ENGINEERING Artificial Intelligence
https://doi.org/10.2738/ENGAI.2026.0012

Review

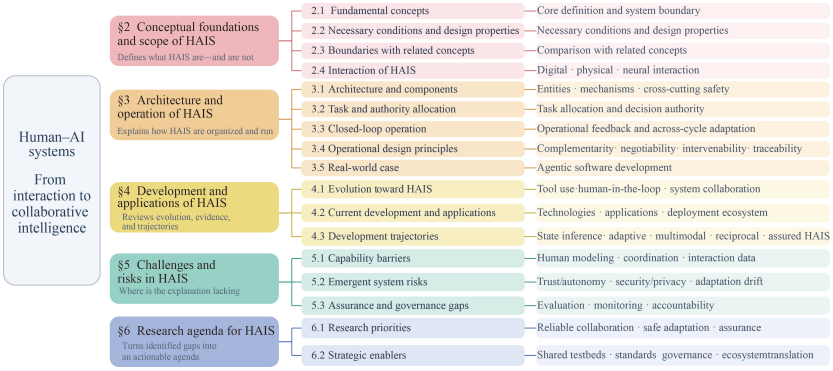
Human–AI Systems: A System-Level Perspective on Collaborative Intelligence

Chao Shen (✉), Hongxin Xie, Le Yang, Chenhao Lin, Bin Shi, Bo Dong, Qinghua Zheng

Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

© The Author(s) 2026. This article is published by Higher Education Press.

GRAPHICAL ABSTRACT



HIGHLIGHTS

- Defines HAIS, its necessary conditions, and boundaries with related paradigms.
- Develops a unified architecture for human–AI interaction, coordination, and governance.
- Synthesizes research and applications to chart trajectories toward deeper human–AI coupling.
- Identifies key technical, safety, and governance challenges in HAIS and establishes priorities for future HAIS research.

ABSTRACT Advances in foundation models, intelligent agents, embodied AI, and multimodal interfaces are enabling humans and AI to collaborate in increasingly complex, dynamic, and high-stakes tasks. However, existing research remains fragmented across human–AI interaction, shared autonomy, intelligent agents, robotics, and brain–computer interfaces, with limited attention to the behavior and performance of the complete Human–AI System. This article presents a system-level perspective on Human–AI Systems (HAIS). We define HAIS as systems in which humans and AI participate in a shared task loop, exchange task-relevant information, and jointly influence decisions or actions toward a system-level objective. We clarify the necessary conditions, design properties, conceptual boundaries, and digital, physical, and neural interaction channels of HAIS. We then propose a reference architecture comprising system entities, operational mechanisms, and cross-cutting safety and governance, and explain task and authority allocation, closed-loop operation, and adaptation across task cycles. Building on this framework, we review the evolution, enabling technologies, representative applications, deployment ecosystem, and future development of HAIS. We further examine capability barriers, emergent risks, and assurance and governance gaps, including unreliable human-state modeling, coordination failures, trust miscalibration, security and privacy threats, co-evolutionary drift, and unclear accountability. Finally, we identify research priorities in uncertainty-aware human modeling, adaptive coordination, safe reciprocal adaptation, runtime assurance, system-level evaluation, and traceability. This perspective provides a unified foundation for designing and evaluating HAIS that preserve human agency while achieving reliable, safe, and accountable collaborative intelligence.

KEYWORDS Human–AI Systems; Human machine integrated intelligence; Collaborative evolution; Deep human−AI coupling

Received 24 Aug 2026; accepted 22 Sep 2026

E-mail: chaoshen@xjtu.edu.cn (Shen C.).

1 Introduction

Recent advances in large language models [1], foundation models [2], multimodal AI [3,4], AI coding agents [5], and embodied AI systems [6] have rapidly expanded the capabilities of artificial intelligence (AI). AI can now combine information processing with perception, reasoning, and action [7,8], allowing it to perform a growing range of tasks once mainly carried out by humans. It is also being deployed in open, dynamic, and high-stakes environments, where conditions may change rapidly and the effects of decisions can be difficult to predict or reverse [9–11]. Such environments, including deep-Earth exploration, offer high scientific and practical value but also involve serious risks and uncertainty. Neither humans nor AI can address these challenges alone. AI excels at large-scale information processing, rapid computation, and precise execution, whereas humans remain essential for understanding context, making value-based judgments, developing creative solutions, and taking responsibility for important decisions. However, humans are limited by attention, physical and cognitive capacity, while AI may be unreliable in unfamiliar situations, under unclear goals, or when decisions depend heavily on context and values [12]. Therefore, the central challenge is to design HAIS as shown in Fig. 1 that combine their complementary capabilities while keeping AI aligned with human goals and ensuring effective oversight, safety, and reliable system-level performance.

Building such systems requires more than simply placing humans and AI in the same task. In this article, we use the term HAIS to describe systems in which humans and AI jointly take part in understanding tasks, making decisions, and carrying out actions. The performance of HAIS depends not only on the individual capabilities of humans and AI, but also on how information, tasks, decision authority, and control are coordinated between them [13,14]. A more capable AI model may not improve the overall result if it is poorly integrated into human decision-making and control processes. Similarly, human involvement alone cannot ensure effective collaboration if people lack the information, time, or authority needed to guide or correct the system. Therefore, the design and evaluation of HAIS should focus on the capability of the complete system, extending the current focus on AI autonomy toward system-level collaborative intelligence. Here, system-level collaborative intelligence does not refer to an independent or single quantifiable intelligence metric. It refers to the overall system capability that emerges from information exchange, task and authority allocation, feedback, and coordination between humans and AI within a shared task. Specifically, both humans and AI participate in task-related judgment, decision-making, or

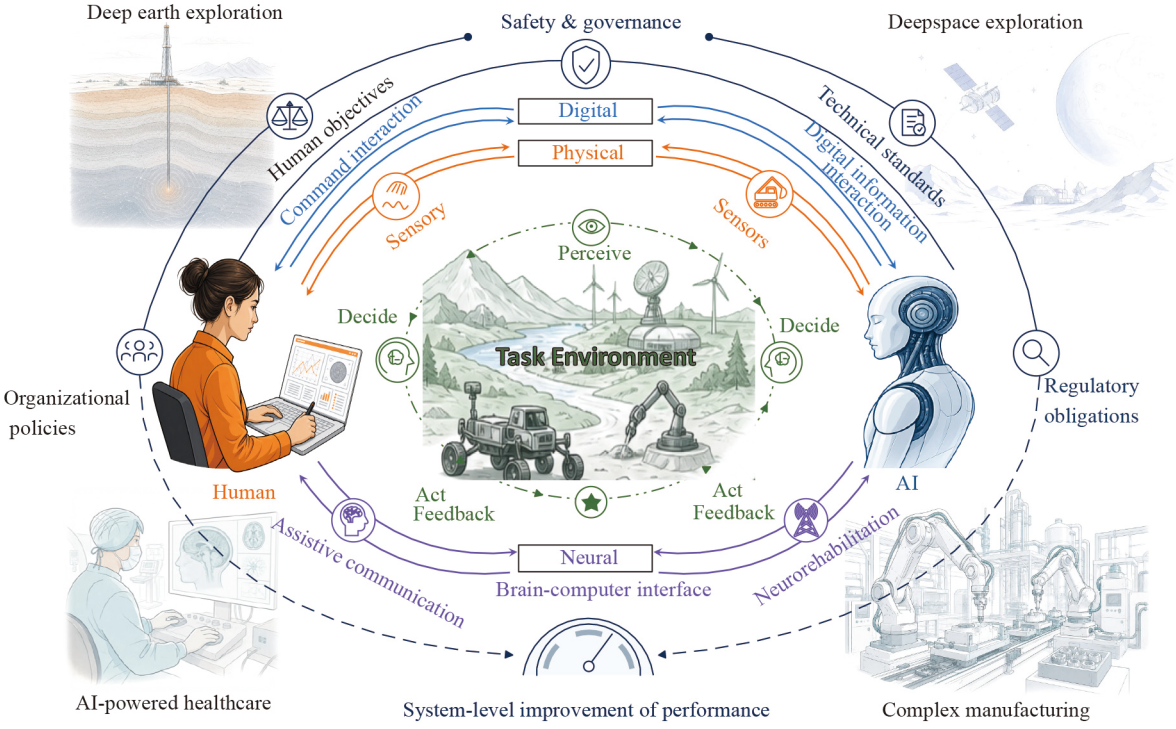


Fig. 1 An illustration of future Human–AI Systems.

action, and each can influence the other's subsequent behavior. Their contributions are connected through continuous feedback and coordinated under explicit task and authority arrangements. The resulting system behavior arises from their ongoing interaction within a shared task loop, rather than from either participant acting alone.

Despite substantial progress in AI capability, human–AI interaction and AI safety, research on coupled HAIS remains fragmented across different fields. Three major limitations can be identified. (1) Lack of system-level evaluation: Many studies still evaluate AI models or individual interaction outcomes, while a common framework for assessing the overall capability and performance of HAIS is lacking. (2) Limited understanding of long-term human–AI adaptation: Humans are often treated as sources of input, feedback, or supervision, with limited attention to how their states, decisions, and behaviors change during long-term interaction with AI. (3) Lack of system-level safety research: Existing safety research often focuses on model alignment and safeguards for individual components, while risks arising from human–AI interaction—such as poorly calibrated trust, conflicts over decision authority, and harmful feedback loops—have received less systematic attention. Together, these limitations make it difficult to explain, evaluate, and manage HAIS as complete and continuously changing systems.

These limitations share a common cause: humans, AI, and safety mechanisms are often studied separately rather than as parts of an integrated system. In HAIS, however, system behavior is shaped by their continuous interaction. AI outputs can influence human decisions and actions, while human responses can change the data, feedback, and operating conditions of the AI. Thus, improving an individual AI model does not necessarily improve the performance of the system, and protecting each component separately does not always prevent risks that arise from their interaction. A system-level perspective is therefore needed to study human and AI capabilities, task and authority allocation, mutual adaptation, and safety as closely connected parts.

In response to these needs, this article presents a Perspective on HAIS rather than a systematic review. Our aim is to connect representative research from AI agent, human–AI interaction, shared autonomy, embodied intelligence, and brain–computer interfaces within a common system-level framework. Although these fields focus on different forms of human–AI interaction, they face several shared questions: how human and AI capabilities should complement each other, how tasks and authority should be allocated, how humans and AI adapt to each in long-term, and how risks to the complete Human–AI System can be assessed, monitored, mitigated, and governed throughout operation.

As a Perspective, this article does not seek to exhaustively review the HAIS literature, but instead selects representative studies in line with the analytical framework developed here. The scope covers human–AI interaction and teaming, shared autonomy, foundation-model agents, embodied intelligence and robotics, brain–computer interfaces, and system safety and governance. Studies are included primarily when they satisfy at least one of four criteria: they establish conceptual or operational foundations for HAIS; represent important technical advances in digital, physical, or neural interaction; describe concrete systems that permit analysis of human and AI roles, shared task loops, authority allocation, feedback and adaptation, or system-level outcomes; or provide representative approaches to safety, evaluation, and lifecycle governance. We also retain several established systems that do not fully satisfy the HAIS criteria as boundary cases. These cases help clarify the continuum from AI tools and human-in-the-loop systems to HAIS. For rapidly evolving areas, we prioritize representative recent work while retaining earlier systems of foundational importance.

This paper makes four main contributions. First, we define HAIS, identify the basic conditions of such systems, and clarify their differences from conventional human–AI interaction and autonomous systems. Second, we describe cognitive, physical, and neural forms of human–AI interaction and present a system architecture that connects humans, AI, task environments, coordination and control, and safety governance. Third, we review representative research, applications, and industrial developments, and discuss how demonstrated advances motivate longer-term directions toward intent understanding, dynamic shared autonomy, mutual adaptation, and deeper neural integration. Fourth, we examine the main technical, safety, and governance challenges of HAIS and identify priorities for future research and ecosystem development. The overall organization and logical relationships among the sections are illustrated in Fig. 2, which focuses on the main content from Sections 2 to 6; the Introduction and Conclusion are omitted because they follow the conventional structure of an academic article.

2 Conceptual Foundations and Scope of HAIS

This section defines the conceptual foundations and scope of HAIS, whose principle is whether humans and

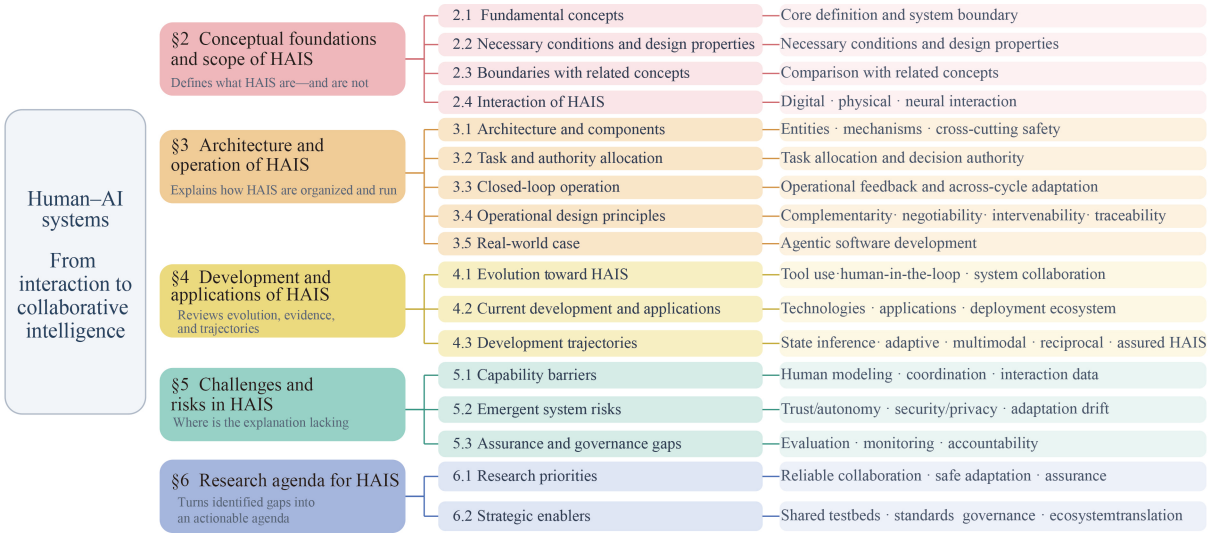


Fig. 2 Overall organization and logical relationships among the main sections of this article.

AI participate in the same task process and jointly influence its decisions or actions. Section 2.1 defines HAIS and establishes its system boundary. Section 2.2 distinguishes the necessary conditions for identifying HAIS from the desirable properties of well-designed systems. Section 2.3 clarifies the relationship between HAIS and related concepts, including conventional AI tools, shared autonomy, hybrid intelligence, and autonomous systems. Section 2.4 describes the digital, physical, and neural channels through which humans and AI exchange information and influence system behavior.

2.1 Fundamental concepts

In this paper, we define a HAIS as a system in which humans and AI participate in a shared task loop, exchange task-relevant information, and jointly influence decisions or actions toward a system-level objective. The unit of analysis is the coupled human–AI–task–environment system, rather than the human, the AI, or their interface in isolation.

Humans contribute task intent, contextual understanding, domain knowledge, value judgment, and responsibility for consequential decisions. AI contributes capabilities in perception, information processing, reasoning, planning, and execution. Their roles need not be symmetric, and their influence need not be equal. What matters is that both can materially affect the task process. Their influence may concern intermediate decisions, subsequent actions, authority arrangements, or final outcomes. Roles and authority may also change with task demands, uncertainty, risk, and operating conditions.

A shared task loop requires more than isolated input and output. Human and AI contributions must be connected within an evolving task process. AI outputs may affect subsequent human judgments or actions, while human decisions, corrections, or interventions may change the information, constraints, or operating conditions available to the AI. Actions by either side may also alter the environment observed in later stages. HAIS are defined by the organization of participation, influence, and feedback rather than by a particular interface, degree of AI autonomy, or learning mechanism.

2.2 Necessary conditions and design properties

The definition above leads to three necessary conditions for identifying an HAIS: (1) humans and AI must participate in the same task process. A single request-response exchange qualifies only when it forms part of a wider process in which human and AI contributions jointly affect the task outcome. (2) Both humans and AI must contribute to the task's decision or control process. Each contribution must be capable of changing an intermediate decision, authority allocation, subsequent action, or final task outcome. If removing either participant's contribution under the same task state leaves all of these unchanged, the system does not satisfy this condition [15]. (3) Information or control feedback must connect their contributions, allowing the decision or action of one participant to affect the subsequent behavior of the other.

Several additional properties support the effective design and operation of HAIS. Human and AI roles, decision authority, and intervention conditions should be identifiable so that control and responsibility remain clear during collaboration. Evaluation should address the performance of the complete Human–AI System,

including task outcomes, human workload, safety, and observable coordination measures such as coordination cost, disagreement frequency, authority-transfer success, and intervention effectiveness. Repeated interaction may also support reciprocal adaptation. AI may adjust to human expertise, preferences, and corrections, while human participants may refine their judgments and task strategies through AI-generated information and feedback. Reciprocal adaptation may occur without continuous online learning or equal changes on both sides, and its effects should be monitored at the system level.

2.3 Boundaries with related concepts

HAIS overlap with several established concepts in human–AI research because these concepts address related aspects of collaboration, cognition, control, and system organization. They should not be treated as mutually exclusive categories. A particular operational system may simultaneously exhibit characteristics associated with more than one framework. The relevant distinction is therefore not whether an individual property—such as complementarity, interdependence, system-level analysis, adaptation, or human control—appears in prior work, but which properties are constitutive, what is taken as the unit of analysis, and which aspects of system operation the concept is intended to explain. Table 1 summarizes these differences.

AI-as-tool primarily provide bounded task assistance under human authority. The human defines the task and directs the use of AI, so the main unit of analysis remains the user–AI tool interaction rather than an integrated Human–AI System. Shared autonomy focuses on the allocation of control authority between humans and autonomous components. Its defining concern is joint or dynamic control within a Human–autonomy control loop, rather than the organization of the complete task system. Hybrid intelligence emphasizes the complementary combination of human and machine intelligence. Complementarity, collective gain, and mutual learning are central concerns, and the resulting socio-technical ensemble is typically considered as the analytical unit. human–AI teaming focuses on team interdependence. Humans and AI assume differentiated and complementary roles, coordinate their activities, and work toward a common goal; accordingly, the human–AI team is the primary unit of analysis. Joint cognitive systems emphasize the cognitive functioning that emerges from functional coupling between humans and technological components. The joint cognitive system, rather than either component alone, is therefore treated as the analytical unit. Meaningful human control addresses a different dimension. It focuses on whether system behavior remains responsive to relevant human reasons and whether responsibility for consequential outcomes can be traced to appropriate human actors. Its primary concern is therefore the human–system control relation rather than collaborative system organization. Autonomous systems emphasize independent machine perception, decision-making, and execution, with human intervention remaining limited or supervisory, while HAIS emphasize how this combination operates within a defined task system. Autonomous systems may also function within HAIS when humans retain meaningful roles in goal setting, supervision, intervention, or shared decision-making. The classification therefore depends on the organization of human and AI participation during actual system operation.

2.4 Interaction of HAIS

Humans and AI exchange task-relevant information and influence decisions or actions through three main channels: digital, physical, and neural. The classification follows the interface through which information crosses the human–AI boundary. Multiple channels can operate within the same HAIS, depending on the task, environment, response time, and coordination requirements, as shown in Fig. 3.

To make this taxonomy operational, the classification is applied to individual human–AI interaction links rather than to the HAIS as a whole. The criterion is the interface through which task-relevant information or control

Table 1 Comparison of HAIS with related human–AI concepts.

Concept	Primary emphasis	Distinguishing characteristic	Primary unit of analysis
AI-as-tool	Task assistance	Human authority; user-directed support	User–AI tool interaction
Shared autonomy	Allocation of control authority	Joint/dynamic control	Human–autonomy control loop
Hybrid intelligence	Complementary intelligence	Complementarity; collective gain; mutual learning	Socio-technical ensemble
Human–AI teaming	Team interdependence	Role complementarity; interdependence; common goal	Human–AI team
Joint cognitive systems	Joint cognitive functioning	Functional cognitive coupling	Joint cognitive system
Meaningful human control	Human control and responsibility	Reason-responsiveness; traceable responsibility	Human–system control relation
Autonomous systems	Autonomous action	Independent execution; limited intervention	Autonomous system
HAIS	**Integrated human–AI functioning**	**Reciprocal coupling; task–feedback–adaptation**	**Complete HAIS**

HAIS denotes Human–AI Intelligent Systems. The listed characteristics summarize the primary emphasis and distinguishing features of each concept; specific implementations may vary across tasks and application contexts

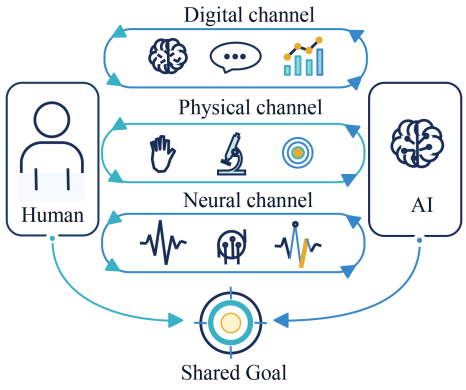


Fig. 3 HAIS realize information and control coupling of varying depths at the digital, physical, and neural channels.

crosses the human–AI boundary. An interaction is digital when this exchange occurs through computational representations, such as language, images, code, or structured data; physical when it occurs through sensorimotor variables, such as motion, force, position, touch, or changes in the physical environment; and neural when signals are acquired directly from, or delivered to, the central or peripheral nervous system. These channels are not mutually exclusive. A single HAIS may combine multiple channels simultaneously or sequentially, and their relative roles may change across task stages and operating conditions. When an interaction pathway involves more than one interface, each relevant channel should be identified rather than assigning the complete system to a single category. For example, a brain-controlled prosthesis may combine neural signal acquisition, digital information processing, and physical actuation. The taxonomy characterizes the composition of human–AI coupling within a system, while the specific combination depends on the task, environment, response time, and coordination requirements.

- Digital channel: digital interaction uses natural language, images, code, structured data, and software interfaces [16]. Humans communicate objectives, constraints, context, and corrections, while AI provides analyses, recommendations, generated content, or executable operations. Digital channels support decision assistance, software development, scientific research, and agent-based collaboration [17]. Their effectiveness depends on consistent task context, clear communication of uncertainty, and the ability of exchanged information to guide subsequent decisions or actions.
- Physical channel: physical interaction occurs through motion, force, position, touch, or changes in the environment [18]. Human actions affect the state observed by AI, while AI-generated commands affect the environment or the human participant. Examples include collaborative robots, wearable devices, intelligent prostheses, surgical robots, and shared-control vehicles [19]. Since physical actions may have immediate and difficult-to-reverse effects, these systems require timely coordination, clear control authority, effective intervention, and safe recovery mechanisms.
- Neural channel: neural interaction uses signals acquired from, or delivered to, the central or peripheral nervous system. Brain–computer interfaces can decode neural activity related to movement intention, attention, or other task-relevant states [20]. Bidirectional interfaces can also provide sensory feedback through neural stimulation. Applications include neuroprosthetics, assistive communication, neurorehabilitation, and sensory restoration [21]. Their practical value depends on signal quality, individual differences, invasiveness, response time, and long-term reliability.

These channels often work together. A neural prosthesis, for example, may acquire neural signals, process them through digital models, and produce physical movement [22]. The next section explains how these interaction channels and other system elements are organized within an operational HAIS architecture.

3 Architecture and Operation of HAIS

3.1 Architecture and components

Section 2 defines the scope and interaction channels of HAIS. This section presents a system-level reference architecture that organizes these elements into three layers, as shown in Fig. 4. The first layer contains three system entities: the human, the AI, and the task and environment. The second layer contains three operational mechanisms: the shared task representation, coordination and control and adaptation and update. The third layer provides cross-cutting safety and governance assurance. Digital, physical, and neural channels connect these layers by carrying task information, control signals, actions, and feedback. Human and AI refer to functional roles, each of which may include one or multiple participants or components. The architecture is intended as a conceptual and design-level reference rather than a claim that all of these mechanisms are already realized in current HAIS.

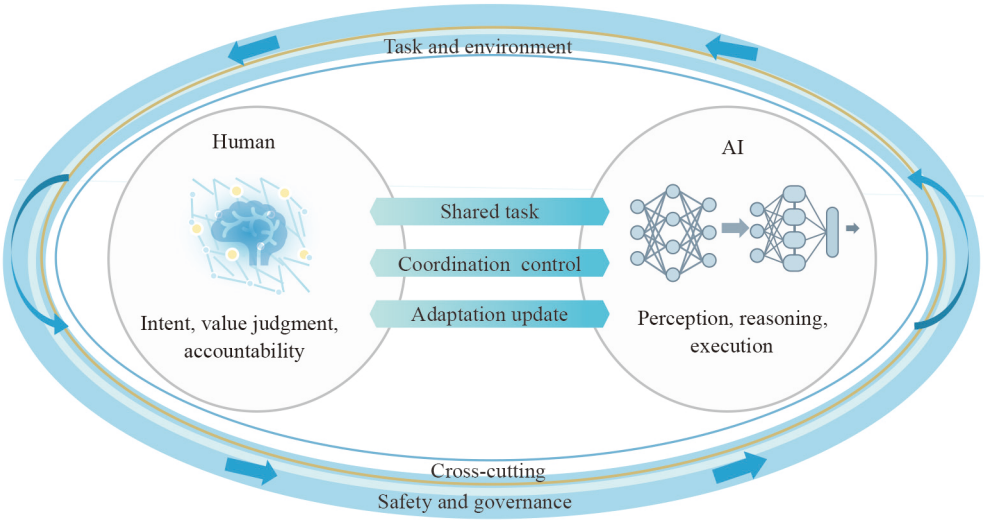


Fig. 4 Architectural components of HAIS. Deeply bridging human cognition and AI intelligence through multimodal interaction interfaces to facilitate reciprocal learning and co-evolution in a closed-loop paradigm.

3.1.1 System entities

(1) Human. The human contributes task intent, contextual understanding, domain knowledge, value judgment, and responsibility for important decisions. Human participation may include defining objectives, reviewing intermediate results, approving proposed actions, or intervening during execution [23]. The required involvement may change with task conditions. The architecture should represent human roles, authority, and intervention mechanisms explicitly [24]. It should also provide the task state, relevant uncertainty, proposed actions, and possible consequences needed for effective judgment.

(2) AI. The AI provides computational capabilities in perception, prediction, reasoning, planning, and execution [25]. Its functions vary across applications. Language and multimodal models may support semantic reasoning, while control models may perform signal decoding, state estimation, motion planning, or low-level control. Agent-based systems may also decompose tasks, use external tools, and revise plans. The architecture should define the AI's operational scope, access permissions, uncertainty, and conditions for requesting human input or transferring authority.

(3) Task and environment. The task defines the objective, constraints, performance requirements, and acceptable outcomes. The environment contains the physical, digital, or social states that affect decisions and actions [26]. Both may change as new information becomes available, disturbances occur, or previous actions alter the system state. Time pressure, uncertainty, and action reversibility may also change the relative reliability of human and AI contributions. These factors provide the context for interpreting decisions, allocating authority, and evaluating outcomes.

3.1.2 Operational mechanisms

(1) Shared task representation. The shared task representation maintains the information required for human–AI coordination. Humans and AI may use different internal representations [27]; therefore, shared information should remain interpretable across the human–AI interface. This information includes the current objective, task state, constraints, uncertainty, proposed actions, and authority arrangement. The representation is updated as actions change the task or environment. When assessments differ, it helps identify the source of disagreement and the information requiring clarification.

(2) Coordination and control. Coordination and control transform human and AI contributions into proposed system actions [28]. The mechanism receives task information, estimates of human and AI states, uncertainty, and risk. Many of these variables are only partially observable and should be represented with appropriate confidence estimates. Explicit policies, safety constraints, or optimization criteria guide proposal integration, conflict resolution, and authority allocation. The mechanism outputs a proposed action, an authority assignment, and any required intervention or clarification. Each proposal proceeds through authority verification and safety checks before execution.

(3) Adaptation and update. Adaptation and update operate across multiple task cycles. Operational feedback may reveal changes in human behavior, AI performance,

task conditions, or observable coordination indicators, such as disagreement frequency, authority-transfer success, intervention delay, and revision cycles. Based on this evidence, the system may update AI models, shared task representations, allocation policies, or interaction strategies. Human participants may also revise their knowledge, expectations, and working practices through repeated interaction with AI. These updates should follow defined validation, approval, monitoring, and rollback procedures. Safety and governance assurance should prevent unverified changes from directly affecting consequential system operation.

3.1.3 Cross-cutting safety and governance

Safety and governance assurance monitors and constrains the complete human–AI task loop, including the coordination mechanism itself. Requirements may arise from human objectives, organizational policies, technical standards, and regulatory obligations. Based on available indicators of human inputs, AI outputs, environmental states, authority transfers, and execution results, the system may issue warnings, request confirmation, restrict actions, transfer control, or enter a safe state.

The specific safeguards depend on the interaction channels involved. Digital systems require access control, input validation, controlled tool execution, runtime monitoring, and auditable logs [29,30]. Physical systems require force and speed limits, collision avoidance, emergency stops, and safe recovery. Neural systems require signal-integrity checks, stimulation limits, privacy protection, informed authorization, and immediate suspension under abnormal conditions. These safeguards support risk control, responsibility assessment, and accountable system operation.

3.2 Task and authority allocation

Allocation in HAIS involves two related decisions. Task allocation determines who performs perception, analysis, planning, execution, and monitoring. Authority allocation determines who may propose, approve, modify, suspend, or terminate an action. The two allocations may differ: AI may perform most of the task while a human retains approval authority, or a human may direct the task while AI holds limited authority for time-critical safety actions.

Allocation strategies may be static, stage-based, or dynamic. Static allocation assigns stable roles under predictable conditions. Stage-based allocation changes roles at predefined points in the task. Dynamic allocation adjusts roles or authority during operation according to changes in task demands, human state, AI reliability, uncertainty, time pressure, risk, and action reversibility [28]. Since many of these factors can only be estimated, allocation policies should define the evidence, confidence thresholds, and safety constraints required for each adjustment.

Human authority generally carries greater importance when objectives are unclear, consequences are difficult to reverse, or value judgment is central. Greater AI authority may be suitable for rapid computation, continuous monitoring, repeated execution, or precise control. Any change in authority should follow predefined conditions, remain visible to the affected participants, and leave enough time for effective human intervention when required.

Allocation should be evaluated at the system level. Relevant measures include task performance, human workload, decision time, coordination cost, intervention effectiveness, safety, and error recovery. The objective is to select an allocation strategy that maintains reliable system performance under changing operating conditions.

3.3 Closed-loop operation

As shown in Fig. 5, HAIS operate through a closed loop connecting goal setting, state perception, assessment, coordination, action, and feedback. The process begins by defining task objectives, constraints, and acceptable outcomes. Humans and AI receive task-relevant information through digital, physical, or neural channels and generate assessments, recommendations, or control inputs according to their assigned roles.

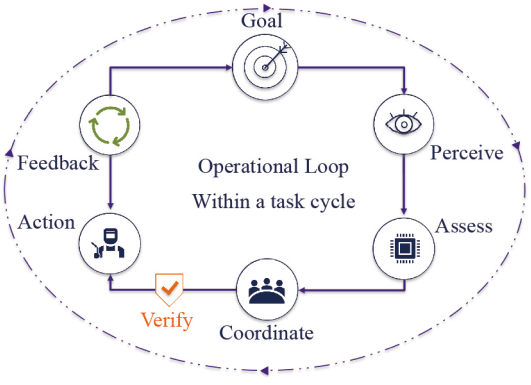


Fig. 5 Operational loop and adaptation across task cycles. The operational loop consists of goal setting, perception, assessment, coordination, verification, action, and feedback within each task cycle.

The coordination mechanism integrates these contributions and determines whether clarification, additional information, proposal revision, or authority transfer is required. The selected action then undergoes authority verification and safety checks before being

communicated or executed through the relevant interface.

Acting changes the task or environment and produces feedback for the human, AI, and shared task representation. Within a task cycle, operational feedback updates the current state, uncertainty, and authority arrangement, providing the basis for the next decision. Across multiple cycles, accumulated evidence may support longer-term adaptation of AI models, allocation policies, interaction strategies, or human working practices. Such changes are managed through the adaptation and update mechanism and remain subject to validation, monitoring, and rollback.

This sequence describes the logical relations among system functions. Practical HAIS may implement these functions through parallel, asynchronous, or hierarchical loops operating at different time scales. High-level reasoning may take seconds or longer, while physical control and neural signal processing may require millisecond-level or continuous responses. Coordination, safety constraints, and traceability should remain effective across these loops.

3.4 Operational design principles

Drawing on research in human-centered AI, human–AI teaming, shared autonomy, meaningful human control, and accountable AI systems [31–34], we synthesize four operational principles for HAIS: complementarity, negotiability, intervenability, and traceability. They respectively address capability allocation, disagreement resolution, control correction, and decision reconstruction. These principles focus on the operational relationship between humans and AI. Broader requirements, including robustness, privacy, fairness, security, and trust calibration, are addressed through the safety and governance layer described in Section 3.1. Human agency is reflected across the four principles through effective participation, explicit authority, and practical intervention.

- Complementarity: tasks and authority should reflect the relative capabilities of humans and AI under current conditions [35]. Their performance may vary with task demands, available information, uncertainty, workload, and environmental change. Complementarity can be assessed by comparing human-only, AI-only, and joint system performance across these conditions.
- Negotiability: the system should support clarification, challenge, proposal revision, and conflict resolution when human and AI assessments differ [36,37]. The basis of the final decision and the resulting authority arrangement should remain explicit before a disputed action proceeds to execution.
- Intervenability: authorized human participants should be able to review, modify, suspend, or terminate consequential actions within an effective time window [34,38]. Intervention mechanisms should remain available during normal and degraded operation, with clear information about system state, risk, and available control options.
- Traceability: the system should preserve sufficient evidence to reconstruct important decisions and actions [39,40]. Relevant records include task states, human and AI inputs, uncertainty estimates, proposed actions, authority changes, safety checks, interventions, and execution outcomes. These records support diagnosis, safety analysis, responsibility assessment, and system improvement.

Together, these principles provide an operational basis for evaluating whether human and AI capabilities are coordinated through suitable allocation, explicit conflict resolution, effective intervention, and accountable decision processes.

3.5 Real-world case

A representative digital HAIS can be found in agentic software development. We use the GitHub Copilot cloud agent as a real-world example because it operates within a shared software-development workflow rather than as an isolated code-generation model. A developer can assign an issue or task to the agent, which then analyzes repository context, modifies code, executes tests or linters in an isolated environment, and creates a pull request for human review [41]. The developer may inspect the proposed changes, request revisions, or approve the final contribution. This workflow forms a digital human–AI task loop in which both parties materially influence the software outcome.

(1) Operational mechanisms. The human defines development goals, constraints, and acceptance criteria, while the AI performs code analysis, planning, modification, and validation. The task environment consists of the repository, issue specification, dependencies, development tools, and CI infrastructure. The shared task representation is distributed across the issue description, repository state, working branch, pull request diff, test results, and review discussion. Coordination occurs through task assignment, agent-generated changes, automated checks, and human review comments. When the developer requests changes, the

updated pull request state becomes new input to the agent, producing an iterative feedback loop.

(2) Authority allocation. The case illustrates the distinction between task execution and decision authority. After a task is delegated, the agent receives bounded authority to modify code, run development tools, and update its working branch. However, this execution authority does not extend to final integration. Agent-generated pull requests require human review, and the agent cannot approve or merge its own contribution [42]. Authority therefore shifts across the workflow: the human specifies and delegates the task, the AI performs bounded execution, and control returns to the human for evaluation and integration. If revisions are requested, execution is delegated back to the AI while final approval remains human-held. This repeated delegation–review cycle provides a concrete instance of dynamic authority allocation without removing human control over consequential repository changes.

(3) Safety, governance, and traceability. Safety constraints are embedded in the execution environment. The agent is restricted to designated repository and branch contexts, operates in an ephemeral environment, and is subject to branch protection and required checks. Generated changes can be examined using CodeQL, secret scanning, and dependency analysis, while outbound network access is restricted by a firewall [42,43]. Human review provides an additional verification layer before integration. Agent-authored commits and session records also preserve the provenance of generated changes and the human who initiated the task, supporting subsequent audit and failure analysis [42].

(4) Closed-loop operation and evaluation. A typical task proceeds from human goal specification to AI analysis, code modification, automated validation, pull request generation, human assessment, and either revision or integration. The pull request and its associated checks therefore function as both an action boundary and a feedback interface. Evaluation should address the complete workflow rather than code-generation quality alone. Relevant measures include task completion and pull request acceptance, test and validation outcomes, development time, number of revision cycles, human review effort, intervention frequency, security findings, recovery from unsuccessful changes, and traceability of the final contribution. Empirical studies of coding agents already use real-world pull request acceptance as one measure of task effectiveness [5]; the HAIS perspective extends such evaluation to coordination quality, authority handling, human effort, and safety.

4 Development and Applications of HAIS

The development of HAIS reflects a broader transition from using AI as an external tool to integrating humans and AI within shared task loops. This transition involves changes in their roles, mutual influence, decision authority, and capacity for adaptation, although different forms of interaction may coexist across applications. Section 4.1 describes the evolution from tool-oriented interaction and human-in-the-loop collaboration to system-level human–AI collaboration. Section 4.2 reviews the enabling technologies, representative applications, and deployment ecosystems supporting current HAIS. Section 4.3 discusses future directions in uncertainty-aware human-state understanding, adaptive coordination and authority allocation, safe reciprocal adaptation, deeper multimodal and neural interaction, and proactive system-level assurance.

4.1 Evolution toward HAIS

The development toward HAIS can be understood through changes in how humans and AI participate in a task, influence intermediate decisions, and share control and responsibility. Three broad stages can be identified: tool-oriented interaction, human-in-the-loop collaboration, and system-level human–AI collaboration. These stages represent different forms of system organization rather than a strict chronological sequence, and they may coexist across applications. The main transition is from using AI as an external capability to integrating humans and AI within a shared and continuously updated task loop.

4.1.1 Tool-oriented human–AI interaction

In tool-oriented interaction, the human defines the task, provides the input, initiates execution, and interprets the result. The AI performs a bounded function, such as information retrieval, prediction, optimization, or automated execution [44]. Expert systems, search engines, conventional decision-support systems, and task-specific automation commonly follow this arrangement. Their division of labor is relatively stable: the human retains decision authority, while the AI supplies computational support through explicit commands and predefined interfaces.

This model improves the speed, scale, and consistency of human work, but provides limited support for joint decision-making. Human knowledge and intentions are expressed mainly through initial inputs, while the AI rarely represents the user's workload, uncertainty, expertise, or changing goals. Intermediate decisions

generally remain outside a shared task loop, and evaluation focuses on model accuracy, output quality, or user performance. These limitations motivated the integration of human feedback into model learning and task execution.

4.1.2 Human-in-the-loop collaboration

Human-in-the-loop approaches incorporate human input at intermediate stages of learning or decision-making. Humans may provide labels, demonstrations, corrections, preferences, approvals, or interventions, while AI systems perform data-intensive analysis, generate alternatives, or execute selected actions. Active learning, interactive machine learning, reinforcement learning from human feedback, and many shared-control systems follow this pattern. Repeated interaction allows human feedback to refine AI outputs and helps reduce errors under uncertainty [45].

This stage establishes an iterative task loop, although collaboration often remains asymmetric. Human input is commonly treated as supervision for improving AI performance, while changes in human judgment, behavior, and workload receive less systematic attention. Roles and authority are frequently specified in advance, and evaluation continues to emphasize the AI component or a single interaction outcome. Consequently, human-in-the-loop participation provides an important foundation for HAIS but does not automatically constitute system-level human–AI collaboration.

4.1.3 System-level human–AI collaboration

System-level human–AI collaboration integrates human and AI participation within the same task loop. Both sides can influence intermediate decisions and subsequent actions according to their capabilities, states, and operating conditions. A shared task representation maintains relevant information about goals, system state, uncertainty, constraints, and available actions. Coordination mechanisms allocate tasks and decision authority, manage disagreements, and support intervention or control transfer when conditions change.

At this stage, the human, AI, task, and environment are treated as parts of one operational system [46]. Roles and authority remain identifiable even when their allocation changes dynamically. Evaluation therefore considers joint task performance, human workload, uncertainty communication, authority transfer, safety, and recovery, in addition to model capability. Interaction records may also support longer-term adaptation of AI models, allocation policies, interaction strategies, or human working practices. Such adaptation remains controlled through validation, monitoring, and rollback. These characteristics distinguish HAIS from conventional AI tools and from human-in-the-loop systems whose primary purpose is to supervise or improve an AI component [47].

4.2 Current development and applications

Current HAIS development is supported by foundation-model agents, embodied intelligence, multimodal interfaces, collaborative control, and runtime assurance. These technologies alone do not define an HAIS. As established in Section 2, humans and AI must participate in a shared task loop, influence intermediate decisions or actions, and operate under identifiable roles and authority arrangements. Current systems satisfy these conditions to different degrees. This subsection therefore reviews their enabling technologies, representative applications, and deployment ecosystem. The discussion focuses on capabilities and system configurations that have been demonstrated in existing research or deployed systems. Their maturity and level of integration, however, vary substantially across applications.

4.2.1 Enabling technologies

Foundation-model agents support collaboration in digital environments by interpreting objectives, decomposing tasks, retrieving information, using tools, and revising plans. ReAct connects language-based reasoning with external actions [48], while AutoGen [49] supports structured interaction among humans, specialized agents, and tools. Reliable use in HAIS further requires shared state representation, uncertainty communication, permission control, and clear intervention conditions.

Embodied intelligence extends collaboration into physical environments. Vision–language–action models such as RT-2 and OpenVLA connect visual observations and semantic instructions with robot actions [50,51]. Their integration with shared control, low-level controllers, and human intervention enables collaborative physical execution. Remaining limitations include scarce interaction data, long-horizon error accumulation, latency, embodiment differences, and unreliable performance under unfamiliar conditions.

Multimodal sensing expands the information available to the task loop. Language, gaze, gesture, movement, physiological signals, and interaction history can support estimates of human attention, workload, and intended action. Brain–computer interfaces provide a specialized neural channel for decoding intended actions and, in some cases, returning sensory information through stimulation [52–55]. Signal variability, invasiveness,

individual differences, long-term stability, and clinical safety continue to limit their deployment.

Runtime assurance connects these capabilities to dependable operation. Digital agents require controlled tool access, source tracing, and protection against adversarial inputs, while embodied systems also require constrained planning, collision avoidance, and safe-state recovery. ToolEmu and AgentDojo provide methods for evaluating unsafe tool use and prompt-injection attacks [56, 57]. Integrating these capabilities into a reliable collaborative loop remains a central system-level challenge.

4.2.2 Representative applications

Current applications range from AI-assisted work to system-level human–AI collaboration. Their classification should depend on the structure of the task loop, mutual influence, authority allocation, and adaptation, rather than the technical capability or application label alone.

In scientific discovery, AlphaFold mainly serves as an advanced prediction tool whose results are interpreted by researchers [58]. More integrated platforms connect prediction, experiment selection, execution, and feedback. GNoME supports large-scale material candidate generation, while the A-Lab combines machine-learning-guided planning, robotic synthesis, measurement, and iterative updating [59,60]. When scientists influence intermediate decisions and experimental outcomes guide subsequent analysis, these platforms approach an HAIS. Their evaluation should consider experimental validity, reproducibility, human contribution, and overall discovery efficiency.

In cyber defense, AI agents can analyze alerts, logs, code, and threat intelligence, while human analysts assess attribution, operational impact, and response priorities. Such systems approach HAIS when analysts can revise intermediate plans, consequential actions follow defined authorization rules, and execution results update the shared operational state. Least-privilege access, source provenance, human confirmation, and auditable execution are essential because external content may contain adversarial instructions.

Health care and assistive robotics illustrate physical collaboration. The da Vinci surgical system maps a surgeon's commands to robotic instruments and improves precision through motion scaling and tremor reduction [61]. Its conventional master–slave structure provides limited AI autonomy and is better regarded as a precursor to advanced HAIS. The HAL exoskeleton estimates intended movement from physiological signals and adjusts assistance through repeated perception–action–feedback cycles [62], although its shared decision-making remains limited to motor support.

Remote systems such as Robonaut 2 extend human action into hazardous environments [63]. Remote operation constitutes teleoperation; qualification as an HAIS additionally requires AI participation in perception, planning, execution, or risk control within the shared task loop. The level of autonomy, conditions for control transfer, and allocation of responsibility should therefore be reported explicitly.

Neural systems provide another form of human–AI collaboration. Lee et al. combined a brain–computer interface with an AI copilot to support shared control of cursor and robotic tasks, illustrating how neural intent decoding and AI assistance can jointly determine task actions [64]. Other systems have enabled real-time robotic hand control from EEG signals [65] and high-dimensional digital control through intracortical decoding [66]. Long-term independent use of intracortical interfaces for communication and computer control further demonstrates the importance of sustained reliability, recalibration, and user control beyond short laboratory sessions [67].

These cases demonstrate a continuum of human–AI integration. Comparative analysis should identify the roles of humans and AI, the task-loop structure, authority allocation, adaptation mechanisms, and system-level outcomes. These properties provide a consistent basis for determining whether an application qualifies as an HAIS.

4.2.3 Deployment and ecosystem development

Deployable HAIS require coordinated sensing, computation, interaction, control, data management, and safety assurance. Their ecosystem includes hardware providers, model developers, system integrators, domain organizations, users, standards bodies, and regulators. The main deployment challenge is to integrate these participants and technologies into systems that remain effective and accountable throughout their lifecycle.

Hardware infrastructure includes multimodal sensors, computing devices, communication modules, actuators, and interaction interfaces. Cloud resources can support training and long-horizon reasoning, while edge and device-side computing provide low-latency perception, control, and safety functions. Software infrastructure includes foundation models, robot operating systems, simulation platforms, digital twins, shared task representations, and coordination modules. Platforms such as ROS 2 and NVIDIA Isaac connect simulation, software, and physical systems [68]. HAIS deployment additionally requires synchronized state, bounded latency, controlled permissions, and traceable execution.

Deployment forms a lifecycle loop involving adaptation, validation, monitoring, incident analysis, and updates. Changes to models, users, tools, or environments may require renewed evaluation. Initiatives such as Open X-Embodiment support data and policy transfer across devices and tasks [69], although differences in hardware, semantics, data quality, privacy, and intellectual-property rules continue to limit interoperability.

Standards and regulations provide further support. ISO/TS 15066 addresses safety requirements for collaborative industrial robots [70], while the European Union Artificial Intelligence Act establishes broader requirements for risk management, human oversight, documentation, and accountability [71]. Further work is required to address system-level evaluation, dynamic authority transfer, interaction-data governance, reciprocal adaptation, and cross-domain failure propagation.

4.3 Development trajectories

Section 3 defines human-state assessment, dynamic coordination, reciprocal adaptation, and safety assurance as core mechanisms of HAIS. Current HAIS, however, remain constrained by uncertain human-state inference, fixed authority arrangements, one-way adaptation, limited interaction channels, and fragmented safeguards. This section therefore outlines the major directions in which HAIS are evolving. As shown in Fig. 6, five development trajectories are particularly important: implicit intent understanding, adaptive coordination and authority allocation, safe reciprocal adaptation, deeper multimodal and neural interaction, and proactive system-level assurance. The directions discussed in this subsection are prospective research directions rather than capabilities assumed to be generally available in current HAIS. They build on demonstrated advances reviewed in Section 4.2, but their system-level reliability, generalizability, and long-term effectiveness remain open research questions.

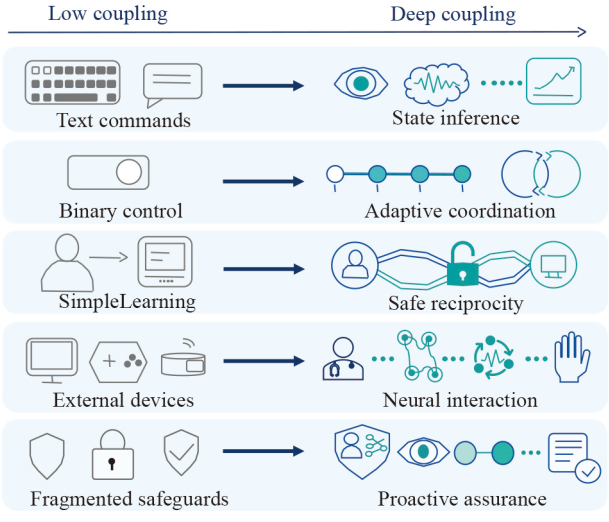


Fig. 6 Conceptual research directions for HAIS. The right-hand side represents prospective capabilities, including intent understanding, adaptive coordination, safe reciprocity, neural interaction, and proactive assurance. These directions denote research objectives rather than capabilities generally demonstrated in current systems.

4.3.1 Human-state inference

HAIS are moving from explicit user commands toward richer inference of task-relevant human states, including goals, attention, workload, and intended actions. Multimodal evidence from language, gaze, gesture, behavior, physiological signals, and interaction history can support such inference [72]. Because these signals are incomplete and context dependent, future systems will need to represent uncertainty rather than treat inferred states as direct or definitive access to human intent. The objective is to support more responsive interaction while keeping inferred intentions distinguishable from explicit human authorization.

4.3.2 Adaptive coordination and authority allocation

Current systems often rely on fixed task assignments or predefined control transfers. HAIS are increasingly moving toward adaptive coordination in which task allocation and decision authority change with human state, AI capability, uncertainty, environmental risk, and available response time [73]. Such arrangements may range from AI assistance under human control to shared execution, conditional autonomy, and transfer of control to a human decision maker. The key shift is from static role assignment toward context-dependent coordination while keeping current authority arrangements visible and understandable to the participants.

4.3.3 Safe reciprocal adaptation

HAIS are also moving from one-sided adaptation toward reciprocal adaptation across repeated interactions. AI systems may adjust to human instructions, corrections, and preferences, while humans may change their strategies and expectations in response to AI behavior [74]. This development requires a distinction between operational feedback within a task cycle and longer-term adaptation across cycles. The former updates the current task state and coordination process, whereas the latter may change models, allocation policies, interaction strategies, or human procedures. Neural interfaces

provide one example, where prolonged use may require decoder recalibration as neural signals and user interaction patterns change [75].

4.3.4 Deeper multimodal and neural interaction

Human–AI interaction is expanding from predominantly explicit digital input toward multimodal and, in specialized applications, neural interaction. Language, vision, gesture, touch, physiological signals, and environmental context can provide complementary information for understanding task state and human intent. Neural interfaces further extend this interaction by decoding intended movement or communication and, in some systems, returning sensory information through stimulation [76]. These developments may support more direct control and feedback in assistive, rehabilitation, and remote-operation settings. Neural interaction, however, represents one specialized interaction channel rather than a universal endpoint of HAIS.

4.3.5 Proactive system-level assurance

As human–AI coupling becomes deeper, assurance is moving from separate component safeguards toward monitoring and constraining the complete task loop. Failures may propagate across models, tools, human decisions, physical controllers, and neural interfaces, making isolated safeguards insufficient. Future HAIS will increasingly incorporate system-level constraints on goals, authority, permitted actions, and acceptable risk before or during execution. Proactive assurance is treated here as a long-term system design objective rather than a claim that complete system safety can be guaranteed. The main transition is from fragmented and reactive protection toward proactive assurance that remains effective across interacting human–AI components and operating conditions.

5 Challenges and Risks in HAIS

Deeper human–AI coupling can improve joint performance, while also creating new paths through which errors, attacks, and harmful adaptations can spread across humans, AI components, tools, and environments. These problems arise at three connected levels. Capability barriers reduce the accuracy of human-state modeling, collective coordination, and interaction-based learning. Emergent risks appear when incorrect estimates, unsafe actions, or malicious inputs propagate through the collaborative loop. Assurance and governance gaps make these failures difficult to detect, contain, and attribute. This section examines how HAIS may fail and why such failures occur. The corresponding research directions are discussed in Section 6.

5.1 Capability barriers

Reliable HAIS require task-relevant models of human states, effective coordination among heterogeneous participants, and interaction data that preserve temporal and causal context. Current methods remain limited in all three areas. These limitations can distort the shared task state, weaken task and authority allocation, and reduce the reliability of system-level adaptation.

5.1.1 Human-state modeling

Effective collaboration depends on the ability of an AI system to estimate task-relevant human states, including goals, knowledge, attention, workload, preferences, and intended actions. These states cannot be observed directly [75]. They must be inferred from incomplete evidence such as language, behavior, gaze, interaction history, and, in specialized applications, physiological or neural signals.

The relationship between observable signals and internal states is uncertain. Hesitation may indicate fatigue, disagreement, confusion, or careful reflection. The same instruction may also express different goals across users and situations. Differences in expertise, culture, physical condition, and task context further reduce the reliability of fixed inference rules. Foundation models can integrate multimodal evidence, yet strong performance on controlled benchmarks does not ensure reliable understanding across unfamiliar users and environments. A model may rely on familiar linguistic or behavioral patterns while appearing to possess a stable representation of human beliefs or intentions [78].

Errors in human-state modeling can propagate directly into system operation. An incorrect estimate of workload may cause the AI to transfer authority at an unsuitable time. A mistaken inference of intention may trigger an unwanted action, while an outdated preference model may bias later recommendations. Model opacity, unsupported outputs, and poorly calibrated confidence make these errors harder to identify [79]. The system may produce a coherent interpretation even when several explanations remain possible. This uncertainty weakens the basis for deciding when the AI should act, seek clarification, or defer to human judgment.

5.1.2 Collective coordination

Many real-world HAIS involve teams of human experts,

multiple AI agents, and external tools. Moving from one human and one AI to a mixed collective introduces differences in expertise, objectives, information access, workload, and authority. Existing coordination methods provide limited support for this many-to-many structure [80].

Failures may originate in task allocation, communication, or shared state representation. Participants may receive overlapping tasks, depend on inconsistent information, or remain unaware of changes made by other members. Shared memory can improve access to information, while also allowing an incorrect claim to influence the whole group. Agents based on similar models may produce correlated errors and create an appearance of agreement without independent evidence. Repeated exchanges may also generate communication loops, duplicated work, and high computational cost.

Authority becomes difficult to trace when several agents act for different people or organizations. An agent may follow one participant's instruction while affecting the interests of others. Conflicting goals can remain hidden until they influence an important decision or action. Current multi-agent evaluations mainly measure task completion and provide limited evidence about human workload, distribution of influence, disagreement resolution, and decision traceability [81]. Consequently, a collective may achieve a strong task score while remaining vulnerable to correlated errors, coordination breakdown, or unclear responsibility [82].

5.1.3 Interaction data

HAIS operate through sequences of interdependent decisions. A useful interaction record may need to include the task objective, environmental state, human behavior, AI estimates and recommendations, tool calls, authority transfers, interventions, and final outcomes. Model versions and previous interactions may also affect later decisions. Conventional datasets built around independent input–label pairs cannot represent this structure adequately.

Available interaction datasets are commonly short, small, and collected under controlled conditions. They often overrepresent healthy adults, trained operators, and a limited range of languages, cultures, devices, and tasks [83]. Rare failures, emergencies, long-term dependence, and gradual changes in human behavior are particularly difficult to observe. Systems trained on such data may treat patterns from a narrow population as general indicators of intention, workload, or preference.

Interaction data are also shaped by the AI system that produces them. Once a user follows an AI recommendation, the observed outcome reflects both the user's preference and the system's influence. Counterfactual outcomes under a different recommendation or without AI assistance are usually unavailable. This makes causal evaluation and preference learning difficult. Detailed behavioral, physiological, and neural records also create privacy constraints that limit collection and sharing [81]. Weak temporal coverage, limited diversity, feedback bias, and missing counterfactuals can therefore lead to poor generalization and unreliable long-term adaptation.

5.2 Emergent system risks

Tightly coupled HAIS can produce system-level risks even when individual components satisfy their local requirements. Human decisions affect AI updates, AI recommendations influence human behavior, and digital outputs may trigger physical actions. This coupling allows small errors or malicious inputs to spread through feedback loops and produce consequences that cannot be explained by one component alone.

5.2.1 Trust and autonomy

Human participants must decide when to accept, question, or reject AI outputs. Their decisions depend on perceived competence, past performance, interface design, explanations, and organizational expectations. These signals may differ from the actual reliability of the AI under current operating conditions.

Fluent language, confident recommendations, and plausible explanations can create excessive trust when the underlying output is uncertain or incorrect. Conversational interfaces may further encourage users to attribute understanding or intention to the AI. Repeated success under routine conditions can produce automation bias, causing users to accept recommendations without sufficient verification. Repeated errors or unexplained changes can produce the opposite response, leading users to reject useful assistance. Both forms of trust mismatch reduce joint performance and can delay appropriate intervention [79].

Autonomy risks increase when the AI infers a goal or initiates an action before receiving clear authorization. An inferred intention may reflect an outdated preference, a temporary response, or an incorrect interpretation of hesitation. Proactive assistance can also narrow the options shown to the user and influence later choices. In organizational settings, AI recommendations may reflect institutional objectives that differ from the interests of the affected person. These mechanisms can weaken meaningful human control even when the user formally retains final authority.

5.2.2 Security and privacy

HAIS often connect AI models to memory, external data, software tools, communication services, sensors, and physical actuators. Each connection creates an additional path for unauthorized access or manipulation. An attack on one component can propagate through the same channels used for legitimate collaboration.

Foundation-model agents are exposed to adversarial inputs, indirect prompt injection, malicious external content, unauthorized tool use, and excessive privileges [56,57]. A harmful instruction embedded in a document or website may influence an agent to disclose information or execute an action outside the user's intended task. In a multi-agent system, a compromised agent may modify shared memory, distort task allocation, or influence other agents. Unsafe or malicious human instructions can also be amplified when AI systems execute them at greater speed and scale.

Physical and neural integration increases the possible consequences. A manipulated digital output may become an unsafe control command, while compromised sensor data may distort the system's estimate of the human or environment. Failures involving robots, vehicles, medical devices, or assistive systems may produce direct physical harm. Neural interfaces create additional concerns because altered decoding or stimulation may affect user control, sensation, or health.

Privacy risks extend beyond conventional identifiers. Interaction histories may reveal expertise, routines, preferences, workload, emotional state, or personal vulnerabilities. Physiological and neural signals may expose health conditions and intended movements. Persistent records can also reveal information about colleagues, patients, or family members who did not directly interact with the system. Longitudinal behavioral data may remain identifiable after direct identifiers are removed.

Consent becomes difficult to maintain as data are reused across personalization, monitoring, evaluation, and model updates. Users may agree to immediate task assistance without expecting their data to support employee assessment or behavioral prediction. The combination of broad system access, sensitive data, and repeated interaction creates security and privacy risks that can spread across individuals, organizations, and physical environments.

5.2.3 Co-evolutionary drift

Reciprocal adaptation can improve collaboration, but it can also move the joint system away from its original objectives. AI recommendations influence human decisions, and these decisions may later become training or personalization data [81]. The system therefore learns from behavior that it has already helped to shape.

Observed acceptance does not always represent genuine preference. Users may follow a recommendation because it is prominent, easy to select, or supported by organizational procedures. Repeated exposure may turn the recommended option into a habit. If the AI treats this behavior as positive feedback, it may reinforce interface effects, authority effects, or learned dependence. Early biases can grow as the human and AI repeatedly respond to one another.

Long-term use can also change human capabilities and organizational practices. Users may lose experience in tasks that are routinely delegated to AI, while workflows and data formats become dependent on a specific system. Short-term performance may continue to improve as independent human capability, alternative procedures, and organizational resilience decline. This creates lock-in and increases the impact of later model failure or system withdrawal.

Co-evolutionary drift develops across many interactions and may remain invisible in short evaluations. No individual update needs to appear unsafe. Harm emerges as user behavior, model policies, data distributions, and institutional practices gradually reinforce one another [84]. The resulting system may continue to perform efficiently while pursuing objectives that differ from those under which it was originally designed and evaluated.

5.3 Assurance and governance gaps

Section 3 describes safety assurance as a cross-cutting part of the HAIS architecture, and Section 4 identifies proactive assurance as an important development direction. In practice, existing evaluation, monitoring, and accountability methods still focus heavily on individual models or fixed products. These gaps limit the ability to identify system-level failures, trace their causes, and determine who could have prevented them.

5.3.1 System-level evaluation gaps

Most AI evaluations measure properties of individual models, such as accuracy, robustness, calibration, fairness, or computational performance. A strong model score does not ensure reliable performance of the complete human–AI–task–environment configuration. Joint failure may result from a misunderstood recommendation, delayed intervention, inconsistent shared state, unsuitable authority allocation, or an incorrect connection between high-level planning and

physical control.

HAIS evaluation must cover joint task performance and the quality of interaction. Relevant properties include human workload, uncertainty communication, trust calibration, conflict frequency, authority-transfer quality, error detection, and recovery. Agentic systems also require evaluation of tool selection, permission compliance, source provenance, and unauthorized actions. Embodied and neural systems introduce further requirements related to latency, physical safety, signal reliability, reversibility, and recovery.

Current benchmarks rarely represent long-term adaptation, rare emergencies, adversarial interference, or multiple interacting failures [85]. They also provide limited evidence about performance across different users and operating conditions. As a result, systems may pass component-level tests while important failure paths within the complete task loop remain unexamined. Existing risk-management and organizational standards provide useful foundations [86], but they do not fully address dynamic authority, reciprocal adaptation, and cross-domain failure propagation.

5.3.2 Lifecycle monitoring gaps

HAIS can change after deployment even when their initial components have been validated. Models and tools may be updated, data distributions may shift, users may alter their working practices, and operating environments may move beyond their original boundaries. Long-term human adaptation can also change the behavior of the joint system without any formal software update [87]. One-time evaluation therefore provides limited evidence about continued system performance. This issue is particularly relevant to neural HAIS, where signal characteristics, decoder performance, and user capabilities may change over extended periods of use, as observed in long-term implanted BCI studies [88].

Post-deployment monitoring often focuses on technical indicators such as model errors or system availability. These measures may miss changes in user reliance, intervention quality, workload, control conflicts, or independent human capability. Slow changes are especially difficult to detect because each interaction may appear acceptable when examined separately. Monitoring can also fail when the same model or communication channel is used both for operation and for detecting its own failures.

Incident analysis faces similar limitations. Logs from models, interfaces, tools, controllers, and human actions may use different time scales or remain stored by different organizations. Missing or unsynchronized records make it difficult to reconstruct how an initial error became a system-level outcome. Near misses may remain unreported, preventing organizations from identifying recurring failure paths before severe harm occurs.

5.3.3 Accountability and oversight gaps

HAIS distribute influence across model providers, data suppliers, software developers, hardware manufacturers, system integrators, deploying organizations, managers, and operators [89]. The person who issues the final command may have limited control over the model, available information, interface design, or time allowed for intervention. Assigning responsibility only at the final decision point can therefore conceal earlier technical and organizational causes.

Formal human oversight may also lack practical meaning. An operator may retain nominal authority while receiving insufficient information, time, training, or control to challenge the system. AI recommendations can shape which options and evidence are visible before the human makes a decision. In such cases, placing a human in the loop may shift liability without providing effective control. Conversely, the use of AI does not remove the responsibility of organizations that select, configure, and deploy it.

Traceability is required to identify how humans and technical components influenced an outcome, yet extensive logging may conflict with privacy and data-minimization requirements. Records of human behavior, physiological state, or neural signals can support incident analysis while also creating sensitive datasets. Incomplete logging weakens accountability, whereas excessive logging increases surveillance and privacy risks.

HAIS also cross regulatory boundaries involving AI, cybersecurity, robotics, medical devices, transportation, privacy, and employment [90]. Compliance with one domain-specific framework cannot demonstrate the safety of the complete system. Dynamic authority allocation and shared decision-making further complicate the division of operational control, causal responsibility, and legal liability. These evaluation, monitoring, and accountability gaps explain why safety mechanisms at the component level cannot fully contain the risks of HAIS. Section 6 develops research priorities and strategic actions for addressing these limitations.

6 Research Agenda for HAIS

Section 4.3 describes the major development trajectories of HAIS. This section focuses instead on the research needed to make these trajectories reliable, safe, and

accountable. The challenges identified in Section 5 provide the basis for this research agenda. Capability barriers require stronger methods for human-state modeling, interaction representation, and coordination. Emergent system risks require mechanisms for preserving human agency, containing cross-domain failures, and controlling long-term adaptation. Assurance and governance gaps require system-level evaluation, lifecycle monitoring, traceability, and responsibility allocation. Accordingly, we organize the research agenda into three priorities: foundations for reliable collaboration, safe interaction and adaptation, and lifecycle assurance and accountability. These priorities are supported by interdisciplinary testbeds, standards and lifecycle governance, and ecosystem collaboration. The following priorities are open research problems derived from the limitations identified in Section 5, rather than capabilities assumed to have been achieved. Fig. 7 connects the capability barriers, emergent risks, and assurance gaps identified in Section 5 with the research priorities and strategic enablers proposed in Section 6.

6.1 Research priorities

The three research priorities correspond to the three categories of challenges introduced in Section 5. Each priority addresses a distinct level of HAIS behavior. The first focuses on the information and coordination capabilities required for effective collaboration. The second addresses risks generated through interaction, authority allocation, and long-term adaptation. The third concerns the evidence and governance mechanisms needed to evaluate, monitor, and manage HAIS throughout their lifecycle.

6.1.1 Foundations for reliable collaboration

This priority addresses human-state modeling, interaction data, and coordination among heterogeneous participants. Reliable collaboration requires task-relevant estimates of human states, representations that preserve the temporal structure of interaction, and coordination mechanisms that account for differences in capability, information, uncertainty, and authority.

Human-state modeling should represent goals, knowledge, attention, workload, preferences, and intended actions from incomplete multimodal evidence. Because the same observation may support several interpretations, models should maintain alternative hypotheses, estimate their uncertainty, and request clarification when the available evidence is insufficient for reliable action [77,78]. They should also distinguish relatively stable preferences from temporary behavior caused by fatigue, time pressure, unfamiliar tasks, or earlier AI recommendations.

Progress also depends on interaction data that preserve complete task trajectories. Relevant records include task objectives, environmental states, human behavior, AI recommendations, tool calls, authority transfers, interventions, and outcomes. Datasets should cover different users, tasks, operating conditions, successful interactions, and failure cases. Longitudinal data are particularly important for studying changes in trust,

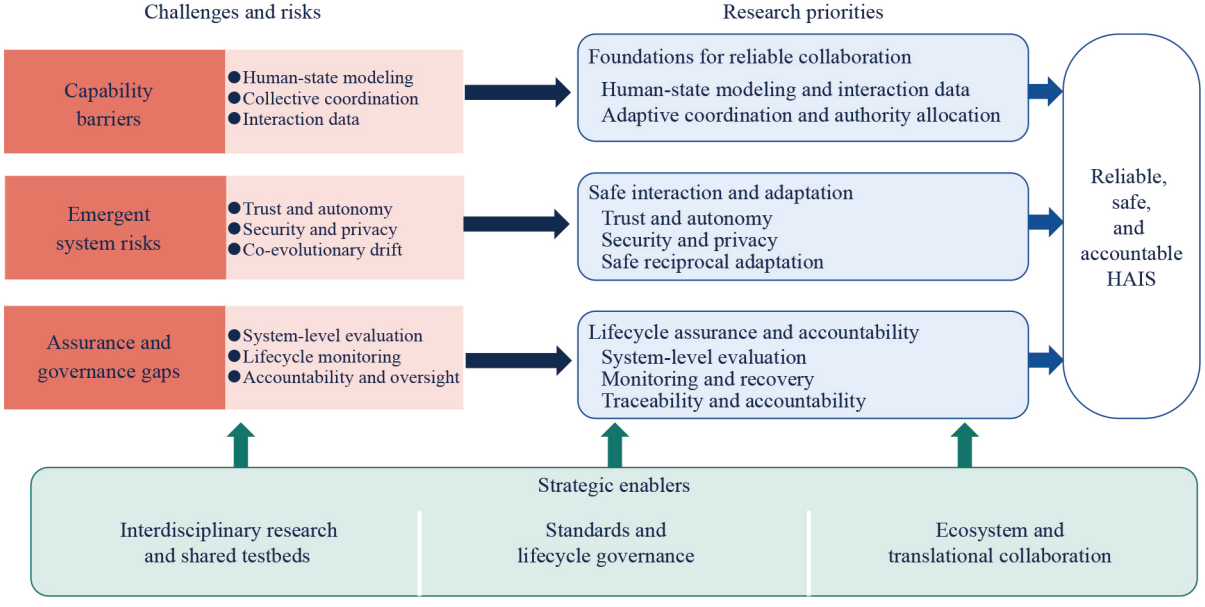


Fig. 7 Correspondence between the major challenges, research priorities, and strategic enablers for HAIS.

working practices, and human capabilities. Because AI recommendations can influence later human behavior, observed acceptance should not automatically be treated as evidence of an independent human preference. Controlled comparisons and causal analysis are needed to separate prior preferences from behavior shaped by the system.

Coordination research should develop methods for allocating tasks, information, and authority among humans and AI. Allocation decisions should account for human expertise and workload, AI capability and uncertainty, environmental risk, action reversibility, available intervention time, and the consequences of error. These methods should also support mixed collectives containing multiple humans, AI agents, and external tools. Evaluation should consider collective decision quality, communication and coordination cost, disagreement-resolution rate, authority-transfer success rate, distribution of influence, resilience to compromised participants, and preservation of predefined human authority and intervention rights [82].

6.1.2 Safe interaction and adaptation

This priority addresses risks that arise through interaction among humans, AI components, tools, and environments. Research is needed to maintain appropriate human reliance, preserve effective intervention and authority, contain failures across interaction channels, protect sensitive data, and keep long-term adaptation within defined operating boundaries.

Trust research should examine whether human reliance matches the actual capabilities and limitations of AI under current conditions. Interfaces should communicate system state, uncertainty, operating boundaries, and changes in authority in forms that support informed human judgment. Evaluation should examine overreliance, underuse, delayed intervention, error detection, and the transfer of trust from familiar tasks to unfamiliar or higher-risk settings.

Human agency also depends on maintaining clear distinctions among assistance, recommendation, approval, shared execution, and autonomous action. Inferred intention should not automatically be treated as authorization. Consequential actions and authority changes should remain perceptible to affected users, with sufficient information and time for intervention. Research is therefore needed on authority constraints, intervention conditions, and fallback mechanisms that remain effective when communication, perception, or control degrades.

Security research should address how attacks and component failures propagate across digital, physical, and neural channels. Digital agents require protection against indirect prompt injection, malicious external content, unauthorized tool use, and excessive execution privileges [56,57]. Relevant mechanisms include identity verification, least-privilege access, source provenance, isolation of untrusted execution, and independent confirmation of high-impact actions. Embodied systems additionally require monitoring of unsafe plans, abnormal sensor inputs, controller failures, and actions outside approved operating boundaries [91]. Neural applications require additional safeguards for signal reliability, stimulation limits, privacy, and reversibility.

Privacy research should address information inferred from long-term interaction, including expertise, routines, workload, health-related states, and personal vulnerabilities. Behavioral, physiological, and neural data should be collected only when relevant to the task and should be subject to purpose limitation, access control, controlled reuse, and appropriate retention policies [81].

Safe adaptation requires a clear separation between operational feedback and persistent system updates. Methods should determine whether observed behavior reflects a stable preference, a temporary response, an interface effect, or adaptation to previous AI recommendations. They should also detect harmful dependence, capability degradation, feedback amplification, and coordination drift. Model and policy updates should remain bounded, traceable, and reversible through validation, staged deployment, monitoring, version control, and rollback [84].

6.1.3 Lifecycle assurance and accountability

This priority addresses system-level evaluation, lifecycle monitoring, traceability, and accountability. Because HAIS may change through model updates, new tools, environmental shifts, and changes in human behavior, assurance must cover the complete human–AI–task–environment configuration across design, deployment, operation, adaptation, and retirement.

System-level evaluation should examine both joint task outcomes and observable interaction processes. Relevant measures include task success, task quality, completion time, human workload, uncertainty communication, disagreement frequency and resolution, authority-transfer success and latency, intervention effectiveness, error detection, recovery, and safety outcomes. Agentic systems additionally require evaluation of tool selection, permission compliance, source provenance, and containment of unauthorized actions. Embodied and neural systems require application-specific measures such as latency, physical safety, signal reliability, reversibility, and recovery.

Benchmark design should represent complete interaction trajectories rather than isolated model outputs. The same or matched task instances should, where applicable, be evaluated under human-only, AI-only, and HAIS conditions, as well as under alternative task and authority allocation strategies. Benchmarks should include normal, uncertain, degraded, and adversarial conditions and record human and AI inputs, disagreements, authority transfers, interventions, actions, and outcomes. Long-duration settings are also needed to examine adaptation and gradual changes in system behavior.

Lifecycle monitoring should track both technical and human changes after deployment. Relevant indicators include task outcomes, AI uncertainty, human reliance, intervention frequency, control conflicts, workload, security incidents, and performance across users and operating conditions. Monitoring should detect gradual drift as well as immediate failures, and should include near misses that reveal failure paths before serious harm occurs.

Changes to models, tools, allocation policies, interfaces, or operating conditions may invalidate earlier assurance evidence. Research is therefore needed on change-impact analysis and risk-based re-evaluation. Significant modifications may require renewed simulation, adversarial testing, controlled trials, or staged deployment. Monitoring and recovery mechanisms should also define when normal operation can continue, when autonomy should be reduced, and when the system should return to a previously validated state.

Traceability should connect task states, model and policy versions, AI recommendations, uncertainty estimates, human inputs, tool calls, authority transfers, interventions, and final outcomes. Records should be sufficient to reconstruct consequential decisions while remaining proportional to risk and avoiding unnecessary collection of sensitive behavioral, physiological, or neural data.

Accountability research should distinguish operational control, causal influence, decision authority, and organizational responsibility. Developers, model providers, hardware manufacturers, system integrators, deploying organizations, managers, and operators influence different stages of HAIS behavior. Responsibility should therefore reflect each actor's practical ability to anticipate, prevent, monitor, or correct a failure.

6.2 Strategic enablers

The three research priorities require common technical, institutional, and organizational support. Three enablers are particularly important: interdisciplinary research and shared testbeds, standards and lifecycle governance, and ecosystem and translational collaboration.

6.2.1 Interdisciplinary research and shared testbeds

HAIS research spans AI, human factors, cognitive science, control engineering, robotics, cybersecurity, neuroscience, ethics, law, and application-specific expertise. These perspectives should be integrated when defining operating boundaries, evaluation criteria, and safety requirements. Shared testbeds should support controlled and repeatable study of long-duration interaction, dynamic authority allocation, adversarial conditions, physical simulation, and controlled system updates. Common experimental protocols and reporting formats can improve reproducibility and comparison across research groups.

6.2.2 Standards and lifecycle governance

Standards should extend beyond model-level properties to cover joint system performance, authority allocation, adaptation, recovery, traceability, and human intervention. They should specify the evidence required to support claims concerning intended use, operating boundaries, autonomous action, human oversight, and safe degradation. Governance should cover system design, data use, integration, testing, deployment, updating, incident response, and retirement. Existing frameworks such as the NIST AI Risk Management Framework, ISO/IEC 42001, and the European Union Artificial Intelligence Act provide useful foundations [71,85,86], while additional guidance is needed for dynamic authority, reciprocal adaptation, neural data, and cross-domain failure propagation.

6.2.3 Ecosystem and translational collaboration

Operational HAIS involve model providers, hardware manufacturers, software developers, system integrators, domain organizations, users, standards bodies, and regulators. Their responsibilities for validation, monitoring, maintenance, data governance, and incident response should be defined before deployment. The transition from research prototypes to operational systems should proceed through simulation, controlled trials, limited deployment, and continued monitoring. Shared datasets, interoperable interfaces, reference implementations, and governed testing facilities can further improve reproducibility and reduce duplicated effort.

7 Conclusions

This article presents HAIS as integrated systems in which humans and AI participate in a shared task loop, influence intermediate decisions and actions, and operate under identifiable roles and authority arrangements. Based on this definition, it establishes the conceptual boundaries of HAIS and proposes a reference architecture comprising system entities, operational mechanisms, and cross-cutting safety and governance. The architecture supports closed-loop collaboration within individual tasks and controlled adaptation across repeated task cycles. The review further shows that current applications range from AI-assisted tools to system-level Human–AI collaboration, with substantial differences in mutual influence, authority allocation, and adaptation. As coupling becomes deeper, limitations in human-state modeling, coordination, and interaction data can propagate into trust miscalibration, security and privacy threats, co-evolutionary drift, and unclear accountability. HAIS should therefore be evaluated as complete human–AI–task–environment configurations throughout their lifecycle. Future progress depends on uncertainty-aware human modeling, adaptive coordination, safe reciprocal adaptation, runtime assurance, system-level evaluation, and traceability, supported by interdisciplinary testbeds, lifecycle standards, and coordinated deployment ecosystems. These developments can enable HAIS to extend complementary human and AI capabilities while preserving meaningful human agency, operational reliability, and accountable control.

Data Availability No datasets were generated or analyzed for this work.

Acknowledgments This work was supported in part by the National Key Research and Development Program of China (No. 2023YFB3107400), in part by the National Natural Science Foundation of China (No. 62688101, No. 62521002, No. U24B20185, No. U2441240, No. 62132011, and No. 62192781), in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM114), and in part by the Key Research and Development Project in Shaanxi Province (No. 2024PT-ZCK-89). The authors would also like to thank the New Cornerstone Science Foundation and the Xplorer Prize for their support.

Competing interests The authors declare no conflicts of interest.

References

1. Naveed H, Khan A U, Qiu S, et al. A comprehensive overview of large language models. ACM Transactions on Intelligent Systems and Technology, 2025, 16(5): 106
2. Zhao W X, Zhou K, Li J Y, et al. A survey of large language models. Frontiers of Computer Science, 2026, 20(12): 2012627
3. Jin Y Z, Li J, Gu T J, et al. Efficient multimodal large language models: a survey. Visual Intelligence, 2025, 3(1): 27
4. Yang L, Tian Y, Li B W, et al. MMaDA: multimodal large diffusion language models. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA, 2026, 138867–138907
5. Pinna G, Gong J Z, Williams D, et al. Comparing AI coding agents: a task-stratified analysis of pull request acceptance. In: Proceedings of the 23rd International Conference on Mining Software Repositories. Rio de Janeiro, Brazil, 2026, 792–796
6. Liu Y, Chen W X, Bai Y J, et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI. IEEE/ASME Transactions on Mechatronics, 2025, 30(6): 7253−7274
7. He R, Cao J, Tan T N. Generative artificial intelligence: a historical perspective. National Science Review, 2025, 12(5): nwaf050
8. Sengar S S, Hasan A B, Kumar S, et al. Generative artificial intelligence: a systematic review and applications. Multimedia Tools and Applications, 2025, 84(21): 23661−23700
9. Hanna M G, Pantanowitz L, Dash R, et al. Future of artificial intelligence—machine learning trends in pathology and medicine. Modern Pathology, 2025, 38(4): 100705
10. Zhang J F, Chen C M, Chu S C, et al. *Advances in intelligent systems and computing*. Singapore: Springer, 2022
11. Wang Q, Li Y F, Li R R. Integrating artificial intelligence in energy transition: a comprehensive review. Energy Strategy Reviews, 2025, 57: 101600
12. Chen J, Sun J, Wang G. From unmanned systems to autonomous intelligent systems. Engineering, 2022, 12: 16−19
13. Henry K E, Kornfield R, Sridharan A, et al. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digital Medicine, 2022, 5(1): 97
14. Lu T, Zhang Y J. 1+ 1> 2? Information, humans, and machines. Information Systems Research, 2025, 36(1): 394−418
15. Pedreschi D, Pappalardo L, Ferragina E, et al. Human-AI coevolution. Artificial Intelligence, 2025, 339: 104244
16. Liu F L, Zhou H J, Gu B Y, et al. Application of large language models in medicine. Nature Reviews Bioengineering, 2025, 3(6): 445−464
17. Cui Z Y, Li N, Zhou H K. A large-scale replication of scenario-based experiments in psychology and management using large language models. Nature Computational Science, 2025, 5(8): 627−634
18. Yang L C, Zhang C, Wang R Q, et al. Dynamic electromechanical co-stimulation based enhancement of skeletal muscle tissues for fast biosyncretic robots actuation. Advanced Functional Materials, 2025, 35(2): 2410334
19. Chen Z K, Chen J F, Jung S, et al. Bioinspired and biohybrid soft robots: principles and emerging technologies. Matter, 2025, 8(4): 102045

20. Khan W U, Shen Z Z, Mugo S M, et al. Implantable hydrogels as pioneering materials for next-generation brain–computer interfaces. Chemical Society Reviews, 2025, 54(6): 2832−2880
21. Patrick-Krueger K M, Burkhart I, Contreras-Vidal J L. The state of clinical trials of implantable brain–computer interfaces. Nature Reviews Bioengineering, 2025, 3(1): 50−67
22. Lorach H, Galvez A, Spagnolo V, et al. Walking naturally after spinal cord injury using a brain–spine interface. Nature, 2023, 618(7963): 126−133
23. Chandra S, Shirish A, Srivastava S C. To be or not to be … human? Theorizing the role of human-like competencies in conversational artificial intelligence agents. Journal of Management Information Systems, 2022, 39(4): 969−1005
24. Agashe S, Han J Z, Gan S Y, et al. Agent S: an open agentic framework that uses computers like a human. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore, 2025, 22924–22946
25. Chan J S, Chowdhury N, Jaffe O, et al. MLE-bench: evaluating machine learning agents on machine learning engineering. In: Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore, 2025, 50466–50494
26. Hermanns L, Teubner T. Under pressure: how time constraints, task complexity, and AI reliability shape human-AI interaction. Behaviour & Information Technology, 2026, 45(11): 2597−2621
27. Edwards J, Nguyen A, Lämsä J, et al. Human-AI collaboration: Designing artificial agents to facilitate socially shared regulation among learners. British Journal of Educational Technology, 2025, 56(2): 712−733
28. Mayer L W, Karny S, Ayoub J, et al. Human–AI collaboration: trade-offs between performance and preferences. Cognitive Research: Principles and Implications, 2026, 11(1): 18
29. Phiri C C. Creating characteristically auditable agentic AI systems. In: Proceedings of the Intelligent Robotics FAIR 2025. Budapest, Hungary, 2025, 1–14
30. Gyevnár B, Kasirzadeh A. AI safety for everyone. Nature Machine Intelligence, 2025, 7(4): 531−542
31. Capel T, Brereton M. What is human-centered about human-centered AI? A map of the research landscape. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg, Germany, 2023, 359
32. Gonzalez C, Donahue K, Goldstein D G, et al. Toward a science of human–AI teaming for decision making: a complementarity framework. PNAS Nexus, 2026, 5(3): pgag030
33. Liu C Q, Wang Z L, Liu Z Y, et al. Multi-agent reinforcement learning framework for addressing demand-supply imbalance of shared autonomous electric vehicle. Transportation Research Part E: Logistics and Transportation Review, 2025, 197: 104062
34. Tsamados A, Floridi L, Taddeo M. Human control of AI systems: from supervision to teaming. AI and Ethics, 2025, 5(2): 1535−1548
35. Gonzalez C, Heidari H. A cognitive approach to human–AI complementarity in dynamic decision-making. Nature Reviews Psychology, 2025, 4(12): 808−822
36. Vaccaro M, Caosun M, Ju H R, et al. Advancing AI negotiations: a large-scale autonomous negotiation competition. Proceedings of the National Academy of Sciences of the United States of America, 2026, 123(23): e2521774123
37. Jahromi M N S, Holm S, Moeslund T B, et al. Operationalizing AI contestability through technical feasibility analysis. Discover Artificial Intelligence, 2026, 6(1): 648
38. Bahurmuz A. Measuring the impact of human–AI collaborative personalized interventions through temporal causal inference. International Journal for Applied Information Management, 2026, 6(2): 325−342
39. Wang F, Xi X F, Cui Z M, et al. Embedding traceability in large language model code generation: towards trustworthy AI-augmented software engineering. In: Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering. Trondheim, Norway, 2025, 1760–1763
40. Radanliev P. Privacy, ethics, transparency, and accountability in AI systems for wearable devices. Frontiers in Digital Health, 2025, 7: 1431246
41. GitHub. About GitHub copilot cloud agent. Available online (accessed 16 Sep 2026).
42. GitHub. Risks and mitigations for GitHub copilot cloud agent. Available online (accessed 16 Sep 2026).
43. GitHub. Customizing or disabling the firewall for GitHub copilot. Available online (accessed 16 Sep 2026).
44. Kadenhe N, Al Musleh M, Lompot A. Human-AI co-design and co-creation: a review of emerging approaches, challenges, and future directions. In: Proceedings of the 2025 AAAI Symposium Series. Dubai, UAE, 2025, 265–270
45. Natarajan S, Mathur S, Sidheekh S, et al. Human-in-the-loop or AI-in-the-loop? Automate or collaborate? In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA, 2025, 28594–28600
46. George M, Binh A D T, Tran L T H, et al. The evolution of human-AI collaboration: a data-driven systematic literature review and multi-level maturity framework. Journal of Intelligent Decision Making and Information Science, 2026, 3(S1): 840−863
47. Zhang S N, Wang H, Yi X. Exploring collaboration patterns and strategies in human-AI co-creation through the lens of agency: a scoping review of the top-tier HCI literature. Proceedings of the ACM on Human-Computer Interaction, 2025, 9(7): CSCW413
48. Yao S Y, Zhao J, Yu D, et al. ReAct: synergizing reasoning and acting in language models. In: Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda, 2023
49. Wu Q Y, Bansal G, Zhang J Y, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversations. In: Proceedings of the First Conference on Language Modeling. Philadelphia, USA, 2024
50. Zitkovich B, Yu T H, Xu S C, et al. Rt-2: vision-language-action models transfer web knowledge to robotic control. In:

Proceedings of the 7th Conference on Robot Learning. Atlanta, USA, 2023, 2165–2183

51. Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model. In: Proceedings of the 8th Annual Conference on Robot Learning. Munich, Germany, 2024, 2679–2713
52. Flesher S N, Downey J E, Weiss J M, et al. A brain-computer interface that evokes tactile sensations improves robotic arm control. Science, 2021, 372(6544): 831−836
53. Wairagkar M, Card N S, Singer-Clark T, et al. An instantaneous voice-synthesis neuroprosthesis. Nature, 2025, 644(8075): 145−152
54. Littlejohn K T, Cho C J, Liu J R, et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. Nature Neuroscience, 2025, 28(4): 902−912
55. Haghi B, Aflalo T, Kellis S, et al. Enhanced control of a brain–computer interface by tetraplegic participants via neural-network-mediated feature extraction. Nature Biomedical Engineering, 2025, 9(6): 917−934
56. Ruan Y J, Dong H H, Wang A, et al. Identifying the risks of lm agents with an LM-emulated sandbox. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria, 2024, 27031–27098
57. Debenedetti E, Zhang J, Balunovic M, et al. AgentDojo: a dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada, 2024, 2636
58. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596(7873): 583−589
59. Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery. Nature, 2023, 624(7990): 80−85
60. Szymanski N J, Rendy B, Fei Y X, et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. Nature, 2023, 624(7990): 86−91
61. DiMaio S, Hanuschik M, Kreaden U. The *da vinci* surgical system. In: Rosen J, Hannaford B, Satava R M, eds. Surgical Robotics: Systems Applications and Visions. Boston: Springer, 2010, 199–217
62. Sankai Y. HAL: hybrid assistive limb based on cybernetics. In: Kaneko M, Nakamura Y, eds. Robotics Research: The 13th International Symposium ISRR. Berlin, Heidelberg: Springer, 2010, 25–34
63. Diftler M A, Mehling J S, Abdallah M E, et al. Robonaut 2-the first humanoid robot in space. In: Proceedings of 2011 IEEE International Conference on Robotics and Automation. Shanghai, China, 2011, 2178–2183
64. Lee J Y, Lee S, Mishra A, et al. Brain–computer interface control with artificial intelligence copilots. Nature Machine Intelligence, 2025, 7(9): 1510−1523
65. Ding Y D, Udompanyawit C, Zhang Y S, et al. EEG-based brain-computer interface enables real-time robotic hand control at individual finger level. Nature Communications, 2025, 16(1): 5401
66. Willsey M S, Shah N P, Avansino D T, et al. A high-performance brain–computer interface for finger decoding and quad copter game control in an individual with paralysis. Nature Medicine, 2025, 31(1): 96−104
67. Card N S, Singer-Clark T, Peracha H, et al. Long-term independent use of an intracortical brain–computer interface for speech and cursor control. Nature Medicine, 2026, 32(7): 2504−2510
68. Al-Batati A, Koubaa A, Gabr K, et al. ROS 2 in a nutshell: a survey. ACM Computing Surveys, 2026, 58(15): 382
69. O'Neill A, Rehman A, Maddukuri A, et al. Open X-embodiment: Robotic learning datasets and RT-X models: open X-embodiment collaboration. In: Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan, 2024, 6892–6903
70. International Organization for Standardization. ISO/TS 15066: 2016 Robots and robotic devices—collaborative robots. Geneva: International Organization for Standardization, 2016
71. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 Laying down harmonised rules on artificial intelligence. Luxembourg: Official Journal of the European Union, 2024
72. Liu J X, Chen S C. TimesURL: self-supervised contrastive learning for universal time series representation learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024, 13918–13926
73. Chib P S, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: a survey. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 103−118
74. Raees M, Meijerink I, Lykourentzou I, et al. From explainable to interactive AI: a literature review on current trends in human-AI interaction. International Journal of Human-Computer Studies, 2024, 189: 103301
75. Wilson G H, Stein E A, Kamdar F, et al. Long-term unsupervised recalibration of cursor-based intracortical brain–computer interfaces using a hidden Markov model. Nature Biomedical Engineering, 2026, 10(7): 1466−1484
76. Sadeh S, Clopath C. The emergence of NeuroAI: bridging neuroscience and artificial intelligence. Nature Reviews Neuroscience, 2025, 26(10): 583−584
77. Liu Y Q, Liu G Y, Wang J C, et al. *LAMeTA*: intent-aware agentic network optimization via a large AI model-empowered two-stage approach. IEEE Journal on Selected Areas in Communications, 2026, 44: 2462−2478
78. Chiossi F, Imamaliyev E, Bleichner M, et al. Anticipation before action: EEG-based implicit intent detection for adaptive gaze interaction in mixed reality. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Barcelona, Spain, 2026, 636
79. Oren Y, Meister N, Chatterji N, et al. Proving test set

contamination in black-box language models. In: Proceedings of the 12th International Conference on Learning Representations. Vienna, Austria, 2024, 16354–16372
80. Ye H C, Gao Z Q, Ma M Y, et al. KVCOMM: online cross-context KV-cache communication for efficient LLM-based multi-agent systems. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA, 2025, 20359–20405
81. Dorna V, Mekala A, Zhao W L, et al. OpenUnlearning: accelerating LLM unlearning via unified benchmarking of methods and metrics. In: Proceedings of the 39th International Conference on Neural Information Processing Systems. San Diego, USA, 2025, 48903–48940
82. Lim H, Choi D, Nam S, et al. Understanding human–multi-agent team formation for creative work. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Barcelona, Spain, 2026, 70
83. Malki L M, Polamarasetty A, Hatamian M, et al. Hoovered up as a data point: exploring privacy behaviours, awareness, and concerns among UK users of LLM-based conversational agents. Proceedings on Privacy Enhancing Technologies, 2025, 2025(4): 838−860
84. Lara-Gutierrez A, Fernandez-Gago C, Onieva J A. A framework for drift detection and adaptation in AI-driven anomaly and threat detection systems. International Journal of Information Security, 2025, 24(5): 199
85. Zheng Y S, Hu Y P, Yu T, et al. AgentSight: system-level observability for AI agents using eBPF. In: Proceedings of the 4th Workshop on Practical Adoption Challenges of ML for Systems. Seoul, Republic of Korea, 2025, 110–115
86. International Organization for Standardization. ISO/IEC 42001: 2023 Information technology—artificial intelligence—management system. Geneva: International Organization for Standardization, 2023
87. Plociennik C, Watjanatepin P, Van Acker K, et al. Life cycle assessment of artificial intelligence applications: research gaps and opportunities. Procedia CIRP, 2025, 135: 924−929
88. Vansteensel M J, Leinders S, Branco M P, et al. Longevity of a brain–computer interface for amyotrophic lateral sclerosis. New England Journal of Medicine, 2024, 391(7): 619−626
89. Chappidi S, Cobbe J, Norval C, et al. Accountability capture: how record-keeping to support AI transparency and accountability (re)shapes algorithmic oversight. In: Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society. Madrid, Spain, 2025, 554–566
90. Abulibdeh R, Agyemang G O, Celi L A, et al. Who's really in the loop? Rethinking oversight in AI-assisted health care. The Lancet, 2026, 407(10545): 2340−2344
91. Mao Z J, Sobolewski C, Ruchkin I. How safe am I given what I see? Calibrated prediction of safety chances for image-controlled autonomy. In: Proceedings of the 6th Annual Learning for Dynamics & Control Conference. Oxford, UK, 2024, 1370–1387



Chao Shen (Fellow, *IEEE*) is currently a Professor with the School of Electronic and Information Engineering, Xi'an Jiaotong University, China, where he is also the Associate Dean with the School of Cyber Security. He is also with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security. From 2011 to 2013, he was a research scholar with Carnegie Mellon University. His research interests include network security, human–computer interaction, insider detection, and behavioral biometrics.



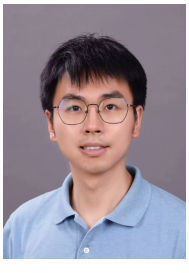
Hongxin Xie received the M.Eng. degree in communication engineering from Shandong Normal University in 2025. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Xi'an Jiaotong University. His current research interests include adversarial machine learning and multi-agent security.



Le Yang (Associate Member, *IEEE*) received the B.S. degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2015, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2021. He is currently a Special Research Fellow with Xi'an Jiaotong University. His research interests include trustworthy AI and computer vision, particularly AI security, multimodal large language models, and dynamic neural networks.



Chenhao Lin (Senior Member, *IEEE*) received the B.Eng. degree in automation from Xi'an Jiaotong University in 2011, the M.Sc. degree in electrical engineering from Columbia University in 2013, and the Ph.D. degree from The Hong Kong Polytechnic University in 2018. He is currently a Professor with Xi'an Jiaotong University. His research interests include artificial intelligence security, identity authentication, biometrics, adversarial attack and robustness, and pattern recognition.



Bin Shi received the Bachelor's degree in 2013 and the Ph.D. degree in 2019 in Computer Software and Theory from Beihang University. He currently serves as a Professor in SPKLBDKE Lab, School of Computer Science and Technology, Xi'an Jiaotong University. His research interests include data mining and domain specific large language model.



Bo Dong received the Ph.D. degree in Computer Science and Technology from Xi' an Jiaotong University in 2014. He is currently a professor in Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University. His research interests include large language model and machine learning.



Qinghua Zheng received the B.S. degree in Computer Software in 1990, the M.S. degree in computer organization and architecture in 1993, and the Ph.D. degree in System Engineering in 1997 from Xi'an Jiaotong University. His research interests include big data knowledge engineering and memory-inspired machine intelligence.