

From Rules to Autonomous Agents: The Evolution of Artificial Intelligence, Human-AI Integration, and the Governance of Agentic Systems-A Review

Loso Judijanto

IPOSS Jakarta, Indonesia

ABSTRACT

Artificial intelligence has evolved from rule-based symbolic systems into learning architectures capable of perception, prediction, content generation, reasoning, planning, tool use, and increasingly autonomous action. This qualitative literature review examines this transition as more than a technological chronology. It interprets AI evolution through five interrelated dimensions: technological evolution, capability expansion, changing human roles, human-AI integration, and governance. Drawing on contemporary peer-reviewed literature published primarily between 2020 and 2026, the review identifies seven thematic developments. AI has progressed from encoded rules toward learned representations and goal-directed action; generative AI has expanded AI into a general-purpose content collaborator; agentic AI adds planning, memory, tool use, and iterative task execution; human roles increasingly shift from operator toward collaborator, delegator, and governor; human-AI complementarity remains conditional rather than automatic; effective deployment requires calibrated rather than maximal trust; and governance requirements increase as AI systems acquire greater operational agency. The article proposes the Evolution-Capability-Role-Integration-Governance (ECRIG) framework to explain these relationships. The framework suggests that sustainable benefits from advanced AI emerge when technological capability, human role allocation, interaction architecture, and governance mechanisms develop jointly. The review concludes that the central challenge of the agentic-AI era is not choosing between humans and machines, but designing accountable systems in which increasingly capable artificial agents remain aligned with meaningful human goals, authority, and oversight.

*Corresponding author

Loso Judijanto, IPOSS Jakarta, Indonesia.

Received: October 01, 2026; **Accepted:** October 14, 2026; **Published:** October 25, 2026

Keywords: Artificial Intelligence, AI Evolution; Generative AI, Agentic AI, Large Language Models, Human-AI Collaboration, Human-AI Integration, AI Augmentation, AI Agents, Responsible AI

JEL Classification: O33; O32; M15; D83; L86

Introduction

Artificial intelligence (AI) has undergone repeated conceptual and technological transformations. Early AI emphasized symbolic representations, explicit rules, formal logic, and expert knowledge. Machine learning subsequently shifted attention toward algorithms capable of inferring relationships from data, while deep learning expanded this capacity through hierarchical representation learning. Transformer architectures further changed the technological trajectory by enabling scalable contextual modeling and transfer across tasks. More recently, generative AI has transformed AI systems from predominantly predictive technologies into systems capable of producing language, images, software code, audio, and other forms of novel content. The newest phase is increasingly described as agentic AI, in which AI does not merely respond to an individual prompt but can interpret goals, decompose tasks, formulate plans, use tools, retain working context, interact with an environment, and execute sequences of actions with varying degrees of autonomy [1-10].

This development suggests that AI evolution cannot adequately be understood as a sequence of models that simply become more accurate. A more consequential transformation concerns what kind of role AI can occupy within a human activity system. Conventional software executes predetermined instructions; machine-learning systems classify or predict; generative models create candidate outputs; and agentic systems can increasingly participate in planning and execution. Contemporary surveys describe LLM-based agents in terms of capabilities such as reasoning, planning, memory, external-tool interaction, feedback processing, and iterative action, indicating a movement from language generation toward goal-directed behavior.

The organizational implications are equally important. AI capability increasingly depends not simply on model quality but on organizational data resources, technical infrastructure, human skills, coordination mechanisms, and the ability to redesign workflows around the technology. Research on organizational AI capability therefore emphasizes complementarities between tangible technological resources, human competencies, and intangible organizational resources rather than treating AI as a standalone asset. AI is simultaneously changing innovation processes, strategic management, knowledge work, and the distribution of organizational control, reinforcing the need for multidisciplinary perspectives on its organizational and societal implications [11-19].

The acceleration of generative AI makes this issue especially urgent. Experimental evidence shows that generative AI can substantially improve performance in some professional tasks. Noy and Zhang (2023), for example, found that access to ChatGPT in professional writing tasks reduced average completion time by approximately 40% while raising evaluated output quality by approximately 18%. Brynjolfsson, Li and Raymond (2025), examining 5,172 customer-support workers, reported an average productivity increase of roughly 15%, with particularly large benefits among less experienced workers. These findings support the augmentation potential of generative AI but also reveal a recurrent pattern: technological benefits vary according to the user, task, existing expertise, and design of the work system.

The same qualification applies to human-AI collaboration more generally. The assumption that adding an AI system to a human decision maker will automatically improve performance is not supported consistently by empirical evidence. Vaccaro, Almaatouq and Malone's (2024) meta-analysis of 106 experimental studies covering 370 effect sizes found that human-AI combinations, on average, performed worse than the best-performing human or AI alone, although outcomes varied considerably across tasks. Content-creation activities showed more promising complementarities than many decision tasks. Human-AI integration therefore cannot be reduced to the availability of stronger models; it depends on whether tasks, authority, information, feedback, and verification are allocated appropriately.

This problem becomes more significant as AI moves from advising humans to acting on their behalf. Baird and Maruping (2021) argued that increasingly agentic information-system artifacts challenge traditional assumptions that the human is always the sole active agent and the technology merely a passive instrument. Their delegation perspective is particularly relevant for modern AI agents because task responsibility can increasingly be distributed between human and artificial actors. In parallel, Raisch and Krakowski (2021) show that automation and augmentation should not be treated as mutually exclusive alternatives. Organizations can simultaneously automate parts of a process while augmenting human judgment in other parts, producing a persistent automation-augmentation paradox.

Greater AI agency also intensifies questions about trust, explanations, responsibility, security, fairness, and oversight. Trust is essential for use, yet excessive trust creates automation bias and overreliance. Explanations may assist appropriate reliance but can also make incorrect recommendations appear more persuasive. Naiseh et al. (2023), for example, found that explanations can improve some aspects of human-AI decision performance but may simultaneously increase reliance on AI recommendations. Meanwhile, large language models remain vulnerable to factual and faithfulness-related hallucinations, while social biases may be learned or reproduced through generated outputs. Connecting such models to external tools means that an erroneous output can potentially become an erroneous action.

Despite extensive research on particular components of this transformation, the literature remains fragmented. Technical studies focus on architectures and capabilities; information-systems research emphasizes organizational adoption and delegation; human-computer-interaction research examines trust and explanations; management research emphasizes augmentation; and responsible-AI research concentrates on accountability, fairness, safety, and control. This article addresses that fragmentation by synthesizing these streams through the

proposed **Evolution-Capability-Role-Integration-Governance (ECRIG) framework**.

Accordingly, the review addresses seven research questions: how AI evolution can be interpreted in terms of changing capabilities and agency; what distinguishes generative from agentic AI; how human roles change across AI paradigms; under what conditions human-AI integration creates complementarity; how trust, explanation, expertise, and workflow design affect collaboration; how governance requirements change with increasing autonomy; and how these elements can be integrated into a coherent framework. The objective is therefore not merely to reconstruct a history of AI, but to explain how technological evolution changes the architecture of human-AI relationships and the requirements for responsible governance.

Literature Review

From Symbolic AI to Learning Systems

The intellectual origins of AI are strongly associated with attempts to represent reasoning in forms that computers can manipulate. Early expert systems encoded domain knowledge through explicit rules and knowledge structures. Their central advantage was transparency: where the rules were explicit, the reasoning process could often be traced. Their weakness was brittleness. Knowledge acquisition was costly, rules were difficult to maintain as environments changed, and performance deteriorated when problems departed from modeled conditions. Contemporary reviews of machine learning illustrate the subsequent transition from explicit programming toward statistical learning from observations, making data rather than manually encoded rules increasingly central to system performance.

Machine learning altered the relationship between software and its environment because behavior could be inferred from patterns rather than exhaustively specified in advance. Classification, prediction, clustering, recommendation, and anomaly detection could increasingly be performed from learned statistical relationships. The development of deep learning intensified this transition by learning complex hierarchical representations from large datasets, reducing dependence on manual feature engineering in many applications. Deep neural architectures became especially powerful in perception-oriented tasks, while reinforcement learning expanded the conceptual focus from static prediction toward sequential decision making and interaction with environments [21-30].

Transformer architectures subsequently introduced another major shift. Their self-attention mechanisms enabled models to represent relationships among elements of sequences at scale, initially producing breakthroughs in language processing and later spreading to vision, audio, multimodal systems, and other areas. Reviews by Lin, Khan and Han document the breadth of this diffusion. The importance of transformers extends beyond improvements in particular benchmarks: they contributed to the architectural foundations of models that can be pretrained at scale and subsequently adapted across many downstream tasks.

Consequently, AI evolution has progressively reduced the requirement that every capability be explicitly programmed. The dominant trajectory can be summarized as movement from **encoded knowledge**, toward **learned statistical relationships**, toward **learned representations**, and eventually toward **general-purpose contextual models**. This process is cumulative rather than strictly substitutive. Rule-based reasoning remains useful in settings requiring deterministic control; conventional machine

learning remains efficient for well-defined prediction problems; deep learning excels in high-dimensional representation; and transformer-based models provide new forms of contextual generality.

Figure 1 summarizes this evolutionary trajectory as a progression in capability rather than merely a sequence of dates. The central observation is that each major paradigm expands the range of tasks that can be delegated to computational systems while simultaneously changing the role humans must perform.

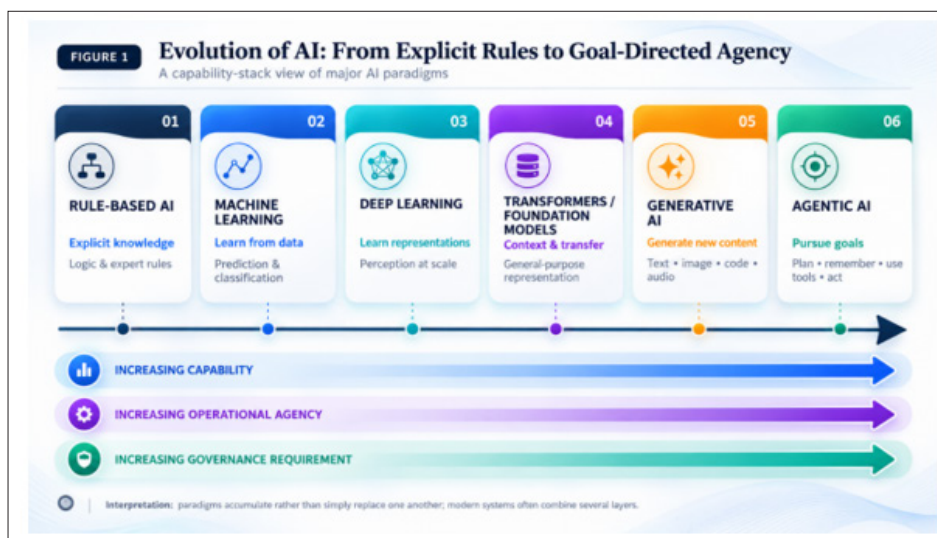


Figure 1: Evolution of AI from explicit rules to goal-directed agency

Source: author synthesis based on Sarker, Alzubaidi, Lin, Feuerriegel, Wang, Xi and Acharya

Figure 1 should not be interpreted as implying the disappearance of earlier technologies. Rather, contemporary AI systems frequently combine paradigms. An agent may use an LLM for language reasoning, deterministic rules to constrain actions, machine-learning classifiers for specialized prediction, retrieval systems for external knowledge, and conventional software for execution. The evolution of AI is therefore better characterized as an expanding **capability stack** than as replacement through successive generations.

Generative AI and Foundation-Model Generality

Generative AI represents a significant change in interaction between humans and computational systems. Whereas conventional analytics commonly returns predictions, classifications, recommendations, or scores, generative systems can produce artifacts that resemble products of knowledge work. These include explanatory text, computer code, images, presentations, summaries, simulations, designs, and synthetic data. Feuerriegel conceptualize generative AI as an important new class of information systems with implications for content creation and decision support, while Floridi and Chiriatti (2020) had already emphasized that large language models combine remarkable linguistic performance with important epistemic limitations.

The rapid adoption of conversational generative AI has made these capabilities accessible through natural-language interfaces. This lowers technical barriers to using AI and expands the potential user population beyond specialist programmers or data scientists. Dwivedi describe broad implications across research, education, policy, organizational practice, and knowledge creation, while van Dis emphasize the need for careful interdisciplinary investigation of transparency, reliability, human behavior, and scientific consequences.

Generative AI also changes creative collaboration. Doshi and Hauser (2024) found that generative AI can improve individual creative output while potentially decreasing collective diversity because different people may become anchored on similar AI-generated ideas. This illustrates a crucial distinction between **local productivity** and **system-level consequences**. An AI tool can make an individual more efficient while simultaneously homogenizing outputs across a population. Similarly, educational literature highlights both opportunities for personalized support and risks related to inaccurate output, dependency, assessment, and academic integrity. Scholarly communication faces parallel issues involving authorship, responsibility, disclosure, and the epistemic status of machine-generated prose [31-40].

Workplace effects are also mediated by perceptions of the technology. Jo and Park (2024) show that perceived AI capabilities and human-like traits can influence information-related behaviors in working environments. This means that employees respond not only to objective model quality but also to perceived competence, social presence, and ease of interaction. Generative AI therefore operates simultaneously as a computational technology and as an interactive actor within organizational communication.

From Generative Models to Agentic Systems

The distinction between generative and agentic AI is central to the present review. A generative model typically transforms an input into an output. Although the output can be highly sophisticated, the interaction often remains request-response oriented. Agentic systems add a control layer that allows a model to pursue goals over multiple steps. Common components include task decomposition, reasoning, planning, short- and long-term memory, access to external tools, action selection, environmental observation, feedback, and iterative adjustment

Acharya, Kuppan and Divya (2025) define agentic AI around autonomous pursuit of complex goals with limited human intervention, while later surveys emphasize heterogeneous architectures ranging from single agents to interacting multi-agent systems similarly highlight reasoning, planning, memory, and external-tool integration, while Bajoria et al. (2026) characterize the field as a transition from passive language models toward autonomous, action-enabled, and collaborative systems.

Agentic systems are therefore distinguished less by whether they use an LLM than by the architecture of action around the model. A language model may be one component within an agent that interprets goals, invokes search engines or databases, calls software APIs, writes and executes code, interacts with other agents, evaluates partial results, and continues until termination criteria are met. This can transform AI from a source of information into a participant in workflow execution. Recent tool-learning research likewise shows that reliable agentic performance depends on task decomposition, appropriate tool retrieval, ordered invocation, and validation of tool outputs rather than on language generation alone

Gao et al. (2024) illustrate another implication: LLM-enabled agents can be embedded in simulations in which multiple artificial actors interact, producing emergent social, economic, cyber, or physical dynamics. This creates research opportunities but also complicates evaluation because system behavior may arise from interactions among multiple adaptive components rather than from a single model output.

Human Roles: Automation, Augmentation, Teaming, and Delegation

As AI capability evolves, human roles evolve as well. Raisch and Krakowski (2021) distinguish automation, in which machines perform tasks previously executed by humans, from augmentation, in which humans and machines combine capabilities. They argue that these processes are intertwined rather than mutually exclusive. Organizations often automate routine components precisely to free human attention for interpretation, judgment, communication, or exception handling.

The teaming literature goes further. Seeber ask what changes when machines are regarded as teammates rather than tools, identifying research problems associated with machine characteristics, collaboration processes, and institutional arrangements. Empirical work indicates that performance is sensitive to the form of AI participation. Flathmann find that dynamically adjusting the influence of an AI teammate can affect human performance, while Hauptman emphasize the importance of adaptive autonomy across changing work cycles. Zhang show that perceived teammate performance is particularly consequential for cooperation, and Harris-Watson demonstrate that perceived competence and social characteristics influence receptivity to AI teammates. Recent synthesis further indicates that human-AI teams can underperform when coordination, shared cognition, and trust are weak, even when the AI component is individually capable [41-50].

Baird and Maruping's delegation framework is particularly valuable for the agentic era. Delegation implies more than using a technology: responsibility for aspects of goal pursuit is allocated to an artifact capable of initiating or adjusting behavior. This perspective enables a more precise vocabulary for examining agentic systems. The relevant questions become: What is delegated? Under what constraints? What authority does the agent possess? When must the agent return control? Who monitors performance? Who remains accountable for outcomes?

Method

Research Design

This study employs a **qualitative, integrative literature-review design rather than a systematic literature review (SLR)**. Its purpose is interpretive synthesis and conceptual development. Accordingly, the method does not claim exhaustive retrieval of every publication within a predefined domain and does not apply PRISMA as though the study were a systematic evidence review. Instead, literature is purposively combined across computer science, information systems, management, human-computer interaction, psychology, and responsible-AI research to develop a cross-disciplinary theoretical account.

This choice is consistent with contemporary methodological literature emphasizing that literature reviews can function as independent research studies whose contribution lies in integration, critique, conceptual clarification, framework development, and agenda setting. The research problem is particularly suitable for such a design because the concepts of generative AI, agentic AI, human-AI teaming, augmentation, delegation, explainability, and AI governance have developed in partly separate scholarly communities.

Literature Identification and Selection

The literature base was developed iteratively around four families of concepts. The first addressed technological evolution using terms such as artificial intelligence evolution, machine learning, deep learning, transformers, and foundation models. The second focused on emerging AI forms using generative AI, large language models, agentic AI, AI agents, autonomous agents, and tool-using LLMs. The third covered human integration using human-AI collaboration, human-AI teaming, AI augmentation, delegation, trust, and human-centered AI. The fourth addressed governance through terms including explainable AI, responsible AI, bias, hallucination, accountability, and human oversight.

Priority was given to peer-reviewed journal literature from 2020 onward. Earlier technological milestones inform the historical interpretation, but the evidentiary foundation of this draft relies on contemporary journal publications. Sources were selected for conceptual relevance, methodological contribution, empirical importance, or capacity to connect previously separate streams of scholarship. Citation chaining and cross-referencing between major surveys and conceptual papers were used to identify related work.

The resulting literature was treated as a heterogeneous theoretical corpus rather than as statistically combinable evidence. This distinction is important because studies differ markedly in task, population, AI architecture, outcome measure, and disciplinary assumptions. For example, productivity experiments with generative AI, clinical decision-support studies, human-AI teaming experiments, AI-governance analyses, and computer-science surveys answer fundamentally different questions. Combining them qualitatively allows those differences to inform the conceptual interpretation rather than treating heterogeneity as statistical noise.

Coding and Thematic Analysis

The review followed an iterative thematic procedure. First-cycle coding identified descriptive categories such as **rule-based reasoning, data-driven learning, contextual modeling, generation, planning, memory, tool use, autonomy, augmentation, delegation, trust, explanation, hallucination,**

bias, oversight, and accountability. Second-cycle analysis examined relationships among those categories, particularly how changes in AI capability were associated with changes in human roles and governance requirements.

Reflexive thematic analysis is appropriate for this theory-building purpose because themes are interpreted as analytic constructions rather than mechanically discovered entities. Braun and Clarke (2021) emphasize methodological coherence and reflexive interpretation, while Byrne (2022) provides a practical illustration of how iterative thematic development can move from coding toward higher-level analytic themes.

The thematic procedure eventually produced seven overarching findings: evolution from encoded rules to goal-directed action; generative AI as content collaborator; agentic AI as execution architecture; transformation of human roles; conditional human–AI complementarity; the importance of calibrated trust; and increasing governance burdens. These themes were subsequently integrated into the ECRIG framework.

Analytical Rigor

Rigor was supported through four practices. First, the review used sources from multiple disciplines to reduce dependence on a single conceptual tradition. Second, empirical evidence was considered alongside conceptual and technical literature. Third, contradictory evidence was retained rather than suppressed; for example, productivity gains from generative AI were interpreted alongside evidence that human–AI combinations may fail to generate synergy. Fourth, claims about emerging agentic systems were distinguished from established empirical findings because the newest agentic-AI literature remains dominated by surveys, architectures, benchmarks, and rapidly evolving applications.

This interpretive transparency is important because qualitative synthesis should not imply an artificial precision that the underlying literature cannot support. The purpose is to produce a conceptually defensible account of relationships among technologies, capabilities, human roles, and governance—not to estimate a universal causal effect of “AI.”

Results

Theme 1: AI Evolution Is a Transition from Encoded Rules toward Goal-Directed Action

The first major theme is that AI evolution can be interpreted through progressive changes in the source of system behavior. In symbolic AI, much of the relevant structure is encoded explicitly. Machine learning shifts part of this structure into patterns learned from data. Deep learning shifts further toward learned representations. Transformers enable contextual modeling across increasingly broad input spaces. Generative AI extends contextual models into production of novel artifacts, while agentic architectures connect model outputs to iterative goal-directed execution.

This capability progression changes what humans must specify. Rule-based systems require humans to specify much of the relevant reasoning structure. Statistical learning requires humans to provide data, objectives, architecture, evaluation criteria, and deployment context. Foundation models can operate over broader domains, allowing humans to communicate through natural-language goals rather than formal instructions. Agentic systems take this transformation further because users can increasingly specify an objective while leaving some decomposition, planning, tool selection, and sequencing to the system.

This should not be mistaken for unlimited autonomy. Agent capabilities remain constrained by training, context windows, tools, permissions, system architecture, computational resources, and the quality of feedback. Agentic autonomy is therefore better conceptualized as **bounded operational discretion** rather than independence from human-designed systems.

Theme 2: Generative AI Creates a General-Purpose Content Collaboration Layer

The second theme concerns the general-purpose nature of generative AI. Generative models lower the cost of producing first drafts, summaries, explanations, translations, software code, illustrations, and alternative formulations. The significance is not limited to automation of existing tasks. Natural-language interaction enables iterative collaboration in which users specify intent, inspect a response, add constraints, request revisions, and use model outputs as intermediate cognitive artifacts.

Empirical productivity evidence supports the importance of this mechanism. The improvements observed by Noy and Zhang in professional writing and by Brynjolfsson, Li and Raymond (2025) in customer support suggest that generative AI can compress production time and diffuse elements of high-performer practices to less experienced workers. These results are consistent with the idea of AI as an augmentation technology, but they should not be generalized indiscriminately across tasks.

Generative AI can also restructure creative processes. Instead of generating every candidate idea independently, users can ask AI systems to expand, critique, recombine, translate, or reframe ideas. Yet Doshi and Hauser show why individual-level gains need not translate into collective gains: generative assistance may improve evaluated creativity for individual users while reducing diversity across outputs. The technology can therefore simultaneously expand access to ideation and concentrate users around similar machine-generated conceptual spaces.

Academic and educational use presents a similar duality. Generative AI can provide immediate feedback, explanations, examples, and personalized interaction, yet unreliable responses and effortless text production can weaken learning if learners outsource activities that are themselves important for cognitive development. Questions of scholarly authorship and accountability likewise become important when AI-generated text enters research workflows

The results therefore support a distinction between **AI as output generator** and **AI as cognitive collaborator**. The latter framing is more consequential because users integrate AI-generated intermediate artifacts into their reasoning process. The quality of interaction then depends not only on model performance but also on the user's ability to ask appropriate questions, verify outputs, understand domain constraints, and recognize when external evidence is required.

Theme 3: Agentic AI Extends Generation into Execution

The third theme marks the most important contemporary transition. Agentic AI integrates generative or reasoning models with mechanisms for persistent goal pursuit. Instead of terminating after content generation, an agent can observe a task state, generate a plan, select a tool, execute an action, inspect the result, revise the plan, and continue.

A useful abstraction of agentic execution is the iterative sequence Goal → decomposition → planning → tool selection → action → observation → evaluation → revision → termination. This sequence emphasizes that agentic behavior is a closed-loop process in which intermediate observations can alter subsequent plans and actions.

This architecture changes the unit of analysis from an individual model response to a **trajectory of actions**. An agent can fail even when most individual outputs are reasonable because small errors may accumulate across a long sequence. Conversely, an agent can recover from an intermediate failure if feedback and verification mechanisms are well designed.

Contemporary surveys identify memory as another essential characteristic. Memory can preserve previous observations, intermediate results, user preferences, or task history, allowing behavior to remain coherent over a longer horizon. Tool use expands the potential consequences further because an agent may retrieve

external information, access databases, manipulate files, call software services, or execute code.

Multi-agent architectures add another layer of complexity. Different agents may be assigned specialized roles such as planner, researcher, critic, executor, or evaluator. Such arrangements can increase modularity and permit internal critique, yet they also create new coordination costs. Gao et al. (2024) demonstrate how LLM-based agents can participate in broader agent-based simulations, highlighting the possibility of emergent system-level behavior that cannot be understood simply by inspecting one isolated model.

The architectural distinction between generative and agentic AI is summarized in **Figure 2**. The figure emphasizes that agentic capability emerges through a feedback loop connecting language or multimodal reasoning to memory, tools, environmental observation, and action.



Figure 2: From generative response to agentic action loop

Source: author synthesis based on Wang et al. (2024), Xi et al. (2025), Acharya et al. (2025), Abou Ali et al. (2026), Chowa et al. (2026) and Bajoria et al. (2026).

Figure 2 clarifies why governance requirements increase when systems become agentic. A generative model may provide an erroneous recommendation that a human can reject. An agent connected to operational tools may convert a similar error into a series of consequential actions before a person reviews the result. Reliability must therefore be assessed not only at the model level but across **action chains, tool permissions, feedback loops, and escalation rules**.

Theme 4: Human Roles Move from Operator toward Collaborator, Delegator, and Governor

The fourth theme concerns the transformation of human roles. At lower levels of AI agency, humans remain close to execution. They program rules, select inputs, interpret predictions, or decide whether to accept recommendations. Generative systems support a more collaborative interaction in which humans formulate objectives, provide contextual knowledge, inspect alternatives, and iterate with the machine. Agentic systems enable further delegation, moving the human increasingly toward setting objectives, specifying boundaries, reviewing exceptions, and monitoring outcomes.

This role transition can be expressed as a progression from programmer/operator to user, AI-assisted worker, co-author/co-creator, augmented expert, teammate, delegator, and ultimately supervisor/governor. The sequence is analytical rather than strictly chronological because the same individual may occupy several of these roles within a single workflow.

These roles coexist rather than replace one another. The same person may be a co-author during idea development, an expert verifier when checking technical accuracy, and a governor when deciding whether an agent is authorized to perform an external action. Role design is therefore contextual.

The human–AI teaming literature demonstrates that collaboration is influenced by social and organizational as well as technical variables. Seeber et al. (2020) identify machine, collaboration, and institutional dimensions of teaming. Harris-Watson et al. (2023) find that perceived competence influences the acceptance and utilization of AI teammates. Zhang et al. (2023) similarly demonstrate the importance of perceived teammate performance and user expertise. Adaptive autonomy may further improve human–AI coordination

by allowing levels of AI influence to change with task conditions (Hauptman et al., 2023; Flathmann et al., 2023).

Figure 3 integrates these findings by relating increasing AI agency to a corresponding shift in the dominant human function. It also shows that increasing AI agency does not remove the human role; instead, it shifts human contribution toward orchestration, boundary setting, and accountability.



Figure 3: Human-role transition under increasing AI agency

Source: author synthesis based on Raisch and Krakowski (2021), Baird and Maruping (2021), Seeber et al. (2020), Flathmann et al. (2023), Zhang et al. (2023), Harris-Watson et al. (2023) and Hauptman et al. (2023).

The governance implication is substantial. The movement toward delegation does not eliminate human responsibility merely because a system executes the task. Instead, responsibility moves upward from individual micro-actions toward **designing objectives, access rights, validation requirements, intervention thresholds, and accountability arrangements**.

Theme 5: Human-AI Complementarity Is Conditional

The fifth theme is that complementarity is not an inherent property of combining a human and an AI. A high-performing AI can reduce human-AI performance if the human adds noise, misinterprets the recommendation, or intervenes at the wrong moment. Conversely, a human may correct model weaknesses when possessing relevant contextual knowledge that the AI lacks.

Vaccaro, Almaatouq and Malone (2024) provide particularly important evidence. Their meta-analysis found that human-AI systems frequently improved performance relative to humans alone but did not reliably outperform the better of human-alone or AI-alone conditions. The results were heterogeneous, and content-generation tasks produced more favorable synergy than many decision tasks. This means that the appropriate benchmark for a human-AI system is not simply whether it improves a human user, but whether the combined arrangement produces value beyond the strongest feasible alternative.

Clinical evidence illustrates why. Tschandl et al. (2020) demonstrated benefits from human-computer collaboration in skin-cancer recognition, but the broader clinical literature shows strong dependence on how predictions are presented and interpreted. Gaube et al. (2021) found that people can be susceptible to incorrect AI recommendations, highlighting risks of overreliance. Kiani et al. (2020) found that a deep-learning assistant improved pathologists under some circumstances, yet incorrect model predictions could also reduce diagnostic performance.

These results suggest several boundary conditions for complementarity. First, **task decomposability** matters: collaboration

is easier when human and machine contributions can be allocated clearly. Second, **relative competence** matters: users need some understanding of whether they or the AI are more reliable in a particular subtask. Third, **error detectability** matters: AI is safer to use where wrong outputs can be recognized before they create irreversible consequences. Fourth, **feedback quality** matters because users and systems require signals allowing performance to be adjusted. Fifth, **workflow integration** matters; a model with excellent standalone accuracy can still perform poorly within a badly designed decision process.

Theme 6: Trust Should Be Calibrated, Not Maximized

Trust is often treated as a prerequisite for AI adoption, but the literature supports a more precise concept: appropriate reliance. Glikson and Woolley (2020) show that trust in AI depends on characteristics of the system, its representation, its capability, the user, and the interaction. Kaplan et al. (2023) similarly identify multiple human-, AI-, and context-related determinants of trust through meta-analysis.

High trust is beneficial only when it corresponds to high reliability. Excessive trust in a poorly calibrated system produces overreliance, while insufficient trust in a reliable system causes users to ignore useful assistance. The optimal objective is therefore calibration: reliance should increase when AI is demonstrably capable and decrease when uncertainty or model limitations are material.

Explainability is frequently proposed as a solution, but the relationship is complex. Barredo Arrieta et al. (2020) describe explainability as a major component of responsible AI, while Vilone and Longo (2021) show that both definitions and evaluation criteria for explainability remain diverse. Mohseni, Zarei and Ragan (2021) similarly argue that explanation design must reflect stakeholder goals rather than assume a universal explanation format. Dazeley et al. (2021) emphasize conversational and human-aligned explanation approaches.

Importantly, explanations do not automatically produce good decisions. Naiseh et al. (2023) show that explanations may influence reliance in ways that can sometimes increase overdependence. An explanation can create a persuasive appearance of reasoning without guaranteeing that the underlying recommendation is correct. Effective explanation should therefore support users in understanding limitations, uncertainty, evidence, and conditions for challenge, not merely make AI outputs appear convincing. Experimental and review evidence further suggests that user decision control, cognitive capacity, and interactive system design can be as important as explanation provision itself, because explanations may increase cognitive load or induce reliance without improving understanding (Westphal et al., 2023; Wang and Ding, 2024; Raees et al., 2024).

Theme 7: Greater Agency Creates a Larger Governance Surface
The seventh theme is that governance burden expands as AI systems acquire more operational authority. Traditional software governance focuses on specification, cybersecurity, quality assurance, access control, and change management. Contemporary AI adds concerns regarding opacity, data bias, probabilistic outputs, drift, and emergent behaviors. Generative and agentic AI add hallucination, prompt sensitivity, tool-use risk, long-horizon error propagation, and uncertainty regarding responsibility for autonomous action.

Large-language-model hallucinations illustrate the reliability problem. Huang et al. (2025) distinguish factuality-related and

faithfulness-related forms of hallucination and review detection and mitigation techniques. Bias presents another challenge: Gallegos et al. (2024) show that LLMs can reproduce or amplify social biases and that fairness must be examined across datasets, model behavior, evaluation metrics, and mitigation stages.

Responsible governance therefore cannot rely exclusively on broad ethical principles. Hagendorff (2020) found significant commonality across AI ethics guidelines but also raised questions about their translation into practice. Human-centered AI provides one route from principles to system design. Shneiderman (2020a) argues that high levels of computer automation can coexist with high levels of human control, while Shneiderman (2020b) emphasizes reliable engineering, organizational safety culture, and broader governance structures. Recent agentic-AI safety research therefore emphasizes bounded autonomy, runtime monitoring, verification, and structured human oversight when agents can act through external tools or within consequential environments

These findings are synthesized in Figure 4, which proposes that governance requirements should be proportionate not to AI sophistication alone but to the interaction between autonomy and consequence. The matrix distinguishes low- and high-consequence settings so that the required intensity of verification, permissions, and human oversight can be interpreted proportionately.

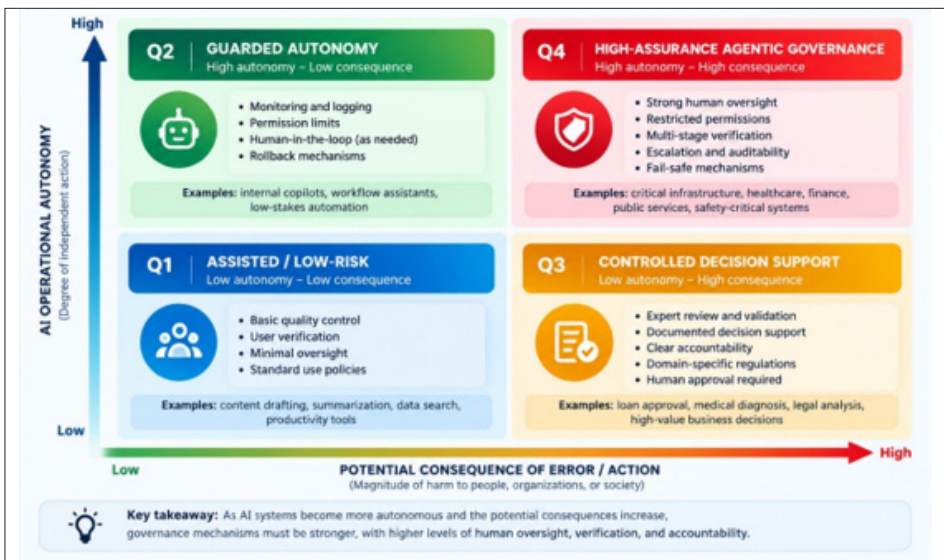


Figure 4: AI autonomy–consequence governance matrix

Source: author synthesis based on Shneiderman (2020a, 2020b), Barredo Arrieta et al. (2020), Hagendorff (2020), Gallegos et al. (2024), Huang et al. (2025), Acharya et al. (2025) and Chowa et al. (2026).

The matrix illustrates why autonomy should never be evaluated in isolation. An autonomous system scheduling low-risk internal activities presents a different governance problem from an autonomous system authorized to make irreversible financial, clinical, infrastructural, or legal actions. Human intervention requirements should increase with uncertainty, irreversibility, scale, and potential harm.

Discussion and Analysis

AI Evolution as Capability Accumulation

The literature supports interpreting AI evolution as **capability accumulation**. New paradigms expand the kinds of information the system can process, the temporal horizon over which it can operate, and the level of task structure that humans must explicitly specify. This interpretation avoids a common technological-determinist mistake: assuming that a new paradigm automatically makes earlier ones obsolete.

Deterministic systems remain preferable when behavior must be strictly bounded. Conventional statistical models may outperform large foundation models when the task is narrow and high-quality structured data are available. Deep learning remains central for specialized perception problems. Generative models are especially useful where outputs are open-ended. Agentic architectures become valuable when tasks require coordinated sequences of actions. Selection should therefore be driven by task architecture rather than fashion.

This perspective also changes how organizations should evaluate AI capability. Mikalef and Gupta (2021) show that organizational AI capability depends on complementary resources, while Enholm et al. (2022) link business value to broader organizational conditions. AI adoption should consequently be interpreted as organizational transformation rather than software procurement. Haefner et al. (2021) similarly show that AI alters innovation processes, while Keding (2021) illustrates its relevance to strategic management.

Generative AI and Agentic AI Are Related but Conceptually Distinct

A key conceptual contribution of the synthesis is the distinction between **generation** and **agency**. Generative capability concerns the production of novel outputs conditioned on inputs. Agency concerns the capacity to pursue a goal over time by selecting and sequencing actions under changing conditions. A generative model can be embedded inside an agent, but the two terms should not be treated as synonyms.

Three differences are particularly important. First is **temporal horizon**. Generative interactions may be single-turn or conversational, while agents are usually designed around multi-step completion. Second is **environmental coupling**. A generative system primarily produces information; an agent can interact with external tools and alter system states. Third is **operational discretion**. Generative AI typically leaves execution to the user, whereas agentic systems may decide how to pursue the objective within delegated boundaries.

This distinction is increasingly visible in the 2024–2026 agent literature. Wang et al. (2024) describe autonomous-agent architectures around planning, memory and tool use. Xi et al. (2025) examine broader agent ecosystems, while Acharya et al. (2025) explicitly focus on autonomous goal pursuit. Recent surveys by Abou Ali et al. (2026), Chowa et al. (2026), and Bajoria et al. (2026) further emphasize action-enabled and collaborative architectures.

The Human Role Moves Upward in the Control Hierarchy

The results suggest that increases in AI agency tend to move the human role upward in a control hierarchy. When a system performs only narrow computation, the human controls the workflow directly. When a system generates content, the human increasingly defines intention and evaluates output. When an agent performs sequential execution, human effort shifts toward setting objectives, constraints, permissions, escalation rules, and evaluation criteria.

This does not necessarily imply reduced human importance. It means that the **location of human contribution changes**. Human work may become less focused on producing every intermediate step and more focused on framing problems, applying contextual judgment, identifying exceptions, verifying evidence, coordinating

stakeholders, and assuming responsibility for consequential decisions.

The delegation perspective of Baird and Maruping (2021) is therefore highly compatible with agentic AI. Delegation requires appraisal of the agent, distribution of tasks and authority, and coordination between actors. An organization should not merely ask whether an agent can technically execute a process; it must determine what rights the agent has within that process and what responsibilities remain non-delegable.

A Theory of Conditional Complementarity

The combined evidence supports a **conditional complementarity** proposition. Human–AI integration creates the greatest value when the comparative strengths of human and artificial actors are matched to appropriate subtasks.

AI systems typically offer advantages in computational scale, consistency, rapid information processing, pattern extraction, and generation of many alternatives. Humans retain advantages in contextual interpretation, normative reasoning, social legitimacy, understanding unstated organizational goals, accountability, and managing ambiguous exceptions. These are general tendencies rather than universal properties, and the boundary between them moves as technology changes.

The meta-analysis by Vaccaro et al. (2024) is particularly instructive because it undermines the simplistic expectation that human plus AI must exceed either component alone. Coordination itself has costs. Humans may disregard good recommendations, accept bad ones, spend excessive time evaluating outputs, or lose task awareness through overdependence. AI may introduce persuasive errors or outputs that require expensive verification [51-63].

The most effective unit of design is therefore not the AI model but the **human–AI workflow**. Designers should ask where human judgment has the highest marginal value, where AI provides superior scale, which errors are easiest to detect, and which stages require independent verification.

Trust Calibration as a Core Integration Mechanism

Trust connects technological capability with human behavior. Without sufficient trust, useful AI remains unused; with excessive trust, unreliable AI becomes dangerous. Glikson and Woolley (2020) and Kaplan et al. (2023) demonstrate that trust is multidimensional and context-sensitive. The design target should consequently be **calibrated reliance**.

This shifts explainability from a cosmetic objective toward a functional one. An explanation is valuable insofar as it improves the user's mental model of when the AI should or should not be relied upon. Providing more explanation is not necessarily beneficial if the explanation is difficult to understand or creates unjustified confidence. The findings of Naiseh et al. (2023) reinforce this point.

For generative and agentic systems, confidence should therefore be supplemented by provenance, external evidence, verification checkpoints, explicit uncertainty where meaningful, and mechanisms for escalating unfamiliar situations. This is especially important because fluent language can create an impression of competence that exceeds factual reliability.

From Output Governance to Action Governance

Agentic AI requires a major shift in responsible-AI thinking. Governance of conventional predictive or generative systems frequently focuses on outputs: Is the prediction accurate? Is the generated text biased? Is the recommendation explainable? Agentic systems require **action governance**.

Action governance concerns what a system may do, not only what it may say. Relevant controls include tool whitelisting, transaction limits, credential scoping, approval gates, logging, reversibility, sandboxing, rate limits, environmental constraints, anomaly detection, and mandatory human escalation. A model hallucination becomes substantially more consequential when it can trigger external actions.

The appropriate degree of autonomy can be represented conceptually as $A = f(C, V, R, E)$, where A denotes permissible autonomy, C is the consequence of error, V is the verifiability of the output or action, R is action reversibility, and E is environmental uncertainty. This expression is a heuristic relationship rather than an empirically estimated predictive equation, and it is intended to clarify how governance conditions should shape delegated autonomy.

Higher consequences, lower verifiability, lower reversibility, and greater uncertainty should reduce permissible autonomous discretion and increase human-control requirements. Accordingly, high-impact deployments should combine restricted permissions with stronger verification, human escalation, and auditable control mechanisms (Gohil et al., 2026; Ramaswamy and Nayak, 2026).

The ECRIG Framework

The findings can be integrated into the proposed ECRIG framework. Evolution describes the technological trajectory from explicit rules toward learned and increasingly action-enabled systems, while Capability captures what AI can do across prediction, perception, generation, reasoning, planning, memory, tool use, and execution.

Role describes how the human position changes as capability expands, including movement from direct operation toward collaboration, delegation, supervision, and governance. Integration then concerns how tasks, information, authority, feedback, and responsibility are allocated across human and artificial actors.

Governance concerns the mechanisms that keep the resulting socio-technical system reliable, accountable, secure, and aligned with legitimate human objectives. Taken together, these five dimensions connect technological change to organizational role design and to the controls needed for responsible deployment.

These dimensions are not independent. Increasing capability changes feasible human roles. Changing human roles creates new integration requirements. Greater delegation expands governance requirements. Governance, in turn, may deliberately constrain technical capability to maintain acceptable risk.

A concise representation of the framework is AI Evolution → Expanded Capability → Reconfigured Human Roles → New Integration Architecture → Increased Governance Requirements. The arrows indicate a dominant analytical direction rather than a one-way causal law, because governance and institutional requirements also feed back to determine which capabilities may legitimately be deployed.

Theoretical Implications

The ECRIG framework extends the automation–augmentation discussion by adding **agency and governance**. The automation–augmentation paradox explains why organizations often need both machine substitution and human complementarity (Raisch and Krakowski, 2021). Agentic AI adds another question: once a task is automated, how much discretion should the artificial actor receive over how the objective is achieved?

The framework also extends delegation theory. Baird and Maruping (2021) conceptualize delegation between humans and agentic information-system artifacts. Contemporary AI agents make this increasingly operational because they can accept high-level goals, determine intermediate steps, and use external tools. Delegation theory may therefore become central to understanding AI-enabled organizations.

Finally, ECRIG provides a bridge between human-centered AI and agentic systems. Shneiderman's human-centered perspective stresses the possibility of simultaneously high automation and meaningful human control (Shneiderman, 2020a, 2020b). This principle is especially important for AI agents because autonomy should not be equated with absence of governance.

Practical Implications for Organizations

Organizations considering generative or agentic AI should avoid beginning with the question, “Where can AI replace employees?” A more analytically useful sequence is: Which tasks can be decomposed? Which steps are information intensive? Which require social legitimacy or expert accountability? Which errors are easily reversible? Which actions should require human approval? Which capabilities should be restricted?

Organizations also need AI literacy beyond prompt engineering. Users should understand model limitations, verification practices, data sensitivity, hallucination risk, and the consequences of connecting AI to external tools. Managers need additional capabilities in workflow design, control architecture, auditability, and responsibility allocation.

Algorithmic management research also warns that AI-enabled control can alter power relations within organizations. Kellogg, Valentine and Christin (2020) show that algorithms can restructure workplace control and contestation. AI governance should therefore address not only technical safety but also the organizational distribution of authority, information, and discretion.

Future Research Agenda

Several research priorities emerge from the synthesis. First, scholarship needs stronger measures of **agentic autonomy**. Describing a system simply as an “AI agent” conceals substantial differences between systems that suggest actions, systems that execute within restricted tools, and systems that autonomously coordinate long-horizon workflows.

Second, longitudinal research is required on human capability. Productivity improvements measured shortly after AI adoption may not reveal whether long-term dependence strengthens learning, transfers expertise, or causes skill deterioration. The distributional findings of Noy and Zhang (2023) and Brynjolfsson et al. (2025) suggest that AI may narrow performance differences in some tasks, but longer-term capability effects remain an important research frontier.

Third, multi-agent systems require investigation at the system level. When multiple specialized agents negotiate, critique, plan, and execute, emergent errors and coordination problems may arise even when each component behaves plausibly. Research should therefore examine system-level robustness and not only single-agent benchmark performance.

Fourth, responsible-AI scholarship should increasingly distinguish **content risk from action risk**. Hallucinations in a chatbot, biased outputs in a recommendation interface, and erroneous actions executed by an autonomous agent are not governance-equivalent events.

Fifth, cross-cultural and institutional research is needed. Trust, acceptable delegation, authority, privacy expectations, and preferences for human oversight may differ across organizations and societies. Human-AI integration theory cannot assume that governance norms derived from one institutional context apply universally.

Conclusion

The evolution of artificial intelligence is best understood not simply as a sequence of technological inventions but as a progressive expansion of computational capability and operational agency. AI has moved from systems dominated by explicitly encoded rules toward technologies that learn representations, generate content, interpret complex goals, construct plans, use external tools, and increasingly execute multi-step activities.

This evolution changes the human role rather than simply eliminating it. Human participation moves from direct operation toward problem framing, collaboration, verification, delegation, supervision, exception management, and governance. The most significant question is therefore no longer whether a human or an AI should perform an entire task. The relevant question is how capabilities, authority, information, responsibility, and verification should be distributed across a human-AI system.

The review further shows that complementarity is conditional. Combining a human with an AI does not guarantee superior outcomes. Successful integration depends on task structure, relative competence, error detectability, workflow design, expertise, feedback, and appropriate reliance. Trust should consequently be calibrated to actual system capability rather than maximized.

The proposed **ECRIG framework—Evolution, Capability, Role, Integration, and Governance**—provides an integrated way of interpreting these developments. It suggests that the governance burden increases as AI moves from generating recommendations toward executing actions with real-world consequences.

Future AI systems will therefore require more than increasingly capable models. They will require well-designed socio-technical architectures in which autonomy is matched to risk, human judgment is preserved where consequential, actions remain auditable and reversible where possible, and accountability remains clearly defined. The central challenge of agentic AI is ultimately not to remove humans from intelligent systems but to determine how increasingly capable artificial agents can operate within systems of meaningful human purpose, control, and responsibility [60-63].

References

1. Abou Ali M, Dornaika F, Charafeddine J (2026) Agentic AI: a comprehensive survey of architectures, applications, and future directions. *Artificial Intelligence Review* 59: 11
2. Acharya DB, Kuppan K, Divya B (2025) Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. *IEEE Access* 13: 18912–18936.
3. Alzubaidi L, Zhang J, Humaidi AJ (2021) Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data* 8: 53.
4. Baird A, Maruping LM (2021) The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly* 45: 315–341.
5. Bajoria S, Ranjan S, MA (2026) From Language Models to Agentic AI: A Survey of Autonomous, Action-Enabled, and Collaborative LLM Agents. *Cognitive Computation* 18: 103.
6. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115.
7. Braun V, Clarke V (2021) One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18: 328–352.
8. Brynjolfsson E, Li D, Raymond L (2025) Generative AI at Work. *Quarterly Journal of Economics* 140: 889–942.
9. Byrne D (2022) A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality & Quantity* 56: 1391–1412.
10. Chowa SS, Alvi R, Rahman SS (2026) From language to action: A review of large language models as autonomous agents and tool users. *Artificial Intelligence Review* 59: 71.
11. Dazeley R, Vamplew P, Foale C, Young C, Aryal S, Cruz F (2021) Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence* 299: 103525.
12. Doshi AR, Hauser OP (2024) Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10: eadn5290.
13. Dwivedi YK (2021) Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 57: 101994.
14. Dwivedi YK (2023) "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management* 71: 102642.
15. Enholm IM, Papagiannidis E, Mikalef P, Krogstie J (2022) Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers* 24: 1709–1734.
16. Feuerriegel S, Hartmann J, Janiesch C, Zschech P (2024) Generative AI. *Business & Information Systems Engineering* 66: 111–126.
17. Flathmann C, Schelble BG, Rosopa PJ, McNeese NJ, Mallick R, Madathil KC (2023) Examining the impact of varying levels of AI teammate influence on human-AI teams. *International Journal of Human-Computer Studies* 177: 103061.
18. Floridi L, Chiriatti M (2020) GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30: 681–694.
19. Gallegos IO, Rossi RA, Barrow J (2024) Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50: 1097–1179.
20. Gao C, Lan X, Li N et al. (2024) Large language models empowered agent-based modeling and simulation: A

- survey and perspectives. *Humanities and Social Sciences Communications* 11: 1259.
21. Gaube S, Suresh H, Raue M et al. (2021) Do as AI say: Susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4: 31.
22. Glikson E, Woolley AW (2020) Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals* 14: 627–660.
23. Gohil V, Shah SB, Rauniyar K, Shiwakoti S, Thapa S, Sheng D, Thapa L, Johnson KT, Naseem U (2026) Towards safe and trustworthy agentic AI: Foundations, taxonomy, technologies, applications, and future directions. *Artificial Intelligence Review*.
24. Haefner N, Wincent J, Parida V, Gassmann O (2021) Artificial intelligence and innovation management: A review, framework, and research agenda. *Technological Forecasting and Social Change* 162: 120392.
25. Hagendorff T (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* 30: 99–120.
26. Han K, Wang Y, Chen H et al. (2023) A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45: 87–110.
27. Harris-Watson AM, Larson LE, Lauharatanahirun N, DeChurch LA, Contractor NS (2023) Social perception in Human-AI teams: Warmth and competence predict receptivity to AI teammates. *Computers in Human Behavior* 145: 107765.
28. Hauptman AI, Schelble BG, McNeese NJ, Madathil KC (2023) Adapt and overcome: Perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior* 138: 107451.
29. Huang L, Yu W, Ma W (2025) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43: Article 42.
30. Jo H, Park D-H (2024) Effects of ChatGPT's AI capabilities and human-like traits on spreading information in work environments. *Scientific Reports* 14: 7806.
31. Kaplan AD, Kessler TT, Brill JC, Hancock PA (2023) Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors* 65: 337–359.
32. Kasneci E et al. (2023) ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103: 102274.
33. Keding C (2021) Understanding the interplay of artificial intelligence and strategic management: Four decades of research in review. *Management Review Quarterly* 71: 91–134.
34. Kellogg KC, Valentine MA, Christin A (2020) Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals* 14: 366–410.
35. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M (2022) Transformers in Vision: A Survey. *ACM Computing Surveys* 54: Article 200.
36. Kiani A, Uyumazturk B, Rajpurkar P (2020) Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digital Medicine* 3: 23.
37. Kraus S, Breier M, Lim WM et al. (2022) Literature reviews as independent studies: Guidelines for academic practice. *Review of Managerial Science* 16: 2577–2595.
38. Lin T, Wang Y, Liu X, Qiu X (2022) A survey of transformers. *AI Open* 3: 111–132.
39. Lund BD, Wang T, Mannuru NR, Nie B, Shimray S, Wang Z (2023) ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74: 570–581.
40. Matsuo Y, LeCun Y, Sahani M (2022) Deep learning, reinforcement learning, and world models. *Neural Networks* 152: 267–275.
41. Mikalef P, Gupta M (2021) Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management* 58: 103434.
42. Mohseni S, Zarei N, Ragan ED (2021) A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* 11: 1–45.
43. Naiseh M, Al-Thani D, Jiang N, Ali R (2023) How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies* 169: 102941.
44. Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381: 187–192.
45. Paul J, Criado AR (2020) The art of writing literature review: What do we know and what do we need to know? *International Business Review* 29: 101717.
46. Raees M, Meijerink I, Lykourantzou I, Khan V-J, Papangelis K (2024) From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies* 189: 103301.
47. Raisch S, Krakowski S (2021) Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review* 46: 192–210.
48. Ramaswamy S, Nayak D (2026) Bounded Autonomy and Verifiable Safety for Agentic AI Enabled Automation. *Journal of Intelligent & Robotic Systems*.
49. Sarker IH (2021) Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science* 2: 160.
50. Schmutz JB, Outland N, Kerstan S, Georganta E, Ulfert A-S (2024) AI-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology* 58: 101837.
51. Seeber I, Bittner E, Briggs RO (2020) Machines as teammates: A research agenda on AI in team collaboration. *Information & Management* 57: 103174.
52. Shneiderman B (2020a) Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction* 36: 495–504.
53. Shneiderman B (2020b) Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction* 12: 109–124.
54. Tschandl P, Rinner C, Apalla Z (2020) Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26: 1229–1234.
55. Vaccaro M, Almaatouq A, Malone T (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8: 2293–2303.
56. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL (2023) ChatGPT: Five priorities for research. *Nature* 614: 224–226.
57. Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76: 89–106.
58. Wang L, Ma C, Feng X (2024) A survey on large language model based autonomous agents. *Frontiers of Computer*

- Science 18: 186345.
59. Wang P, Ding H (2024) The rationality of explanation or human capacity? Understanding the impact of explainable artificial intelligence on human-AI trust and decision performance. *Information Processing & Management* 61: 103732.
60. Westphal M, Vössing M, Satzger G, Yom-Tov GB, Rafaeli A (2023) Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior* 144: 107714.
61. Xi Z, Chen W, Guo X (2025) The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68: 121101.
62. Xu W, Huang C, Gao S (2025) LLM-Based Agents for Tool Learning: A Survey. *Data Science and Engineering* 10: 533–563.
63. Zhang G, Chong L, Kotovsky K, Cagan J (2023) Trust in an AI versus a Human teammate: The effects of teammate identity and performance on Human-AI cooperation. *Computers in Human Behavior* 139: 107536.