RITAM ISH PUBLICATIONS

Journal of Progress in Artificial Intelligence
https://www.jpairitamish.org/

ISSN: 3139-4957 (Online)

Vol. 2, Iss. 3, July 2026 – September 2026

Human-AI Collaboration and the Renegotiation of Managerial Authority: An Integrative Critical Review with a Multi-Case Analytical Framework for Agentic AI Adoption in Industry

Dasaradha Arangi[1], Sreemanthula Jhansi[2], Dr. Minhaz Husain[3], Dr. Yuvraj Khare[4], Balveer Singh[5]

[1]Department of CSE (AI & ML), Aditya Institute of Technology and Management, Tekkali, Srikakulam, Andhra Pradesh, India

[2]Department of Management, Sun International Institute College, Andhra Pradesh, India

[3]Department of Economics, Khwaja Moinuddin Chishti Language University, Lucknow, Uttar Pradesh, India

[4]Associate Professor, J.S. College of Law, Prithvipur, Niwari, Madhya Pradesh, India

[5]Research Scholar, School of Studies in Law, Jiwaji University, Gwalior, Madhya Pradesh, India

Received: 18-July-2026, Manuscript No. JPAI-2026-0021; **Editor assigned:** 19-July-2026, PreQC No. JPAI-2026-0021 (PQ); **Reviewed:** 30-July-2026, QC No. JPAI-2026-0021; **Revised:** 06-Aug-2026, Manuscript No. JPAI-2026-0021 (R); **Published:** 20-Aug-2026

Citation: Arangi, D., Jhansi, S., Husain, M., Khare, Y., & Singh, B. (2026). Human-AI Collaboration and the Renegotiation of Managerial Authority: An Integrative Critical Review with a Multi-Case Analytical Framework for Agentic AI Adoption in Industry. J Prog Artif Intell.

Corresponding: Dasaradha Arangi, Email: dasaradha9999@gmail.com

ABSTRACT

The spread of agentic Artificial Intelligence (AI) systems that can autonomously plan, use tools, and perform multi-step tasks instead of offering single recommendations challenges long-held views on

OPEN ACCESS | REVIEW ARTICLE

where managerial authority lies within organisations. Using a PRISMA-guided search across databases like Scopus, Web of Science, ABI/INFORM, Google Scholar, and industry-specific repositories, this comprehensive review synthesises research on agentic and generative AI, human-AI cooperation, algorithmic management, trust and explainability, decision-rights theory, and technology adoption. It explores how increasing AI autonomy shifts decision-making, accountability, and control from human managers to computational agents. Rather than merely summarising, the review critically organises findings around a five-stage transition from AI as a tool to an assistant, collaborator, decision agent, and finally an organisational actor, highlighting areas of consensus, debate, and gaps, especially in accountability, managerial discretion, and industry differences in autonomy governance. Drawing on theories such as agency, sociotechnical systems, and adoption models (TOE, TAM, UTAUT), the paper introduces a conceptual multi-case framework comparing agentic AI adoption in sectors such as healthcare, finance, manufacturing, retail, logistics, professional services, and IT. It proposes a model linking key factors influencing authority shifts and posits ten propositions for future testing. The review provides a shared vocabulary for authority shifts in AI contexts, extending control theories into AI environments, and guiding managers, boards, and policymakers in developing delegation, escalation, and accountability strategies. While the framework is conceptual and propositional without empirical validation, it outlines a future research agenda aimed at understanding how organisations can maintain meaningful human oversight as AI agents play a larger decision-making role.

KEYWORDS: Agentic Artificial Intelligence, managerial authority, human-AI collaboration, algorithmic management, decision rights, organizational control, human oversight, AI governance

1. INTRODUCTION

Over the past decade, artificial intelligence has evolved through several distinct organisational roles: a back-office analytical tool, a decision-support layer embedded in dashboards, a generative assistant capable of drafting and summarising, and, most recently, an agentic system capable of planning multi-step tasks, invoking external tools, and executing actions with limited direct human supervision [1,2]. Agentic AI is commonly defined by its capacity for autonomy, reactivity, proactivity, and learning attributes that distinguish it from earlier generations of decision-support and generative systems, which primarily responded to discrete prompts rather than pursuing goals across extended sequences of action [1]. Industry surveys suggest this shift is no longer speculative: enterprise adoption of AI agents has accelerated sharply, with large panels of executives reporting active deployment or near-term plans across customer service, supply chain, finance, and knowledge-work functions [3,4].

This evolution matters for management theory because it challenges a premise that much organizational and leadership research has taken for granted: that decision rights, discretion, and accountability are held by human actors situated within a formal hierarchy. Conventional AI assistance, such as spell-checkers, recommendation engines, and dashboards, left this premise largely intact because the human manager remained the point of initiation, interpretation, and final action. Agentic AI complicates this premise

because the system itself increasingly initiates action, interprets context, and adapts its own plan of execution, sometimes without a discrete moment of human authorisation for each step [1,2,5]. The distinction is not merely technical; it is organisational and normative, because authority in classical management theory is inseparable from the capacity to decide and to be held accountable for the consequences of that decision. A growing but fragmented body of research has begun to document the organisational consequences of this shift. Studies of algorithmic management show that when prescriptive algorithms take over managerial functions such as task assignment, performance evaluation, and scheduling, workplace control begins to operate through mechanisms that restrict, recommend, record, rate, replace, and reward, which differ qualitatively from traditional supervisory control [6]. Parallel work argues that algorithmic outputs can function not only as tools of surveillance but also as legitimate bases for decision-making in their own right, giving rise to hybrid regimes of managerial authority in which discretion is distributed unevenly across human managers, data infrastructures, and computational models [7]. At the same time, labour-law and governance scholarship has proposed rights-based responses, including explanation, contestation, and human review, intended to preserve a floor of human oversight as algorithmic and, increasingly, agentic systems take on managerial functions [8]. A second, closely related body of literature examines the psychological and interactional dynamics of human-AI collaboration. This work documents two opposing failure modes: Automation bias, in which humans over-rely on AI outputs and defer even when the system is wrong, and algorithm aversion, in which humans discount accurate AI recommendations after observing even minor errors [9,10]. Neither failure mode is a simple function of individual psychology; both are shaped by task design, explanation quality, and the extent to which humans retain meaningful authority to contest or modify AI outputs [10,11]. In this literature, explainability and human oversight are increasingly treated not as substitutes but as complementary and mutually reinforcing components of accountable AI governance, since oversight without explanation risks collapsing into rubber-stamping, while explanation without effective oversight authority does little to change outcomes [12].

A third stream draws on economic and organisational theories of delegation. Agency theory, originally developed to explain the separation of ownership and control when firms delegate decision rights to professional managers [13,14], has recently been extended to characterise AI systems as prospective "agents of the firm," raising the question of how principal-agent alignment mechanisms monitoring, incentive design, bounded discretion must be reconceived when the agent is not human and cannot be motivated by conventional incentives [15]. A related reformulation, delegated agency theory, focuses on the fidelity between the delegation a user believes they authorised and the behaviour an agentic AI system subsequently exhibits, arguing that user trust depends on perceived alignment with authorised goals, constraints, and remedies rather than on outcome quality alone [16]. Information systems scholarship has similarly proposed frameworks for delegation to and from "agentic IS artifacts" that treat AI systems as capable of exercising a bounded form of agency within organisationally defined limits [17]. A fourth stream, situated in innovation and information-systems research, addresses organizational readiness and adoption. The Technology-Organization-Environment

(TOE) framework [18], the Technology Acceptance Model (TAM) [19], and the Unified Theory of Acceptance and Use of Technology (UTAUT) [20] have each been applied, individually and in combination, to explain why some organizations adopt AI more readily than others. However, recent work argues that these frameworks, developed for earlier waves of information technology, require adaptation for example, through integration with digital-maturity and dynamic-capabilities perspectives to capture the distinctive uncertainty and autonomy characteristics of agentic AI [21].

Despite the depth of these separate literatures, they remain only loosely connected. Research on algorithmic management is grounded primarily in labour process theory and organisational control; research on human-AI teaming is grounded in cognitive science and human factors; research on AI delegation is grounded in agency theory and information systems; and research on adoption is grounded in innovation diffusion. Few studies integrate these strands to ask a single question directly: as AI systems become more autonomous, how and under what conditions is managerial authority itself being renegotiated, rather than simply assisted or automated? Reviews of agentic AI to date have tended to be technical or sector-specific, cataloguing architectures, tools, and applications rather than developing an organisation-theoretic account of authority, accountability, and power [1,2]. Conversely, reviews of algorithmic management have tended to focus on worker-level outcomes surveillance, precarity, procedural justice rather than on the higher-order question of how the institutional feature of authority itself is reconfigured when computational systems participate in its exercise [7]. This fragmentation leaves both scholars and practitioners without an integrative vocabulary for describing the stages through which AI moves from assisting managerial judgement to substituting for it, and without a comparative framework for understanding why this transition proceeds at different speeds and with different governance requirements across industries. This article addresses that gap through an integrative critical review structured around six research questions: (1) how is agentic AI changing the nature and boundaries of managerial authority; (2) how does human-AI collaboration redistribute decision-making power within organizations; (3) what organizational, technological, and institutional factors influence managerial acceptance of and delegation to agentic AI; (4) how do accountability, trust, transparency, and human oversight change as AI systems become more autonomous; (5) what differences emerge across industries in the adoption and governance of agentic AI; and (6) what theoretical framework can explain the renegotiation of managerial authority in AI-enabled organizations. In pursuing these questions, the review makes four contributions. First, it develops an integrative critical synthesis that traces a five-stage transition in the locus of managerial authority from AI as a tool to AI as an organisational actor and shows how each stage implicates managerial discretion, decision rights, control, accountability, expertise, and trust in different ways. Second, it develops an original multi-case analytical framework comparing agentic AI adoption across seven industry contexts, using publicly available, non-fabricated evidence rather than invented organizational case studies. Third, it proposes a conceptual model linking antecedents, mediating mechanisms, moderators, and outcomes of authority renegotiation, together with ten theoretically derived propositions to guide future empirical research. Fourth, it identifies the boundary

conditions under which delegation to agentic AI is more or less likely to be accepted by managers, and translates these into practical implications for AI governance, human oversight design, and organizational policy. The review does not claim that any element of the proposed framework has been empirically validated; it is offered explicitly as a conceptual and propositional contribution intended to organize and stimulate future empirical work.

1.1 Examination of the Methodology for the Review and the Strategy for Literature Search

Because the credibility of an integrative review rests on the transparency of how its evidence base was assembled, this section documents the search, screening, and synthesis procedures used to identify the literature underlying Sections 2 through 5, before those findings are presented. Consistent with Torraco's and Snyder's methodological guidance for integrative and narrative-synthesis reviews, this article is explicitly not a systematic review in the Cochrane or PRISMA sense its purpose is conceptual integration and theory building across disconnected literatures rather than an exhaustive, protocol-registered retrieval of every study meeting narrow inclusion criteria [38,39]. Nonetheless, the search and screening steps below were conducted using PRISMA-informed reporting conventions, adapted for an integrative-review purpose, so that readers can evaluate the scope and possible boundaries of the evidence base rather than take its comprehensiveness on faith [40]. Databases and search period. We collected literature through systematic searches of platforms such as Scopus, Web of Science Core Collection, ABI/INFORM, and Google Scholar. These were complemented by targeted searches of SSRN and publicly accessible research repositories from MIT Sloan Management Review, Boston Consulting Group, McKinsey, and Deloitte, focusing on industry-survey evidence discussed in Section 4. The search covered publications from January 2018 to July 2026, with an emphasis on 2023–2026, since agentic AI, as explained in Section 2.1, is a recent development. Key foundational theories such as agency theory, TAM, TOE, UTAUT, and organizational-trust theory, were identified via citation tracing from more recent work rather than through date-limited searches, due to their status as established theoretical frameworks rather than emerging empirical results. Search strings. Search terms combined three concept blocks using Boolean AND/OR logic: (a) an AI-system block ("agentic AI" OR "AI agent*" OR "autonomous AI" OR "generative AI" OR "algorithmic management" OR "algorithmic decision-making"); (b) an organizational-authority block ("managerial authority" OR "decision rights" OR "human-AI collaboration" OR "human-AI teaming" OR "organizational control" OR "delegation" OR "accountability"); and (c) a governance/adoption block ("human oversight" OR "explainability" OR "trust" OR "technology adoption" OR "agency theory"). Blocks (a) and (b) were required to co-occur in the title, abstract, or keywords of retrieved records; block (c) was used to narrow high-volume results and to identify literature for Sections 2.5 and 2.7 specifically.

Screening and eligibility. Records were screened in two stages. Title/abstract screening excluded purely technical computer-science literature with no organizational, managerial, or governance dimension (for example, papers concerned solely with agent architectures, planning algorithms, or benchmark performance), non-English-language sources, and sources lacking identifiable authorship or peer review, with the partial exception of clearly labelled industry-

survey reports used specifically for the descriptive adoption statistics in Section 4. Their non-peer-reviewed status is flagged throughout the text and in the Limitations section, rather than treated as equivalent evidence to peer-reviewed sources. Full-text screening then assessed the remaining records against three inclusion criteria: (i) substantive engagement with managerial authority, decision rights, control, accountability, trust, or adoption in an organisational context; (ii) sufficient methodological or theoretical transparency to support citation (an identifiable method, argument, or dataset); and (iii) relevance to at least one of the six research questions stated at the end of Section 1. Records failing any of these three criteria were excluded. Figure 1, described immediately below, summarises this flow.

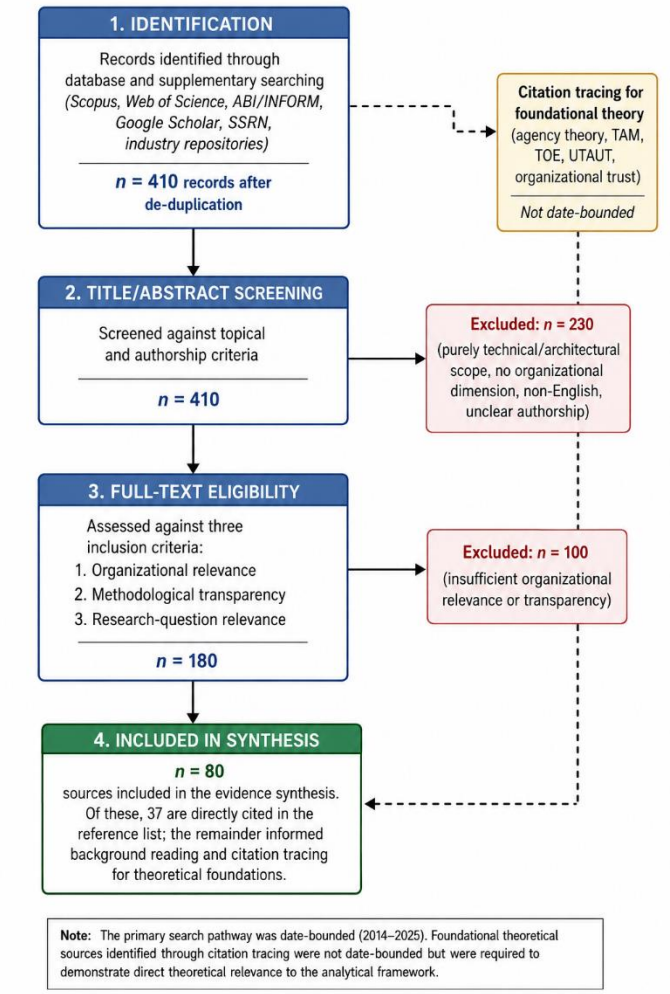

Figure 1. Literature Identification and Screening Process

The flow diagram outlines the procedures for identifying, screening, assessing full-text eligibility, and including literature in the integrative review.

The main search process comprised database searches and supplementary scholarly searches, followed by screening titles, abstracts, and full texts against predefined eligibility criteria. A separate citation-tracing method was used to identify key foundational theories, including agency theory, the Technology Acceptance Model (TAM), the Technology–Organization–Environment (TOE) framework, the Unified Theory of Acceptance and Use of Technology (UTAUT), and literature on organisational trust. These foundational sources identified through citation tracing were not constrained by the empirical search period but had to demonstrate direct relevance to the analytical framework.

Quality appraisal and synthesis. Because the reviewed literature spans heterogeneous methodologies systematic reviews, empirical experiments, survey research, conceptual/theoretical articles, and industry reports no single quantitative quality-appraisal instrument (such as the Mixed Methods Appraisal Tool) was applied uniformly across source types. Instead, each source's evidentiary weight was assessed qualitatively against source-appropriate criteria (peer review status, sample and method transparency for empirical work, argumentative rigour for conceptual work, and methodological disclosure for industry reports). This weighting is made explicit in-text wherever a claim rests primarily on industry rather than peer-reviewed evidence. Synthesis followed thematic, narrative-synthesis procedures consistent with integrative-review methodology [38,39]: extracted material was coded against the six research questions and the constructs, then organised into the five-stage framework (Section 3) and the ten analytical dimensions of the multi-case comparison (Section 4). Coding categories were developed iteratively as recurring constructs (decision rights, trust basis, accountability attribution, oversight mechanism) emerged across literatures that do not otherwise share a common vocabulary. This coding process facilitated integration across the literature, instead of providing a side-by-side summary that sets Sections 3 to 5 apart from the discipline-specific literature review in Section 2.

To improve bibliographic accuracy and reproducibility, all references in the final list were independently verified against authoritative publisher records, indexing databases (such as Scopus, Web of Science, and PubMed where applicable), or recognized preprint repositories. The verification aimed to confirm correct author names, publication titles, journal or venue, publication year, volume and issue details when available, and DOI or other persistent identifiers. Any duplicate records or discrepancies found during verification were addressed before finalizing the list. Specifically, one duplicate citation was replaced with a relevant, independent source, and an author attribution issue in another reference was corrected after checking against the publisher record. These steps help to reduce bibliographic errors, ensure sources are retrievable, and ensure consistency between in-text citations and the reference list.

2. LITERATURE REVIEW

2.1 Evolution from traditional AI to generative and agentic AI

Organisational applications of AI have evolved through several overlapping generations. Early rule-based and expert systems encoded predefined decision logic and offered limited adaptability. Machine-learning-based decision-support systems introduced statistical prediction but generally required a human to

interpret outputs and initiate action. Deep learning expanded the range of tasks that could be automated, particularly in pattern recognition, but decision-support systems of this kind remained fundamentally reactive: they responded to a query or triggered an alert, and a human retained the discretion to act. Generative AI, associated with large language and multimodal models, extended this reactivity into open-ended content production, but a single generative exchange still typically concludes with a human reviewing and acting on the output. Agentic AI differs from earlier types by four main features identified in the literature: autonomy, reactivity, proactivity, and ongoing learning within a task environment [1]. Unlike systems that respond to a single prompt, agentic systems can plan multiple steps, utilise external tools and APIs, observe their own results, and adjust their subsequent actions [2,5]. This ability for autonomous planning and action distinguishes AI assistance initiated and ended by humans from AI agency, a sequence of system-driven or sustained actions aimed at a goal. Recent reviews and narrative surveys highlight this shift across various technical architectures (like planning modules, memory systems, tool-use interfaces, multi-agent coordination) and application areas, including customer service, supply-chain exception handling, and financial workflows [1,2,4]. For business and management audiences, the transition is framed more in organisational terms, viewing agentic AI as altering the boundary between human judgment and automation, creating tensions regarding oversight, workforce roles, and value capture that leaders must actively manage rather than considering it just a technical change [3]. It is useful, for the purposes of this review, to distinguish four organisational postures toward AI along a rough continuum of autonomy: (a) AI assistance, in which the system supplies information but the human performs all interpretation and action; (b) AI augmentation, in which the system actively shapes the human's judgment (for example, through recommendation or explanation) but the human retains and exercises final discretion; (c) AI automation, in which the system executes a bounded, well-specified task without human involvement in each instance, but within narrow parameters set in advance by humans; and (d) AI agency, in which the system exercises discretion over how to pursue a broader goal, including decisions about sub-goals, tool use, and sequencing, typically subject to ex post rather than ex ante human review. This continuum is not strictly linear in practice; organisations frequently operate several postures simultaneously across different functions, but it provides an organising vocabulary for the authority-renegotiation argument developed later in this review.

2.2 Human-AI collaboration

A significant body of literature explores how humans and AI systems collaborate in decision-making, often using terms like human-in-the-loop, human-on-the-loop, and human-in-command to describe varying levels of human participation in automated processes. More recent research redefines this relationship as human-AI teaming, where one or more humans and AI agents work together toward common goals, treating it as an active, adaptable process rather than a static setup [22]. This shift carries normative implications: calling AI a teammate instead of a simple tool does not suggest human-like agency but emphasises the importance of clearly defining authority and responsibility between humans and machines. It also highlights the necessity of designing explicit protocols for escalation and disagreement such as confidence

thresholds that prompt human oversight or requirements for consensus in high-stakes decisions instead of relying on implicit divisions of labor [22]. The central theoretical construct in this literature is complementarity: the condition under which a human-AI team outperforms either the human or the AI system operating alone. Achieving complementarity is not automatic. It depends on team composition, calibrated trust, shared mental models of each party's competence, and task structures that assign responsibilities based on comparative strength rather than convenience or habit [22]. Trust itself is not static; empirical work tracking trust over time within human-AI teams shows that trust calibration is a dynamic process shaped by the pattern and salience of errors, not a fixed disposition formed at first exposure to a system [23]. Two opposing failure modes recur throughout this literature. Algorithm aversion is the tendency of decision-makers to discount or abandon an algorithm after observing it err, even when it remains statistically superior to unaided human judgment. Experimental work shows that this aversion can be substantially reduced when users are given even limited ability to adjust or override the algorithm's recommendations, suggesting that perceived control over a decision not merely its accuracy shapes willingness to rely on it [9]. Automation bias, the opposite failure, is uncritical over-reliance on automated outputs, including a documented tendency for even highly experienced professionals to adopt incorrect AI-generated conclusions in high-stakes domains such as clinical diagnosis: a controlled reader study found that incorrect AI-generated suggestions measurably degraded the diagnostic accuracy of inexperienced, moderately experienced, and very experienced radiologists alike, with even very experienced readers' correct-classification rate falling from roughly 82% to 46% when the AI suggestion was wrong, demonstrating that automation bias is not simply a function of inexperience [11,24]. Work on perceived controllability suggests that the two failure modes are related manifestations of the same underlying problem: when users cannot see, and cannot meaningfully act on, the boundary between what the AI decided and what remains open for human judgment, they either over-defer or under-defer, but rarely calibrate appropriately [25]. Explanation quality mediates this relationship: experimental evidence indicates that AI explanations improve complementary team performance only under specific conditions, and can, in some configurations, increase over-reliance rather than reduce it, underscoring that explainability is not an unconditional good but a design variable that must be evaluated against its effect on human decision behaviour, not merely against technical interpretability criteria [26]. Collectively, this literature establishes that human-AI collaboration is not a fixed state achieved once trust is established, but an ongoing governance challenge that requires continuous calibration of authority boundaries, explanation strategies, and escalation protocols a conclusion directly relevant to how managerial authority is renegotiated as the AI component of the collaboration becomes more autonomous.

2.3 Managerial authority and organizational power

Classical organization theory treats managerial authority as a formally sanctioned right to direct subordinates' actions and to make binding decisions on behalf of the organization, typically legitimated by hierarchical position, expertise, or a combination of both. The introduction of algorithmic and, increasingly, agentic systems into managerial functions unsettles this account in at least three ways,

as documented in recent literature. First, algorithmic systems have begun to perform core managerial tasks, including assigning work, evaluating performance, scheduling, and disciplining, functions historically understood as constitutive of managerial authority rather than merely supportive of it. A widely cited synthesis of this literature identifies six mechanisms through which algorithmic control operates in the contested terrain between employers and workers, arguing that algorithmic technologies do not simply automate existing supervisory practices but introduce qualitatively new forms of control with distinct implications for worker autonomy and resistance [6]. Second, recent conceptual work explicitly distinguishes algorithmic authority from algorithmic control, arguing that algorithmic outputs can serve as a legitimate basis for organizational decisions not merely as instruments through which human managers monitor or discipline workers and proposing a typology of hybrid managerial-authority regimes defined by the balance struck between human managerial discretion and algorithmic decision-making within a given organization [7]. This distinction matters because it reframes the debate away from a binary of human control versus algorithmic surveillance and toward a spectrum of hybrid authority configurations that organizations must deliberately design rather than allow to emerge by default. Third, a growing body of work links algorithmic management to broader organizational and labour consequences, including the erosion of interpersonal managerial relationships, dehumanisation, and the intensification of precarious work arrangements in gig and platform contexts, even as some workers report a preference for the perceived objectivity and consistency of "algorithmic bosses" over inconsistent human supervisors [27]. This ambivalence the simultaneous erosion of relational authority and the appeal to algorithmic objectivity is a recurring and unresolved tension in the literature, and one that the integrative synthesis in Section 3 revisits explicitly.

2.4 Algorithmic and AI-enabled decision-making

A parallel stream distinguishes Algorithmic Decision-Making (ADM) as a broader category encompassing descriptive, predictive, and prescriptive algorithms, with algorithmic management, the automation of specifically managerial tasks, as a subset [10]. A systematic review following PRISMA guidelines and examining ninety articles on ADM in organizations identifies four theoretical strands underpinning this literature: organizational and management theory, trust-in-algorithms research, justice and ethics theories, and decision theory grounded in computational rationality. It also develops an integrated tension-alignment framework intended to reconcile the frequently mixed and contradictory findings reported across these strands [10]. This fragmentation across theoretical traditions helps explain why the literature has produced inconsistent conclusions about whether ADM improves or degrades decision quality, procedural fairness, and employee trust: studies operating from different theoretical premises often measure different outcomes and apply different normative standards. A complementary lifecycle perspective considers algorithmic decision-making systems as sociotechnical organisational entities, where their effects develop through design, deployment, and use stages. Ethical and organizational issues such as accountability and managerial responsibility cannot be fully addressed if each phase is examined in isolation [28]. This approach corrects views that treat algorithmic authority as a static feature of the deployed system,

instead emphasising it as an evolving organizational relationship. It also anticipates the multi-stage process of authority renegotiation discussed in Section 3. Early foundational work saw algorithmic management as a sociotechnical process that results from ongoing interactions between platform algorithms and the humans who use, resist, or bypass them, rather than as a fixed technical artifact imposed unilaterally on an organization [29].

2.5 Trust, explainability, transparency, and accountability

Trust in automated and algorithmic systems has long been theorized using models originally developed for interpersonal and organizational trust. The influential integrative model of organizational trust identifies perceived ability, benevolence, and integrity as the three antecedents of trust in an exchange partner [30]. This framework has since been adapted to human-automation interaction through the construct of process trust, defined as confidence that an automated system followed an appropriate procedure, distinct from outcome trust, which is confidence based purely on the correctness of results [31]. This distinction is directly relevant to agentic AI, because agentic systems increasingly make procedural choices which sub-goals to pursue, which tools to invoke, and in what order that are not fully visible to the humans nominally overseeing them, making process trust harder to establish through observation alone. Explainability and human oversight are often treated as substitutes, yet recent expert commentary argues they are best understood as complementary and mutually reinforcing components of accountable AI governance: oversight without adequate explanation risks reducing human overseers to rubber-stamping machine-generated decisions, while explanation without genuine oversight authority the practical ability to contest, modify, or halt a decision does little to alter organisational outcomes [12]. This conclusion is reinforced by regulatory developments; the European Union's framework legislation for artificial intelligence explicitly requires that high-risk AI systems incorporate mechanisms for human oversight, record-keeping, and meaningful explanation of decisions, treating these as jointly necessary rather than independently sufficient safeguards [32]. Governmental and standards-setting bodies have converged on a broadly similar view: risk-management guidance for trustworthy AI identifies accountability, transparency, explainability, and human oversight as interdependent, rather than freestanding, characteristics of responsible AI systems [33]. A persistent theme across this literature is the responsibility gap, the concern that as AI systems exercise greater autonomous discretion, it becomes progressively harder to attribute moral and organisational responsibility for outcomes to any single human actor, since neither the system designer, the deploying manager, nor the AI itself straightforwardly satisfies the conditions traditionally required for accountability. Addressing this gap is less a matter of increasing technical transparency alone than of designing organisational mechanisms, such as audit trails, escalation thresholds, and defined override authority, that specify, in advance, where human accountability attaches as AI autonomy increases. This concern connects directly to the agency-theoretic literature discussed in Section 2.7 and motivates the accountability dimension of the multi-case framework developed in Section 4.

2.6 Agentic artificial intelligence and organizational autonomy

As agentic systems move beyond narrow task automation toward planning and executing multi-step workflows, their organisational footprint extends into areas historically reserved for managerial judgement: task and resource allocation, exception handling, and increasingly, elements of strategic and operational coordination. Industry analyses describe organisations moving through discernible maturity stages from defining an AI vision and piloting narrow use cases to operationalising governed agentic systems embedded in core workflows such as supply-chain exception management, where AI agents shift human roles from manual, case-by-case intervention to supervision and optimisation of an AI-driven process [4]. This shift illustrates a broader pattern documented across the agentic AI literature: autonomy does not simply eliminate the managerial role but relocates it, from direct execution to the design, calibration, and periodic audit of the autonomous system's operating parameters a form of "management of management," in which human managers increasingly manage the conditions under which AI agents exercise a delegated form of managerial discretion [3,4]. Expert panels convened to study responsible AI implementation report that accountability for agentic systems is widely regarded by senior leaders as harder to operationalize than accountability for earlier, narrower automation, precisely because the chain of causation between a managerial decision (to deploy or configure an agent) and an organisational outcome (an action the agent takes autonomously) is longer and less directly observable [34]. This concern is echoed in survey-based industry research, which indicates that governance and control rather than technical capability are increasingly identified as the primary constraint on scaling agentic AI deployment beyond pilot projects [35].

2.7 Organizational readiness, adoption theories, and agency-theoretic perspectives

Several established theoretical frameworks have been applied, with varying degrees of adaptation, to explain organizational readiness for AI adoption. The Technology-Organization-Environment (TOE) framework examines adoption as a function of technological characteristics (relative advantage, compatibility, complexity), organizational characteristics (leadership support, resources, readiness), and environmental characteristics (competitive pressure, regulatory context, vendor support) [18]. TOE has been widely applied to AI adoption, including in regulated and public-sector settings, where technological, organizational, and environmental challenges have been shown to interact systemically rather than operate as independent barriers [36]. Its principal limitation, repeatedly noted in the literature, is that TOE operates at the organizational level and does not capture individual managers' psychological readiness to accept and rely on AI-generated recommendations or actions [37]. The Technology Acceptance Model addresses this individual-level gap by explaining adoption through perceived usefulness and perceived ease of use [19], while the Unified Theory of Acceptance and Use of Technology (UTAUT) extends this individual-level account by incorporating performance expectancy, effort expectancy, social influence, and facilitating conditions as joint predictors of technology acceptance and use [20]. TAM has been criticized for insufficient attention to organizational and environmental context, and TOE has been criticized in turn for insufficient behavioral and psychological depth, motivating a

growing number of hybrid TOE-UTAUT and TOE-TAM models intended to explain both organizational-level and individual-level adoption of AI systems simultaneously [37]. Recent work further argues that static adoption frameworks, however hybridized, are poorly suited to capturing the continuous, iterative nature of agentic AI adoption, and proposes integrating adoption constructs with dynamic-capabilities theory and digital-maturity models to represent adoption as an ongoing process of sensing, seizing, and reconfiguring organizational resources rather than a discrete accept/reject decision [21]. None of these frameworks, in their original or hybridized forms, was designed to address a technology capable of exercising discretion after adoption; this is a gap the present review returns to in Section 3.

Agency theory offers a distinct and, for the purposes of this review, especially consequential lens. In its classical formulation, an agency relationship exists when a principal delegates decision rights to an agent expected to act on the principal's behalf. This delegation reduces the cost of transferring decision-relevant knowledge but simultaneously introduces agency costs arising from misaligned interests and imperfect monitoring [13,14]. Recent scholarship extends this framework directly to AI, arguing that the integration of AI into firm decision-making parallels the historical emergence of the professional manager, the original impetus for agency theory, and theorizing a developmental point at which an AI system achieves sufficient autonomy and self-determination to be considered an agent of the firm in a meaningful, rather than metaphorical, sense. At that point, conventional alignment mechanisms designed for human agents (incentive contracts, monitoring, reputational sanctions) become inapplicable and must be reconceived [15]. A complementary reformulation, delegated agency theory, shifts the analytical focus from outcome quality to delegation fidelity the degree of alignment between the delegation contract a human principal believes they authorized and the behaviour the AI agent subsequently exhibits and identifies contract-specification quality, capability-context fit, governance visibility, and recovery capacity as the design dimensions that determine whether delegation to agentic AI is experienced by users as legitimate and trustworthy [16]. Information-systems research has proposed a parallel theoretical vocabulary for delegation to and from "agentic IS artifacts," treating the degree of agency exercised by a system as a variable to be designed and bounded rather than a fixed technical given [17].

Taken together, Sections 2.1 through 2.7 establish that the literature has separately documented (a) the technical and organisational emergence of agentic AI, (b) the psychological dynamics and failure modes of human-AI collaboration, (c) the reconfiguration of managerial authority under algorithmic management, (d) the theoretical fragmentation of algorithmic decision-making research, (e) the interdependence of trust, explainability, and accountability, (f) the extension of managerial autonomy into agentic systems, and (g) both adoption-theoretic and agency-theoretic accounts of organisational readiness for AI delegation. What remains underdeveloped is an integrative account that connects these strands into a single explanatory structure describing how and through what mechanisms managerial authority itself is renegotiated as AI systems move along the autonomy continuum described in Section 2.1. Section 3 develops this synthesis directly.

3. Integrative Critical Synthesis: From AI Assistance to Renegotiated Managerial Authority

The literature reviewed above, considered collectively, supports the argument that no single stream of research fully articulates: organizations are progressing through a five-stage transition in the locus and configuration of managerial authority. Across these stages, the growing role of AI progressively reshapes managerial discretion, decision rights, control mechanisms, accountability, expertise, leadership, trust, and organizational power. The following framework integrates these previously fragmented perspectives into a sequential analytical model of this transition. Figure 1 presents the proposed five-stage progression and illustrates how the locus of managerial authority evolves as AI moves from a predominantly assistive role to increasingly agentic forms of organizational intelligence.

Stage 1 — AI as Tool. AI serves as a passive instrument invoked at the manager's discretion (spreadsheets, dashboards, basic analytics). Authority remains entirely with the human manager; the system has no capacity for initiative. This stage is well described by early decision-support and management-information-systems research and is not contested in the literature.

Stage 2 — AI as Assistant. Generative and predictive systems offer recommendations, drafts, or forecasts based on human queries, with explainability research gaining importance because managers increasingly base decisions on the assistant's outputs, while still holding final discretion [12,26]. At this stage, automation bias first appears, as even advisory systems can lead to over-reliance [11].

Stage 3 — AI as Collaborator. In this stage, the system engages in an iterative and adaptive process within a shared task, aligning with the human-AI teaming literature that highlights interdependent and flexible collaboration over mere request-response exchanges [22]. Decision-making rights start to be divided rather than kept fully in one entity, with the aim of complementarity as the primary design principle. Trust evolves to become dynamic, requiring continuous adjustment rather than a one-time assessment [23].

Stage 4 — AI as Decision Agent. The system exercises bounded autonomy executing multi-step workflows, invoking tools, and adapting its own plan within parameters set by managers but without step-by-step authorization [1,2,4]. This is the stage at which algorithmic authority, as distinct from algorithmic control, becomes organizationally salient [7] the system's outputs function as an operative basis for action, not merely as an input to human judgment. Managerial work correspondingly shifts from execution toward calibration, exception review, and periodic audit [4].

Stage 5 — AI as Organizational Actor. The system is treated, functionally if not legally, as an agent of the firm in the agency-theoretic sense, capable of initiating and managing decisions on the firm's behalf without direct human supervision of each instance. This requires alignment mechanisms that are not merely technical but organizational and governance-based [15,16]. This stage remains largely prospective in the literature reviewed here: most empirical accounts of current practice describe organizations operating at Stages 3–4, with Stage 5 treated as a theorized horizon rather than a documented empirical condition. This review does not claim that Stage 5 is currently realised in practice; it is presented as the endpoint of a

trajectory the literature increasingly anticipates rather than as an established fact.

Table 1, presented immediately below, summarises this five-stage evolution and its managerial implications, synthesising the constructs discussed in Sections 2.1, 2.3, 2.6, and 2.7.

Table 1. The Evolution of Artificial Intelligence from Automation to Agentic AI.

AI Stage	Core Capability	Degree of Autonomy	Human Role	Managerial Implication
Tool	Data processing, retrieval, calculation	None — fully reactive	Sole initiator and actor	Authority unaffected; AI is instrumental only
Assistant	Recommendation, drafting, forecasting	Low — advisory only	Evaluator and final decision-maker	Trust calibration and explainability become salient [12,26]
Collaborator	Iterative, adaptive joint task execution	Moderate — shared initiative	Partner in a partitioned division of labor	Decision rights are explicitly partitioned; complementarity becomes a design goal [22,23]
Decision Agent	Multi-step planning, tool use, bounded execution	High — autonomous within set parameters	Calibrator, auditor, exception-handler	Algorithmic authority emerges; managerial work shifts to oversight design [1,4,7]
Organizational Actor	Independent initiation and management of decisions	Very high — approaching functional agency	Governance designer and principal	Agency-theoretic alignment mechanisms required; largely prospective [15,16]

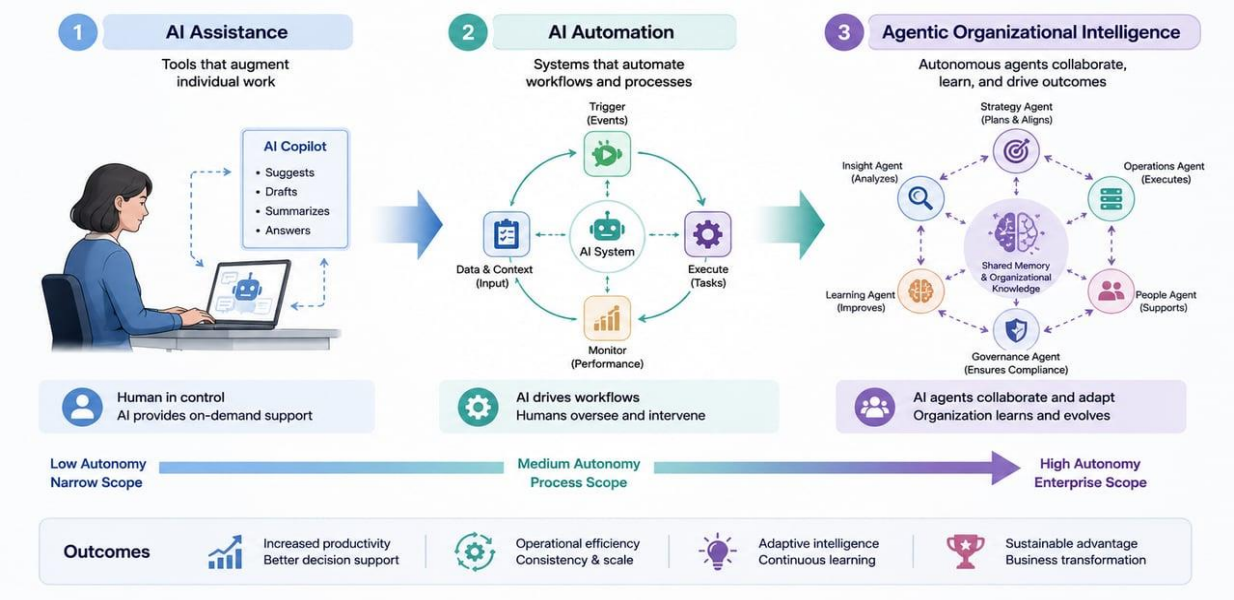

Figure 2. Evolution from AI Assistance to Agentic Organizational Intelligence

Design Specification. A horizontal five-stage continuum diagram runs left to right, mirroring Table 1. Five equally sized rounded rectangles are arranged left to right and connected by a single bold horizontal arrow beneath them, labeled "Increasing Autonomy →". Boxes, in order: (1) "Tool" (icon: gear), (2) "Assistant" (icon: speech bubble), (3) "Collaborator" (icon: two overlapping circles), (4) "Decision Agent" (icon: branching path), (5) "Organizational Actor" (icon: building/org-chart node). Beneath each box, a smaller secondary label states the corresponding human role from Table 1 ("Sole actor," "Evaluator," "Partner," "Calibrator/Auditor," "Governance designer"). Above the continuum, a shaded gradient band (light to dark) visually reinforces increasing autonomy from left to right. A vertical dashed line is placed between boxes 4 and 5, labeled "Empirically documented ↔ Largely prospective," explicitly marking that Stage 5 is theorized rather than established in current practice, consistent with the caveat in Section 3. Small citation markers [1,2,7,15] are placed beneath the relevant stages to anchor each stage in the reviewed literature.

3.1 Positioning the five-stage model against prior staged accounts of human–automation interaction

A five- or ten-point staged continuum of human–machine interaction is not, in itself, a novel device in this literature, and claiming novelty for staging alone would overstate the contribution. Sheridan and Verplank's original ten-level scale of human–computer control and Parasuraman, Sheridan, and Wickens' widely cited four-function, ten-level model of types and levels of automation established the underlying logic of staged autonomy continua decades before the emergence of agentic AI and remain important reference points against which newer staged models should be evaluated [43,44]. Similarly, the human-in-the-loop, human-on-the-loop, and human-in-command taxonomy used in regulatory and human-factors literature captures different configurations of human responsibility and system autonomy, whereas Raisch and Krakowski's automation–augmentation paradox frames the tension between replacing and augmenting human judgment rather than the degree of system autonomy [45]. The five-stage model developed here therefore differs from these precedents not merely because it has stages, but through three cumulative distinctions. First, the model stages organizational authority rather than task-level automation. Classical automation models primarily ask how much of a discrete function is performed by the system versus the human. By contrast, the stages in Table 1 are defined by configurations of decision rights, the basis for trusting system outputs, and the attribution of accountability, as operationalized in Table 3. A system may therefore exhibit a high level of automation for a narrowly defined function while remaining at Stage 2 (Assistant) on the organizational-authority dimension developed here. The two scales measure related but analytically distinct constructs; this decoupling is an important conceptual implication of the present synthesis rather than an assumption that all forms of automation necessarily entail equivalent shifts in authority. Second, the model explicitly links stages of authority to agency-theoretic and sociotechnical constructs that are particularly relevant to organizational decision-making with AI. These include algorithmic authority, distinct from algorithmic control [7], delegation fidelity [16], and the responsibility gap [12]. These constructs enable the model to move beyond the question of how much work a system performs to the organizational question of when system outputs begin to reshape human decision

rights, discretion, and accountability. This distinction is particularly important for agentic AI, whose organizational significance may arise not simply from increased task execution but from its capacity to initiate, coordinate, recommend, and execute decisions across interconnected workflows. Third, and most importantly, the model establishes an explicit empirical–prospective boundary. Stages 1–4 are grounded in configurations documented in the existing literature, whereas Stage 5 is deliberately theorized as a prospective configuration rather than presented as an already established empirical condition (Section 3; Figure 1). The dashed boundary in Figure 1 therefore signals a transition from empirically documented forms of algorithmic involvement to a hypothesized configuration of organizational authority. The model is consequently intended not merely as a taxonomy of already-deployed systems, but as a framework for locating and evaluating an emerging transition in the distribution of decision rights between humans and increasingly agentic systems. Taken together, these distinctions define the model's contribution to the staged-model tradition: it stages the transformation of organisational authority rather than the degree of task automation; integrates agency-theoretic, accountability, and sociotechnical constructs into that staging; and explicitly distinguishes empirically documented configurations from a prospective form of agentic authority. The contribution therefore lies in the analytical dimension staged and in the integration of previously separate theoretical constructs, rather than in the five-part structure itself.

What is established, contested, missing, and contributed.

Reading the literature through this staged lens reveals substantial convergence alongside important unresolved questions. What is established is that AI systems are increasingly incorporated into organisational decision processes [1,2,3]; that automation bias and algorithm aversion are opposing but persistent failure modes that require active management rather than passive reliance on either trust or distrust [11,24,25]; that explainability and meaningful human oversight are complementary safeguards rather than straightforward substitutes [12]; and that algorithmic systems already perform functions historically associated with managerial authority, particularly within the Stage 2–4 configurations identified here [6,7]. What remains contested is equally important. The literature has not yet established whether algorithmic and agentic decision-making systematically improves or degrades decision quality and procedural fairness. A systematic review of ninety studies on automated decision-making reports heterogeneous findings that reflect differences in theoretical assumptions, contexts, and outcome measures rather than a settled empirical consensus [10]. Similarly, workers and managers may experience algorithmic authority either as a source of consistency and objectivity or as a source of dehumanization and alienation, with evidence supporting both responses and limited agreement on the conditions under which each predominates [27]. Finally, agency-theoretic accounts raise an unresolved problem about how alignment and governance mechanisms developed for human agents should be adapted when the delegated actor is a non-human system that cannot be motivated through conventional incentive structures [15]. What is missing is an integrated account of mid-level managerial authority. Much of the algorithmic-management literature focuses on frontline worker monitoring and control, while much of the emerging agentic-AI literature

concentrates either on technical architecture or on senior-executive and strategic applications. The resulting literature leaves comparatively under-theorized the ways in which middle managers renegotiate discretion, judgment, escalation, and accountability as algorithmic systems move from decision support toward delegated decision authority. Three additional gaps follow: limited comparative analysis of how authority transitions vary across industries with different risk profiles and regulatory environments; insufficient integration of adoption-theoretic and agency-theoretic explanations of these transitions; and a lack of empirically validated measures of constructs such as delegation fidelity and algorithmic authority. What this article contributes is therefore fourfold: an explicit five-stage vocabulary for conceptualizing the transition in organizational authority (Table 1); a multi-case comparative framework for examining industry variation (Section 4); a conceptual model integrating adoption-theoretic and agency-theoretic constructs to address the mid-level managerial authority gap (Section 5); and a set of falsifiable propositions intended to make delegation fidelity and algorithmic authority empirically tractable (Section 6). The model should consequently be understood as a framework for organising, integrating, and extending existing literatures around a specific unresolved organisational transition, rather than as a claim that staged representations of human–automation interaction are themselves novel.

4. Multi-case analytical framework

To address the industry-variation gap identified above, this section presents an original, analytically structured comparison of agentic AI adoption across seven industry contexts: healthcare, finance and banking, manufacturing, retail, logistics and supply chain, professional services, and information technology. Consistent with the instructions guiding this review, these are presented as industry-level analytical cases, synthesised from published industry and survey evidence [3,4,34,35], rather than as verified organisational case studies; no company-specific claims are asserted, and no findings are fabricated. Industry survey research indicates substantial cross-sector variation in adoption maturity: financial services and healthcare are frequently identified as leading sectors in piloting activity, while regulatory scrutiny is identified as the primary brake on moving agentic systems into production in finance specifically [35]; manufacturing and logistics are identified as leading sectors for supply-chain and operational agent deployment, reflecting well-defined, rule-governed processes that are comparatively easier to bound and audit [4,35]. These patterns are treated here as illustrative of broader dynamics rather than as precise, generalisable statistics, given the fast-moving and heterogeneous nature of the underlying survey evidence.

Two industry rows required additional, rigorously sourced evidence beyond what industry-survey material alone could provide. In healthcare, a peer-reviewed scoping review of 43 studies on AI agents supports the pattern in Table 2: most systems are categorised as workflow/automation assistants and multimodal decision-support agents, rather than autonomous diagnostic agents. The review's authors noted that most studies were exploratory, methodologically limited, and lacked strong clinical validation, with autonomous clinical decision-making remaining with clinicians in nearly all systems-providing direct peer-reviewed support for the Stage 2–3 characterizations in Table 2, rather than relying solely on industry surveys [41]. For professional

services, independent sector-specific survey data as of early 2026 shows about 15% of organizations have adopted some form of agentic AI, with another 53% actively planning or evaluating adoption. Additionally, governance and approval processes in partnership-led firms progress more slowly than in corporate legal departments, aligning with the' professional liability norms slow progression to Stage 4" description in Table 2. This slower pace is linked to the governance structure- partnership-based, consensus-driven- which provides an institutional explanation for the delay, rather than viewing it as an unexplained sector-level trend [42].

For each industry case, Table 2 (introduced next) synthesizes ten analytical dimensions: AI autonomy, decision complexity, human oversight requirements, managerial delegation, organisational readiness, risk level, accountability requirements, trust requirements, regulatory constraints, and the expected transformation of managerial authority, drawing on the theoretical constructs developed in Sections 2 and 3.

Table 2. Multi-Case Industry Comparison

Industry	Typical AI Applications	Autonomy Potential	Decision Risk	Human Oversight	Authority Implication
Healthcare	Diagnostic support, workflow triage, clinical	Moderate — high-stakes decisions remain	Very high (patient safety)	Mandatory, structured, and clinically accountable [34,41]	Authority stays formally human-held;
Finance / Banking	Fraud detection, credit-risk scoring, reconciliation, compliance reporting	Moderate-to-high in back-office functions; constrained in customer-facing credit decisions	Very high (regulatory, systemic)	Extensive, audit-driven, regulator-facing [35]	Regulatory pressure accelerates back-office authority shift (Stage 3–4) while constraining front-office autonomy
Manufacturing	Predictive maintenance, quality control, production scheduling agents	High in bounded operational domains	Moderate-to-high (safety, downtime)	Engineering-defined thresholds and escalation rules	Authority shifts toward calibration and exception management (Stage 4) within well-specified operational envelopes
Retail	Demand forecasting, dynamic pricing, customer-service agents	High in customer-facing and inventory domains	Low-to-moderate	Business-rule-based, less regulator-driven	Rapid Stage 3–4 transition; managerial authority reallocates toward exception handling and

Industry	Typical AI Applications	Autonomy Potential	Decision Risk	Human Oversight	Authority Implication
					brand-risk oversight
Logistics / Supply Chain	Exception handling, route optimization, procurement agents	High — identified as a leading domain for agentic deployment [4,35]	Moderate (cost, service-level)	Workflow-embedded, with human review of high-value exceptions [4]	Clearest empirical example of Stage 3-to-4 transition; managers shift from manual triage to supervising autonomous workflows [4]
Professional Services	Legal/consulting research agents, document drafting and review	Moderate — output-facing autonomy, advisory authority remains human	High (reputational, legal liability)	Professional and ethical accountability regimes	Authority renegotiation concentrated in Stage 2–3; professional liability norms slow progression to Stage 4 [42]
Information Technology	Code generation and review agents, DevOps and incident-response agents	Very high — technically mature tooling ecosystem [1,2]	Moderate-to-high (system reliability, security)	Technical review, staged autonomy, rollback mechanisms	Among the most advanced sectors for Stage 4 adoption, given mature tool-use infrastructure [1,2]

This comparison shows that autonomy potential and actual authority transformation are distinct: sectors with high technical autonomy potential (finance, IT) may still show constrained authority transformation in specific functions due to regulatory or liability exposure, while sectors with moderate technical sophistication (logistics) may show comparatively advanced authority transformation because decision risk is lower and processes are more amenable to bounding. This distinction between technical capability and organizationally sanctioned authority is a central analytical contribution of the multi-case comparison and is developed further in the conceptual model below, where it is captured through the moderating role of risk and regulation.

5. Conceptual framework

Building on the integrative synthesis (Section 3) and the multi-case comparison (Section 4), this section introduces an original conceptual model of how managerial authority is renegotiated when agentic AI is involved. Reflecting its conceptual and propositional nature, the model has not been empirically validated; instead, it combines constructs from the reviewed literature into a framework designed to guide future empirical research.

The model categorises constructs into four groups. Antecedents include AI capability- covering the technical sophistication and reliability of the agentic system [1,2]; organizational readiness and digital maturity- based on TOE and dynamic-capabilities

frameworks [18,21]; task complexity and environmental uncertainty; regulatory pressure, especially relevant in finance and healthcare [32,35]; and managerial AI literacy, which influences individual acceptance as described by TAM and UTAUT [19,20]. Mediating mechanisms involve pathways through which antecedents lead to authority renegotiation, including trust- measured through ability-benevolence-integrity and process-trust constructs [30,31]; perceived AI competence; decision delegation, viewed *via* agency theory and delegated-agency perspectives [15,16]; human-AI coordination quality, drawing from the literature on complementarity and teaming [22,23]; and perceived control, identified as a key factor in willingness to rely on a system by the algorithm-aversion research [24,25]. Moderators elements that influence the strength of the antecedent-mediator-outcome relationships include risk level and task criticality, which vary significantly by industry as shown in the multi-case analysis in Section 4; industry regulation, which in finance and healthcare cases can disconnect technical autonomy from actual authority changes; AI explainability, seen in trust and oversight studies as essential alongside human oversight [12]; organizational culture; and managerial experience. Outcomes involve decision quality, managerial effectiveness, organizational agility, innovation, employee autonomy, accountability, and overall organizational performance. Figure 2, introduced here, depicts this model as a path structure: antecedents flow into mediating mechanisms, which are moderated by contextual factors, and jointly determine organizational outcomes, with an explicit feedback loop from outcomes back to organizational readiness, reflecting the dynamic-capabilities argument that adoption is iterative rather than a single decision point [21].

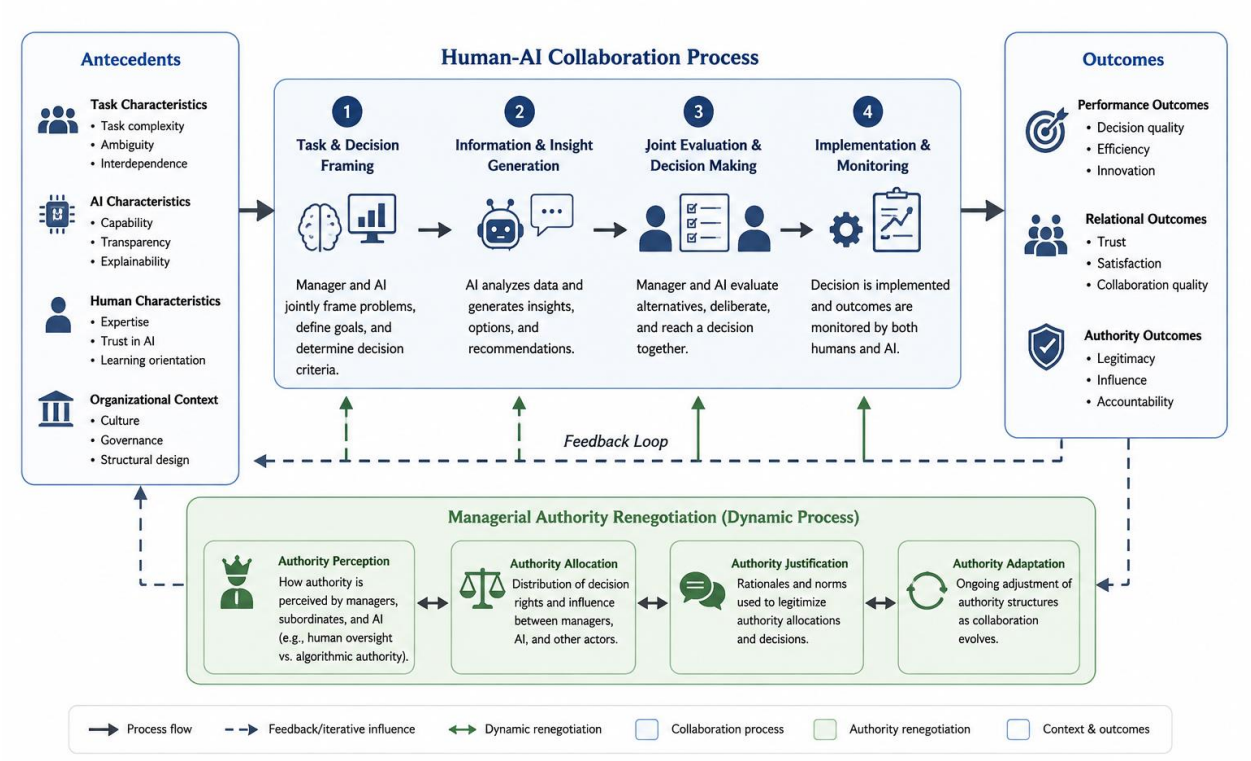

Figure 3. Conceptual Model of Human-AI Collaboration and Managerial Authority Renegotiation

A second figure disaggregates the mechanisms specifically responsible for redistributing authority, since the model above treats "managerial authority renegotiation" as a single focal construct that, examined more closely, decomposes into several distinct sub-mechanisms documented across the literature review.

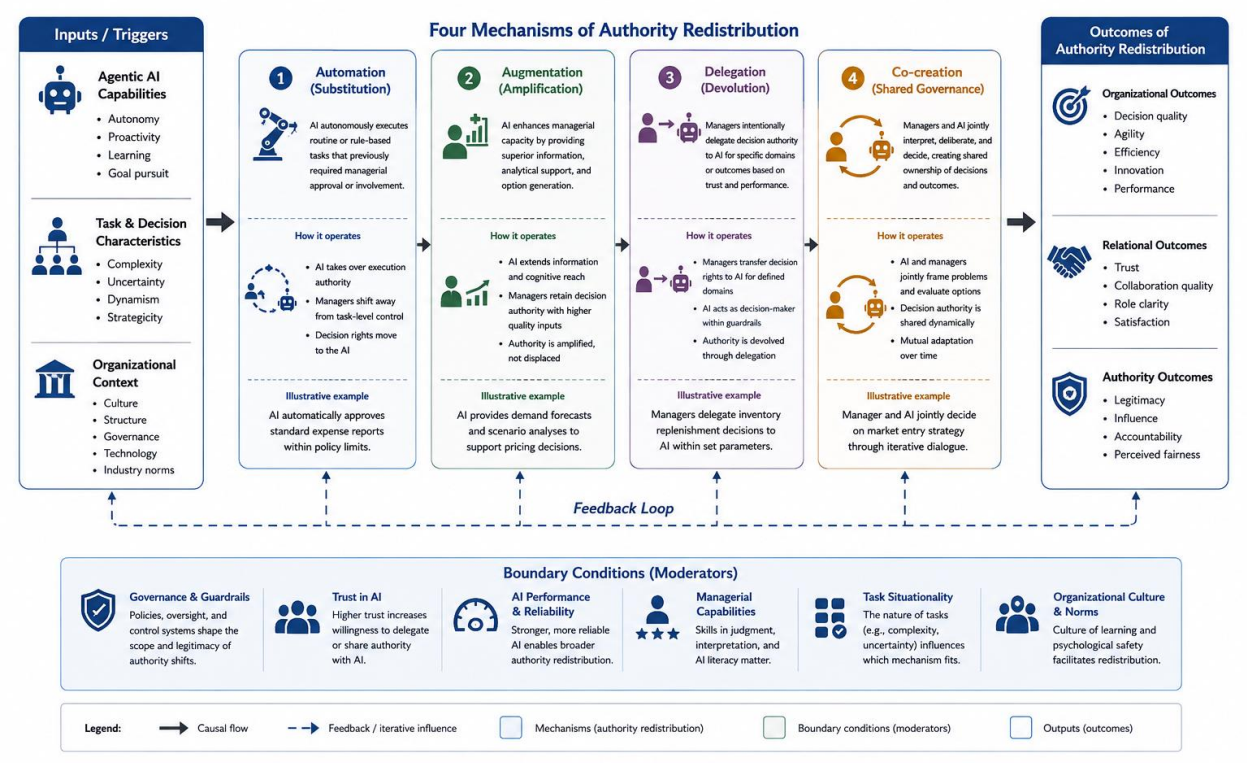

Figure 4. Mechanisms of Managerial Authority Redistribution Under Agentic AI

The proposed figure uses a radial (hub-and-spoke) structure with a central hub labelled "Managerial Authority Renegotiation" and four spokes, each terminating in a labelled mechanism box with a two- to three-word sub-label drawn from the literature review. Spoke 1 represents "Decision-Rights Partitioning", with the sub-label "who decides which sub-goals" [17,22]. Spoke 2 represents "Algorithmic Authority Legitimation", with the sub-label "outputs as a basis for action, not just an input" [7]. Spoke 3 represents "Delegation Fidelity Monitoring", with the sub-label "contract specification, governance visibility" [16]. Spoke 4 represents "Accountability Re-attachment", with the sub-label "closing the responsibility gap via audit and escalation design" [12,32,33]. Each spoke is represented by a double-headed arrow, indicating that authority may flow toward the AI system through delegation or be reclaimed by the human manager through escalation, override, or audit. This bidirectional structure reflects the bounded and reversible nature of delegation emphasized in the agency-theoretic literature. A thin outer ring connects the four mechanism boxes and is annotated "jointly determine the organization's position on the five-stage continuum (Figure 1 / Table 1)", thereby visually

linking the mechanisms to the staged model developed in Section 3.

Table 3, introduced here, translates the conceptual model into a comparative statement of how key dimensions of managerial authority differ across traditional, AI-augmented, and agentic AI contexts, synthesizing constructs from Sections 2.3, 2.6, and 3.

Table 3. Principal Aspects of Managerial Authority in Agentic Artificial Intelligence

Dimension	Traditional Management	AI-Augmented Management	Agentic AI Context	Emerging Risk
Decision initiation	Human-initiated	Human-initiated, AI-informed	Frequently system-initiated within bounded goals	Loss of visibility into initiation triggers [4]
Basis of legitimacy	Hierarchical position, expertise	Position augmented by algorithmic input	Algorithmic output treated as legitimate basis for action [7]	Legitimacy conferred without commensurate accountability
Locus of discretion	Concentrated in the manager	Shared, manager retains final say	Partitioned; discretion bounded by design parameters [17,22]	Ambiguity over where discretion actually resides
Trust basis	Interpersonal, reputational	Ability-based trust in tool accuracy	Process trust in procedural fidelity of autonomous action [30,31]	Difficulty verifying procedural fidelity without full transparency
Accountability attribution	Clear — the deciding manager	Generally clear — manager acts on AI input	Contested — responsibility gap between designer, deployer, system [12]	Diffusion of accountability; "many hands" problem
Oversight mechanism	Direct supervision	Review of AI-informed recommendations	Escalation thresholds, audit trails, periodic review [4,32]	Oversight mechanisms may lag behind system autonomy

6. Propositions

These ten propositions, drawn directly from the literature and the conceptual model outlined in Section 5, serve as a foundation for future empirical testing. However, they have not been empirically validated within this review.

Proposition 1 suggests that when AI systems are perceived as more competent, managers are more inclined to delegate decision-making authority to them, especially if accountability measures are well-established [15,16,24].

Proposition 2. The relationship between AI capability and managerial delegation is likely to be mediated by process trust, specifically confidence in the procedural fidelity of the AI's actions, rather than by outcome trust alone, especially in high-stakes decision contexts [30,31].

Proposition 3. Organisations with higher digital maturity and stronger dynamic capabilities are more likely to progress along the five-stage authority continuum Table 1 at a faster rate than organisations with comparable AI capability but lower organisational readiness [18,21].

Proposition 4. Industry regulation and task criticality are likely to moderate the relationship between AI capability and authority renegotiation, such that, holding technical autonomy potential constant, more heavily regulated or higher-risk industries will progress more slowly toward Stage 4–5 authority configurations [32,35].

Proposition 5. Explainability and human oversight authority are likely to interact such that the positive effect of explainability on managerial trust and effective delegation is significantly weaker when the overseeing managers lack genuine authority to contest or override AI-generated decisions [12].

Proposition 6. Perceived control over an AI system's actions is likely to reduce algorithm aversion and increase sustained managerial reliance more than improvements in the AI system's average accuracy alone [24,25].

Proposition 7. As organisations move from Stage 3 (Collaborator) to Stage 4 (Decision Agent), the primary focus of managerial work is likely to shift measurably from direct task execution to calibration, exception handling, and periodic auditing of autonomous system behaviour [4,7].

Proposition 8. Delegation fidelity the perceived alignment between the delegation a manager believes was authorised and the AI system's subsequent behaviour is likely to be a stronger predictor of sustained managerial trust in agentic AI than the AI system's task-completion accuracy considered alone [16].

Proposition 9. Organisations that design explicit accountability-attachment mechanisms (audit trails, escalation thresholds, defined override authority) prior to deploying agentic AI at Stage 4 are more likely to sustain both regulatory compliance and employee trust than organisations that deploy comparable systems without such mechanisms [12,32,33].

Proposition 10. The relationship between agentic AI adoption and employees' perceived procedural justice is likely to be conditional on whether algorithmic authority is explicitly legitimated and communicated by management, rather than allowed to emerge informally. Unexplained legitimation is associated with greater perceived dehumanization and alienation [7,27].

Two further figures support the multi-case and future-research contributions of this review and are introduced here in sequence.

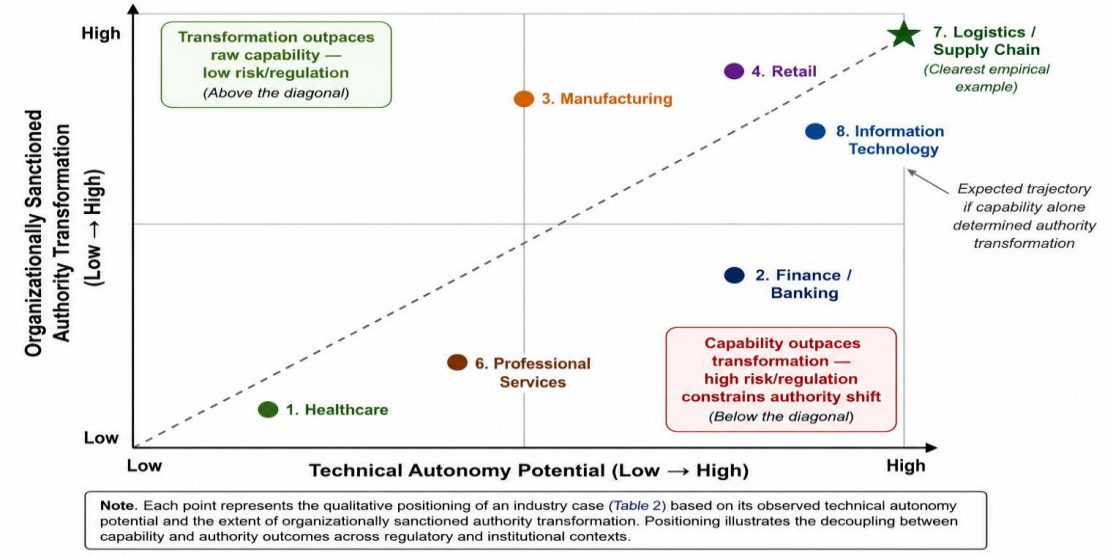

Figure 5. Multi-Case Analytical Framework for Agentic AI Adoption Across Industries

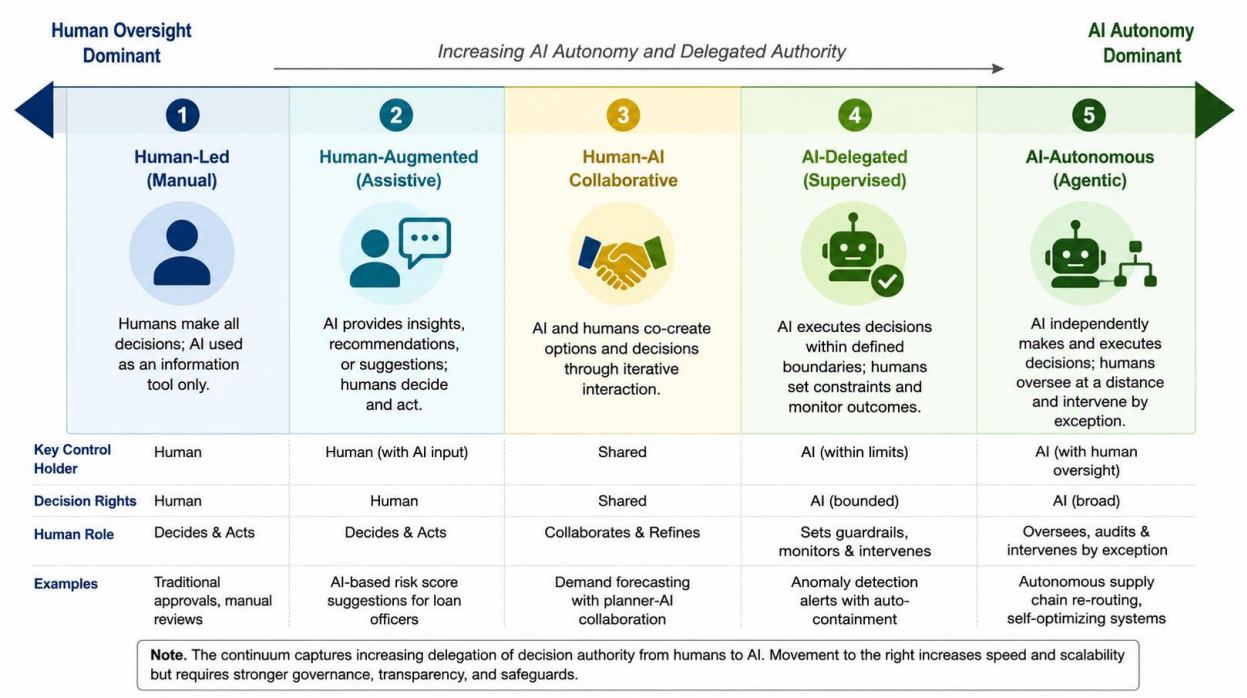

Figure 6. Continuum of Human Oversight and AI Autonomy

8. DISCUSSION

The synthesis developed in this review indicates that the central organizational consequence of agentic AI is not automation in the conventional sense the elimination of a task but the renegotiation of who, or what, holds the discretion to decide how a task is accomplished. This distinction matters because much of the popular and even scholarly discourse on agentic AI still frames its organizational significance primarily in terms of efficiency and productivity [3,4]. The literature reviewed here suggests a different, and in

some respects more consequential, framing: agentic AI is reshaping the institutional feature of authority itself, in the sense used by organization theory the socially and organizationally sanctioned right to direct action and to be held accountable for its consequences [7].

Traditional management theories, centred on the human manager as the locus of discretion and accountability, are only partially sufficient for this environment. Agency theory offers real purchase because it was designed precisely to analyse delegation under imperfect monitoring and divergent interests [13,14], and its recent extension to AI-as-agent-of-the-firm captures an important intuition: that the organisational problem posed by agentic AI resembles, in structure if not in substance, the classical problem of aligning a professional manager's actions with the interests of the firm's owners [15]. But this analogy is imperfect in a way the literature has not yet fully resolved: human agents can be motivated through incentive contracts, reputational concerns, and legal liability, none of which straightforwardly apply to an AI system. Delegated agency theory's shift towards delegation fidelity alignment with an authorised contract rather than alignment of interests is a promising response to this imperfect fit, but it has not yet been extensively tested empirically [16], and this review is unable to resolve, only to flag, the open theoretical question of what an "agency cost" means when the agent has no utility function of its own.

Human-AI collaboration differs from simple automation in another respect, as documented in Sections 2.2 and 3: it is not a stable equilibrium but an ongoing calibration problem. Trust must be continuously recalibrated as systems and tasks evolve [23]; explanation strategies must be evaluated based on their actual effect on complementary performance, not merely on interpretability criteria [26]; and the boundary between appropriate reliance and both automation bias and algorithm aversion must be actively managed through system design (for example, allowing limited user modification of algorithmic outputs) rather than assumed to self-correct [24,25]. This has a direct implication for how decision rights are redistributed: redistribution is not a single event (a deployment decision) but a continuous process requiring ongoing governance attention, consistent with the "iterative adoption feedback" loop in the conceptual model Figure 2.

Accountability becomes increasingly important, not less, as autonomy increases, a conclusion that runs somewhat against the intuitive assumption that more autonomous systems require less human involvement. The literature on the responsibility gap and on explainability-oversight complementarity indicates the opposite: as the causal chain between a managerial decision and an organizational outcome lengthens, the design of accountability-attachment mechanisms (audit trails, escalation thresholds, defined override authority) becomes more, not less, consequential [12,32-34]. Organizations can preserve meaningful human oversight not by resisting the shift towards Stage 4–5 configurations, which industry evidence suggests is already substantially underway in sectors such as logistics and IT [1,2,4,35], but by redesigning oversight from an instance-level to a system-level activity, the central claim visualised in Figure 5's U-shaped oversight-intensity curve.

Finally, industry context changes the appropriate level of AI autonomy in ways that are not fully captured by technical capability alone. The multi-case comparison in Section 4 and its visualization in Figure 4 indicate

that regulatory exposure and decision risk can decouple an industry's technical capacity for autonomous action from the degree of authority transformation that is organizationally and socially sanctioned. This decoupling is itself a finding of theoretical interest: it suggests that future research on agentic AI adoption should treat "capability" and "authorized autonomy" as analytically distinct constructs rather than conflating them, a distinction existing adoption frameworks such as TOE and TAM/UTAUT do not clearly draw [18,19,20,37].

9. THEORETICAL CONTRIBUTIONS

This review offers four principal theoretical contributions, each avowedly incremental rather than claiming unprecedented novelty.

First, it extends theories of managerial authority into AI-mediated organizational environments by proposing an explicit five-stage vocabulary (Table 1, Figure 1) that connects the algorithmic-authority literature's concern with legitimacy [7] to the agentic-AI literature's concern with technical autonomy [1,2], providing a shared framework that neither literature offers on its own.

Second, it connects human-AI collaboration research with organizational power and decision-rights literature by treating complementarity, trust calibration, and perceived control [22-25] not merely as interactional or psychological constructs but as mechanisms through which organizational authority is redistributed, linking micro-level human factors research to macro-level organization theory.

Third, it advances the conceptualization of agentic AI as a prospective organizational decision actor by integrating agency-theoretic [13-15] and delegated-agency [16] perspectives with adoption-theoretic frameworks [18-21], producing a conceptual model (Figure 2) that treats organizational readiness and agentic delegation as jointly determined rather than sequential processes.

Fourth, it identifies boundary conditions for AI delegation across industries through the multi-case comparison (Table 2, Figure 4), contributing the specific analytical distinction between technical autonomy potential and organizationally sanctioned authority transformation a distinction with direct relevance for comparative organizational research beyond the specific industries examined here.

10. PRACTICAL IMPLICATIONS

For senior managers and boards, the central implication is that authority renegotiation should be treated as a governance design problem to be managed proactively rather than an emergent byproduct of technology deployment to be addressed reactively. This includes establishing explicit AI delegation policies that specify, for each functional domain, which stage of the continuum in Table 1 the organization intends to operate at, and why.

For HR managers and people leaders, the algorithmic-management literature's findings on dehumanization, alienation, and the ambivalent worker reception of algorithmic authority [7,27] suggest that communication and legitimation of algorithmic decision-making rather than its mere technical accuracy materially shapes employee trust and procedural-justice perceptions (Proposition 10). Training that builds calibrated trust, emphasizing when to rely on AI pattern recognition versus human contextual judgment, is likely to outperform training focused solely on system accuracy [24,25].

For AI governance teams and technology leaders, the U-shaped oversight-intensity pattern in Figure 5 implies that oversight mechanisms designed for Stage 2–3 (instance-level review) will not transfer effectively to Stage 4–5 deployments; governance investment should shift toward system-level audit design, escalation-threshold calibration, and delegation-fidelity monitoring as autonomy increases [4,16].

For employees, the shift in managerial work from execution toward calibration and audit (Proposition 7) implies that AI literacy the ability to interpret, question, and escalate AI-generated outputs is likely to become a core competency requirement across a widening range of roles, not a specialized technical skill.

For policymakers, the regulatory-constraint pattern observed in finance and healthcare (Table 2, Figure 4) suggests that explanation and human-review requirements of the kind already codified in the European Union's AI Act [32] function, in practice, as accelerants of Stage 2–3 authority configurations and brakes on Stage 4–5 configurations a policy lever whose organizational consequences merit direct empirical study.

For AI developers, the finding that explanation quality affects complementary team performance conditionally, and can in some configurations increase over-reliance [26], implies that explainability features should be evaluated against their downstream effect on human decision behavior in situ, not solely against static interpretability benchmarks.

Practical mechanisms consistent with these implications include tiered AI delegation policies mapped to the five-stage continuum; structured human-oversight protocols with defined escalation thresholds; ongoing AI-literacy training tied to specific decision domains rather than generic AI awareness; explicit decision-rights redesign accompanying any Stage 3-to-4 transition; audit-trail and accountability-attachment infrastructure built prior to, not after, Stage 4 deployment [12]; periodic AI risk-classification reviews aligned with the risk dimension in Table 2; and formal escalation procedures specifying who may override an autonomous agent's action and under what conditions.

11. Ethical and Governance Considerations

The literature synthesized in this review raises several ethical and governance concerns that merit explicit treatment. Accountability remains the most consequential and least resolved of these concerns: the responsibility-gap literature indicates that autonomy without a correspondingly designed accountability architecture risks diffusing responsibility across designers, deployers, and the system itself, a "many hands" problem that existing organizational accountability structures were not designed to address [12]. Bias and fairness concerns, extensively documented in the broader algorithmic-decision-making literature, persist and in some respects intensify under agentic configurations, since a system that plans and executes multi-step sequences may propagate or compound biased intermediate judgments in ways that are harder to audit than single-decision outputs [10]. Transparency and explainability, while necessary, are explicitly not sufficient on their own to guarantee accountability or substantive fairness, a point emphasized in recent trustworthy-AI scholarship and reflected in regulatory frameworks that pair explanation requirements with human-

oversight requirements rather than treating either as independently adequate [32,33].

Privacy and security concerns arise because agentic systems that invoke external tools and access organisational data across multiple steps present a broader attack surface and a more complex data-governance challenge than single- turn systems, though a detailed technical treatment of this concern is beyond the organisational- authority focus of this review. Responsibility allocation should, per the discussion above, be designed prospectively through audit trails, escalation thresholds, and defined override authority, rather than reconstructed retrospectively after an adverse outcome. Human autonomy and employee well- being are implicated by findings on dehumanisation and alienation in the algorithmic-management literature [27], suggesting that governance frameworks should explicitly consider employee experience, not only managerial efficiency and regulatory compliance, as a governance criterion. Surveillance concerns, well documented in algorithmic- management research on gig and platform work [6], extend into agentic contexts insofar as monitoring an autonomous agent' s behaviour may require monitoring the human employees who interact with it. Employment implications changes in the composition and content of managerial and non-managerial work are addressed at the level of managerial role transformation in this review (Proposition 7) but merit dedicated empirical study of broader workforce effects, which is outside this review' s organisational- authority scope. Regulatory compliance considerations are most acute in finance and healthcare, per the multi- case comparison, where existing sectoral regulation already constrains autonomous decision- making independent of any AI-specific legislation, and where AI- specific frameworks such as the EU AI Act add explanation and oversight requirements to existing compliance regimes [32]. Finally, organisational control considerations connect directly back to the algorithmic- authority- versus- algorithmic- control distinction developed in Section 2. 2.3 [7]: governance frameworks should explicitly decide, and communicate, whether a given AI deployment is intended to legitimately hold decision authority or merely to support human decision- making, since ambiguity on this point appears to be a source of both managerial confusion and employee mistrust.

12. FUTURE RESEARCH AGENDA

Table 4 and Table 5, introduced below, synthesize the major research gaps identified across this review and translate them into a structured agenda spanning theoretical perspectives and specific future research directions, complementing the methodological breakdown visualized in Figure 6.

Table 4. Major Theoretical Perspectives Relevant to Human-AI Collaboration

Theory	Core Assumption	Relevance to Agentic AI	Contribution	Limitation
Agency theory [13,14,15]	Principals delegate decision rights to agents under imperfect monitoring	Frames AI as a prospective non-human agent of the firm	Explains delegation, monitoring, and alignment-mechanism design	Alignment mechanisms assume a motivable agent; AI has no independent utility function

Theory	Core Assumption	Relevance to Agentic AI	Contribution	Limitation
Delegated agency theory [16]	Trust depends on delegation fidelity, not outcome quality alone	Directly addresses agentic AI's procedural, multi-step autonomy	Introduces contract-specification and governance-visibility constructs	Recently proposed; limited empirical validation to date
Sociotechnical systems / algorithmic-authority perspective [6,7,29]	Algorithms are embedded, negotiated organizational processes, not neutral tools	Explains how algorithmic outputs acquire legitimacy as a basis for action	Distinguishes algorithmic authority from algorithmic control	Focused historically on frontline/gig work; less developed for managerial/strategic contexts
Human-AI teaming / complementarity [22,23]	Human-AI teams can outperform either party alone under the right design	Provides design principles for Stage 3 (Collaborator) configurations	Identifies trust calibration, role partitioning, escalation protocols	Largely developed in high-stakes, expert-driven domains (medicine, security); generalizability to routine managerial work less tested
TOE / TAM / UTAUT adoption frameworks [18-21]	Adoption depends on technological, organizational, environmental, and individual factors	Explains organizational readiness to adopt AI generally	Well-validated across many prior IT adoption contexts	Not designed for technologies that exercise post-adoption discretion; static rather than iterative in original form

Table 5. Major Research Gaps

Research Theme	Existing Knowledge	Limitation	Research Gap	Future Direction
Mid-level managerial authority	Frontline algorithmic control [6] and C-suite strategic framing [3] both well studied	Middle-management discretion under-examined	How does agentic AI specifically reshape middle-manager decision rights?	Multi-level field studies spanning frontline, middle, and senior management simultaneously
Delegation fidelity measurement	Concept theoretically defined [16]	No validated survey or behavioral measure identified in the reviewed literature	Absence of an operationalized delegation-fidelity construct	Scale development and validation studies (survey + behavioral trace data)
Cross-industry comparative theory	Sector-specific adoption statistics available [1,4,35]	Comparative theorizing across sectors remains largely descriptive	Why does authority transformation decouple from technical capability in some sectors (Figure 4)?	Comparative multi-industry case research explicitly testing the capability–authority decoupling proposition
Accountability-mechanism design	Responsibility-gap concept well theorized [12]; regulatory	Limited empirical evidence on which specific	Which accountability-attachment	Field experiments and longitudinal studies comparing

Research Theme	Existing Knowledge	Limitation	Research Gap	Future Direction
	requirements emerging [32,33]	mechanisms (audit trails, thresholds, override rights) are effective in practice	mechanisms actually close the responsibility gap?	governance-mechanism configurations
Explainability–oversight interaction	Complementarity argued conceptually [12]; explanation effects on team performance studied experimentally [26]	Interaction rarely tested directly with organizational authority outcomes	Does explainability's effect on delegation depend on genuine override authority (Proposition 5)?	Controlled experiments varying both explanation quality and overseer authority
Employee experience of algorithmic authority	Both dehumanization and preference-for-objectivity documented [27]	Conditions determining which reaction predominates are unclear	What organizational and communicative factors determine employee reception of algorithmic authority?	Mixed-method studies combining survey measures with qualitative accounts of legitimation communication

Table 6, introduced here, translates the conceptual model's constructs (Section 5) into a structured basis for future measurement development.

Table 6. Proposed Constructs and Relationships

Construct	Definition	Theoretical Basis	Expected Relationship	Potential Measurement
AI capability	Technical sophistication and reliability of the agentic system	Agentic AI literature [1,2]	Positive antecedent of trust and delegation	Task-completion benchmarks; expert technical assessment
Delegation fidelity	Alignment between authorized delegation and observed AI behavior	Delegated agency theory [16]	Positive predictor of sustained trust (P8)	Behavioral trace analysis; post-hoc alignment scoring
Process trust	Confidence in the procedural correctness of an autonomous action	Trust-in-automation literature [30,31]	Mediates capability-to-delegation relationship (P2)	Adapted process-trust survey instruments
Perceived control	Belief that one can modify or override AI outputs	Algorithm-aversion literature [24,25]	Reduces aversion; increases sustained reliance (P6)	Experimental manipulation of override access; survey measures
Algorithmic authority	Degree to which algorithmic output functions as a legitimate basis for organizational action	Algorithmic-authority literature [7]	Increases with Stage progression (Table 1)	Organizational document analysis; manager interviews
Accountability attachment	Degree to which responsibility for an	Responsibility-gap literature [12]	Moderates trust and regulatory-	Governance-artifact audit (presence/absence

Construct	Definition	Theoretical Basis	Expected Relationship	Potential Measurement
	AI-mediated outcome is clearly assigned to a human actor		compliance outcomes (P9)	of escalation protocols, audit trails)

Table 7, introduced here, closes the future research agenda by specifying methodological pathways for each research area, corresponding to the methodological tier of Figure 6.

Table 7. Future Research Agenda

Research Area	Research Question	Suggested Method	Expected Contribution
Quantitative	Does organizational readiness moderate the AI-capability-to-delegation relationship as predicted (P3)?	Cross-sectional survey with structural equation modelling across multiple industries	Empirical test of the antecedent-mediator path in the conceptual model (Figure 2)
Quantitative	Does perceived control causally reduce algorithm aversion more than accuracy improvements alone (P6)?	Randomized controlled experiment manipulating override access and system accuracy	Causal evidence on the control-versus-accuracy debate in the trust literature
Qualitative	How do middle managers narrate and experience the shift from execution to calibration/audit (P7)?	In-depth interviews and process-tracing case studies within organizations at Stage 3–4	Address the mid-level managerial authority gap identified in Table 5
Qualitative	How is algorithmic authority legitimated (or not) through managerial communication?	Ethnographic observation combined with discourse analysis of internal communications	Empirical grounding for Proposition 10 on legitimation and procedural justice
Computational	Can delegation fidelity be operationalized from system logs and behavioral traces?	Development of computational metrics comparing authorized parameters to observed agent behavior	Operational measure supporting Table 6's delegation-fidelity construct
Computational	How does human-AI team decision quality vary across escalation-threshold configurations?	Agent-based simulation and controlled human-subject experiments with AI teammates	Design guidance for the oversight-mechanism recommendations in Section 10
Mixed-method	Which accountability-attachment mechanisms most effectively close the responsibility gap across industries?	Triangulated field experiments, governance-artifact audits, and longitudinal outcome tracking	Direct empirical test of Proposition 9 and refinement of Table 2's accountability dimension
Mixed-method	Does the capability–authority decoupling pattern in Figure 4 replicate across additional industries and regulatory regimes?	Comparative multi-country, multi-industry mixed-method case research	Extension and stress-test of the multi-case framework developed in Section 4

13. LIMITATIONS

This review is subject to several limitations that should inform how its contributions are read and used. First, despite the documented search and screening protocol reported in Section 1.1, this remains an integrative critical review rather than a registered systematic review: no protocol was pre-registered (for example, with PROSPERO), no inter-rater reliability was calculated for screening decisions, which were conducted using the reviewing process described in Section 1.1 rather than by multiple independent coders, and the reported screening counts in Figure 1 are approximate rather than derived from a fully automated, auditable database export; readers seeking systematic-review-grade reproducibility should treat Section 1.1 as a transparency device rather than as a claim of systematic-review methodology. Second, the multi-case analytical framework (Section 4) and the conceptual model (Section 5) are explicitly conceptual and propositional; no primary empirical data were collected or analysed, and no element of the framework should be read as empirically validated. Third, the literature underlying this review is heterogeneous in disciplinary origin, spanning organization theory, information systems, human factors, labour process theory, and law, and studies within these traditions frequently operate from differing theoretical premises and measure different constructs under similar labels (for example, “trust” and “authority” are operationalised differently across the algorithmic-management and human-AI-teaming literatures); this heterogeneity is made explicit rather than resolved prematurely. Fourth, agentic AI is evolving rapidly, and industry-adoption evidence cited in Section 4 reflects a fast-moving, commercially produced evidence base whose precision and generalisability are inherently limited; figures describing adoption rates and sector patterns should be read as indicative of broad directional trends rather than as stable, precise estimates. Fifth, this review may be subject to publication bias common to rapidly emerging technology topics, in which enthusiastic accounts of transformative potential may be over-represented relative to null or negative findings, particularly in industry-produced sources. Sixth, the industry cases developed in Section 4 are necessarily heterogeneous in the depth and quality of available evidence; logistics and IT are comparatively well documented in the peer-reviewed and industry literature, while professional services and retail are comparatively under-documented from an organisational-authority perspective specifically; a limitation reflected in the differing citation density across rows of Table 2. Seventh, the review lacks longitudinal evidence: because agentic AI adoption is recent, no study reviewed here tracks the same organisations through multiple stages of the five-stage continuum over time, meaning the “progression” implied by Table 1 and Figure 2 is inferred from cross-sectional and comparative evidence rather than observed directly. Finally, direct comparison of AI systems and governance practices across organisations remains difficult given the proprietary nature of many agentic AI deployments and the resulting reliance, in parts of Section 4, on aggregated survey rather than case-verified evidence.

14. CONCLUSION

This review aims to examine the ways in which the advent of agentic AI is transforming human- AI collaboration, with particular emphasis on how it is renegotiating managerial authority within organizations. By synthesising literature on agentic AI, algorithmic management, human- AI teaming,

trust and explainability, agency theory, and technology adoption, the review develops an integrative framework organized around a five- stage transition from AI as a mere tool to AI as a prospective organizational actor. This framework connects previously isolated literatures under a unified vocabulary. The central argument posits that the organisational significance of agentic AI resides more in the redistribution of decision rights, discretion, and accountability-elements that define managerial authority in traditional organizational contexts- than solely in task automation. This redistribution, however, is uneven: a multi- case analysis suggests that the potential for technical autonomy and sanctioned authority transformation can become decoupled, especially in regulated, high- risk sectors. This indicates that the pace and nature of authority renegotiation are influenced as much by governance structures and regulatory environments as by AI capabilities. The conceptual framework and ten propositions presented herein are intended as a foundation for future empirical research, consistent with the review' s explicit focus on conceptual and critical synthesis. Their primary practical implication is that responsible human- AI collaboration concerning agentic AI involves organizations viewing oversight not as a diminishing constant as autonomy increases but as a dynamic activity evolving from instance- level review to system- level governance design, delegation- fidelity monitoring, and accountability infrastructure- developed proactively rather than retroactively reconstructed. An essential avenue for future research, as identified in this review, is the development of validated, empirically testable measures of delegation fidelity and algorithmic authority (see Table 6). Without such measures, the propositions put forward and the broader theoretical claim that managerial authority is being renegotiated rather than simply automated cannot be rigorously evaluated. As agentic AI systems assume increasingly active roles in organizational decision- making, the ability of organizations to maintain meaningful, well-structured human oversight will depend less on resisting this transition and more on governing it with deliberate intent.

REFERENCES

[1] Hosseini S, Seilani H. The role of agentic AI in shaping a smart future: a systematic review. Array. 2025;26:100399.

[2] Ali MA, Dornaika F, Charafeddine J. Agentic AI: a comprehensive survey of architectures, applications, and future directions. Artif Intell Rev. 2025;59(1). doi:10.1007/s10462-025-11422-4

[3] Ransbotham S, Kiron D, Khodabandeh S, Iyer S, Das A. The emerging agentic enterprise: how leaders must navigate a new age of AI. Cambridge (MA): MIT Sloan Management Review and Boston Consulting Group; 2025.

[4] Premalatha MAK. AI agents are transforming decision making: what leaders should know. Harv Data Sci Rev. 2026;8(2).

[5] Aral S. Agentic AI, explained. Cambridge (MA): MIT Sloan School of Management, Ideas Made to Matter; 2026

[6] Kellogg KC, Valentine MA, Christin A. Algorithms at work: the new contested terrain of control. Acad Manag Ann. 2020;14(1):366-410. doi:10.5465/annals.2018.0174.

[7] Nouamani S, Asri HE. Is managerial authority still human? Algorithmic management and the reconfiguration of authority. J Knowl Manag Pract. 2026;26(3).

[8] Adams-Prassl J, Abraha H, Kelly-Lyth A, Silberman MS, Rakshita S. Regulating algorithmic management: a blueprint. Eur Labour Law J.

[9] Dietvorst BJ, Simmons JP, Massey C. Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them. Manage Sci. 2018;64(3):1155-1170.

[10] Wang Y. Algorithmic decision-making in organizations: a systematic review toward an integrated tension alignment framework. Organ Manag J. 2025. doi:10.1108/OMJ-11-2024-2342.

[11] Romeo G, Conti D. Exploring automation bias in human-AI collaboration: a review and implications for explainable AI. AI Soc. 2025. doi:10.1007/s00146-025-02422-7.

[12] MIT Sloan Management Review. AI explainability: how to avoid rubber-stamping recommendations. Cambridge (MA): MIT Sloan Management Review; 2025.

[13] Jensen MC, Meckling WH. Theory of the firm: managerial behavior, agency costs and ownership structure. J Financ Econ. 1976;3(4):305-360.

[14] Fama EF, Jensen MC. Separation of ownership and control. J Law Econ. 1983;26(2):301-325.

[15] Humberd BK, Latham SF. When AI becomes an agent of the firm: examining the evolution of AI in organizations through an agency theory lens. J Manag Stud. 2025;63(2):668-694. doi:10.1111/joms.13274.

[16] Helal M, Dwivedi Y, Ismagilova E, et al. Reinterpreting agency theory in the era of artificial intelligence: conceptualizing delegated agency. AI Soc. 2026;41:7063-7074. doi:10.1007/s00146-026-03034-5.

[17] Baird A, Maruping LM. The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. MIS Q.2021;45(1):315-341. doi:10.25300/MISQ/2021/15882.

[18] Tornatzky LG, Fleischer M. The processes of technological innovation. Lexington (MA): Lexington Books; 1990.

[19] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 1989;13(3):319-340.

[20] Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: toward a unified view. MIS Q. 2003;27(3):425-478.

[21] Dey D, Ghose D. Integrating AI adoption frameworks with digital maturity models: strategic pathways in a rapidly evolving technological landscape. 2025,

[22] Gonzalez C, Donahue K, Goldstein DG, Heidari H, Jalali MS, Schelble B, Singh A, Woolley AW. Toward a science of human-AI teaming for decision making: a complementarity framework. PNASNexus.2026;5(3).doi:10.1093/pnasnexus/pgag030.

[23] Duan W, Zhou S, Scalia MJ, Yin X, Weng N, Zhang R, Freeman G, McNeese N, Gorman J, Tolston M. Understanding the evolvement of trust over time within human-AI teams. Proc ACM Hum Comput Interact. 2024;8(CSCW2):1-31.

[24] Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, Baeßler B, Sauer S, Maintz D, Pinto dos Santos D. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4). doi:10.1148/radiol.222176.

[25] Jones-Jang SM, Park YJ. How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. J Comput Mediat Commun. 2023;28(1).

[26] Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York: ACM; 2021. p.1-16.

[27] Zhang Y. The rise of algorithmic management and implications for work and organisations. New Technol Work Employ. 2025. doi:10.1111/ntwe.12343.

[28] Marabelli M, Newell S, Handunge V. The lifecycle of algorithmic decision-making systems: organizational choices and ethical challenges. J Strateg Inf Syst. 2021;30(3):101683. doi:10.1016/j.jsis.2021.101683.

[29] Jarrahi MH, Sutherland W. Algorithmic management and algorithmic competencies: understanding and appropriating algorithms in gig work. In: Proceedings of iConference 2019. Cham: Springer; 2019.

[30] Mayer RC, Davis JH, Schoorman FD. An integrative model of organizational trust. Acad Manage Rev. 1995;20(3):709-734.

[31] Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004;46(1):50-80.

[32] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union; 2024.

[33] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Gaithersburg (MD): NIST; 2023.

[34] MIT Sloan Management Review, Boston Consulting Group. Agentic AI at scale: redefining management for a superhuman workforce. Cambridge (MA): MIT Sloan Management Review; 2025.

[35] First Page Sage. Agentic AI adoption statistics for 2026. 2026.

[36] Pulkkinen J, Huttu K, Suhonen M. Systemic challenges in AI adoption in public social and health organizations in Finland: A technology-organisation-environment perspective. J Health Organ Manag. 2025;39(9):435-456. doi:10.1108/JHOM-06-2025-0309.

[37] Mpanza SS. Revisiting the Technological-Organizational-Environmental (TOE) Framework and Diffusion of Innovation (DOI): a theoretical review for artificial intelligence (AI) adoption. Int J Appl Res Bus Manag. 2025;6(5). doi:10.51137/wrp.ijarbm.441.

[38] Torraco RJ. Writing integrative literature reviews: Guidelines and examples. Hum Resour Dev Rev. 2005;4(3):356-367. doi:10.1177/1534484305278283.

[39] Snyder H. Literature review as a research methodology: an overview and guidelines. J Bus Res.2019;104:333-339. doi:10.1016/j.jbusres.2019.07.039.

[40] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372. doi:10.1136/bmj.n71.

[41] Njei B, Al-Ajlouni YA, Kanmounye US, Boateng S, Nguefang GL, Njei N, Hamouri S, Al-Ajlouni AF. Artificial intelligence agents in healthcare

research: a scoping review. PLoS One. 2026;21(2). doi:10.1371/journal.pone.0342182.

[42] Thomson Reuters Institute. 2026 AI in Professional Services Report. New York: Thomson Reuters; 2026.

[43] Sheridan TB, Verplank WL. Human and computer control of undersea teleoperators. Cambridge (MA): MIT Man-Machine Systems Laboratory; 1978. Report No.: N00014-77-C-0256.

[44] Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybern A Syst Hum. 2000;30(3):286-297. doi:10.1109/3468.844354.

[45] Raisch S, Krakowski S. Artificial intelligence and management: the automation-augmentation paradox. Acad Manage Rev. 2021;46(1):192-210. doi:10.5465/amr.2018.0072.