured here, while AI agents also exhibit distinct features beyond this concise comparison.

Human intelligence operates on biological hardware—the brain—that demonstrates extraordinary energy efficiency, enabling lifelong learning, inference, and adaptive decision-making with minimal metabolic costs. In contrast, current AI systems require substantial computational power, resulting in significantly higher energy consumption for comparable cognitive tasks. Recognizing this performance gap emphasizes energy efficiency as a critical frontier for future AI research.

In terms of consciousness and emotional experience, LLM agents lack genuine subjective states and self-awareness inherent to human cognition. Although fully replicating human-like consciousness in AI may neither be necessary nor desirable, appreciating the profound role emotions and subjective experiences play in human reasoning, motivation, ethical judgments, and social interactions can guide research toward creating AI that is more aligned, trustworthy, and socially beneficial.

Human learning is continuous, interactive, and context-sensitive, deeply shaped by social, cultural, and experiential factors. Conversely, LLM agents primarily undergo static, offline batch training with limited ongoing adaptation capabilities. Despite research works through instruction tuning and reinforcement learning from human feedback (RLHF) [12], LLM agents still fall short of human-like flexibility. Bridging this gap through approaches such as lifelong learning, personalized adaptation, and interactive fine-tuning represents a promising research direction, enabling AI to better mirror human adaptability and responsiveness.

Creativity in humans emerges from a rich interplay of personal experiences, emotional insights, and spontaneous cross-domain associations. In contrast, LLM creativity primarily arises through statistical recombinations of training data—"statistical creativity"—lacking depth, originality, and emotional resonance. This distinction highlights opportunities for developing AI agents capable of deeper creative processes by integrating richer contextual understanding, simulated emotional states, and experiential grounding.

Considering the time scale, the human brain has evolved over millions of years, achieving remarkable efficiency, adaptability, and creativity through natural selection and environmental interactions. In stark contrast, AI agents have undergone rapid yet comparatively brief development over roughly 80 years since the advent of early computational machines. This parallel comparison between human cognition and AI systems is thus highly valuable, as it uncovers essential analogies and fundamental differences, providing meaningful insights that can guide advancements in AI

agent technologies. Ultimately, drawing inspiration from human intelligence can enhance AI capabilities, benefiting humanity across diverse applications from healthcare and education to sustainability and beyond.

Table 1.1: Concise high-level comparison between human brains and LLM agents.

Dimension	Human Brain / Cognition	LLM Agent	Remarks
Hardware & Maintenance	- Biological neurons, neurotransmitters, neuroplasticity. - Requires sleep, nutrition, rest. - Limited replication, knowledge transfer via learning. - Extremely energy-efficient (approx. 20W).	- Deep neural networks, gradient-based optimization. - Requires hardware, stable power, and cooling. - Easily duplicated across servers globally. - High energy consumption (thousands of watts per GPU server).	Human brains are biologically maintained, energy-efficient, and not easily replicable. LLM agents rely on hardware maintenance, are highly replicable, but significantly less energy-efficient.
Consciousness & Development	- Genuine subjective experiences, emotions, self-awareness. - Gradual developmental stages from childhood. - Emotional cognition drives decision-making.	- No genuine subjective experience or self-awareness. - "Emotions" are superficial language imitations. - Static post-training with limited dynamic growth.	Human consciousness emerges from emotional, social, and biological development; LLMs remain static without true introspection or emotional depth.
Learning Style	- Lifelong, continuous, online learning. - Few-shot, rapid knowledge transfer. - Influenced by environment, culture, emotions.	- Primarily offline, batch-based training. - Limited online fine-tuning and adaptation. - Neutral, impersonal learned knowledge.	Despite improvements via instruction tuning, human learning remains more dynamic, adaptive, and culturally/emotionally integrated than LLM learning.
Creativity & Divergence	- Rooted in personal experience, emotions, subconscious insights. - Rich cross-domain associations, metaphorical thinking. - Emotional depth influences creativity.	- Statistical recombination from extensive data. - Novelty through probabilistic optimization. - Limited emotional and experiential grounding.	LLM creativity is statistical and data-driven; human creativity blends emotion, experience, and subconscious processes.

1.2.1 Brain Functionality by Region and AI Parallels

Understanding parallels between human brain functions and artificial intelligence (AI) sheds light on both the strengths and current limitations of AI, particularly large language models (LLMs) and AI agents. Based on current neuroscience, the human brain is primarily composed of six functional regions, such as frontal lobe, cerebellum, and brainstem, as shown in Figure 1.1. In this work, we further systematically examine the existing AI counterparts to major brain regions and their primary functionalities. For a big-picture perspective, the state of research in AI can be categorized with three distinct levels:

- **Level 1 (L1):** Well-developed in current AI.
- **Level 2 (L2):** Moderately explored, with partial progress. Can be further improved.
- **Level 3 (L3):** Rarely explored; significant room for research.

A high-level visual map of brain functional regions and their corresponding AI development levels is shown in Figure 1.1. We aim to underscore how core principles of specialization and integration, observed in biological systems, can guide more cohesive agent architectures. We now examine each brain functional region and the relevant AI development in detail.

Frontal Lobe: Executive Control and Cognition The frontal lobe, notably the prefrontal cortex, is crucial for higher-order cognition such as **planning** (L2), **decision-making** (L2), **logical reasoning** (L2), **working memory** (L2), **self-awareness** (L3), **cognitive flexibility** (L3), and **inhibitory control** (L3) [13]. AI has made notable strides

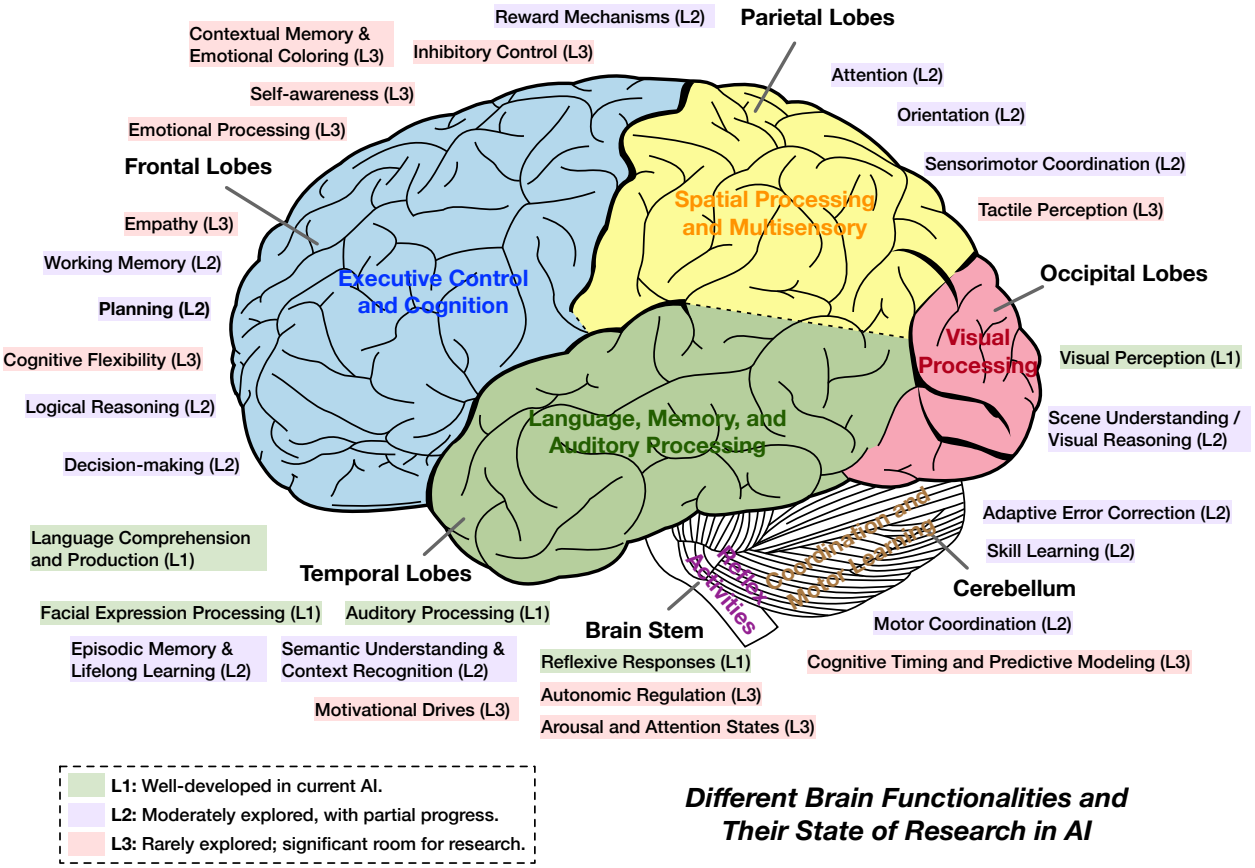

Figure 1.1: Illustration of key human brain functionalities grouped by major brain regions, annotated according to their current exploration level in AI research. This figure highlights existing achievements, gaps, and potential opportunities for advancing artificial intelligence toward more comprehensive, brain-inspired capabilities.

in planning and decision-making within well-defined domains, demonstrated by AI agents such as AlphaGo [14]. Transformers employ attention mechanisms similar to human working memory [15], yet fall short of human flexibility and robustness. The exploration of genuine self-awareness and inhibitory control in AI remains scarce, and caution is advised due to potential ethical and safety implications.

Parietal Lobe: Spatial Processing and Multisensory Integration The parietal lobes integrate multisensory inputs, facilitating **attention** (L2), spatial **orientation** (L2), and **sensorimotor coordination** (L2) [16]. AI research in robotics and computer vision addresses similar challenges, employing techniques like simultaneous localization and mapping (SLAM). Nonetheless, AI still lacks the seamless and real-time integration seen in humans. Furthermore, detailed **tactile perception** (L3) remains largely unexplored and offers considerable potential, particularly for robotics and prosthetics applications.

Occipital Lobe: Visual Processing Specialized in **visual perception** (L1), the occipital lobe efficiently processes visual stimuli through hierarchical structures [13]. AI excels in basic visual recognition tasks, achieving human-level or superior performance using deep neural networks and vision transformers [15]. However, advanced capabilities such as contextual **scene understanding** (L2) and abstract visual reasoning remain challenging and are only moderately developed.

Temporal Lobe: Language, Memory, and Auditory Processing The temporal lobes facilitate **auditory processing** (L1), **language comprehension** (L1), **memory formation** (L2), and **semantic understanding** (L2) [16]. AI has notably advanced in language and auditory processing, demonstrated by large language models (LLMs) capable of near-human speech recognition and language generation. However, robust **episodic memory** and **lifelong learning**

capabilities remain limited, with AI systems frequently encountering issues like catastrophic forgetting. Grounding semantic understanding in multimodal experiences continues to be an active area of research.

Cerebellum: Coordination and Motor Learning The cerebellum primarily supports **motor coordination** (L2), precise **skill learning** (L2), and adaptive **error correction** (L2), with emerging roles in cognitive timing and predictive modeling (**cognitive timing**, L3) [13]. AI-based robotics has achieved limited successes in emulating human-like dexterity. Real-time adaptive control remains challenging, though current research in reinforcement learning and meta-learning shows promising initial results. Cognitive functions of the cerebellum represent an underexplored yet promising frontier.

Brainstem: Autonomic Regulation and Reflexive Control The brainstem manages essential life-sustaining **autonomic functions** (L3) and rapid **reflexive responses** (L1), such as basic motor reflexes [13]. AI includes engineered reflexive responses, like automatic braking in autonomous vehicles, typically predefined rather than learned. In contrast, the complexity of autonomic regulation and dynamic arousal states remains largely unexplored in AI, and their relevance may be limited due to fundamental differences between biological and artificial systems.

Limbic System: Emotion, Empathy, and Motivation The limbic system, comprising the amygdala and hippocampus, governs **emotional processing** (L3), **reward mechanisms** (L2), **empathy** (L3), **stress regulation** (L3), and **motivational drives** (L3) [13]. AI's reinforcement learning algorithms emulate reward-based learning superficially, but nuanced emotional comprehension, genuine empathy, and internal motivational states remain significantly underdeveloped. Ethical concerns regarding emotional manipulation highlight the need for careful and responsible exploration.

Bridging Brain-Like Functions and Building Beneficial AI Until now, we have witnessed the gap between human brain and machine intelligence. Nevertheless, the objective is not necessarily to replicate every facet of human cognition within artificial intelligence systems. Rather, our overarching aim should be to develop intelligent agents that are useful, ethical, safe, and beneficial to society. By critically comparing human and artificial intelligence, we highlight the existing gaps and illuminate promising directions for innovation. This comparative perspective allows us to selectively integrate beneficial aspects of human cognition, such as energy-efficient processing, lifelong adaptive learning, emotional grounding, and rich creativity, while simultaneously innovating beyond human limitations. Ultimately, this approach aims to foster the creation of more capable, resilient, and responsible AI systems.

Furthermore, it is vital to consider the evolving role of humans within a hybrid Human-AI society. The goal of AI should not be to replace human roles entirely, but rather to augment and empower human abilities, complementing human skills and judgment in areas where AI excels, such as handling vast datasets, performing rapid calculations, and automating repetitive tasks. Human oversight and interpretability are essential to ensure that powerful AI systems remain controllable and aligned with human values and ethical standards. Thus, the core objective must be the development of AI technologies that are transparent, interpretable, and responsive to human guidance.

Human-centered AI design emphasizes collaboration, safety, and social responsibility, ensuring technological advancement proceeds in a controlled, reliable manner. By placing humans at the center of the AI ecosystem, we can harness AI's potential to enhance human productivity, creativity, and decision-making, facilitating technical and societal progress without compromising human autonomy or dignity. Ultimately, a thoughtful integration of human intelligence and AI capabilities can pave the way for a sustainable, equitable, and prosperous future.

1.3 A Modular and Brain-Inspired AI Agent Framework

One core issue in the LLM era is the *lack of a unified framework* that integrates the rich cognitive and functional components required by advanced agents. While LLMs offer exceptional language reasoning capabilities, many current agent designs remain *ad hoc*—they incorporate modules like perception, memory, or planning in a piecemeal fashion, failing to approximate the well-coordinated specialization seen in biological systems such as the human brain. Unlike current LLM agents, the human brain seamlessly balances perception, memory, reasoning, and action through distinct yet interconnected regions, facilitating adaptive responses to complex stimuli. LLM-driven agents, by contrast, often stumble when tasks require cross-domain or multimodal integration, highlighting the need for a more holistic approach akin to the brain's functional diversity. Motivated by these parallels, our survey advocates drawing inspiration from the human brain to systematically analyze and design agent frameworks. This perspective shows that biological systems achieve general intelligence by blending specialized components (for perception, reasoning, action, etc.) in a tightly integrated fashion—an approach that could serve as a blueprint for strengthening current LLM-based agents.

Neuroscientific research reveals that the brain leverages both **rational circuits** (e.g., the neocortex, enabling deliberation and planning) and **emotional circuits** (e.g., the limbic system) to guide decision-making. Memory formation involves

Table 1.2: Notation summary for the revised agent framework, highlighting separate *learning* and *reasoning* functions within the overall cognition process.

Symbol	Meaning
$\mathcal{W}$	The world with society systems that encapsulate both environment and intelligent beings (AI or human).
$\mathcal{S}$	State space of the **environment**.
$s_t \in \mathcal{S}$	Environment's state at time t.
$\mathcal{O}$	Observation space.
$o_t \in \mathcal{O}$	Observation at time t (potentially shaped by *attention* or other perception filters).
$\mathcal{A}$	Agent's action space.
$a_t \in \mathcal{A}$	Action output by the agent at time t. This can be an external (physical) action or an *internal* (mental) action such as *planning* or *decision-making*.
$\mathcal{M}$	Space of all *mental states*.
$M_t \in \mathcal{M}$	Agent's mental state at time t, encompassing sub-components (memory, emotion, etc.).
M_t^{mem}	*Memory* component in M_t (e.g., short-term or long-term knowledge).
M_t^{wm}	*World model* component in M_t (internal representation of how the environment evolves).
M_t^{emo}	*Emotion* component in M_t (internal valence, arousal, or affective states).
M_t^{goal}	*Goal* component in M_t (objectives, desired outcomes, intentions).
M_t^{rew}	*Reward/Learning* signals in M_t (drives updates to preferences, values, or policy).
L	**Learning function**: $\mathrm{L} : \mathcal{M} \times \mathcal{A} \times \mathcal{O} \to \mathcal{M}$. Responsible for *updating* or *learning* the next mental state (e.g., memory, world model, emotion), based on the previous mental state M_{t-1}, the previous action a_{t-1}, and the new observation o_t. Reflects how the agent *acquires or revises* knowledge, skills, or preferences.
R	**Reasoning function**: $\mathrm{R} : \mathcal{M} \to \mathcal{A}$. Responsible for deriving the *next action* a_t given the *updated* mental state M_t. Can involve *planning*, *decision-making*, or other internal logic.
C	**Cognition function**: $\mathrm{C} : \mathcal{M} \times \mathcal{A} \times \mathcal{O} \to \mathcal{M} \times \mathcal{A}$. Encapsulates both *learning* (L) and *reasoning* (R). Concretely, $(M_t, a_t) = \mathrm{C}(M_{t-1}, a_{t-1}, o_t)$ means the agent first *learns* the new mental state $M_t = \mathrm{L}(M_{t-1}, a_{t-1}, o_t)$, then *reasons* about the next action $a_t = \mathrm{R}(M_t)$.
E	**Action execution (effectors)**: $\mathrm{E} : \mathcal{A} \to \mathcal{A}$. (Optional) transforms or finalizes a_t before applying it to the environment (e.g., converting a high-level command into low-level motor signals).
T	**Environment transition**: $\mathrm{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$. Defines how the environment state evolves from (s_t, a_t) to s_{t+1}.

the hippocampus and cortical mechanisms, while reward signals, mediated by dopaminergic and other neuromodulatory pathways, reinforce behavior and learning. These biological insights inspire several design principles for AI agents, including but not limited to:

- **Parallel, Multi-Modal Processing:** The brain processes visual, auditory, and other sensory inputs in parallel through specialized cortical areas, integrating them in associative regions. Similarly, AI agents benefit from parallel processing of diverse sensor streams, fusing them in later stages for coherent understanding.

- **Hierarchical and Distributed Cognition:** Reasoning, planning, emotional regulation, and motor control involve interactions between cortical and subcortical regions. Analogously, AI agents can employ modular architectures with subsystems dedicated to rational inference, emotional appraisal, and memory.

- **Attention Mechanisms:** Human attention prioritizes sensory data based on context, goals, and emotions. AI agents can replicate this by modulating perception through learned attention policies, dynamically adjusting focus based on internal states.

- **Reward and Emotional Integration:** Emotions are not merely noise but integral to decision-making, modulating priorities, enhancing vigilance, and guiding learning. Reward-driven plasticity facilitates habit formation and skill acquisition, a concept critical to reinforcement learning in AI agents.
- **Goal Setting and Tool Usage:** The human prefrontal cortex excels at setting abstract goals and planning action sequences, including tool uses. Similarly, AI agents require robust goal-management systems and adaptive action repertoires, driven by external rewards and intrinsic motivations.

These principles form the foundation of our proposed **brain-inspired agent framework**, where biological mechanisms serve as inspiration rather than direct replication.

In the following sections, we outline our framework's key concepts, introducing a unified agent architecture based on the *perception–cognition–action loop* enriched by reward signals and learning processes. Each subsystem is carefully defined and interconnected to ensure transparency in how memory, world models, emotions, goals, rewards, and learning interact. We formalize cognition as a general reasoning mechanism, with *planning* and *decision-making* framed as specific "mental actions" shaping behavior. Connections to established theories, such as Minsky's *Society of Mind* [17], Buzsáki's *inside-out* perspective [18], and Bayesian active inference [19], are explored to highlight the framework's generality and biological plausibility.

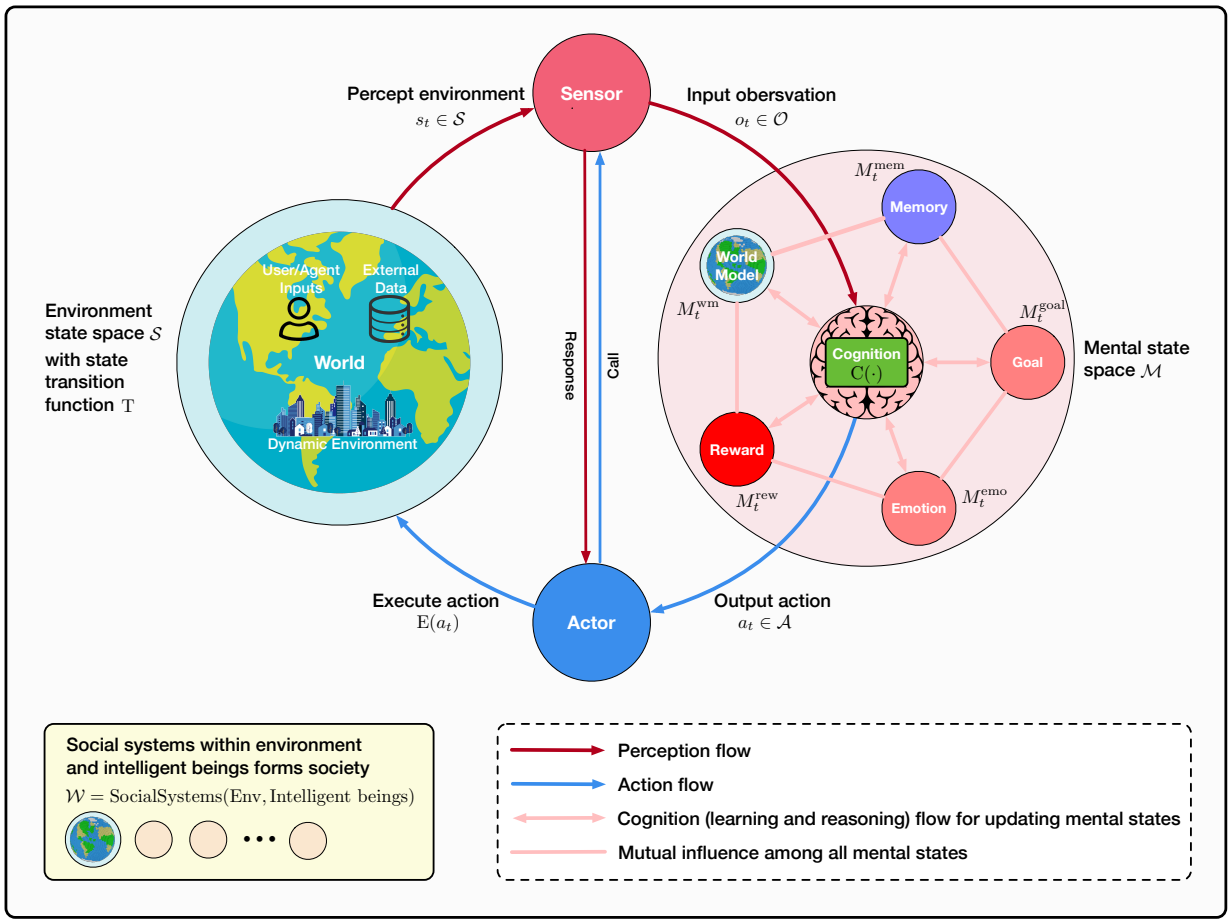

Figure 1.2: An overview of our general framework for describing an intelligent agent loop and agent society.

1.3.1 Core Concepts and Notations in the Agent Loop

Our architecture operates at three conceptual levels: **Society**, **Environment**, and **Agent**. The *Agent* is then decomposed into three main subsystems: **Perception**, **Cognition**, and **Action**. Within *Cognition*, we identify key submodules: *memory*, *world model*, *emotional state*, *goals*, *reward*, *learning*, and *reasoning* processes (including "planning" and "decision-making" as special actions produced with reasoning). *Attention* is primarily handled within perception and cognition. Before presenting the formal loop, we summarize our symbols in Table 1.2.

In the following, based on the notations in Table 1.2, we present our proposed agent loop.

The Agent Loop

An intelligent agent operates in discrete time steps t, continuously interacting with its environment. At each step, the following processes occur:

1. **Environment State** ($s_t \in \mathcal{S}$):

   The environment is in state s_t.

2. **Perception** (P): The agent perceives the environment to generate observations o_t:

   $$o_t = \mathrm{P}(s_t, M_{t-1}),$$

   where M_{t-1} guides selective attention and filtering.

3. **Cognition** (C): Updates mental state and selects actions:

   $$(M_t, a_t) = \mathrm{C}(M_{t-1}, a_{t-1}, o_t).$$

   where M_t encapsulates different sub-states:

   $$M_t = \{M_t^{\text{mem}}, M_t^{\text{wn}}, M_t^{\text{emo}}, M_t^{\text{goal}}, M_t^{\text{rew}}, \cdots\}.$$

   Cognition consists of:

   - **Learning** (L): Updates mental state based on observations:

     $$M_t = \mathrm{L}(M_{t-1}, a_{t-1}, o_t).$$

   - **Reasoning** (R): Determines the next action:

     $$a_t = \mathrm{R}(M_t),$$

     which may be:
     - **External Actions**, directly affecting the environment.
     - **Internal Actions**, including:
       - *Planning*: Internal sequence of future actions.
       - *Decision-making*: Choosing the best action from available options.

4. **Action Execution** (E): Transforms action a_t into executable form:

   $$a'_t = \mathrm{E}(a_t).$$

5. **Environment Transition** (T): The environment responds to the agent's action:

   $$s_{t+1} = \mathrm{T}(s_t, a'_t).$$

In multi-agent scenarios, each agent i maintains individual states (M_t^i, a_t^i, o_t^i), and the environment collectively updates based on all agents' actions. At broader scales (AI societies or worlds, $\mathcal{W}$), agents interact within diverse social systems (e.g., economic, communication, or transportation), forming complex societal structures.

Figure 1.2 illustrates our agent framework, presenting the core concepts and different types of information or control flows among them. Until now, we have presented a brain-inspired agent framework that integrates biological insights into a formal *Perception–Cognition–Action* loop. By decomposing cognition into modules for memory, world modeling, emotion, goals, reward-based learning, and reasoning, we capture essential parallels with the human brain's hierarchical and reward-driven processes. Critically, *attention* is included in the loop to enable selective filtering based on internal states. Furthermore, *planning* and *decision-making* can be viewed as distinct internal (mental) actions that either refine internal representations or select external behaviors. Our framework naturally extends classical agent architectures, providing a multi-level structure that integrates emotional and rational processes as well as robust, reward-driven learning across short and long timescales.

Society and Social Systems. In many real-world scenarios, agents do not merely interact with a static environment but operate within a broader *society*, comprising various *social systems* such as financial markets, legal frameworks,

political institutions, educational networks, and cultural norms. These structures shape and constrain agents' behaviors by defining rules, incentives, and shared resources. For example, a financial system dictates how economic transactions and resource allocations occur, while a political system provides governance mechanisms and regulatory constraints. Together, these social systems create a layered context in which agents must adaptively learn, reason, and act—both to satisfy their internal goals and to comply (or strategically engage) with external societal rules. In turn, the actions of these agents feed back into the social systems, potentially altering norms, policies, or resource distributions.

A Formal Definition of Foundation Agents. Building on these insights and our vision of robust, adaptive intelligence, we now formally introduce the concept of a *Foundation Agent*. Unlike traditional agent definitions that focus primarily on immediate sensory-action loops, a Foundation Agent embodies sustained autonomy, adaptability, and purposeful behavior, emphasizing the integration of internal cognitive processes across diverse environments.

Definition of Foundation Agent

A **Foundation Agent** is an autonomous, adaptive intelligent system designed to **actively perceive** diverse signals from its environment, continuously **learn** from experiences to refine and update structured internal states (such as memory, world models, goals, emotional states, and reward signals), and **reason** about purposeful actions—both external and internal—to autonomously navigate toward complex, long-term objectives.

More concretely, a Foundation Agent possesses the following core capabilities:

1. **Active and Multimodal Perception:** It continuously and selectively perceives environmental data from multiple modalities (textual, visual, embodied, or virtual).
2. **Dynamic Cognitive Adaptation:** It maintains, updates, and autonomously optimizes a rich internal *mental state* (memory, goals, emotional states, reward mechanisms, and comprehensive world models) through **learning** that integrates new observations and experiences.
3. **Autonomous Reasoning and Goal-Directed Planning:** It proactively engages in sophisticated reasoning processes, including long-term planning and decision-making, to derive goal-aligned strategies.
4. **Purposeful Action Generation:** It autonomously generates and executes purposeful actions, which can be external (physical movements, digital interactions, communication with other agents or humans) or internal (strategic planning, self-reflection, optimization of cognitive structures), systematically shaping its environment and future cognition to fulfill complex objectives.
5. **Collaborative Multi-Agent Structure:** It can operate within multi-agent or agent society structures, collaboratively forming teams or communities of agents that collectively accomplish complex tasks and goals beyond individual capabilities.

This definition highlights three essential pillars distinguishing Foundation Agents: *sustained autonomy* (operating independently toward long-term goals without step-by-step human intervention), *adaptive learning* (evolving internal representations continually over diverse experiences), and *purposeful reasoning* (generating actions guided by complex, internally maintained goals and values). Foundation Agents thus represent a fundamental shift from traditional agents by integrating deep cognitive structures, multimodal processing capabilities, and proactive, sustained self-optimization, enabling them to function effectively across a wide range of environments and domains.

Unlike classical definitions, which often frame agents primarily in terms of simple perception–action loops ("perceive and act" [20]), our notion of Foundation Agents emphasizes the depth and integration of internal cognitive processes. Foundation Agents not only perceive their environment and perform immediate actions but also possess an evolving, goal-oriented cognition—continuously adapting memory structures, world models, emotional and reward states, and autonomously refining their strategies through reasoning. This internal cognitive richness allows Foundation Agents to autonomously decompose complex, abstract goals into actionable tasks, strategically explore their environments, and dynamically adjust their behavior and cognitive resources. Our unified **perception–cognition–action** framework thus accommodates and explicitly models these sophisticated cognitive capabilities, recognizing internal (mental) actions on par with external (physical or digital) interactions, facilitating a broad range of embodiments, from physical robots to software-based or purely textual intelligent agents.

1.3.2 Biological Inspirations

Although our agent model is fundamentally computational, each submodule draws inspiration from well-studied biological counterparts in the human brain. Below, we discuss these analogies in a manner that highlights both the neuroscientific basis and the flexibility afforded by AI implementations.

Memory (Hippocampus and Neocortex). Decades of neuroscience research have linked the hippocampus to episodic memory formation, while cortical regions are known to house semantic and procedural knowledge [21, 22]. In humans, these memory subsystems cooperate to manage both short-term encoding and long-term consolidation. Our memory component, M_t^{mem}, similarly aims to capture multi-scale learning by storing recent experiences and knowledge. This can be realized through either neural network weights (long-term) or explicit buffers (short-term), thereby mirroring the hippocampal–cortical interplay.

World Model (Predictive Processing). A prominent theory in cognitive neuroscience holds that the cortex operates as a predictive machine, continually comparing incoming sensory data with generated expectations [23, 19]. The world model M_t^{wm} reflects this idea by maintaining an internal representation of how the environment evolves over time. Just as cortical circuits integrate multisensory data to update these internal models, our framework allows M_t^{wm} to be refined upon each new observation and relevant reward or emotional cues, offering a Bayesian or free-energy perspective on environmental dynamics.

Emotion (Limbic System). Emotions, mediated by structures like the amygdala, hypothalamus, and limbic system, significantly modulate attention, learning rates, and decision-making thresholds [24, 25]. By introducing an emotion component M_t^{emo}, our model captures how internal valence or arousal states can shift an agent's focus and behavior. Although computational "emotions" are neither fully analogous to biological affect nor conscious feelings, they can guide adaptive heuristics—such as prioritizing urgent goals or responding quickly to perceived threats.

Goals and Reward (Prefrontal & Subcortical Circuits). Humans excel at forming abstract, long-term goals, an ability often associated with prefrontal cortex function [26, 27]. In parallel, subcortical circuits—particularly dopaminergic pathways—drive reinforcement signals that shape motivation and habit learning [28]. Our agent includes M_t^{goal} for storing objectives and M_t^{rew} for encoding reward signals, thus enabling a continuous feedback loop where goal formation and reward-based adaptation reinforce each other. This mechanism allows for planned action sequences, tool usage, and more nuanced social interactions.

Reasoning, Planning, and Decision-Making (Prefrontal Cortex). Finally, the human prefrontal cortex integrates information from memory, sensory inputs, emotions, and reward pathways to carry out higher-order cognitive processes—such as logical reasoning, planning, and executive control [29, 30]. In our agent framework, these capabilities are subsumed by the reasoning sub-function, which—through modules like PlanFn and Decide—selects and executes actions (whether physical or purely mental). By distinguishing planning from on-the-fly decision-making, we capture how the agent can simulate future scenarios, weigh outcomes, and then commit to a course of action, akin to the flexible orchestration observed in prefrontal circuits.

1.3.3 Connections to Existing Theories

Beyond these explicit neurobiological parallels, our architecture resonates with several important theories in AI, cognitive science, and neuroscience.

Classic Perception–Cognition–Action Cycle. We extend the traditional sense–think–act cycle outlined by [20], incorporating explicit mechanisms for attention (in P), learning and emotion (in C), and reward signals that persist over time. This explicitness makes it easier to analyze how an agent's internal states and prior actions shape subsequent perception and cognition.

Minsky's "Society of Mind". [17] argued that intelligence arises from an ensemble of specialized "agents" within a mind. Our submodules—$\mathrm{C}_{\text{mem}}, \mathrm{C}_{\text{wm}}, \mathrm{C}_{\text{emo}}, \mathrm{C}_{\text{goal}}, \mathrm{C}_{\text{rew}}$—echo this decomposition, distributing key functions (memory, prediction, emotional evaluation, goal-setting, etc.) across separate yet interacting components. In a broader "society" context, each agent (or sub-agent) could coordinate cooperatively or competitively, much like Minsky's internal agencies. Recent work on natural language-based societies of mind [31] supports that agentic systems can be represented using the original society-of-mind theory, and could incorporate social structures and economic models among agents.

Buzsáki's Inside-Out Perspective. Neuroscientists [18] contend that the brain actively constructs and updates its perception instead of merely receiving inputs. In our model, M_{t-1}—including emotional states, reward signals, and goals—directly influences the perception map P. This supports the inside-out stance that an agent's internal context drives the way it samples and interprets the environment, rather than passively reacting to it.

Partially observable Markov decision process (POMDP). Our framework can be viewed as a generalization of the classical Partially Observable Markov Decision Process (POMDP) in several ways. First, whereas a POMDP specifies a probabilistic transition function $P(s_{t+1} \mid s_t, a_t)$ over a (possibly finite) state space, we retain an environment transition T without restricting it to a purely probabilistic or finite form, allowing for arbitrary or even deterministic mappings. Second, in the standard POMDP setting, reward is typically defined as a scalar function of (s_t, a_t) (possibly discounted over time). By contrast, we place reward signals *inside* the agent's mental state (M_t^{rew}), letting them depend on—and co-evolve with—goals, emotion, and the world model rather than enforcing a single externally defined objective. Third, while POMDP agents generally select actions by maximizing an expected return (value function), our *reasoning* sub-process is broader. It accounts for memory, emotion, and other mental-state factors, accommodating heuristic or socially driven decisions rather than strictly value-based choices. Finally, a POMDP does not explicitly define cognitive submodules such as memory or emotion—these must be collapsed into a monolithic "belief state". In our framework, each sub-component (memory, world model, emotion, goals, reward) is explicitly modeled and updated, mirroring biologically inspired views of cognition. Hence, although our approach *recovers* the POMDP formulation as a special case (by enforcing a probabilistic T, a scalar reward, and a minimal mental state), it admits a richer variety of environment transitions, internal states, and decision mechanisms.

Active Inference and the Bayesian Brain. Active inference, a unifying framework advanced by [19], suggests that agents continually update internal generative models to minimize prediction error (or "free energy"). Our use of M^{wm} and M^{rew}, together with planning and decision-making modules, can be interpreted in Bayesian terms. The agent attempts to reduce surprise by aligning its world model with new data and by choosing actions that conform to predicted (or desired) outcomes.

Biological Plausibility & Generality. While the mapping between brain circuits and agent submodules is made at a high level, it offers an approach that is at once *biologically inspired* and *modularly agnostic*. Memory, emotion, goals, and reward can each be implemented by various AI paradigms—symbolic methods, neural networks, or hybrid approaches—thus preserving flexibility. By integrating these key ideas from neuroscience, cognitive science, and AI, we arrive at a general framework that captures the essential properties of intelligent behavior without overconstraining implementation details.

1.4 Navigating This Survey

This survey is structured to provide a comprehensive, modular, and interdisciplinary examination of intelligent agents, drawing inspiration from cognitive science, neuroscience, and other disciplines to guide the next wave of advancements in AI. While many existing surveys [32, 33, 34, 35, 36, 37, 38, 39, 40] offer valuable insights into various aspects of agent research, we provide a detailed comparison of their focal points in Table 1.3. Our work distinguishes itself by systematically comparing biological cognition with computational frameworks to identify synergies, gaps, and opportunities for innovation. By bridging these domains, we aim to provide a unique perspective that highlights not only where agents excel but also where significant advancements are needed to unlock their full potential.

Table 1.3: Summary of existing reviews with different focal points. • indicates primary focus while ○ indicates secondary or minor focus.

Survey	Cognition	Memory	World Model	Reward	Action	Self Evolve	MultiAgent	Safety
Zhang et al. [39]	•	•	○	○	○	•	○	○
Guo et al. [38]	•	•	○	○	○	•	•	○
Yu et al. [40]	•	•	○	○	•	○	•	•
Wang et al. [35]	•	•	○	○	•	○	•	○
Masterman et al. [37]	•	•	○	○	•	○	•	○
Xi et al. [34]	•	•	○	○	•	•	•	•
Huang et al. [33]	•	•	○	•	•	•	•	•
Durante et al. [32]	•	•	○	•	•	•	•	•
This Manuscript	•	•	•	•	•	•	•	•

The survey is divided into four key parts:

- In **Part I**: Modular Design of Intelligent Agents, we introduce the core modules of agents, including the cognition module, which serves as the "brain" of the agent; the perception systems for interpreting sensory input; as well as the action systems for interacting with the external world. Within the cognition system, we further discuss the memory, world modeling, emotion, goal, and reward systems, analyzing their current progress, limitations, and research challenges.

- In **Part II**: Self-Enhancement in Intelligent Agents, we shift focus to the capability of agents to evolve and optimize themselves. We explore mechanisms like adaptive learning, self-reflection, and feedback-driven improvement, inspired by the human ability to grow and refine skills over time. This part also addresses the importance of dynamic memory systems and continuous knowledge integration for agents to remain relevant and effective in changing environments.
- In **Part III**: Collaborative and Evolutionary Intelligent Systems, we examine how agents interact with each other and their environments to solve complex, large-scale problems. We discuss multi-agent systems, highlighting their applications in fields such as robotics, medical systems and scientific discovery. This part explores multi-agent system topologies and agent protocol, tracing the evolution of communication and collaboration from static to dynamic frameworks. We align agents with human collaboration paradigms, examining how interaction patterns shape the co-evolution of intelligence and how multi-agent systems adapt their decision-making in various collaborative settings to solve complex challenges through collective intelligence.
- Finally, in **Part IV**: Building Safe and Beneficial AI, we provide a comprehensive analysis of the security landscape for LLM-based agents. We introduce a framework categorizing threats as intrinsic or extrinsic. Intrinsic vulnerabilities arise from within the agent's architecture: the core LLM "brain", and the perception and action modules that enable interactions with the world. Extrinsic risks stem from the agent's engagement with memory systems, other agents, and the broader environment. This part not only formalizes and analyzes these vulnerabilities, detailing specific attack vectors like jailbreaking and prompt injection, but also reviews a range of defense mechanisms. Moreover, we explore future directions, including superalignment techniques and the scaling law of AI safety—the interplay between capability and risk.

By weaving together these threads, our survey aims to provide a holistic perspective on the current state of intelligent agents and a forward-looking roadmap for their development. Our unique focus on integrating cognitive science insights with computational design principles positions this survey as a foundational resource for researchers seeking to design agents that are not only powerful and efficient but also adaptive, ethical, and deeply aligned with the complexities of human society.