

Governable agentic intelligence: a lifecycle framework linking capability, assurance, and oversight in large language model agents

Harry Udoh

hudoh@arden.ac.uk

Arden University

Research Article

Keywords: Agentic AI, Large language model agents, Agent evaluation, AI safety, Human oversight, Trustworthy AI

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11168127/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Agentic artificial intelligence (AI) embeds large language models (LLMs) in action loops in which they plan, retrieve evidence, invoke tools, maintain memory, coordinate with other agents, and change external state. Reviews of this field have proliferated, yet they remain fragmented: architecture surveys catalogue components, evaluation surveys catalogue benchmarks, and safety surveys catalogue risks, while deployed agents couple all three. Scholars have yet to sufficiently examine how capability, evaluation, and governance interact within the lifecycle of a single agentic system. We address this gap through a critical review that codes 19 reviews and surveys published in *Artificial Intelligence Review* between 2024 and 2026 and situates them within foundational primary studies and cross-venue surveys. We argue that agency is a configurable property of a socio-technical control loop, not an attribute of a model. From the corpus, we develop the Agentic Intelligence Lifecycle (AIL), a six-layer framework spanning foundation, cognition, grounding, orchestration, evaluation, and governance. Our central finding is that every mechanism that expands capability also incurs an *assurance debt*: a gap between the risks the mechanism introduces and the controls that keep those risks observable and bounded. We theorize this relationship in five propositions, propose a five-question assurance model that shifts evaluation from outputs to trajectories and deployment conditions, and derive a ten-point research agenda for agentic systems that are observable, attributable, interruptible, and governable.

1 Introduction

Large language models (LLMs) have shifted the center of gravity of artificial intelligence (AI) from narrowly specified prediction toward general-purpose, language-mediated computation. Reviews published in *Artificial Intelligence Review* (AIR) trace this progression from technical frameworks for LLMs to multimodal, retrieval-augmented, tool-using, and increasingly autonomous systems (Kumar 2024; Li et al. 2025; Wu et al. 2026; Waqas et al. 2026), and they record how rapidly agentic AI has moved from research prototype onto industrial and entrepreneurial agendas (Patel et al. 2026). The technical foundations of this shift are well documented beyond the journal: prompting models to produce intermediate reasoning (Wei et al. 2022), interleaving reasoning with actions (Yao et al. 2023), teaching models to call external tools (Schick et al. 2023), and equipping agents with memory, reflection, and planning (Park et al. 2023; Shinn et al. 2023). Cross-venue surveys have consolidated these advances into general architectures for LLM-based agents (Wang et al. 2024; Xi et al. 2025). The most consequential transition, however, is not larger models or better generation. It is the embedding of models in closed or semi-closed action loops in which they interpret goals, decompose tasks, invoke external resources, maintain memory, evaluate progress, coordinate roles, and revise subsequent actions. In such loops, the model becomes one component of a wider agentic system rather than the complete unit of analysis.

The review literature has responded with unusual speed. Within 2025 and 2026 alone, AIR published surveys of agentic architectures (Abou Ali et al. 2026), LLMs as autonomous agents and tool users (Chowa et al. 2026), agent evaluation and benchmarking (Kehkashan et al. 2026; Nageshwaran et al.

2026), LLM-based judging (Nadăș 2026), reward-guided reasoning (Liu et al. 2026), agentic safety (Gohil et al. 2026), LLM safety (Chen et al. 2026), privacy (Elsheikh et al. 2026), toxicity (Villate-Castillo et al. 2026), factuality (Rahman et al. 2026), and trust in human–autonomous teams (Zamecnik et al. 2026). This density creates a second-order problem: the literature is rich in taxonomies but fragmented across levels of analysis. Architecture surveys emphasize components and workflows; evaluation surveys emphasize benchmarks and metrics; safety, privacy, and explainability surveys emphasize cross-cutting requirements. Deployed agents, by contrast, couple all of these concerns. A retrieval decision can alter factuality; a memory mechanism can alter privacy exposure; a planner can amplify an erroneous premise across many steps; a tool call can convert a harmless textual error into an external action; and a multi-agent architecture can distribute capability and accountability at the same time.

Scholars have yet to sufficiently examine how capability, evaluation, and governance interact within the lifecycle of a single agentic system. Existing reviews typically treat one of these as the object of analysis and the others as context. This matters because the failures that most concern deployers—an unsupported claim that triggers an irreversible tool call, a memory that carries sensitive data across sessions, a multi-agent consensus that merely echoes a shared error—arise precisely at the interfaces between layers that separate reviews examine in isolation. Scholars of algorithmic harm have argued that increasingly agentic systems pose distinctive risks precisely because they act with less direct human specification (Chan et al. 2023), yet this insight has not been integrated with the architectural and evaluation literatures. It also matters for measurement: AIR’s evaluation literature documents a persistent disconnect between benchmark performance and deployment viability (Kehkashan et al. 2026), yet the field lacks a theoretical account of *why* capability gains so often fail to translate into trustworthy operation.

We argue that agency is a *configurable property of a socio-technical control loop*, not an attribute of a model. On this view, the relevant question is not whether a model “is” agentic, but how much observation, decision, and action authority a system is granted, under which controls, and with what recoverability. This reframing moves the analytical focus from what agents can do to what must be true for their doing to remain governable. Given these theoretical conundrums, we ask:

RQ1 How does the AIR literature characterize the transition from generative models to agentic AI systems?

RQ2 Which architectural and cognitive mechanisms enable reasoning, planning, memory, retrieval, tool use, and multi-agent coordination, and which failure pathways do they open?

RQ3 How should agentic systems be evaluated when task completion is insufficient to represent deployment quality?

RQ4 Which safety, privacy, factuality, and alignment concerns intensify as autonomy increases?

RQ5 How can human and agent oversight preserve useful autonomy while maintaining accountability and control?

RQ6 Which research directions are most important for governable agentic intelligence?

We make three important contributions. First, we develop the *Agentic Intelligence Lifecycle* (AIL), a six-layer framework that integrates the fragmented AIR literature on foundations, cognition, grounding, orchestration, evaluation, and governance into a single system-level view (RQ1, RQ2). Unlike prior taxonomies, which organize the field by component or by risk, the AIL organizes it by the assurance obligations each layer creates for the others. Second, we introduce the concept of *assurance debt*—the gap between the risk exposure a capability mechanism introduces and the control coverage that keeps that exposure observable and bounded—and theorize its dynamics in five propositions (RQ2, RQ4). Assurance debt offers a parsimonious explanation for why benchmark gains coexist with disappointing deployment reliability. Third, we derive and conceptualize a *deployment-oriented assurance model*—five questions spanning observation, decision, action, evaluation, and recovery—and translate it into nested evaluation targets, an authority-allocation scheme for human oversight, and a ten-point research agenda (RQ3, RQ5, RQ6).

The remainder of the article proceeds as follows. Section 2 describes the review methodology and evidence boundary. Section 3 formalizes the transition from generative to agentic AI. Section 4 presents the AIL framework layer by layer. Section 5 develops the concept of assurance debt and our propositions. Sections 6 to 9 translate the synthesis into an assurance model, an evaluation architecture, a safety and privacy analysis, and a scheme for human–agent authority allocation. Section 10 sets out the research agenda, Section 11 draws implications, and Sections 12 and 13 address limitations and conclude.

2 Review methodology

2.1 Review type and rationale

We adopt a structured critical-review design with an explicitly bounded evidence base. Because nearly all corpus articles are themselves reviews or surveys, the study is a second-order synthesis: its unit of analysis is not an individual model, benchmark, or system, but a recurring concept or design problem as it appears across the review literature. Critical reviews of this kind aim to integrate and reinterpret an established body of work and to generate new conceptual frameworks, rather than to estimate effect sizes or exhaustively enumerate primary studies (Grant and Booth 2009; Snyder 2019). The design is appropriate for our purpose—building an integrative lifecycle theory—because review articles already encode explicit taxonomies, comparative structures, research gaps, and future directions that can be cross-read and recombined.

We distinguish two bodies of literature. The *coded corpus*—the evidence from which the framework is inductively derived—is bounded by venue and consists of articles published in AIR. The *contextual*

literature—foundational primary studies, cross-venue surveys, and governance and human-factors sources—is not coded, but is used to ground the constructs we derive, to check that the corpus’s claims are consistent with the primary evidence, and to connect our propositions to established theory. We adopt the venue boundary for the coded corpus deliberately. AIR is a leading venue dedicated to review and survey articles in AI, and its recent output on LLMs and agents is large enough to support a substantive synthesis while remaining small enough to be read in full and coded consistently. A venue-bounded corpus therefore allows a coherent, internally comparable account of how one field-shaping review venue has organized the transition from generative to agentic intelligence. We do not claim that it substitutes for a field-wide systematic review; Section 12 discusses the consequences of this choice. As review methodology itself is being reshaped by AI tooling (Bolaños et al. 2024), we document our procedures so that the synthesis can be audited and extended, and we report search and screening in line with the PRISMA 2020 items applicable to a non-systematic review (Page et al. 2021).

2.2 Search strategy

We searched Springer Nature Link, restricting results to the journal *Artificial Intelligence Review* and to articles published from January 2024 to [final search date]. We combined terms describing the object of study with terms describing capabilities and assurance concerns: (“large language model” OR LLM OR “generative AI” OR “agentic AI” OR “autonomous agent” OR “multi-agent”) AND (“tool use” OR “retrieval-augmented” OR reasoning OR planning OR memory OR “reward model” OR evaluation OR benchmark OR “LLM-as-a-judge” OR factuality OR hallucination OR safety OR privacy OR toxicity OR explainab* OR multilingual OR multimodal OR oversight OR trust). We supplemented database searching with backward citation checking within AIR to identify reviews that the search strings did not retrieve.

2.3 Eligibility criteria

We included articles that (a) were published in AIR; (b) were reviews, surveys, systematic reviews, or scoping reviews; (c) addressed LLMs, LLM-based agents, or a cross-cutting requirement—evaluation, factuality, privacy, safety, toxicity, explainability, or trust—with direct relevance to LLM-based systems; and (d) had a verifiable DOI and publication record. We excluded primary empirical studies that did not synthesize a literature, application-domain reviews in which LLMs or agents were peripheral, and editorials or commentaries. Where overlapping reviews addressed the same topic, we retained all that met the criteria, because convergence and divergence across reviews are themselves analytically informative.

2.4 Screening, verification, and corpus

The search returned [n] records. After title and abstract screening, [n] articles were assessed in full text, and 19 met all criteria. For every included article, we verified title, authorship, journal membership, volume and article number (where assigned), and DOI against Springer Nature Link records. One additional AIR article on AI-assisted literature review (Bolaños et al. 2024) informs our methodological reflection but is not coded as part of the substantive corpus. Table 1 summarizes the corpus. Eighteen

of the 19 articles were published in 2025 or 2026, confirming that the agentic turn in the review literature is recent and fast-moving.

Table 1

Corpus of reviews in *Artificial Intelligence Review* and their coverage of the AIL layers (● primary focus; ○ secondary coverage). L1 foundation; L2 cognition; L3 grounding; L4 orchestration; L5 evaluation; L6 governance

Article	Type	Primary focus	L1	L2	L3	L4	L5	L6
Kumar (2024)	Survey	LLM architectures, technical frameworks, and challenges	●					
Li et al. (2025)	Systematic review	Multimodal LLMs in domain-specific applications	●				○	
Mindlin et al. (2025)	Systematic literature review	Dialogue-based explainable AI						●
Abou Ali et al. (2026)	Survey	Agentic AI architectures, applications, and directions	○	○	○	●		
Chowa et al. (2026)	Review	LLMs as autonomous agents and tool users		●	●	●		
Patel et al. (2026)	Systematic review	Generative and agentic AI: industry, startups, and ethics	○			○		○
Liu et al. (2026)	Analytical survey	Reward models for LLM reasoning		●			○	
Wu et al. (2026)	Survey	Retrieval-augmented generation			●			
Rahman et al. (2026)	Review	Hallucination, fact-checking, and factuality evaluation			○		●	
Kehkashan et al. (2026)	Comprehensive review	Agentic AI evaluation from benchmarks to deployment					●	○
Nageshwaran et al. (2026)	Systematic survey	LLM-agent evaluation and benchmarking meta-taxonomy					●	
Nadăş (2026)	Review	LLMs as judges: evaluation, critique, and feedback					●	
Gohil et al. (2026)	Survey	Safe and trustworthy agentic AI				○		●
Chen et al. (2026)	Survey	AI safety landscape for LLMs						●

Article	Type	Primary focus	L1	L2	L3	L4	L5	L6
Elsheikh et al. (2026)	Systematic review	Privacy preservation in foundation models		○				●
Villate-Castillo et al. (2026)	Systematic review	Toxicity in LLMs: detection and detoxification					○	●
Huang et al. (2026)	Survey	Multilingual LLMs	●					
Waqas et al. (2026)	Review	LLM architectures and cross-domain impact	●					
Zamecnik et al. (2026)	Scoping review	Agent architectures and trust in human–autonomous teams					○	●

2.5 Analytical coding and synthesis

We coded the corpus in two passes. In a deductive first pass, we assigned each article’s principal contributions to six provisional categories derived from the lifecycle of an LLM-based system: foundations; reasoning and cognition; grounding and external interaction; orchestration and autonomy; evaluation and assurance; and governance and oversight. In an inductive second pass, we recorded, for each mechanism an article discussed, the capabilities it enabled, the failure pathways it opened, and the controls it proposed. We treated a concept as *cross-cutting* when it affected at least two layers—for example, retrieval affects grounding but also factuality and evaluation, while memory affects cognition, privacy, personalization, and auditability. Iterative comparison of capability–failure–control triples across articles yielded the assurance-debt construct (Section 5). The resulting framework is interpretive and integrative: it reorganizes published AIR insights into a lifecycle model rather than reproducing any single prior taxonomy. In a final step, we compared each derived construct with the contextual literature, retaining a construct only where the primary evidence was consistent with it and citing that evidence alongside the corpus.

2.6 Use of generative AI tools

[Required by the journal: describe here any use of large language models in preparing this manuscript beyond AI-assisted copy editing—for example, the tool and version used, the tasks it supported (drafting, restructuring, or summarizing), and how the author verified all content, analysis, and citations. State that the author takes full responsibility for the final text.]

3 From generative AI to agentic intelligence

3.1 A functional distinction

A generative model maps an input to an output that a human then interprets, accepts, or discards. An agentic system closes that loop: it converts outputs into actions, observes the consequences of those actions, treats them as new evidence, and repeats the cycle until a goal is reached or control is withdrawn. AIR’s agentic reviews converge on this functional distinction, characterizing modern agents by their capacity for multi-step reasoning, planning, memory, tool invocation, and interaction with an environment (Abou Ali et al. 2026; Chowa et al. 2026; Gohil et al. 2026). Cross-venue surveys reach a similar conclusion, describing LLM-based agents in terms of a model-centered core extended by perception, memory, planning, and action components (Wang et al. 2024; Xi et al. 2025). To make the distinction precise, we define an agentic system as a tuple

$$\mathcal{A} = \langle M, \Omega, K, T, \Pi, E, H \rangle \quad (1)$$

where M is the foundation model or models; Ω the observation channels through which information enters the loop (user inputs, retrieval, sensors, inter-agent messages); K the memory and knowledge stores; T the tool set together with its permissions; Π the orchestration policy governing planning, delegation, and role assignment; E the evaluators that score or select intermediate and final states (reward models, judges, deterministic checks); and H the human oversight controls. At each step t , the system draws an observation o_t from Ω and K , forms a decision d_t under Π using M , executes an action a_t via T , and updates K , subject to E and H . The object that matters for deployment is therefore not an individual output but the *trajectory* $\tau = (o_1, d_1, a_1, \dots, o_t, d_t, a_t)$.

This distinction has major consequences for risk. A passive model generally produces an output for human judgment; an agent can transform that output into an action before any human sees it. The system’s effective behavior emerges from the interaction of the base model, prompts, memory, retrieval, tools, orchestration logic, permissions, environment, evaluators, and human controls. Agentic intelligence is therefore better understood as a socio-technical control loop than as a property of a single neural network.

3.2 Agency as a configurable property

The AIR literature warns against collapsing historical agent concepts and modern LLM agents into a single undifferentiated category. Abou Ali et al. (2026) distinguish contemporary neural agentic systems from earlier symbolic traditions while recognizing that planning, goals, beliefs, and coordination remain useful conceptual anchors. Gohil et al. (2026) situate modern agency across autonomy, cognitive capability, modality, and environmental interaction. Human-factors research has long made the same point about automation, distinguishing the types of functions automated from the levels at which they are automated (Parasuraman et al. 2000). Building on these perspectives, we treat autonomy not as a binary label but as a vector α of configurable dimensions:

$$\alpha = (s, h, r, \iota, u, \rho) \quad (2)$$

where s denotes action scope, h time horizon, r reversibility of actions, ι external impact (financial, physical, legal, or reputational), u uncertainty about user intent and world state, and ρ human

recoverability—the ease with which a human can detect, interrupt, and correct the system. An agent that drafts a report for review and an agent that executes financial transactions may share a foundation model and an orchestration framework, yet occupy very different points in this space. Eq. (2) makes explicit that agency is expressed through system-level affordances granted by designers and deployers, rather than inferred from fluent language generation. It also implies that agency can be *dialed*: the same system can be made more or less autonomous by changing permissions, horizons, and escalation rules without changing the model.

3.3 Positioning relative to prior reviews

Table 1 shows how the corpus distributes across the AIL layers. Three observations position our contribution. First, foundation-level reviews (Kumar 2024; Li et al. 2025; Huang et al. 2026; Waqas et al. 2026) characterize capability but rarely extend to the control problems that capability creates once embedded in an action loop. Second, the agentic reviews (Abou Ali et al. 2026; Chowa et al. 2026) cover cognition, grounding, and orchestration in depth, while safety-oriented reviews (Gohil et al. 2026; Chen et al. 2026) cover governance; few articles span both. Third, evaluation reviews (Kehkashan et al. 2026; Nageshwaran et al. 2026; Nadăș 2026) are the most explicit about the gap between measured and deployed performance, but they do not theorize that gap in terms of the architectural mechanisms that produce it. The AIL framework connects these three bodies of work.

4 The Agentic Intelligence Lifecycle framework

The AIL framework integrates the corpus into six connected layers (Fig. 1; Table 2). The layers are analytical rather than strictly sequential: deployed systems revisit them repeatedly, and controls at one layer can constrain behavior at another. Two properties distinguish the framework from existing taxonomies. First, each layer is defined by an *assurance question* as well as a function, so that capability and control are specified together. Second, the framework makes explicit that capability expands upward—each layer increases the action surface of the system—whereas controls must propagate downward, constraining what lower layers are permitted to contribute to consequential actions.

Table 2
Layers of the Agentic Intelligence Lifecycle (AIL) framework

AIL layer	Primary function	Representative mechanisms	Primary assurance question
L1 Foundation and representation	Provide model capability and representations	LLMs; multimodal and multilingual models; domain models	What capability and limitation profile does the foundation introduce in the deployment context?
L2 Cognition and reasoning	Transform goals into intermediate decisions	Reasoning; planning; reflection; reward models; memory	Are intermediate decisions coherent, verifiable, and calibrated?
L3 Grounding and tool interaction	Connect model states to external information and actions	Retrieval-augmented generation (RAG); search; application programming interfaces (APIs); tools; external memory	Are evidence and actions authorized, attributable, and provenance-aware?
L4 Orchestration and autonomy	Coordinate sequential and multi-agent behavior	Single- and multi-agent workflows; role assignment; delegation	How much autonomy is granted, and where can control be recovered?
L5 Evaluation and assurance	Measure quality across outputs and trajectories	Benchmarks; LLM judges; factuality checks; deployment metrics	Does evaluation predict real-world reliability and failure cost?
L6 Governance and oversight	Align operation with human and institutional requirements	Human and agent oversight; privacy; explainability; safety controls	Who is accountable, what is auditable, and when must humans intervene?

4.1 Layer 1: foundation and representation

Agent behavior inherits both strengths and weaknesses from its foundation model. AIR reviews describe the evolution of LLM architectures, training frameworks, and application breadth, while multimodal and multilingual surveys show that representation quality varies across modalities, languages, and domains (Kumar 2024; Li et al. 2025; Huang et al. 2026; Waqas et al. 2026). For agent design, foundation-model selection is therefore not a neutral implementation choice. It sets the ceiling for instruction following, reasoning, language coverage, perception, and tool-interface competence, while shaping cost, latency, privacy, and reproducibility.

Multimodality expands an agent’s perceptual surface, enabling it to integrate text, images, audio, and other signals. Yet it also expands the verification problem, because evidence can be inconsistent across modalities and domain-specific performance can differ substantially (Li et al. 2025). Multilingual capability introduces a parallel concern: an agent that appears robust in a high-resource language may display different reasoning, retrieval, or safety behavior in other languages (Huang et al. 2026). The assurance implication is that model-level capability claims should be *conditioned* on the actual

languages, modalities, and domains of deployment rather than reported as unconditional properties of the model.

4.2 Layer 2: cognition, reasoning, planning, and memory

Agentic systems require mechanisms that convert broad goals into executable sequences. Chowa et al. (2026) organize this problem around reasoning, planning, memory, self-reflection, prompting, and fine-tuning, building on primary work on elicited intermediate reasoning (Wei et al. 2022) and verbal self-reflection (Shinn et al. 2023). Reward models add a further mechanism by supplying signals that guide candidate selection, iterative improvement, or reinforcement-learning-based optimization (Liu et al. 2026). The resulting cognition stack is heterogeneous: some decisions are generated directly by the LLM, some are constrained by explicit workflows, some are retrieved from memory, and some are selected by evaluators.

The critical issue is not whether an agent can produce a plausible chain of intermediate decisions, but whether those decisions remain anchored to evidence and objectives. Sequential reasoning creates *compounding-error risk*: a weak early assumption can propagate through planning and tool use. Reflection can improve performance, but a self-evaluation loop is not independent verification; primary evidence indicates that LLMs struggle to correct their own reasoning without external feedback (Huang et al. 2024). Reward-guided selection, including process-level supervision of intermediate steps (Lightman et al. 2024), can strengthen reasoning while introducing dependence on the quality and generalization of the reward signal (Liu et al. 2026). For high-impact deployments, cognition therefore requires *externalized checkpoints*: intermediate states should be inspectable, provenance should be retained, and irreversible actions should be separated from speculative reasoning.

Memory is similarly double-edged. Persistent memory supports continuity, personalization, and long-horizon tasks, but it creates risks of stale information, contamination, privacy leakage, and self-reinforcing feedback loops. The privacy literature on foundation models demonstrates that sensitive information can be exposed at multiple stages of the model lifecycle (Elsheikh et al. 2026), including through extraction of memorized training data (Carlini et al. 2021). Agent memory extends that concern from training data to *operational* data: what the system remembers, for how long, and which future actions it may influence become governance questions as much as performance questions.

4.3 Layer 3: grounding, retrieval, and tool interaction

RAG is the central bridge between parametric model knowledge and external evidence (Lewis et al. 2020). Wu et al. (2026) synthesize RAG architectures, retrieval components, and application patterns, while Rahman et al. (2026) connect retrieval to factuality and fact-checking, a problem the broader literature frames as hallucination (Ji et al. 2023). In agentic settings, retrieval is not merely a method for improving an answer: it determines which evidence enters a plan, which tool is selected, and which action is subsequently executed.

Grounding therefore requires provenance-aware design. A robust agent should preserve the chain linking claims, retrieved evidence, transformations, and actions. Retrieval quality should be evaluated not only for relevance but for source suitability, freshness, coverage, and susceptibility to adversarial or misleading content. The threat is concrete: instructions embedded in retrieved content can hijack LLM-integrated applications through indirect prompt injection (Greshake et al. 2023). The same principle applies to tools. Chowa et al. (2026) identify tool integration as a defining feature of modern LLM agents, a capability that primary work has shown can be learned by the model itself (Schick et al. 2023) and interleaved with reasoning (Yao et al. 2023). Tool access increases competence but converts language-model errors into operational risk. Permission boundaries, typed interfaces, argument validation, sandboxing, rate limits, confirmation steps, and action logs should therefore be treated as core architectural components rather than post-deployment safeguards.

We propose *minimum necessary agency* as a design principle for this layer: an agent should receive the narrowest tool permissions and action scope needed for its task. The principle is analogous to least privilege in security engineering, but the agentic context adds a second dimension of uncertainty—the interpretation of intent. A system must therefore control both what an agent is technically permitted to do and how confidently it should infer that a user intended a given action before doing it.

4.4 Layer 4: orchestration, multi-agent systems, and autonomy

AIR reviews distinguish single-agent from multi-agent systems and document the growing use of role specialization, communication, collaboration, and delegation (Abou Ali et al. 2026; Chowa et al. 2026). Multi-agent designs can decompose complex tasks, introduce critique roles—as in multi-agent debate, which has been reported to improve factuality and reasoning (Du et al. 2024)—and parallelize work. They can also create coordination overhead, duplicated errors, hidden dependencies, and ambiguous responsibility. Crucially, more agents do not imply more independent evidence when all agents share similar base models, prompts, or information sources: agreement among correlated agents may reflect a shared bias rather than a verified conclusion.

Autonomy should consequently be treated as the configurable vector of Eq. (2) rather than a binary label. An agent that autonomously drafts a report has a different risk profile from one that executes transactions or modifies production systems. Gohil et al. (2026) emphasize the need to connect agentic capability to safety and trustworthiness; the AIL framework extends this by making *autonomy calibration* a lifecycle control. Higher-impact, less reversible actions should require stronger evidence, narrower permissions, more robust evaluation, and clearer escalation paths.

4.5 Layer 5: evaluation and assurance

Evaluation is the layer at which agentic research remains conceptually weakest. Widely used agent benchmarks measure success on interactive tasks in operating-system, web, software-engineering, and tool–user settings (Liu et al. 2024; Zhou et al. 2024; Jimenez et al. 2024; Yao et al. 2025), but they are

predominantly framed around task completion. Conventional model evaluation asks whether an output is correct; agent evaluation must ask whether a trajectory was appropriate, efficient, safe, reproducible, and recoverable. Kehkashan et al. (2026) identify a disconnect between benchmark performance and deployment viability, noting that task-completion metrics can obscure cost, safety compliance, maintainability, and workflow integration. Nageshwaran et al. (2026) systematize LLM-agent evaluation through a meta-taxonomy and research roadmap; outside the journal, Kapoor et al. (2024) similarly argue that agent evaluations should jointly account for accuracy and cost; and Nadăș (2026) documents both the promise of LLM-as-a-judge approaches for scalable critique and feedback and systematic concerns about judge bias and reliability. Section 7 develops the implications in detail; here we note that evaluation occupies a distinctive position in the framework. It is the layer through which the assurance obligations of all other layers become visible—or remain hidden.

4.6 Layer 6: governance, safety, and human oversight

As agentic systems gain autonomy, governance becomes an architectural requirement rather than an external constraint. AIR’s safety literature organizes risks across trustworthy, responsible, and ecosystemic perspectives, while agentic reviews emphasize alignment, security, reliability, and human control (Chen et al. 2026; Gohil et al. 2026). Toxicity, privacy, and explainability are not peripheral ethical themes: they determine whether an agent can be safely delegated real tasks (Villate-Castillo et al. 2026; Elsheikh et al. 2026; Mindlin et al. 2025). Governance also has a relational dimension. Zamecnik et al. (2026) link agent architectures to the antecedents of trust in human–autonomous teams, underscoring that design choices at lower layers shape whether human oversight is well calibrated or misplaced. Sections 8 and 9 elaborate these themes.

5 Cross-cutting synthesis: capability creates assurance debt

5.1 The assurance-debt pattern

Across the corpus, a recurring pattern emerges: *each mechanism that expands capability also creates an assurance debt*. Retrieval expands knowledge access but introduces provenance and freshness problems. Memory extends temporal continuity but introduces privacy and contamination risks. Tool use expands action capability but creates permission and side-effect risks. Multi-agent coordination enables decomposition and critique but complicates attribution and the validity of consensus. Reward models can improve reasoning but create dependence on reward quality. Automated judges scale evaluation but can reproduce systematic biases. Multimodality broadens perception but multiplies failure channels. Table 3 consolidates these capability–debt–control triples.

Table 3
Capability mechanisms and the assurance debt they incur

Mechanism	Capability gained	Assurance debt incurred	Required controls	Layers
Retrieval (RAG)	Access to current and external knowledge	Provenance loss; stale or adversarial evidence	Source vetting; provenance chains; freshness checks	L3, L5
Persistent memory	Continuity; personalization; long horizons	Privacy leakage; contamination; feedback loops	Retention and deletion policies; provenance tags	L2, L6
Tool use	Ability to change external state	Unauthorized or irreversible side effects	Least privilege; sandboxing; confirmation; rollback	L3, L4
Multi-agent coordination	Decomposition; parallelism; critique	Diffuse accountability; correlated errors	Responsibility boundaries; evidence independence	L4, L6
Reward models	Improved reasoning and selection	Dependence on reward quality and generalization	Independent verification; reward auditing	L2, L5
Automated judges	Scalable evaluation and feedback	Systematic judge bias; false assurance	Judge calibration; evaluator diversity; human audit	L5
Multimodality and multilingualism	Broader perception and reach	More failure channels; uneven safety behavior	Cross-modal verification; localized benchmarks	L1, L5

We define assurance debt, by analogy with the hidden technical debt that accumulates in machine learning systems (Sculley et al. 2015), as the accumulated shortfall between the risk exposure introduced by a system’s capability mechanisms and the control coverage that renders that exposure observable and bounded. Heuristically, for a set of mechanisms $m \in \mathcal{M}$,

$$D(\mathcal{A}) = \sum m \max(0, R_m - C_m) \quad (3)$$

where R_m denotes the risk exposure introduced by mechanism m in a given deployment context and C_m the coverage of the controls applied to it. We do not claim that these quantities are directly measurable today; Eq. (3) is a conceptual device that makes two points explicit. First, debt is incurred mechanism by mechanism, so a strong control at one layer does not offset an uncontrolled exposure at another. Second, debt depends on context: the same retrieval pipeline carries little debt when drafting an internal summary and substantial debt when its outputs drive external actions. This framing explains why benchmark gains can coexist with disappointing real-world reliability: benchmark tasks typically reward capability while externalizing assurance costs to the deployment environment.

5.2 Propositions

The synthesis supports five propositions that specify how assurance debt accumulates and where it is realized. Each is derived from convergent observations across the corpus and is stated in a form amenable to empirical testing.

Action authority determines whether upstream errors remain textual or become operational. Tool integration makes language-model errors consequential (Chowa et al. 2026), agentic safety requires linking capability to control (Gohil et al. 2026), and harms from algorithmic systems scale with the degree of agency they exercise (Chan et al. 2023). Therefore,

Proposition 1

The higher the action authority granted to an agentic system—in scope, impact, and irreversibility—the more likely are reasoning and grounding errors to become consequential external harms, unless permissioning, reversibility, and recovery controls scale proportionally.

Multi-agent designs are frequently justified by the value of critique and cross-checking (Abou Ali et al. 2026; Chowa et al. 2026). Yet agents built on similar models, prompts, and sources share failure modes, and LLM judges exhibit systematic biases (Nadăș 2026), including a preference for their own generations (Panickssery et al. 2024). We theorize:

Proposition 2

The higher the correlation among agents' base models, prompts, and information sources, the less likely is multi-agent agreement to constitute independent evidence of correctness.

Evaluation reviews show that task-completion benchmarks underrepresent cost, safety, and workflow integration (Kehkashan et al. 2026; Nageshwaran et al. 2026). Therefore,

Proposition 3

The higher the share of an evaluation that rests on output-level task completion, the more likely is measured benchmark performance to overestimate deployment reliability.

Privacy risks arise across the model lifecycle (Elsheikh et al. 2026), and agent memory extends them into operational data that persist across tasks. We theorize:

Proposition 4

The higher the persistence and cross-context reach of agent memory, the more likely are privacy exposure and error contamination to propagate across tasks, unless provenance and retention controls are enforced at each interface.

Finally, oversight quality depends on the calibration of human trust and on the salience of consequential decisions (Zamecnik et al. 2026; Mindlin et al. 2025). Human-factors research shows that operators of highly reliable automation tend toward complacency and misuse (Parasuraman and Riley 1997), and uniform confirmation requirements further dilute the salience of the decisions that matter. Therefore,

Proposition 5

The higher the frequency of low-consequence approval requests presented to human overseers, the more likely are those overseers to approve high-consequence actions without meaningful scrutiny.

5.3 From propositions to architectural co-design

The propositions converge on a practical implication: *architectural co-design*. Safety, evaluation, and governance should be specified alongside planning, memory, retrieval, and tool use rather than retrofitted. A tool interface should define not only functionality but authorization, logging, reversibility, and error handling. A memory component should define not only retrieval accuracy but retention, deletion, provenance, and conflict resolution. A multi-agent workflow should define not only roles but responsibility boundaries and evidence independence. Evaluation should test these controls under realistic failure conditions. Agentic progress, on this view, should be measured not solely as increasing task success but as the ratio between useful capability and unmanaged assurance debt: a system becomes deployment-ready when additional capability is accompanied by controls that keep residual risk observable and bounded.

6 A deployment-oriented assurance model

To operationalize the synthesis, we propose a five-question assurance model to be applied to each agentic task: (1) What can the agent observe? (2) What can it decide? (3) What can it change? (4) How is its behavior evaluated? (5) How can control be recovered? The questions map onto the components of Eq. (1)—observation channels and memory, orchestration policy, tools, evaluators, and human controls—and onto the AIL layers. Their purpose is to prevent assurance from collapsing into final-answer accuracy. Table 4 specifies, for each dimension, the evidence an assurance case should contain, the typical failure it guards against, and the corresponding design response.

Table 4
A deployment-oriented assurance model for agentic tasks

Assurance dimension	Core question	Example evidence	Typical failure	Design response
Observation	What information can enter the agent loop?	Retrieval logs; source provenance; input scope	Poisoned, stale, irrelevant, or private context	Source controls; provenance; filtering; freshness checks
Decision	How are goals decomposed and choices made?	Plans; intermediate states; reward and evaluator traces	Compounding reasoning error; objective drift	Checkpoints; constraint validation; uncertainty gates
Action	What external state can the agent alter?	Tool permissions; action logs; reversibility	Unauthorized or irreversible side effects	Least privilege; sandboxing; confirmations; rollback
Evaluation	How are success and failure measured?	Task and trajectory metrics; judge agreement	Benchmark overfitting; biased automated judges	Multi-method evaluation; human audit; deployment tests
Recovery	How is autonomy interrupted or corrected?	Escalation rules; stop controls; incident logs	Late detection; no accountable owner	Interruptibility; escalation; incident response; post-action review

The model is deliberately task-level rather than model-level. Two deployments of the same agent can yield different answers to all five questions, and an assurance case built for one cannot be transferred to the other without re-examination. The model also makes assurance debt inspectable: any question that cannot be answered with concrete evidence marks a location where debt has accumulated.

7 Evaluation beyond task completion

The evaluation literature in AIR supports a move toward multidimensional assessment. Agent benchmarks should measure not only whether a task was completed but how the agent reached the outcome and whether that path would remain acceptable under deployment constraints (Kehkashan et al. 2026; Nageshwaran et al. 2026). This requires explicit metrics for tool efficiency, unnecessary actions, constraint violations, recovery from errors, evidence use, cost, latency, and human intervention.

We propose four nested evaluation targets. *Output quality* concerns the final answer or action. *Trajectory quality* concerns the sequence of reasoning, retrieval, tool calls, and revisions. *System quality* concerns robustness, privacy, safety, cost, latency, and observability. *Deployment quality* concerns whether the system performs acceptably under real workflows, user behavior, distribution shift, and organizational controls. The targets are nested because each presupposes the one before it but is not implied by it. A

benchmark that measures only the first target should therefore not be used as evidence for the fourth—the precise inference that Proposition 3 predicts will be misleading.

Trajectory evaluation is especially important because two agents can reach the same correct outcome through very different risk profiles. One may retrieve authoritative evidence, make a minimal set of reversible tool calls, and preserve a clear audit trail. Another may reach the same result after speculative reasoning, repeated tool calls, exposure of unnecessary data, and accidental actions that happened to be recoverable. Output-only evaluation treats these systems as equivalent; deployment-oriented evaluation should not.

Automated judges can reduce evaluation cost, and strong LLM judges have been shown to approximate human preferences on open-ended tasks (Zheng et al. 2023), but judge calibration should itself be measured. Nadăș (2026) shows that LLM judging can be affected by systematic biases, and primary evidence documents self-preference when a model evaluates its own outputs (Panickssery et al. 2024). High-stakes evaluation should therefore triangulate among deterministic checks, reference-based measures where available, multiple and diverse evaluators, adversarial tests, and human review. Judge disagreement should be treated as information about uncertainty rather than resolved automatically by majority vote—a direct application of Proposition 2 to the evaluation layer.

Factuality should likewise be integrated into trajectory assessment. Rahman et al. (2026) show that hallucination and fact-checking remain central problems for LLMs. For agents, the relevant question is not only whether a final claim is factual but whether unsupported claims influenced intermediate actions. A factuality failure corrected before external action is materially different from one that triggers a consequential tool call, and evaluation protocols should record which of the two occurred.

8 Safety, privacy, and trustworthiness in autonomous loops

Safety risks change qualitatively when a model can act. Chen et al. (2026) frame LLM safety across trustworthy, responsible, and ecosystemic dimensions, and Gohil et al. (2026) adapt safety concerns to agentic AI. Their combined implication is that agent safety must be evaluated across both *component* failures and *interaction* failures. A safe base model does not guarantee a safe agent if orchestration, memory, tools, or permissions create new pathways to harm—a concern anticipated in early work on accidents arising from misspecified objectives and unsafe exploration (Amodei et al. 2016).

Privacy illustrates this interaction clearly. Foundation-model privacy risks include sensitive-data exposure and privacy–utility trade-offs (Elsheikh et al. 2026). An agent may additionally collect operational context, store memories, retrieve private documents, transmit data to third-party tools, and expose information across agent boundaries. Privacy engineering for agents therefore requires data minimization at each interface of the loop, not merely privacy controls at training time.

Toxicity and harmful content likewise matter beyond conversational quality. Villate-Castillo et al. (2026) review definitions, datasets, detectors, and detoxification approaches for LLM toxicity. In agentic

systems, harmful language can become an instruction, a retrieved artifact, a coordination message, or a trigger for an external action. Detection should therefore cover inputs, intermediate states, inter-agent messages, and outputs, with thresholds adjusted to the consequences of the action domain rather than fixed globally.

Across these risks, the unifying principle is that trustworthiness is a property of the loop, not of its components taken separately. This is why the AIR framework locates safety, privacy, and toxicity controls at the interfaces between layers—where observations enter, where decisions become actions, and where agents exchange messages—rather than solely at the model.

9 Human–agent collaboration and authority allocation

Human oversight is often invoked as the final safeguard for agentic systems, yet it is rarely specified with the precision applied to other system components. The AIR literature suggests why this is a problem. Trust in human–autonomous teams depends on identifiable architectural antecedents rather than on the mere presence of a human (Zamecnik et al. 2026), and explanations that support meaningful intervention tend to be interactive and iterative rather than one-shot (Mindlin et al. 2025). Oversight must therefore be designed as an allocation of authority between humans and agents across time, a view consistent with practitioner guidance that emphasizes constraining action spaces, requiring approval for consequential actions, and preserving the ability to interrupt agents (Shavit et al. 2023).

Three oversight modes are useful. In *ex ante* oversight, humans define goals, permissions, constraints, and approval thresholds before execution. In *concurrent* oversight, humans are invoked when the agent encounters uncertainty, conflict, high-impact actions, or policy exceptions. In *ex post* oversight, logs and outcomes are reviewed to improve policies, detect drift, and support accountability. High-risk deployments generally require all three modes, whereas low-risk tasks may rely primarily on *ex ante* controls and sampled *ex post* review.

A mature oversight model must also distinguish routine autonomy from exceptional escalation. Requiring human confirmation for every low-risk step makes systems unusable and encourages rubber-stamping (Proposition 5); allowing unrestricted autonomy makes intervention come too late. The design goal is *calibrated oversight*: human attention should concentrate where uncertainty, irreversibility, external impact, or normative judgment is highest—that is, where the autonomy vector of Eq. (2) takes its most demanding values.

Authority allocation also requires interface design. Users need to understand whether the system is suggesting, deciding, or acting. Confirmation prompts should describe the actual external consequence rather than merely ask whether the user wishes to “continue.” Where possible, the system should preview intended actions, identify affected resources, and distinguish reversible from irreversible steps. These features make human oversight operational rather than symbolic.

10 Research agenda for governable agentic intelligence

The synthesis identifies ten research priorities. Each targets a location where assurance debt currently accumulates faster than controls. Table 5 summarizes the agenda and links each priority to the ALL layers and corpus anchors.

Table 5
Research agenda for governable agentic intelligence

Priority	Open problem	Illustrative directions	Layers	Anchors
1 Verifiable reasoning	Plausible reasoning is not verified reasoning	Structured intermediate states; proof-carrying actions; selective verification	L2, L5	Liu et al. (2026); Nadăș (2026)
2 Calibrated autonomy	Fixed permissions ignore risk and context	Dynamic permissioning parameterized by the autonomy vector	L4, L6	Gohil et al. (2026); Chen et al. (2026)
3 Deployment-valid evaluation	Benchmarks overestimate field reliability	Long-horizon, cost-aware, perturbed tasks; severity reporting	L5	Kehkashan et al. (2026); Nageshwaran et al. (2026)
4 Provenance-aware memory	Evidence lineage is lost across steps	Lineage and temporal validity for RAG and memory	L2, L3	Wu et al. (2026); Rahman et al. (2026)
5 Privacy-preserving agency	Privacy controls stop at training	Trajectory-level privacy-utility evaluation	L2, L6	Elsheikh et al. (2026)
6 Multilingual and multimodal assurance	Capability and safety vary by language and modality	Localized benchmarks; cross-modal verification	L1, L5	Huang et al. (2026); Li et al. (2025)
7 Plural evaluation	Judges share biases with generators	Evaluator diversity; disagreement as uncertainty	L5	Nadăș (2026)
8 Governable multi-agent systems	Accountability diffuses across agents	Evidence independence; communication integrity	L4, L6	Abou Ali et al. (2026); Chowa et al. (2026)
9 Interactive explainability	Explanations rarely support intervention	Dialogue-based, action-previewing explanations	L6	Mindlin et al. (2025); Zamecnik et al. (2026)
10 Lifecycle monitoring	Assurance ends at deployment	Drift detection; incident taxonomies; near-miss learning	L5, L6	Gohil et al. (2026); Kehkashan et al. (2026)

10.1 Verifiable rather than plausible reasoning

Future work should connect reasoning quality to externally checkable intermediate states. Reward models and self-reflection are promising, but verification should not rely solely on the same model family that generated the reasoning (Liu et al. 2026; Nadăș 2026). Research is needed on proof-carrying actions, structured intermediate representations, uncertainty-aware planning, and selective verification that concentrates computation on consequential decisions.

10.2 Calibrated autonomy and dynamic permissioning

Autonomy should adapt to task risk, confidence, user intent, and environmental conditions. Research should move beyond fixed permission sets toward dynamic policies that reduce autonomy when uncertainty rises and expand it when evidence is strong and actions are reversible. Such policies require interpretable risk signals and robust recovery mechanisms (Gohil et al. 2026; Chen et al. 2026), and Eq. (2) offers a candidate parameterization for them.

10.3 Deployment-valid evaluation

Benchmarks should be designed to predict operational quality. This requires realistic tool interfaces, long-horizon tasks, hidden perturbations, cost accounting, privacy constraints, failure recovery, and integration with human workflows. Evaluation research should report not only success rates but failure severity and detectability (Kehkashan et al. 2026; Nageshwaran et al. 2026), and should test Proposition 3 directly by comparing benchmark and field performance for the same systems.

10.4 Provenance-aware memory and retrieval

RAG and memory should preserve evidence lineage and temporal validity. Agents need mechanisms to distinguish authoritative from weak evidence, current from stale information, user-provided from inferred facts, and durable preferences from transient context (Wu et al. 2026; Rahman et al. 2026).

10.5 Privacy-preserving agency

Privacy research should explicitly model agent workflows, including cross-tool data movement, persistent memory, inter-agent communication, and third-party service boundaries. Privacy–utility trade-offs should be evaluated at task and trajectory levels, not only at training time (Elsheikh et al. 2026).

10.6 Multilingual and multimodal assurance

Capability, safety, and evaluation should be tested across the languages and modalities actually used in deployment. Multilingual and multimodal agents may encounter uneven data quality and inconsistent safety behavior, requiring localized benchmarks and cross-modal verification (Huang et al. 2026; Li et al. 2025).

10.7 Independent and plural evaluation

Automated evaluation should incorporate evaluator diversity and explicit uncertainty. Future systems should quantify judge disagreement, detect self-preference, and combine model-based judging with

deterministic checks and human audit where consequences warrant it (Nadăș 2026). Measuring evaluator correlation would provide a direct test of Proposition 2.

10.8 Governable multi-agent systems

Multi-agent research should examine accountability, evidence independence, communication integrity, coalition failures, and conflict resolution. Role specialization is useful only when responsibilities and escalation paths remain visible (Abou Ali et al. 2026; Chowa et al. 2026).

10.9 Interactive explainability for intervention

Explanations for agentic systems should be designed to support intervention, not merely understanding. Dialogue-based explainability approaches, which allow users to query, probe, and refine explanations iteratively, offer a promising foundation (Mindlin et al. 2025). Research is needed on explanations that expose the evidence, assumptions, and pending actions behind an agent’s next step at the moment a human must decide whether to allow it, and on how such explanations shape calibrated trust in human–autonomous teams (Zamecnik et al. 2026).

10.10 Lifecycle monitoring and incident learning

Agent evaluation should continue after deployment. Systems need drift detection, action logging, near-miss analysis, incident taxonomies, and mechanisms for updating permissions and policies, embedded in end-to-end internal auditing practices of the kind proposed for algorithmic accountability (Raji et al. 2020). The central research question is how to learn from failures without normalizing unsafe autonomy—how, in the terms of Eq. (3), to retire assurance debt rather than merely record it.

11 Implications

11.1 Implications for researchers

The AIL framework suggests that contributions to agentic AI should specify where they sit in the lifecycle and which assurance obligations they create. A new planner should be evaluated not only for task success but for error propagation and recoverability. A new memory architecture should report its privacy and provenance properties. A new multi-agent protocol should address accountability and evaluator independence. Reporting of this kind would make comparisons more meaningful and counter the tendency to treat safety as a separate, downstream research topic. The five propositions also offer testable hypotheses for empirical work.

11.2 Implications for practitioners

For practitioners, the framework yields a deployment sequence: characterize the foundation model in the intended languages, modalities, and domain; externalize critical reasoning checkpoints; constrain retrieval and tool interfaces according to minimum necessary agency; calibrate autonomy to

consequence; evaluate trajectories as well as outcomes; and define human authority, escalation, and recovery before deployment. The five-question assurance model (Table 4) can serve as the template for a task-level assurance case. These practices are consistent with the direction of AIR's recent literature, which increasingly connects technical capability to deployment validity, trustworthiness, and oversight (Kehkashan et al. 2026; Gohil et al. 2026; Chen et al. 2026). They also map onto emerging governance instruments: the AI Risk Management Framework of the US National Institute of Standards and Technology organizes risk management into govern, map, measure, and manage functions (NIST 2023), and the European Union's AI Act requires high-risk AI systems to be designed for effective human oversight (European Union 2024). The AIL framework offers a system-level vocabulary for meeting such requirements in agentic deployments.

11.3 Implications for the review literature

Venue-bounded syntheses can reveal conceptual convergence within a field-shaping journal, but they must be transparent about their evidence boundaries. The AIR corpus provides a coherent and internally comparable body of work and supports a distinctive meta-perspective on how the review literature has framed agentic AI. Grounding the derived constructs in primary and cross-venue evidence, as we do here, guards against the risk that a venue-bounded synthesis merely reflects the preoccupations of one journal. The strongest contribution of this article is accordingly its integrative framework and theory of assurance debt, which can be tested against, and extended with, the wider literature.

12 Limitations and future research

Four limitations qualify our conclusions. First, the coded corpus is restricted to AIR. This boundary secures coherence and comparability; although we ground our constructs in primary studies and cross-venue surveys, these were not systematically coded, and future work should test whether the AIL layers and the assurance-debt pattern hold in a field-wide systematic corpus. Second, agentic AI is changing rapidly, and taxonomies and benchmark landscapes date quickly. Third, because the corpus consists largely of reviews, our synthesis is second-order and inherits the selection choices of the underlying reviews. Fourth, the AIL framework and the propositions are conceptual. Their value should be established empirically, through case studies that map real agent deployments to lifecycle controls and through studies that test whether the five-question assurance model improves failure detection, governance, and deployment decisions. We regard these limitations as the natural boundaries of a theory-building review and as a roadmap for the empirical program it invites.

13 Conclusion

The evolution from generative AI to agentic intelligence changes the unit of analysis from model output to system behavior. AIR's recent literature shows that reasoning, retrieval, tools, memory, multi-agent coordination, automated evaluation, safety, privacy, explainability, and human oversight are converging

into a common deployment problem. The Agentic Intelligence Lifecycle integrates these concerns into six layers spanning foundation, cognition, grounding, orchestration, assurance, and governance, and the concept of assurance debt explains why capability and trustworthiness so often diverge.

Our central conclusion is that capability and trustworthiness must be co-designed. Every increase in agency expands both what a system can accomplish and what can go wrong. Reliable deployment therefore depends on verifiable intermediate decisions, provenance-aware grounding, constrained tools, calibrated autonomy, trajectory-level evaluation, privacy and safety controls, and recoverable human authority. The next stage of agentic AI research should not ask only whether agents can reason and act. It should ask whether their reasoning and actions remain observable, attributable, interruptible, and governable under real conditions.

Declarations

Competing interests

The author declares no competing interests. [Confirm before submission.]

Ethics approval

and consent Not applicable. This article reviews published literature and involves no human participants or personal data.

Funding

[Insert funding information, or state: “The author received no specific funding for this work.”]

Author Contribution

H.U. conceived and designed the study, conducted the literature search, screening and coding, developed the Agentic Intelligence Lifecycle framework and propositions, prepared Figure 1 and Tables 1–5, and wrote the manuscript. H.U. reviewed and approved the final manuscript.

Data availability

No new datasets were generated. The review corpus consists of published articles in *Artificial Intelligence Review* listed in Table 1 and the References. [If a screening log is prepared, state its repository or supplementary-material location.]

References

1. Abou Ali M, Dornaika F, Charafeddine J (2026) Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artif Intell Rev* 59:11. <https://doi.org/10.1007/s10462-025-11422-4>
2. Amodei D, Olah C, Steinhardt J et al (2016) Concrete problems in AI safety. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1606.06565>
3. Bolaños F, Salatino A, Osborne F et al (2024) Artificial intelligence for literature reviews: opportunities and challenges. *Artif Intell Rev* 57:259. <https://doi.org/10.1007/s10462-024-10902-3>
4. Carlini N, Tramèr F, Wallace E et al (2021) Extracting training data from large language models. In: *Proceedings of the 30th USENIX Security Symposium*. USENIX Association, pp 2633–2650
5. Chan A, Salganik R, Markelius A et al (2023) Harms from increasingly agentic algorithmic systems. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 651–666. <https://doi.org/10.1145/3593013.3594033>
6. Chen C, Gong X, Liu Z et al (2026) AI safety landscape for large language models: taxonomy, state-of-the-art, and future directions. *Artif Intell Rev* 59:212. <https://doi.org/10.1007/s10462-026-11590-x>
7. Chowa SS, Alvi R, Rahman SS et al (2026) From language to action: a review of large language models as autonomous agents and tool users. *Artif Intell Rev* 59:71. <https://doi.org/10.1007/s10462-025-11471-9>
8. Du Y, Li S, Torralba A et al (2024) Improving factuality and reasoning in language models through multiagent debate. In: *Proceedings of the 41st International Conference on Machine Learning*. PMLR 235, pp 11733–11763
9. Elsheikh AMA, Rokne J, Alhaji R (2026) Privacy-preserving in foundation models: a systematic review of techniques, threats, and trade-offs. *Artif Intell Rev* 59:145. <https://doi.org/10.1007/s10462-026-11535-4>
10. European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union, L series*, 12 July 2024
11. Gohil V, Shah SB, Rauniyar K et al (2026) Towards safe and trustworthy agentic AI: foundations, taxonomy, technologies, applications, and future directions. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-026-11682-8>
12. Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J* 26:91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
13. Greshake K, Abdelnabi S, Mishra S et al (2023) Not what you’ve signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. In: *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*. ACM, New York, pp 79–90. <https://doi.org/10.1145/3605764.3623985>

14. Huang J, Chen X, Mishra S et al (2024) Large language models cannot self-correct reasoning yet. In: Proceedings of the 12th International Conference on Learning Representations (ICLR)
15. Huang K, Mo F, Zhang X et al (2026) A survey on large language models with multilingualism: recent advances and new frontiers. *Artif Intell Rev* 59:146. <https://doi.org/10.1007/s10462-026-11534-5>
16. Ji Z, Lee N, Frieske R et al (2023) Survey of hallucination in natural language generation. *ACM Comput Surv* 55:248. <https://doi.org/10.1145/3571730>
17. Jimenez CE, Yang J, Wettig A et al (2024) SWE-bench: can language models resolve real-world GitHub issues? In: Proceedings of the 12th International Conference on Learning Representations (ICLR)
18. Kapoor S, Stroebel B, Siegel ZS et al (2024) AI agents that matter. arXiv preprint. <https://doi.org/10.48550/arXiv.2407.01502>
19. Kehkashan T, Abdullah M, Al-Shamayleh AS et al (2026) From benchmarks to deployment: a comprehensive review of agentic AI evaluation. *Artif Intell Rev* 59:167. <https://doi.org/10.1007/s10462-026-11571-0>
20. Kumar P (2024) Large language models (LLMs): survey, technical frameworks, and future challenges. *Artif Intell Rev* 57:260. <https://doi.org/10.1007/s10462-024-10888-y>
21. Lewis P, Perez E, Piktus A et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst* 33:9459–9474
22. Li S, Wong KW, Wang G et al (2025) A systematic review of multi-modal large language models on domain-specific applications. *Artif Intell Rev* 58:383. <https://doi.org/10.1007/s10462-025-11398-1>
23. Lightman H, Kosaraju V, Burda Y et al (2024) Let's verify step by step. In: Proceedings of the 12th International Conference on Learning Representations (ICLR)
24. Liu Q, Xu H, Chen X et al (2026) Enhancing large language model reasoning with reward models: an analytical survey. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-026-11657-9>
25. Liu X, Yu H, Zhang H et al (2024) AgentBench: evaluating LLMs as agents. In: Proceedings of the 12th International Conference on Learning Representations (ICLR)
26. Mindlin D, Beer F, Sieger LN et al (2025) Beyond one-shot explanations: a systematic literature review of dialogue-based xAI approaches. *Artif Intell Rev* 58:81. <https://doi.org/10.1007/s10462-024-11007-7>
27. Nadăș MD (2026) Large language models as judges: recent advances in LLM-based evaluation, critique, preference modeling, and feedback for text and code. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-026-11652-0>
28. Nageshwaran V, Ezekiel S, Tran TT et al (2026) Large language model agent evaluation and benchmarking: a systematic survey, meta-taxonomy, and critical research roadmap. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-026-11678-4>
29. NIST (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology, Gaithersburg. <https://doi.org/10.6028/NIST.AI.100-1>

30. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. <https://doi.org/10.1136/bmj.n71>
31. Panickssery A, Bowman SR, Feng S (2024) LLM evaluators recognize and favor their own generations. *Adv Neural Inf Process Syst* 37. <https://doi.org/10.52202/079017-2197>
32. Parasuraman R, Riley V (1997) Humans and automation: use, misuse, disuse, abuse. *Hum Factors* 39:230–253. <https://doi.org/10.1518/001872097778543886>
33. Parasuraman R, Sheridan TB, Wickens CD (2000) A model for types and levels of human interaction with automation. *IEEE Trans Syst Man Cybern Syst Hum* 30:286–297. <https://doi.org/10.1109/3468.844354>
34. Park JS, O'Brien J, Cai CJ et al (2023) Generative agents: interactive simulacra of human behavior. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, pp 1–22. <https://doi.org/10.1145/3586183.3606763>
35. Patel K, Shah M, Qureshi KM et al (2026) A systematic review of generative AI: importance of industry and startup-centered perspectives, agentic AI, ethical considerations & challenges, and future directions. *Artif Intell Rev* 59:7. <https://doi.org/10.1007/s10462-025-11435-z>
36. Rahman SS, Islam MA, Alam MM et al (2026) Hallucination to truth: a review of fact-checking and factuality evaluation in large language models. *Artif Intell Rev* 59:70. <https://doi.org/10.1007/s10462-025-11454-w>
37. Raji ID, Smart A, White RN et al (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, pp 33–44. <https://doi.org/10.1145/3351095.3372873>
38. Schick T, Dwivedi-Yu J, Dessì R et al (2023) Toolformer: language models can teach themselves to use tools. In: *Adv Neural Inf Process Syst* 36
39. Sculley D, Holt G, Golovin D et al (2015) Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst* 28:2503–2511
40. Shavit Y, Agarwal S, Brundage M et al (2023) Practices for governing agentic AI systems. OpenAI white paper. <https://openai.com/research/practices-for-governing-agentic-ai-systems>. Accessed [date]
41. Shinn N, Cassano F, Gopinath A et al (2023) Reflexion: language agents with verbal reinforcement learning. In: *Adv Neural Inf Process Syst* 36
42. Snyder H (2019) Literature review as a research methodology: an overview and guidelines. *J Bus Res* 104:333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
43. Villate-Castillo G, Del Ser J, Sanz Urquijo B (2026) A systematic review of toxicity in large language models: definitions, datasets, detectors, detoxification methods and challenges. *Artif Intell Rev*. <https://doi.org/10.1007/s10462-026-11548-z>
44. Wang L, Ma C, Feng X et al (2024) A survey on large language model based autonomous agents. *Front Comput Sci* 18:186345. <https://doi.org/10.1007/s11704-024-40231-1>

45. Waqas M, Rajab R, Laghari AA et al (2026) Exploring the architectures of large language models and impact across multiple domains: a review. *Artif Intell Rev* 59:218. <https://doi.org/10.1007/s10462-026-11609-3>
46. Wei J, Wang X, Schuurmans D et al (2022) Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst* 35:24824–24837
47. Wu S, Xiong Y, Cui Y et al (2026) Retrieval-augmented generation for natural language processing: a survey. *Artif Intell Rev* 59:192. <https://doi.org/10.1007/s10462-026-11605-7>
48. Xi Z, Chen W, Guo X et al (2025) The rise and potential of large language model based agents: a survey. *Sci China Inf Sci* 68:121101
49. Yao S, Zhao J, Yu D et al (2023) ReAct: synergizing reasoning and acting in language models. In: *Proceedings of the 11th International Conference on Learning Representations (ICLR)*
50. Yao S, Shinn N, Razavi P, Narasimhan K (2025) τ -bench: a benchmark for tool-agent-user interaction in real-world domains. In: *Proceedings of the 13th International Conference on Learning Representations (ICLR)*
51. Zamecnik A, Febriantoro W, Cukurova M et al (2026) Linking AI agent architectures to trust antecedents in human-autonomous teams: a scoping review. *Artif Intell Rev* 59:178. <https://doi.org/10.1007/s10462-026-11572-z>
52. Zheng L, Chiang WL, Sheng Y et al (2023) Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: *Adv Neural Inf Process Syst* 36 (Datasets and Benchmarks Track)
53. Zhou S, Xu FF, Zhu H et al (2024) WebArena: a realistic web environment for building autonomous agents. In: *Proceedings of the 12th International Conference on Learning Representations (ICLR)*

Figures

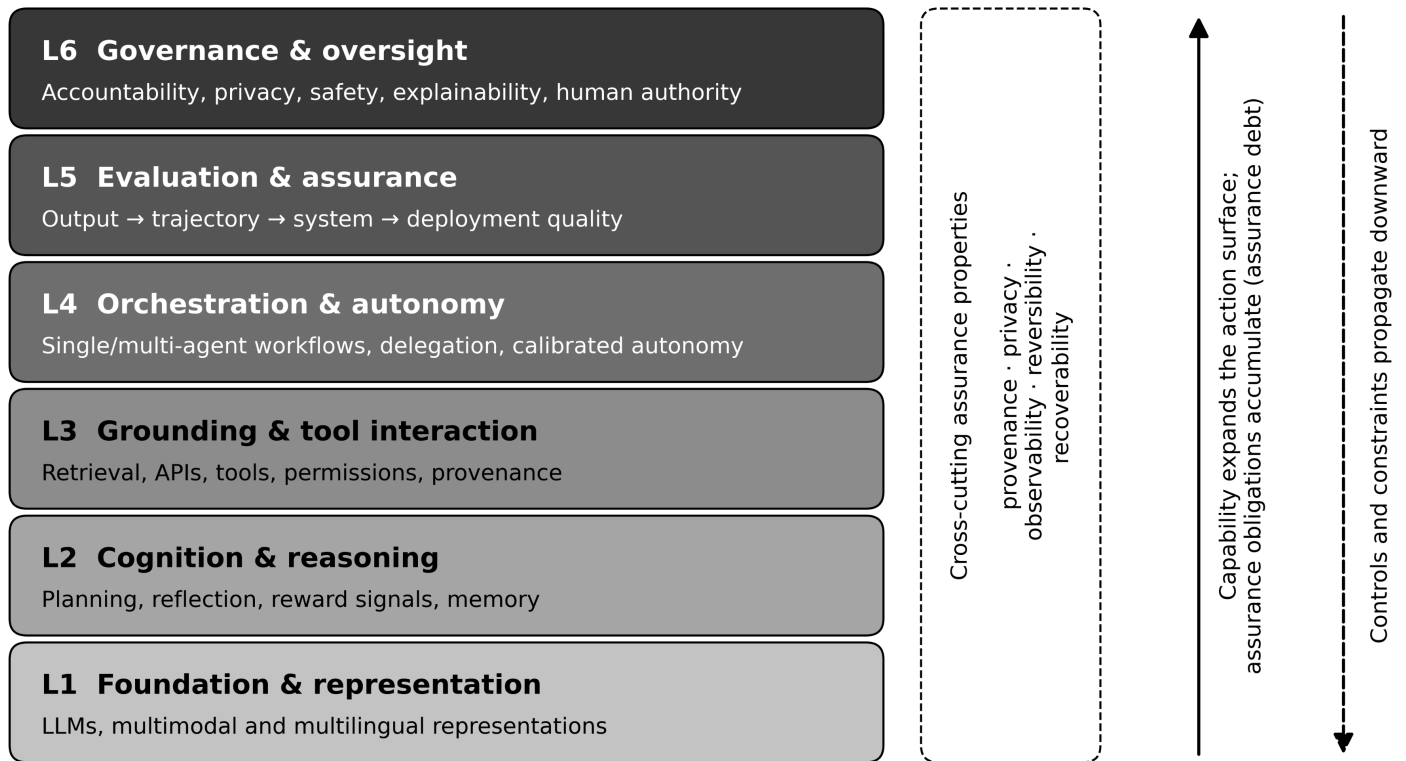


Figure 1

The Agentic Intelligence Lifecycle (AIL) framework. Six analytical, non-sequential layers are paired with cross-cutting assurance properties (dashed box). Capability expansion at each layer increases the system's action surface and accumulates assurance obligations, while controls and constraints must propagate downward from evaluation and governance to lower layers