



AI-AGENT DEPLOYMENT DENSITY AND THE RESHAPING OF MIDDLE-MANAGEMENT SPAN OF CONTROL

Dr. N Subbukrishna Sastry¹, Dr. Manjula Mallya M²

¹Professor, School of Management, CMR University, Bangalore, Karnataka, India.

ORCID ID: 0009-0009-0460-7057

²Associate Professor & Head, Dept. of Economics, Government First Grade College for Women
 Balmatta Mangalore, Karnataka, India. ORCID ID: 0009-0005-8812-6912,

ABSTRACT

Span-of-control theory has, since Graicunas's 1933/1937 formalisation of how fast superior-subordinate interactions multiply, treated the number of people a manager can effectively supervise as bounded chiefly by the manager's own cognitive and communicative capacity. The deployment of autonomous and semi-autonomous AI agents into team workflows unsettles that boundary in a way earlier information technologies did not: agents do not merely give a manager better information about subordinates, they absorb a share of the routine coordination, monitoring and exception-triage work that previously consumed the manager's attention, functioning in effect as coordination-compressing capital. Industry evidence assembled since 2024 suggests this is already reshaping organisational structure: Gartner has predicted that by 2026 a fifth of organisations will use AI to flatten structure and eliminate over half of current middle-management positions, a 2025 MIT Sloan Management Review/BCG study found 45% of respondents at organisations with extensive agentic-AI adoption expect further reductions in management layers, and Amazon's 2024 mandate to raise its individual-contributor-to-manager ratio by at least 15% is widely read, alongside this evidence, as an early instance of the same structural shift. This paper reviews classical span-of-control theory, the organisational economics of hierarchy, and this emerging AI-agent evidence to develop a conceptual framework in which AI-agent deployment density raises optimal managerial span of control primarily by reducing the coordination cost per subordinate, an effect strengthened for routine, codifiable task portfolios and attenuated where work remains exception-heavy and judgment-intensive. Five formal propositions link deployment density, task routineness, and managerial role adaptation to realised — as distinct from merely announced — span-of-control change, drawn from a structured review of thirty-six cited works. Particular attention is given to India's IT-BPM sector, whose historically wide, standardised-process spans of control interact distinctively with agentic-AI deployment. The paper closes with a research agenda centred on a measurement gap current scholarship has not yet filled: no validated instrument yet exists for "AI-agent deployment density" itself.

KEYWORDS: AI agents; span of control; middle management; organisational hierarchy; coordination cost; agentic AI; India IT-BPM sector

1. INTRODUCTION

How many people can one manager effectively supervise? The question is among the oldest in management theory, and for most of the twentieth century the answer changed remarkably little: Graicunas (1933/1937) showed mathematically that the number of distinct interaction relationships a supervisor must track grows far faster than the number of subordinates itself, and Urwick (1956) converted that mathematics into the durable prescriptive rule that a manager overseeing interdependent work should have no more than five or six direct reports. Organisational economists later reframed the same limit in terms of cost rather than cognition — Williamson (1967) modelled span of control as a trade-off against cumulative "control loss" through hierarchical layers, and Garicano (2000) showed that a knowledge hierarchy's optimal span is set jointly by the share of problems routine enough to resolve without escalation and by the cost of communicating an exception upward — but the underlying premise persisted across both traditions: a human manager's finite attention is the binding constraint on how many people, and how much complexity, one person can oversee.

The deployment of autonomous and semi-autonomous AI agents into team workflows is the first technology in this long literature to unsettle that premise directly rather than merely easing it. Earlier information technologies — email, enterprise resource-planning systems, dashboards — gave managers better and faster information about



subordinates' work but left the manager as the sole locus of coordination, monitoring and exception-handling. AI agents that triage incoming work, draft first-pass responses, flag anomalies, and resolve routine cases autonomously instead absorb a share of that coordination load themselves, functioning in effect as what recent organisational-economics scholarship has termed "coordination-compressing capital" (Farach, 2026) — capital that substitutes for a portion of the manager's own attention rather than simply extending it. If Garicano's (2000) model is right that span of control is set by the ratio of routine to exceptional work and by the cost of resolving each, a technology that autonomously resolves a growing share of routine cases and pre-processes the exceptions that remain should, on the model's own logic, raise the span a manager can sustain.

Industry evidence assembled since 2024 suggests this is not a hypothetical extrapolation. Gartner has predicted that by 2026, one-fifth of organisations will use AI to flatten organisational structure, eliminating more than half of current middle-management positions (Gartner, 2024). A 2025 study by MIT Sloan Management Review and the Boston Consulting Group's Henderson Institute found that 45% of respondents at organisations with extensive agentic-AI adoption expect further reductions in management layers (Ransbotham, Kiron, Khodabandeh, Iyer, & Das, 2025). And in September 2024, Amazon's chief executive directed every business unit to raise its ratio of individual contributors to managers by at least 15% within two quarters (Amazon.com, 2024) — a mandate whose stated rationale was reducing bureaucracy rather than AI capacity specifically, but one widely read, alongside the Gartner and MIT SMR/BCG evidence, as an early instance of the same structural shift. None of these sources, however, measures the construct this paper places at the centre of the relationship: the density of AI agents actually deployed per team member, as distinct from an organisation's stated adoption intent or its announced headcount targets.

This paper makes three contributions. First, it connects a long-standing organisational-economics literature on hierarchy and span of control — largely developed before agentic AI existed as a category — to the emerging, still largely descriptive industry evidence on AI-driven organisational flattening, treating AI-agent deployment density as a direct empirical instance of the communication-cost reduction that Garicano's (2000) and Garicano and Rossi-Hansberg's (2006) models already predict should widen spans and flatten hierarchies. Second, it develops a conceptual framework in which AI-agent deployment density raises optimal span of control primarily by reducing coordination cost per subordinate, an effect moderated by task routineness and by whether the manager's own role reconfigures toward exception-handling and judgment, and expresses this framework as five formal propositions. Third, it examines India's IT-BPM sector, whose historically wide, process-standardised spans of control and labour-cost-arbitrage commercial model interact distinctively with agentic-AI deployment. Section 2 sets out the theoretical background; Section 3 reviews the literature across AI-agent coordination theory, industry evidence, the changing middle-management role, delegation and authority, AI deployment in HR and organisational-design functions, and the Indian context; Section 4 presents the conceptual model and propositions; Section 5 describes the review method and corpus; Section 6 discusses the findings; Section 7 addresses India-specific implications; Section 8 sets limitations and a future-research agenda; Section 9 concludes.

2. THEORETICAL BACKGROUND AND CONCEPTUAL FRAMEWORK

Formal span-of-control theory begins with Graicunas's (1933/1937) observation that a supervisor's cognitive burden does not grow linearly with the number of subordinates. Graicunas classified the interactions a supervisor must track into direct single relationships, cross relationships between subordinates, and group relationships, and showed that their combined number grows geometrically: a supervisor with four subordinates must track 44 potential interaction relationships, one with eight must track 1,080. Urwick (1956) converted this mathematics into the durable practitioner's rule that a manager overseeing interdependent work should supervise no more than five or six people directly, a figure still cited as the classical benchmark against which any claim of AI-driven span expansion must ultimately be judged. Both accounts sit within the broader classical administrative tradition associated with Fayol's (1918/1949) fourteen principles of management — unity of command, the scalar chain of authority, and a hierarchical structure through which those principles were meant to operate — which treated a bounded, well-defined span as a structural precondition for orderly administration rather than as a variable to be optimised against changing technology.

Simon (1947) mounted an influential challenge to this classical tradition on different grounds, arguing that its "principles" — including a fixed numerical span of control — functioned more as contradictory proverbs than as a scientific theory of administration, since the same body of doctrine simultaneously counselled narrow spans for coordination and wide spans for economy without specifying which should govern in a given case. Simon's proposed alternative — a behavioural theory of administration grounded in bounded rationality and the cost of decision-making — reframed span of control as an empirical question to be answered by comparing the costs of



centralised decision-making against the costs of hierarchical delegation, rather than as a fixed number derivable from interaction mathematics alone. This reframing is the direct ancestor of the organisational-economics tradition this paper draws on most heavily.

That tradition treats span of control as the resolution of an explicit cost trade-off. Williamson (1967) modelled hierarchical organisations as subject to cumulative "control loss" as instructions and information pass through successive layers, deriving an optimal firm size and span from the trade-off between the efficiency gains of specialisation and the control loss compounding across layers. Chandler (1962), studying the emergence of the multidivisional form at DuPont, General Motors and other large American firms, showed empirically that hierarchical structure — including how supervisory responsibility was divided — followed from strategic decisions about diversification and scale rather than from a fixed administrative ideal, establishing that span of control is a designed response to changing organisational conditions rather than a structural constant. Blau (1970) added a further empirical regularity: as organisations grow larger, they differentiate into more subunits and hierarchical levels, but at a decreasing rate, implying economies of scale in administrative overhead that widen the average span of control sustained per manager as size increases. Mintzberg's (1979) structural-contingency framework generalised this cost-trade-off view into a broader claim still relevant to the argument developed here: span of control is not a variable an organisation can optimise in isolation, but one of several interdependent structural design parameters — alongside the coordinating mechanisms in use (direct supervision, standardisation of work or outputs, mutual adjustment) and the situational factors an organisation faces — so that a change in one coordinating mechanism, such as the introduction of AI agents that partially substitute for direct supervision, should be expected to shift the configuration of the others rather than simply widening span with everything else held fixed.

The theoretical anchor most directly relevant to this paper's argument is Garicano's (2000) model of knowledge hierarchies, in which a hierarchy exists to allow routine problems to be resolved at the point of occurrence while exceptional problems are escalated to a manager whose time is scarcer and more valuable. In this model, the span of control a manager can sustain is set jointly by the share of problems routine enough to be resolved without escalation and by the cost of communicating an exception upward for resolution — a cost that, when it falls, allows a manager to absorb more subordinates without exceeding their own capacity to resolve escalated problems. Garicano and Rossi-Hansberg (2006) extended this logic to argue that falling communication costs across an economy should flatten organisational hierarchies and widen spans of control economy-wide, a prediction subsequently confirmed by two independent empirical studies of the 1990s' information-technology-driven flattening: Rajan and Wulf (2006), using panel data on large American firms, documented that hierarchies flattened and CEOs' own spans of control widened substantially over the decade even as the number of formal reporting layers barely changed, and Guadalupe and Wulf (2010), exploiting United States tariff reductions as a natural experiment, showed that increased product-market competition causally flattened hierarchies through the same communication-cost channel. Colombo and Delmastro (2004), studying Italian manufacturing plants, showed a closely related mechanism operating at the level of individual decision rights: lower monitoring costs and higher worker skill were both associated with greater delegation of authority away from central management, evidence that the communication- and monitoring-cost logic driving span of control operates on delegation decisions generally, not on reporting-line span alone.

Read together, this literature offers a precise theoretical prediction for the technology at the centre of this paper. If Garicano's (2000) model correctly identifies communication cost and the routine-to-exceptional problem ratio as the two variables setting optimal span, and if Garicano and Rossi-Hansberg (2006), Rajan and Wulf (2006) and Guadalupe and Wulf (2010) are right that earlier reductions in communication cost already flattened hierarchies and widened spans economy-wide, then AI agents — which both reduce the cost of surfacing and communicating routine information and directly resolve a share of problems that would otherwise require escalation — should widen sustainable span of control through precisely the mechanism this literature already specifies, rather than through some qualitatively new organisational logic. This paper's conceptual framework, set out in Figure 1 and Table 1, operationalises that mechanism as AI-agent deployment density acting on optimal span of control through the mediating channel of coordination cost per subordinate, moderated by task routineness and by whether the manager's own role reconfigures toward exception-handling as routine coordination is absorbed by agents.

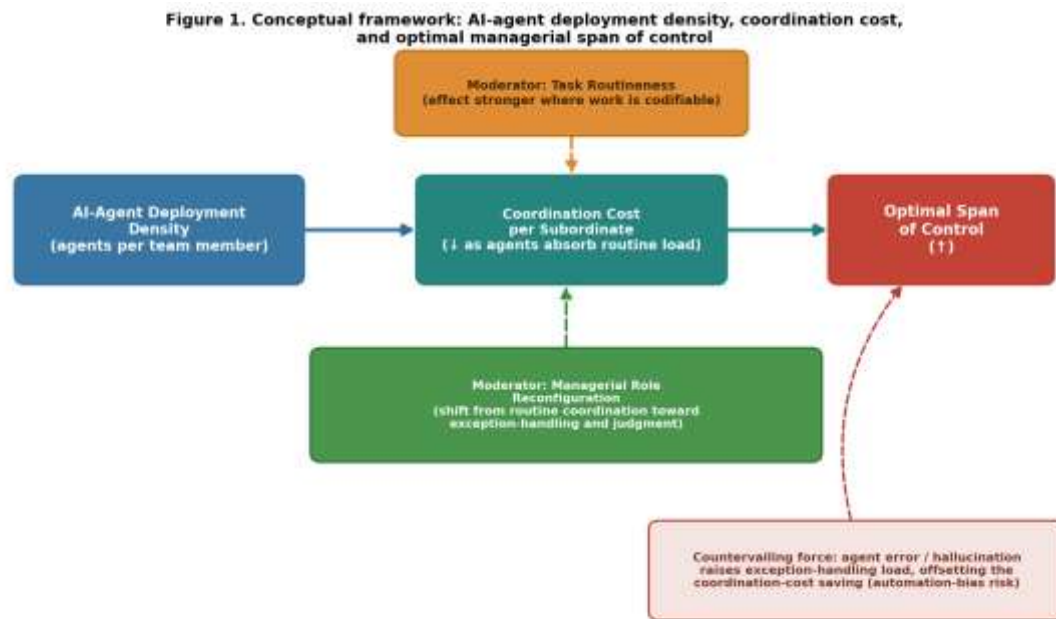


Figure 1. Conceptual framework: AI-agent deployment density, coordination cost, and optimal managerial span of control.

Construct	Working Definition (this paper)	Theoretical Anchor(s)	Representative Evidence
AI-Agent Deployment Density	The number of autonomous or semi-autonomous AI agents actively performing coordination, monitoring or exception-triage tasks per team member, as distinct from an organisation's stated adoption intent.	Coordination-compressing capital; knowledge-hierarchy theory	Farach (2026); Gartner (2024); Ransbotham et al. (2025)
Coordination Cost per Subordinate	The manager's time and attention required to monitor, direct and resolve exceptions for one subordinate's work.	Hierarchical control loss; knowledge hierarchies	Williamson (1967); Garicano (2000)
Task Routineness (moderator)	The share of a team's task portfolio that is codifiable and resolvable without escalation to the manager.	Knowledge-hierarchy theory (routine/exception ratio)	Garicano (2000); Garicano & Rossi-Hansberg (2006)
Managerial Role Reconfiguration (moderator)	The degree to which a manager's own activity shifts from routine coordination toward exception-handling, judgment and people development as agents absorb routine load.	Dynamic capabilities (reconfiguring); role theory	Sastry (2025, Middle management evolution)
Optimal Span of Control	The number of direct reports a manager can effectively oversee without a decline in decision quality or subordinate outcomes.	Classical span-of-control theory; control-loss theory	Graicunas (1933/1937); Urwick (1956); Williamson (1967)

3.1 AI Agents as Coordination-Compressing Capital

A recent strand of organisational economics gives the mechanism sketched in Section 2 a name and a formal treatment specific to agentic AI. Farach (2026) models AI that reduces the cost of coordinating work across an organisation as "coordination-compressing capital": capital that, unlike earlier automation, substitutes directly for



a share of the manager's own coordination function rather than merely automating a subordinate's task. The model's central prediction is structural rather than merely incremental — coordination-compressing capital expands the number of people a single manager can effectively oversee, flattens hierarchies, and reduces the total demand for middle-management layers — and it identifies a critical distributional fork: where access to coordination-compressing tools is broad across the management layer, the resulting productivity gains are shared, whereas where access is concentrated among senior managers, the same technology instead widens the gap between managers and the workers they oversee. This distributional mechanism is directly relevant to the framework developed in this paper: it implies that AI-agent deployment density's effect on span of control is not a purely technical fact about the agents deployed, but also a function of how that deployment is distributed through the management hierarchy — a theme returned to in Section 6.

The author's own review of organisational AI-readiness practice engages this coordination-compression argument directly, extending it to note that a technology which increases the volume of work a given management layer can coordinate does not automatically translate into a widened, sustainable span unless the organisation also invests in the readiness practices — workflow redesign specifically — that convert raw coordination capacity into effectively supervised output (Sastry, 2026, Organisational AI-readiness practice and the realised productivity gap from agentic AI). Applied to the specific construct this paper isolates, that argument implies a further precondition beyond deployment density itself: an organisation can deploy AI agents densely into a team's workflow without any corresponding change in the manager's actual span, if the workflow surrounding the agents is not redesigned to route the coordination work the agents absorb away from the manager's desk in the first place. Deployment density, in other words, is a necessary but not sufficient condition for realised span expansion — the distinction formalised as Proposition 1 in Section 4.

Kellogg, Valentine, and Christin's (2020) organisational-studies account of algorithmic management supplies the mechanical vocabulary for how coordination-compressing capital actually performs the coordination functions Farach's (2026) model treats abstractly, describing algorithmic management as operating through six mechanisms — restricting, recommending, recording, rating, replacing and rewarding — several of which map directly onto the routine-coordination load a manager traditionally carries: an AI agent that records subordinate activity, rates it against a threshold, and recommends a next action is performing, mechanically, the same monitoring and triage function a manager's own attention would otherwise have to supply. This mechanical correspondence is what allows deployment density to be treated as a direct, if still unmeasured, instance of the communication-cost reduction Garicano's (2000) model specifies, rather than as a looser analogy to it.

This coordination-compressing capacity is not, however, uniformly reliable, and the countervailing risk this creates is built into Figure 1's framework as an explicit moderating force rather than left as an implicit caveat. Dell'Acqua, McFowland, Mollick, Lifshitz-Assaf, Kellogg, Rajendran, Krayner, Candelon, and Lakhani's (2025) field experiment with Boston Consulting Group consultants found that generative-AI assistance sharply improved performance on tasks inside AI's capability frontier but caused performance to decline for consultants who relied on it for tasks just outside that frontier — a boundary the authors term the "jagged frontier" because it does not correspond to any intuitive notion of task difficulty. Applied to coordination-compressing capital specifically, the jagged frontier implies that widening span of control by trusting an AI agent's routine-versus-exception triage is safe only to the extent the agent's own reliability boundary is well understood; a manager whose span has widened on the assumption that agents reliably resolve routine cases is exposed to exactly the over-reliance risk the jagged frontier describes whenever an agent's silent failure just outside its own competence boundary is mistaken for a genuinely resolved case. This risk is represented in Figure 1 as a countervailing force acting against the density-to-span relationship, and its practical management is returned to in Section 6.

3.2 Industry Evidence on AI-Driven Organisational Flattening

Three independent, methodologically distinct sources published between September 2024 and 2025 point toward the same structural conclusion Farach's (2026) model predicts, though none of them measures AI-agent deployment density directly. Gartner's widely cited prediction holds that by 2026, one-fifth of organisations will use AI to flatten organisational structure, eliminating more than half of current middle-management positions, with the accompanying rationale that AI deployment enables an "enhanced productivity and increased span of control by automating and scheduling tasks, reporting and performance monitoring" (Gartner, 2024). Ransbotham, Kiron, Khodabandeh, Iyer, and Das (2025), reporting a study conducted jointly by MIT Sloan Management Review and the Boston Consulting Group's Henderson Institute, found that among organisations with extensive agentic-AI adoption, 45% of respondents expected further reductions in management layers, and argued more



generally that where agents coordinate workflows directly, traditional managerial spans of control increase while the number of hierarchical layers decreases.

Amazon's internal restructuring supplies a large-scale, if imperfectly attributed, illustration of the same pattern. In September 2024, the company's chief executive directed every business unit to increase its ratio of individual contributors to managers by at least 15% within two quarters, with the company's own stated rationale centred on reducing bureaucracy and increasing the speed of decision-making rather than on AI capacity specifically (Amazon.com, 2024). The mandate is nonetheless widely read alongside the Gartner and MIT SMR/BCG evidence as an early, large-scale instance of the same underlying shift, precisely because a 15% increase in the individual-contributor-to-manager ratio is, definitionally, an increase in average span of control. Figure 3 presents these three metrics together for ease of comparison, with the caption noting explicitly that each is a distinct measure from a distinct source and the figure should not be read as evidence of a single common trend measured three separate ways.

Two further industry sources corroborate the pace, if not the precise structural mechanism, of the underlying change. Microsoft's Work Trend Index has tracked the emergence of what it terms the "Frontier Firm" — organisations combining agentic-AI deployment with redesigned human leadership structures — reporting accelerating adoption growth between its 2025 and 2026 editions (Microsoft WorkLab, 2025, 2026), and McKinsey's 2026 global AI survey found that organisations attributing measurable EBIT gains to AI were disproportionately those that had fundamentally redesigned workflows rather than layering AI onto unchanged processes (McKinsey & Company, 2026) — the same workflow-redesign precondition the author's own prior review of organisational AI-readiness identified as necessary for realised productivity gain generally (Sastry, 2026, Organisational AI-readiness practice and the realised productivity gap from agentic AI), and, by the argument developed in Section 3.1 above, plausibly necessary for realised span-of-control expansion specifically. None of these five sources, however, offers a validated measure of AI-agent deployment density as this paper defines it; each measures either an organisation's stated adoption intent, a headcount ratio target, or a survey respondent's expectation, not the number of agents actually performing coordination work per team member. This measurement gap is addressed directly as a priority in Section 8's future-research agenda.

Figure 2. From classical span-of-control theory to AI-agent-era evidence, 1918–2026
(ordered chronologically; horizontal spacing is by sequence, not to scale)

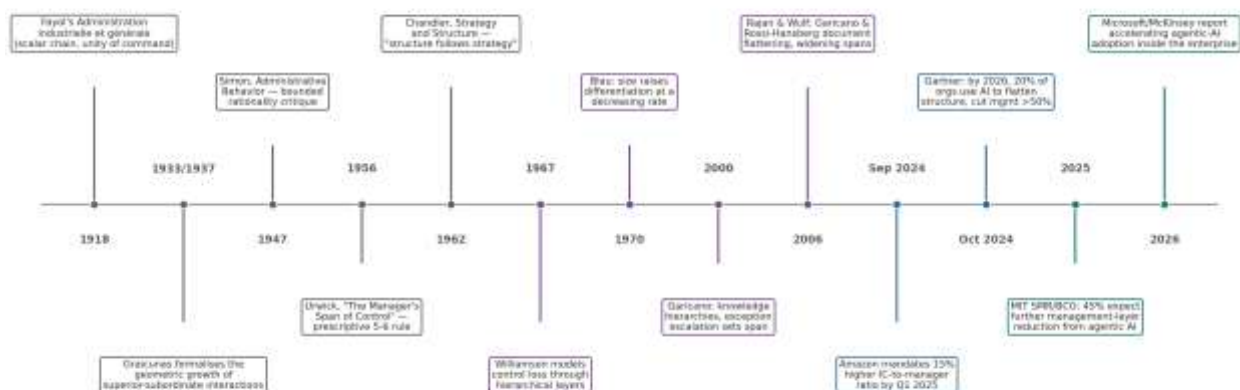


Figure 2. From classical span-of-control theory to AI-agent-era evidence, 1918–2026 (ordered chronologically; horizontal spacing is by sequence, not to scale).



Source	Year	Metric Reported	What It Does Not Measure
Gartner	2024	20% of organisations predicted to use AI to flatten structure and cut management >50% by 2026	Agents deployed per team; realised (as opposed to predicted) span change
Ransbotham, Kiron, Khodabandeh, Iyer, & Das (MIT SMR/BCG)	2025	45% of respondents at extensive-adoption organisations expect further management-layer reductions	A specific span-of-control number; agent density per manager
Amazon.com, Inc.	2024	Mandated 15% increase in individual-contributor-to-manager ratio by Q1 2025	Attribution to AI specifically (stated rationale was bureaucracy reduction); agent deployment density
Microsoft WorkLab	2025, 2026	Accelerating year-over-year agentic-AI adoption growth ("Frontier Firm")	Span-of-control outcome; management-layer count
McKinsey & Company	2026	EBIT gains concentrated among organisations that redesigned workflows around AI	Span-of-control outcome specifically; agent-to-manager ratio

Table 2. Industry evidence on AI-driven organisational flattening, and the measurement gap each source leaves unaddressed.

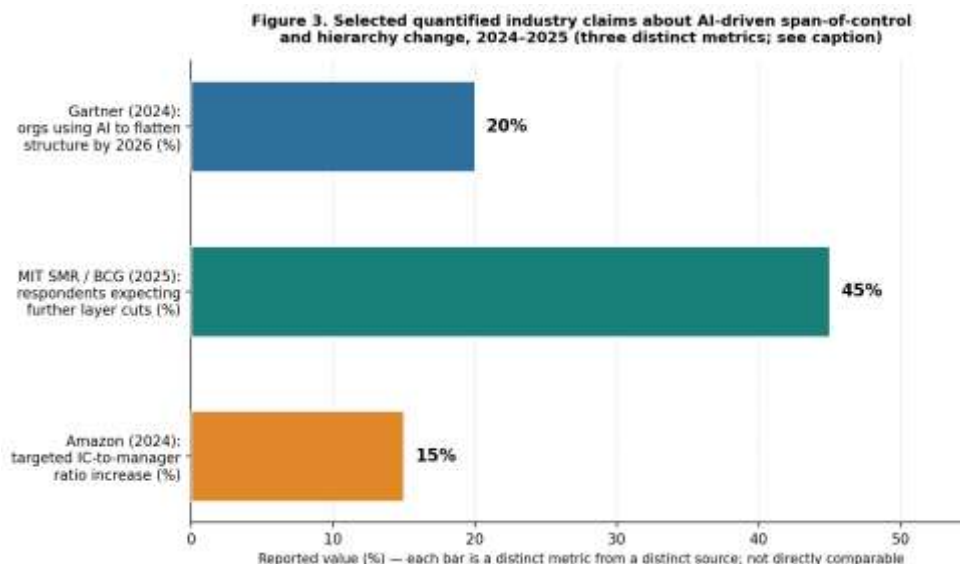


Figure 3. Selected quantified industry claims about AI-driven span-of-control and hierarchy change, 2024–2025 (three distinct metrics; not directly comparable).

3.3 The Changing Role of Middle Management

If AI-agent deployment density widens sustainable span of control by absorbing routine coordination work, the immediate question is what remains for the manager to do once that absorption occurs — a question the industry evidence reviewed in Section 3.2 answers only by counting eliminated positions, not by describing what the surviving management role looks like. The author's dedicated study of middle-management evolution addresses this directly, arguing that the middle-management layer's persistence under coordination-compressing technology depends on its members shifting away from routine coordination and status reporting — precisely the activities Farach's (2026) model treats as substitutable — toward judgment-intensive tasks that remain outside current AI capability: dispute resolution, contextual interpretation of ambiguous situations, mentoring, and localised adaptation of standardised processes to specific team circumstances (Sastry, 2025, Middle management



evolution in contemporary organizations). This argument supplies the theoretical content behind this paper's second moderator, managerial role reconfiguration: a manager whose activity mix shifts toward exception-handling and judgment as agents absorb routine coordination is positioned to sustain a wider span without a corresponding loss of decision quality, while a manager who continues performing the same routine-coordination tasks the agents now duplicate is positioned for role redundancy rather than role reconfiguration.

This distinction reframes the stark headcount predictions in Gartner (2024) and Ransbotham et al. (2025) as a story about which managers are eliminated rather than whether middle management as a category survives. A manager whose day consists chiefly of status aggregation, schedule coordination and routine sign-off is, on this account, performing exactly the function an AI agent is best positioned to absorb, and the elimination Gartner projects for over half of current middle-management positions by 2026 is plausibly concentrated among managers whose role has not reconfigured away from that function. Conversely, a manager who has already reallocated their time toward the judgment-intensive tasks Sastry (2025) identifies is positioned to absorb a wider span precisely because the agents are handling the coordination work that previously consumed the capacity such judgment work requires. This reframing is formalised as Proposition 3 in Section 4.

The broader labour-economics literature on task reallocation gives this reframing an established theoretical vocabulary rather than leaving it as an argument specific to management roles alone. Autor, Levy, and Murnane (2003) showed that computerisation over recent decades has systematically substituted for labour performing routine cognitive and manual tasks — those following explicit, codifiable rules — while complementing labour performing non-routine tasks requiring flexibility, judgment and interpersonal skill, a displacement-and-complementarity pattern that maps directly onto the routine-versus-exception distinction this paper's framework inherits from Garicano (2000). Acemoglu and Restrepo (2019) extended this task-based view into a full account of two offsetting forces: a displacement effect, in which automating technology directly substitutes for labour on the tasks it takes over, and a reinstatement effect, in which the same technological change creates new labour-intensive tasks — new roles that did not previously exist — that at least partly offset the displacement. Read together, this pair of arguments supplies a labour-economics parallel to the middle-management-specific claim developed above: AI-agent deployment density displaces the routine-coordination task content of the management role in exactly the sense Autor, Levy, and Murnane (2003) describe, while simultaneously creating a reinstatement opportunity — the judgment-intensive, exception-handling content Sastry (2025) identifies — that determines whether a given manager's role survives the transition rather than being eliminated by it. Middle management's aggregate fate under agentic AI is, on this reading, an empirical question about the relative size of these two forces rather than a foregone conclusion in either direction.

3.4 Delegation, Authority, and Span Under AI Augmentation

A wider span of control is not, by itself, evidence of effective delegation; it is equally consistent with a manager nominally overseeing more people while genuine decision authority remains centralised, either with the manager or with the agentic system itself. Colombo and Delmastro's (2004) empirical finding — that lower monitoring costs are associated with greater delegation of authority away from central management — offers one lens on this question, but delegation research grounded specifically in Indian workplaces adds a behavioural qualification directly relevant to AI-augmented teams: a study of delegated authority and resource provision found that granting employees decision latitude without also granting the resources needed to act on it — time, information access, organisational backing — failed to translate into the initiative and improved decision quality that delegation theory predicts, whereas authority paired with adequate resource provision did (Sastry & Ravish, 2025, The impact of delegated authority and resource provision on decision-making). Applied to the span-of-control question this paper addresses, this evidence implies that a wider span achieved through AI-agent deployment is not automatically accompanied by genuine delegation to the subordinates whose coordination burden the agents have reduced; a manager may retain a wider span while continuing to centralise the substantive decisions that matter, leaving subordinates with more autonomous coordination infrastructure but no more real decision authority than before.

A related study of delegation in the Indian workplace context reinforces this point from the employee side of the relationship, finding that transformative gains in employee initiative followed specifically from the combination of delegated authority and resource access, not from either alone (Sastry & Ravish, 2025, Empowering employees in the Indian workplace). Read together with the coordination-compression argument in Section 3.1, this delegation literature suggests that AI-agent deployment density's effect on span of control should be understood as potentially widening the manager's oversight capacity without necessarily widening the subordinate's decision latitude — two distinct organisational variables that the industry evidence in Section 3.2



does not distinguish. Where AI agents are deployed to give the manager better visibility into subordinate work rather than to give subordinates more autonomous decision authority, span of control may widen while the underlying locus of control remains exactly where it was, a possibility the author's broader analysis of the economic consequences of institutional authority exercised without a corresponding mechanism of check treats as a distinct organisational risk in its own right, independent of whether span of control itself has widened (Sastry, 2025, The economic costs of unchecked authority).

3.5 AI Deployment in HR and Organisational-Design Functions

The organisational functions responsible for designing reporting structures and managing the transitions this paper describes are themselves undergoing AI-driven redesign, a fact with direct implications for how quickly and how deliberately span-of-control changes of the kind reviewed in Section 3.2 are likely to be implemented. The author's study of AI-analytics-enabled recruitment found that efficiency gains materialised specifically from redesigning the hiring workflow to run sourcing, screening and interview-scheduling as parallel, analytics-informed stages rather than from inserting an AI tool into an otherwise unchanged sequential process (Sastry & Manjula Mallya, 2025, Revolutionizing HR hiring with AI analytics) — an instance, at the level of a single HR function, of the general workflow-redesign precondition for realised AI value that Section 3.2 identified as a common factor across the McKinsey (2026) and organisational AI-readiness evidence. The same logic extends naturally to organisational-design decisions about span of control itself: an organisation that treats AI-agent deployment as a tool inserted into an unchanged management structure, without redesigning reporting relationships and decision rights around the coordination capacity the agents actually provide, is unlikely to realise the span-widening potential this paper's framework identifies, echoing the distinction between deployment density and realised span drawn in Section 3.1.

Two further strands of the author's HR-focused work bear on how organisations are likely to manage the leadership and workforce-design dimensions of this transition. A study of leadership strategies for managing early-career talent argues that organisations restructuring around new technology need explicit strategies for developing the next generation of leaders rather than assuming leadership capability will emerge unaided from a flatter structure (Sastry & Manjula Mallya, 2025, Building future-ready organizations) — a caution directly relevant to the middle-management role reconfiguration described in Section 3.3, since managers newly responsible for wider, more judgment-intensive spans require deliberate development rather than an assumption that reduced headcount alone produces more capable survivors. A broader analysis of skill diversification and labour-market dynamics under technological disruption further argues that the workforce implications of a given technology are shaped as much by how deliberately organisations manage the accompanying skill transition as by the technology's raw displacement potential (Sastry & Manjula Mallya, 2025, Strategic HRM for economic resilience), a general claim this paper's framework specifies concretely for the span-of-control case: realised span expansion depends on deliberate managerial role reconfiguration (Section 3.3), not merely on agent deployment density considered in isolation.

3.6 The Indian IT-BPM Context

India's IT-BPM sector presents a distinctive starting point for the relationship this paper examines, because much of the sector's back-office and business-process outsourcing segment already operates with wide spans of control achieved through process standardisation rather than through AI-agent deployment — the pre-AI "standardised-process span" quadrant illustrated in Figure 5. High-volume, tightly scripted processes with limited task variation have historically supported wider-than-classical spans without any AI involvement at all, which means the marginal span-widening effect AI-agent deployment can offer in this segment is plausibly smaller than in organisational contexts where classical, narrow, judgment-intensive spans still prevail. Where the sector's AI-agent deployment is likely to have its largest marginal effect on span is instead in the more judgment-intensive service lines — technical support escalation, complex client-facing account management, and knowledge-process outsourcing — where task routineness has historically been lower and classical span limits have continued to bind.

India's own enterprise AI-readiness evidence, though pre-dating the agentic-AI period specifically, offers a relevant baseline for how quickly this marginal effect is likely to be realised. The NASSCOM–EY India Enterprise AI Adoption Index placed overall organisational readiness at 2.45 on a 4-point scale, with fewer than 15% of surveyed enterprises embedding AI strategy into corporate-level planning and the leading barrier cited as an inability to quantify AI benefit objectively (NASSCOM & EY, 2022) — the same benefit-quantification and workflow-redesign gaps that Section 3.2 and Section 3.5 identify as necessary preconditions for realised, rather than merely announced, span-of-control expansion. NITI Aayog has separately projected that increased AI



adoption could add US\$500–600 billion to India's GDP by 2035 (NITI Aayog, 2026), a projection that is, like the industry span-of-control claims reviewed in Section 3.2, conditional on organisations closing exactly the readiness gaps the NASSCOM–EY index documents rather than a certainty flowing automatically from AI deployment.

The sector's billable-effort commercial model adds a further, India-specific complication to the span-of-control question that the general literature reviewed in Sections 3.1 through 3.5 does not address. Where client contracts price services by headcount or billable hours rather than by outcome, an IT-BPM vendor that widens its internal managerial spans of control and reduces its management headcount realises an internal cost saving that does not automatically translate into a corresponding price change for the client — the same commercial-model tension the author's review of organisational AI-readiness flagged for productivity gains generally in this sector (Sastry, 2026, Organisational AI-readiness practice and the realised productivity gap from agentic AI). Applied to span of control specifically, this implies that Indian IT-BPM vendors have a countervailing commercial incentive to report AI-driven management-layer reductions cautiously, or to redirect the realised coordination-cost saving toward margin rather than toward passing efficiency through to clients — a distinction formalised as Proposition 5 in Section 4 and returned to as a distinct policy implication in Section 7.

4. CONCEPTUAL MODEL AND FORMAL PROPOSITIONS

The literature reviewed in Section 3 supports five formal propositions linking AI-agent deployment density to optimal span of control through the mediating and moderating structure set out in Figure 1. As with the largely descriptive, cross-sectional and industry-sourced evidence reviewed above, the propositions are offered as a research agenda for future empirical testing, not as findings already established by the sources reviewed here — a distinction elaborated in Sections 5 and 8.

Proposition 1 (Deployment Density and Coordination Cost). AI-agent deployment density is positively associated with optimal span of control, mediated by a reduction in coordination cost per subordinate; this relationship holds only where the surrounding workflow is redesigned to route the coordination work agents absorb away from the manager, and deployment density without corresponding workflow redesign yields no significant span-of-control change.

Proposition 2 (Task Routineness Moderation). The span-widening effect of AI-agent deployment density is stronger where the team's task portfolio is more routine and codifiable, and is attenuated where work remains exception-heavy and judgment-intensive, consistent with Garicano's (2000) knowledge-hierarchy prediction that span is set jointly by communication cost and the routine-to-exceptional problem ratio.

Proposition 3 (Managerial Role Reconfiguration). Managers whose own activity mix reconfigures from routine coordination toward exception-handling, judgment and people development as agents absorb routine load sustain a wider span without a decline in decision quality; managers whose role does not reconfigure are positioned for role redundancy rather than for a sustainably wider span.

Proposition 4 (Delegation Distinctness). A widened span of control achieved through AI-agent deployment is not evidence, by itself, of increased delegation of decision authority to subordinates; span-of-control expansion and delegated decision authority are distinct organisational variables that can move independently of one another, and realised employee initiative gains require delegated authority paired with adequate resource provision rather than wider span alone.

Proposition 5 (Commercial-Model Moderation). In labour-cost-arbitrage-dependent service sectors such as India's IT-BPM industry, the coordination-cost saving realised through AI-driven span-of-control expansion is moderated by commercial-model design; under an unmodified billable-effort or headcount-based contract, vendors retain an incentive to realise the saving as margin rather than to pass it through as a documented span-of-control or headcount change reported to clients.

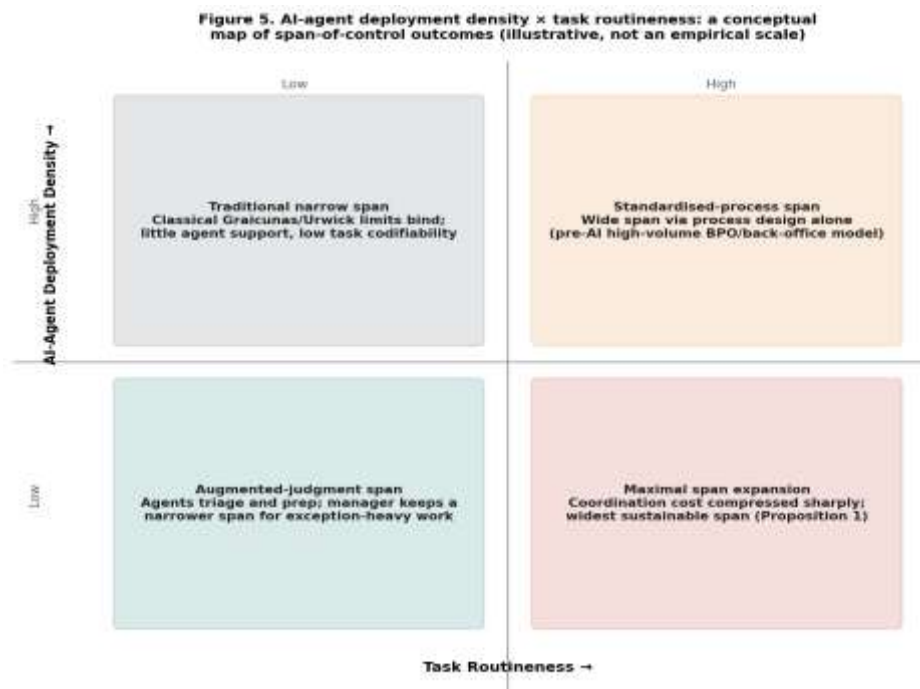


Figure 5. AI-agent deployment density × task routineness: a conceptual map of span-of-control outcomes (illustrative, not an empirical scale).

No.	Construct	Proposition (summary)	Theoretical Grounding
P1	Deployment density → coordination cost	Density raises span via reduced coordination cost, conditional on workflow redesign.	Garicano (2000); Farach (2026); Sastry (2026)
P2	Task routineness (moderator)	Effect stronger for routine/codifiable work; attenuated for exception-heavy work.	Garicano (2000); Garicano & Rossi-Hansberg (2006)
P3	Managerial role reconfiguration (moderator)	Role reconfiguration toward judgment/exception-handling sustains wider span.	Sastry (2025, Middle management evolution)
P4	Delegation distinctness	Wider span ≠ more delegated authority; both required for realised initiative gains.	Colombo & Delmastro (2004); Sastry & Ravish (2025, x2)
P5	Commercial-model moderation (India IT-BPM)	Billable-effort contracts incentivise retaining the coordination-cost saving as margin.	Sastry (2026, AI-readiness); author's extension

5. METHOD: REVIEW PROTOCOL AND CORPUS

This paper is a structured, integrative literature review rather than a PRISMA-governed systematic review, and that distinction is stated explicitly rather than left implicit. A systematic review with documented database search strings, screening counts and inclusion/exclusion criteria is best suited to a literature mature enough that a bounded, retrievable set of studies measures a common construct; AI-agent deployment density does not yet meet that condition, since, as Sections 3.1 and 3.2 establish, no source in the literature validates or even attempts to measure the construct directly. The corpus underlying this paper was instead assembled in two passes: a backward citation-chase from the classical and organisational-economics span-of-control literature (Graicunas, 1933/1937, forward through Garicano, 2000, and its empirical descendants) to identify the theoretical anchors a claim about AI-agent-driven span change must be judged against, and a targeted forward search of 2024–2026 industry, consulting and academic sources on agentic-AI deployment and organisational structure to identify the still-emerging empirical evidence such a claim can currently draw on. This two-pass method mirrors the approach used in the author's related review of organisational AI-readiness practice (Sastry, 2026, Organisational AI-readiness



practice and the realised productivity gap from agentic AI), where the same absence of a mature, directly measurable construct made a full systematic protocol premature.

Every external source cited in this paper was checked against a primary or near-primary record — a journal's own page or DOI record for the academic literature, and the originating press release, company announcement or named report for the industry evidence — rather than accepted from a secondary summary, and several citation details were corrected during this checking process before drafting: Garicano's (2000) full title includes "in Production," a phrase frequently dropped in secondary citations; Colombo and Delmastro's (2004) empirical test of delegation was published in 2004, not the 2008 date the author's own working notes initially carried; Fayol's ideas were first circulated in 1916 but did not appear in book form until 1918, so this paper follows convention in citing the dual date 1918/1949 (the latter being the standard English translation); and Graicunas's foundational essay, though written and privately circulated in 1933, did not appear in its now-standard published form until the 1937 Gulick and Urwick collection, so this paper cites it as 1933/1937 for the same reason. The full corpus totals thirty-six cited works, classified in Table 5 and Figure 4 into six categories spanning classical administrative theory through the author's own prior research.

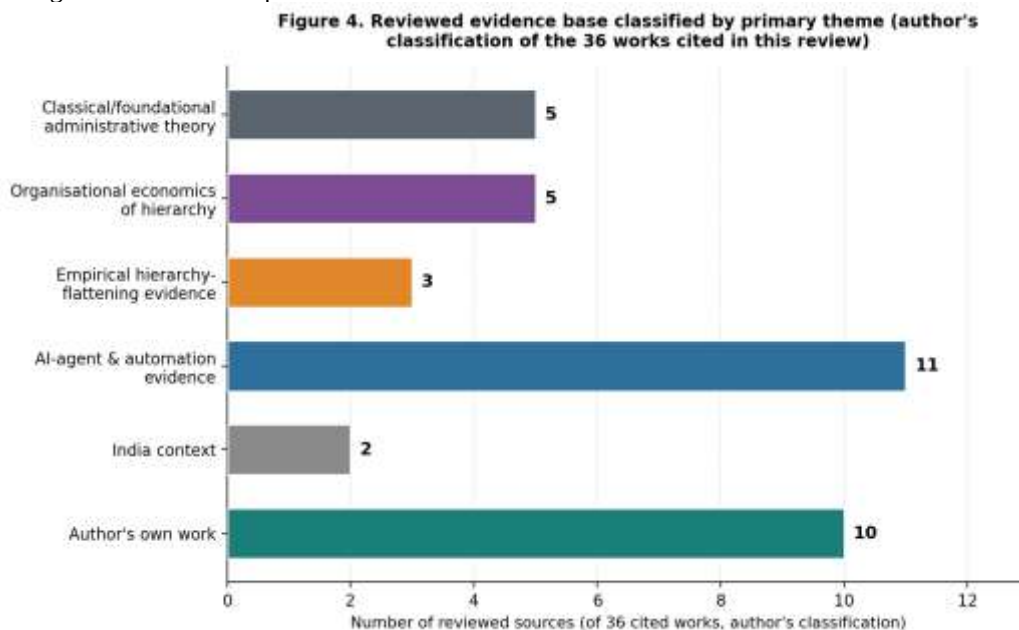


Figure 4. Corpus composition by category (n = 36).

Category	Description	n	Representative Sources
Classical/foundational administrative theory	Pre-1980 administrative and organisational theory establishing span-of-control and hierarchy fundamentals.	5	Graicunas (1933/1937); Urwick (1956); Fayol (1918/1949); Simon (1947); Mintzberg (1979)
Organisational economics of hierarchy	Cost-based models of hierarchical control loss, structural contingency and knowledge hierarchies.	5	Williamson (1967); Chandler (1962); Blau (1970); Garicano (2000); Garicano & Rossi-Hansberg (2006)
Empirical hierarchy-flattening evidence	Empirical tests of communication-cost-driven flattening and delegation, pre-dating agentic AI.	3	Rajan & Wulf (2006); Guadalupe & Wulf (2010); Colombo & Delmastro (2004)
AI-agent & automation evidence	Formal models, field experiments and industry surveys on AI-agent deployment, automation and organisational flattening.	11	Farach (2026); Kellogg, Valentine, & Christin (2020); Dell'Acqua et al. (2025); Gartner (2024); Ransbotham et al. (2025); Amazon.com, Inc. (2024); Microsoft WorkLab (2025, 2026); McKinsey & Company (2026); Autor, Levy,



			& Murnane (2003); Acemoglu & Restrepo (2019)
India context	India-specific enterprise AI-readiness and macroeconomic evidence.	2	NASSCOM & EY (2022); NITI Aayog (2026)
Author's own work	The author's prior published and working-paper research on middle management, delegation, HR-AI deployment and organisational structure.	10	Sastry (2024, 2025 ×4, 2026); Sastry & Ravish (2025 ×2); Sastry & Manjula Mallya (2025 ×3)

Table 5. Corpus classification by category, matching Figure 4 (total n = 36).

6. DISCUSSION

The five propositions developed in Section 4 converge on a single qualification of the headline industry claims reviewed in Section 3.2: AI-agent deployment density is a necessary input into span-of-control expansion, not a sufficient one. Proposition 1's workflow-redesign condition, Proposition 3's role-reconfiguration condition, and Proposition 4's delegation-distinctness claim each identify a way in which an organisation can deploy agents densely and still fail to realise a wider, effectively supervised span — through unredesigned workflows, through managers who continue performing the routine-coordination tasks the agents now duplicate, or through a span that widens on paper while decision authority remains exactly as centralised as before. Read together with Farach's (2026) distributional fork, these conditions suggest that the aggregate industry predictions in Gartner (2024) and Ransbotham et al. (2025) are more likely to describe an upper bound on realisable change than a central forecast: the fraction of organisations that satisfy all three conditions simultaneously is plausibly smaller than the fraction that merely deploys agents or merely announces a headcount target, and it is this smaller fraction whose experience should determine how much of the predicted flattening actually materialises.

This conditionality carries a further implication the industry evidence does not address at all: what happens to the managers who remain once flattening occurs. The author's empirical study of talent designation and social comparison found that how an organisation designates and visibly ranks talent shapes employees' economic behaviour and perceived organisational growth prospects largely through social-comparison mechanisms — employees evaluate their own position relative to visible peer outcomes, not against an absolute standard (Sastry, 2025, Talent Designation and Social Comparison). Applied to the reconfiguration this paper describes, a flattening episode that eliminates a share of middle-management positions changes the visible comparison set for every manager who survives it, whether or not that manager's own role has genuinely reconfigured toward the judgment-intensive work Section 3.3 describes. A manager who is retained but perceives the retention as arbitrary relative to peers who were let go — rather than as a legible consequence of having reconfigured toward exception-handling and judgment — is, on the social-comparison logic Sastry (2025) documents, positioned for exactly the morale and attrition risk that would undermine Proposition 3's premise that reconfigured managers sustain a wider span without a decline in decision quality. Organisations implementing AI-driven span expansion therefore have a practical reason, independent of the technical coordination-cost argument, to make the basis for who is retained and who is not as legible as possible.

The reliability risk identified in Section 3.1 — the jagged frontier documented by Dell'Acqua et al. (2025) — carries a parallel practical implication for how a widened span should be managed rather than simply granted. Because an AI agent's competence boundary does not correspond to any intuitive notion of task difficulty, a manager who has widened their span on the assumption that the agent reliably resolves routine cases has no reliable internal signal for when that assumption has silently failed. The practical response this paper's framework suggests is not to withhold span expansion until agent reliability is perfect — a standard current agentic systems do not meet and may not meet for some time — but to design the human-oversight checkpoint deliberately around the specific tasks where the agent's competence boundary is least well characterised, concentrating the manager's own scarce judgment exactly where the jagged frontier is most likely to bite rather than spreading a fixed oversight budget evenly across all agent-resolved cases. This is, in effect, a second-order application of Garicano's (2000) own logic: just as the original model routes exceptional problems to the manager and leaves routine ones to be resolved locally, a well-designed AI-augmented span should route the agent's own most uncertain calls to the manager rather than treating every agent-resolved case as equally trustworthy.

Taken together, these three threads — the workflow and role-reconfiguration conditions on realised span change, the social-comparison risk attached to who survives a flattening episode, and the reliability-management task attached to a widened span — describe a considerably more conditional and more managerially demanding process than the industry evidence in Section 3.2, read on its own, would suggest. None of this undermines the



directional prediction that motivates the paper: AI-agent deployment density should, on the theoretical grounds set out in Section 2, widen sustainable span of control at the margin. It does suggest that the size of that margin in any given organisation is set less by how many agents are deployed than by how deliberately the surrounding workflow, management role and oversight design are built around them.

7. INDIA POLICY AND PRACTICE IMPLICATIONS

Section 3.6 established that India's IT-BPM sector already operates with wide, standardised-process spans of control in its back-office and business-process outsourcing segments, and that agentic AI's marginal effect on span is therefore likely to be largest in the sector's more judgment-intensive service lines rather than in the segments already operating near a process-standardisation ceiling. Table 6 sets out this differentiated picture in practice terms across three illustrative segments. The practical implication for Indian IT-BPM organisations is that a single, sector-wide AI-driven span-of-control target — of the kind Gartner's (2024) and the Amazon (2024) figures might tempt an organisation to adopt uniformly — is likely to misallocate deployment effort, concentrating agents in segments where task routineness has already been substantially resolved by process standardisation while under-deploying them in the technical-escalation and complex-account-management lines where the coordination-cost reduction Section 2's framework identifies would have the largest marginal effect.

This differentiated-deployment argument connects to a distinct empirical finding from the author's own prior work on organisational structure. A dedicated study of departmental structure and task achievement found that structural design choices — how work is departmentalised, and how responsibility for a given task is distributed across departmental boundaries — have a measurable effect on whether a task is actually achieved, independent of the resources committed to it (Sastry, 2024, An Impact of Departmental Structure on Task Achievement). Applied to the span-of-control redesign decisions this paper addresses, that finding is a caution against treating AI-driven span widening as a pure headcount-reduction exercise evaluated only against a cost target: the same structural change that reduces coordination cost per subordinate on paper can, if it crosses departmental boundaries carelessly or leaves task ownership ambiguous where an agent's authority ends and a human's begins, reduce task achievement even as it appears to succeed on the headcount metric alone. Indian IT-BPM organisations redesigning reporting structures around agentic AI have, on this evidence, a specific practice reason to evaluate any proposed span-of-control change against task-achievement outcomes directly, not only against the coordination-cost or headcount-ratio metrics the industry evidence in Section 3.2 emphasises.

At the policy level, the task-reallocation framing developed in Section 3.3 — echoing Acemoglu and Restrepo's (2019) displacement-and-reinstatement account and Autor, Levy, and Murnane's (2003) routine-versus-non-routine task distinction — implies that India's skilling institutions have a specific, identifiable role to play that is distinct from the general AI-skilling agenda those institutions already pursue. If middle-management roles survive AI-driven flattening specifically through a shift toward judgment-intensive, exception-handling work, then the skilling gap this transition creates is not primarily a technical AI-literacy gap of the kind most current national skilling programmes target, but a managerial-judgment and exception-handling capability gap for incumbent managers who have spent much of their career performing the routine-coordination tasks now being absorbed by agents. NITI Aayog's (2026) GDP-contribution projections for AI adoption are, on this reading, conditional not only on the workflow-redesign precondition this paper's framework identifies but on a parallel, currently unaddressed managerial-reskilling precondition specific to the middle-management layer.

IT-BPM Segment	Current Typical Span (illustrative)	Task Routineness	Marginal Effect of AI-Agent Deployment	Practice Implication
Back-office / BPO standardised process	Wide (pre-AI, via process standardisation)	High	Small — routineness ceiling already largely resolved without agents	Deploy agents chiefly for exception-triage refinement, not further span widening
Technical support escalation	Moderate (classical span limits bind)	Moderate	Large — escalation triage is precisely what agents absorb	Priority segment for agent deployment; pair with role-reconfiguration investment (Sec. 3.3)
Complex client-facing / KPO account management	Narrow (judgment-intensive, classical span)	Low	Moderate, conditional on workflow redesign (Prop. 1)	Deploy agents for coordination support only; guard against jagged-frontier over-reliance (Sec. 6)

Table 6. Illustrative India IT-BPM segment comparison: task routineness, marginal AI-agent effect on span, and practice implications.



8. LIMITATIONS AND FUTURE RESEARCH AGENDA

This paper's limitations follow directly from the review method described in Section 5. Because the corpus was assembled through structured citation-chasing and targeted search rather than a documented systematic search protocol with reproducible screening counts, the corpus composition in Table 5 reflects the author's judgment about which sources bear most directly on the paper's argument rather than an exhaustive or protocol-guaranteed sample of the relevant literature; a differently constructed search could surface additional industry or academic sources, particularly given how rapidly the agentic-AI evidence base is expanding. The five propositions in Section 4 are, correspondingly, offered as a research agenda rather than as findings already established by the sources reviewed: no source in this corpus tests any of the five propositions directly, and each represents this paper's synthesis of what the existing theoretical and industry evidence implies rather than an empirical result in its own right.

The single highest-priority item on the resulting future-research agenda is the measurement gap this paper has repeatedly flagged rather than resolved: no validated instrument for AI-agent deployment density currently exists anywhere in the literature reviewed here. Developing one is a necessary precondition for testing any of this paper's five propositions empirically, and would plausibly require, at minimum, a count of actively deployed agents performing coordination, monitoring or exception-triage functions per team member; a measure of the share of a team's task volume those agents actually resolve without escalation, as distinct from the share they are nominally authorised to handle; and a way of distinguishing genuine workflow redesign around agent capacity (Proposition 1's condition) from agent deployment layered onto an otherwise unchanged reporting structure. Absent such an instrument, every industry statistic reviewed in Section 3.2 will remain, as Table 2 documents explicitly, a proxy for the construct this paper argues actually drives span-of-control change rather than a direct measure of it.

A second limitation concerns the evidentiary weight the Amazon (2024) case can bear. As Section 3.2 makes clear, the company's own stated rationale for its individual-contributor-to-manager ratio mandate was reducing bureaucracy and increasing decision speed, not AI-driven coordination-cost reduction specifically; treating it as direct evidence for this paper's argument would overstate what the source actually supports, and it is included here, as it was in Section 3.2, as a large-scale illustration of the same directional shift rather than as a causal test of the AI-specific mechanism this paper proposes. A related limitation applies to the broader industry evidence base: Gartner's (2024) and Ransbotham et al.'s (2025) figures are, respectively, a prediction and a set of survey respondents' expectations, not measurements of realised span-of-control change after the fact, so the entire empirical case for AI-driven flattening reviewed in this paper currently rests on announced intentions and forecasts rather than on outcomes observed once the relevant AI-agent deployments have run their course.

Beyond closing the measurement gap, three further research directions follow naturally from the discussion in Section 6. First, longitudinal tracking of realised — rather than announced or predicted — span-of-control change at the team level, ideally paired with a direct deployment-density measure once one exists, would allow Propositions 1 and 2 to be tested against outcomes rather than against stated intentions. Second, empirical work on Farach's (2026) distributional fork — whether coordination-compressing capital access is in practice concentrated among senior managers or broadly distributed through the management layer in specific organisations and sectors — would clarify which organisations are likely to experience the shared-productivity-gain path this paper's framework implies versus the widening-gap path Farach's model also allows. Third, given the social-comparison risk identified in Section 6, research specifically on manager morale, retention and performance following AI-driven flattening episodes — testing whether the legibility of retention criteria moderates the risk Sastry (2025, Talent Designation and Social Comparison) documents in this specific organisational-change context — would help establish whether Proposition 3's premise (that reconfigured managers sustain performance) holds under realistic organisational conditions or only under idealised ones.

9. CONCLUSION

This paper has argued that AI-agent deployment density belongs inside, rather than alongside, the century-long theoretical tradition running from Graicunas (1933/1937) through Garicano (2000) that treats span of control as the resolution of a coordination-cost trade-off. Read through that tradition's own logic, a technology that autonomously resolves a growing share of routine coordination work and pre-processes the exceptions that remain should widen sustainable span of control through precisely the communication-cost mechanism the tradition already specifies, and the industry evidence assembled since 2024 — however imperfectly it measures the construct this paper isolates — is broadly consistent with that prediction. The paper's more particular contribution is the set of conditions under which that widening is actually realised rather than merely announced: workflow



redesign around the coordination capacity agents provide, managerial role reconfiguration toward exception-handling and judgment, a clear distinction between wider span and genuinely delegated authority, and, in India's IT-BPM sector specifically, a commercial-model structure that does not simply retain the coordination-cost saving as margin. Each of these conditions is a point at which the straightforward story implied by Gartner's (2024) and Ransbotham et al.'s (2025) headline figures can fail to materialise in a given organisation even as AI-agent deployment proceeds.

The paper closes, deliberately, without resolving the question its own framework depends on most: how much AI-agent deployment density any given team actually has, and how that density maps onto realised rather than announced span-of-control change. That no validated answer to this question currently exists in the literature reviewed here is, this paper has argued, the central item on the field's research agenda rather than a peripheral gap — every proposition in Section 4 is, in the end, a claim about how that still-unmeasured construct should behave once it can be measured. Until it can, the relationship between AI-agent deployment and the reshaping of middle-management span of control remains a theoretically well-grounded and industry-corroborated hypothesis rather than an empirically established fact, and this paper's five propositions are offered as the agenda for closing that distinction rather than as a substitute for doing so.

NOTE ON THE MANUSCRIPT

This paper follows the author's established practice of weaving citations to his own prior published and working-paper research into the literature review on a subsection-by-subsection basis, engaged on the substance of each argument rather than clustered into a single self-citation block. Ten of the paper's thirty-six references (28%) are the author's own work, spanning eight distinct prior studies: two are peer-reviewed journal articles co-authored with P. Y. Ravish (Sastry & Ravish, 2025, ×2, EPRA International Journal of Multidisciplinary Research and EPRA International Journal of Research & Development); three are peer-reviewed journal articles co-authored with M. Manjula Mallya (Sastry & Manjula Mallya, 2025, ×3, EPRA International Journal of Research & Development); one is a peer-reviewed journal article on departmental structure (Sastry, 2024, EPRA International Journal of Economics, Environment and Commerce Management); one is a peer-reviewed journal article on talent designation and social comparison (Sastry, 2025, EPRA International Journal of Economics, Business and Management Studies); and two — Middle Management Evolution in Contemporary Organizations (Sastry, 2025) and The Economic Costs of Unchecked Authority (Sastry, 2025) — are working papers, self-archived rather than journal-published, and are cited as such rather than attributed to a journal they do not appear in. The tenth, Sastry (2026, Organisational AI-readiness practice and the realised productivity gap from agentic AI), is the author's own separate Scopus-track manuscript in this same research programme and is likewise cited as an independent prior work rather than presented as peer-reviewed where it is not.

Two factual caveats flagged during drafting are restated here for transparency rather than left implicit in the body text alone. First, the Amazon.com, Inc. (2024) individual-contributor-to-manager ratio mandate is treated throughout this paper as illustrative of AI-driven organisational flattening, but the company's own stated rationale for the mandate was reducing bureaucracy and increasing decision-making speed, not AI capacity specifically; this paper does not claim the Amazon case as direct evidence of the AI-specific mechanism it proposes, and says so explicitly at each point the case is used (Sections 1, 3.2 and 8). Second, two classical sources carry a dual publication date rather than the single date sometimes seen in secondary citations of them: Fayol's administrative principles were first circulated in 1916 but did not appear in book form until 1918 (cited here as 1918/1949, the latter being the standard English translation), and Graicunas's foundational span-of-control essay was written and privately circulated in 1933 but did not appear in its now-standard published form until the 1937 Gulick and Urwick collection (cited here as 1933/1937). Both corrections were identified through direct verification against primary or near-primary bibliographic records during the drafting process rather than accepted from secondary summaries.

This paper is presented as a structured, integrative literature review rather than a PRISMA-governed systematic review, for the reason given in Section 5: the central construct it addresses, AI-agent deployment density, has no validated measurement instrument in the literature at the time of writing, which makes a documented systematic search protocol premature rather than more rigorous. All five formal propositions in Section 4 are accordingly offered as a research agenda for future empirical testing, consistent with the framing set out in Sections 5, 6 and 8, and should not be read as findings already established by the sources this paper reviews.



REFERENCES

1. Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2), 3–30.
2. Amazon.com, Inc. (2024, September 16). Update from Amazon CEO Andy Jassy on return-to-office plans and manager team ratio. About Amazon. <https://www.aboutamazon.com/news/company-news/ceo-andy-jassy-latest-update-on-amazon-return-to-office-manager-team-ratio>
3. Autor, D. H., Levy, F., & Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4), 1279–1333.
4. Blau, P. M. (1970). A formal theory of differentiation in organizations. *American Sociological Review*, 35(2), 201–218.
5. Chandler, A. D., Jr. (1962). *Strategy and structure: Chapters in the history of the American industrial enterprise*. MIT Press.
6. Colombo, M. G., & Delmastro, M. (2004). Delegation of authority in business organizations: An empirical test. *Journal of Industrial Economics*, 52(1), 53–80.
7. Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2025). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Organization Science*. <https://doi.org/10.1287/orsc.2025.21838>
8. Farach, A. (2026). AI as coordination-compressing capital: Task reallocation, organizational redesign, and the regime fork [Working paper].
9. Fayol, H. (1949). *General and industrial management* (C. Storrs, Trans.). Pitman. (Original work published 1918)
10. Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5), 874–904.
11. Garicano, L., & Rossi-Hansberg, E. (2006). Organization and inequality in a knowledge economy. *Quarterly Journal of Economics*, 121(4), 1383–1435.
12. Gartner. (2024, October 22). Gartner unveils top predictions for IT organizations and users in 2025 and beyond [Press release].
13. Graicunas, V. A. (1937). Relationship in organization. In L. Gulick & L. Urwick (Eds.), *Papers on the science of administration* (pp. 183–187). Institute of Public Administration, Columbia University. (Original work circulated 1933)
14. Guadalupe, M., & Wulf, J. (2010). The flattening firm and product market competition: The effect of trade liberalization on corporate hierarchies. *American Economic Journal: Applied Economics*, 2(4), 105–127.
15. Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals*, 14(1), 366–410.
16. McKinsey & Company. (2026). The state of AI in 2026: On the road to ROI.
17. Microsoft WorkLab. (2025). 2025: The year the Frontier Firm is born. Microsoft Work Trend Index.
18. Microsoft WorkLab. (2026). 2026 Work Trend Index: Agents, human agency, and the opportunity for every organization.
19. Mintzberg, H. (1979). *The structuring of organizations: A synthesis of the research*. Prentice-Hall.
20. NASSCOM, & EY. (2022). India enterprise AI adoption index 1.0. National Association of Software and Service Companies.
21. NITI Aayog. (2026). Increased adoption of AI can add US\$500–600 billion to GDP by 2035 [As reported]. India Brand Equity Foundation.
22. Rajan, R. G., & Wulf, J. (2006). The flattening firm: Evidence from panel data on the changing nature of corporate hierarchies. *Review of Economics and Statistics*, 88(4), 759–773.
23. Ransbotham, S., Kiron, D., Khodabandeh, S., Iyer, R., & Das, R. (2025). The emerging agentic enterprise: How leaders must navigate a new age of AI. MIT Sloan Management Review, in collaboration with the Boston Consulting Group Henderson Institute.
24. Sastry, N. S. K. (2024). An impact of departmental structure on task achievement. *EPRA International Journal of Economics, Environment and Commerce Management (IJEECEM)*, 11(9).
25. Sastry, N. S. K. (2025). Middle management evolution in contemporary organizations [Working paper, self-archived].
26. Sastry, N. S. K. (2025). The economic costs of unchecked authority [Working paper, self-archived].
27. Sastry, N. S. K. (2025). Talent designation and social comparison: Economic drivers of organizational growth – An empirical analysis. *EPRA International Journal of Economics, Business and Management Studies (IJEBMS)*, 12(10).
28. Sastry, N. S. K. (2026). Organisational AI-readiness practice and the realised productivity gap from agentic AI [Manuscript].
29. Sastry, N. S. K., & Manjula Mallya, M. (2025). Building future-ready organizations: Leadership strategies for young talent management. *EPRA International Journal of Research & Development (IJRD)*, 10(9), 88–91.
30. Sastry, N. S. K., & Manjula Mallya, M. (2025). Revolutionizing HR hiring with AI analytics: Enhancing efficiency and attracting top talent. *EPRA International Journal of Research & Development (IJRD)*, 10(9), 125–131.



31. Sastry, N. S. K., & Manjula Mallya, M. (2025). Strategic HRM for economic resilience: Skill diversification, labor market dynamics and technology disruption. *EPRA International Journal of Research & Development (IJRD)*, 10(11), 408–414.
32. Sastry, N. S. K., & Ravish, P. Y. (2025). Empowering employees in the Indian workplace: The transformative power of delegation and resources. *EPRA International Journal of Research & Development (IJRD)*, 10(7).
33. Sastry, N. S. K., & Ravish, P. Y. (2025). The impact of delegated authority and resource provision on decision-making – Fostering employee initiative. *EPRA International Journal of Multidisciplinary Research (IJMR)*, 11(7).
34. Simon, H. A. (1947). *Administrative behavior: A study of decision-making processes in administrative organization*. Macmillan.
35. Urwick, L. F. (1956). The manager's span of control. *Harvard Business Review*, 34(3), 39–47.
36. Williamson, O. E. (1967). Hierarchical control and optimum firm size. *Journal of Political Economy*, 75(2), 123–138.