Louvain School of Management

What Makes Human-AI Collaboration Work in Organisations: An Empirical Framework of Four Coordinated Dimensions

Author(s): Maxime Kohnen
Supervisor(s): Philippe Chevalier
Academic year 2026-2027
Dissertation for the master of CEMS
Master subject and focus in Master ingénieur de gestion à finalité spécilalisée
Daytime schedule

DECLARATION

The declaration below is **mandatory and must appear on the first page (after the front page) of your master's thesis**.

STATEMENTS ON YOUR USE OF GENERATIVE AI

Confirm the following statements:

Statements on responsible use of generative AI	**Your answer**
The content of my/our master's thesis is my/our original out~~in~~put, even if I/we used generative AI to help prepare it.	☒Yes ☐No
I/we master the theories, tools and methods that I/we use for the thesis.	☒Yes ☐No
I/we verified that the output presented in the thesis is accurate and valid, and that it contains no hallucination, no false reference, etc.	☒Yes ☐No
I/we acknowledge that I/we master the final output and that I am/we are able to explain it and answer questions about it.	☒Yes ☐No
I/we made sure not to provide generative AI tools with access to confidential data or information	☒Yes ☐No

SIGNATURE

Sign your declaration:

	Author last name, first name(s)	**Date**	**Signature**
1	Maxime Kohnen	28/05/2026	

I would like to thank my supervisor, Professor Philippe Chevalier, for the guidance and thoughtful feedback that shaped this work and Justin Loroy for his support throughout.

Abstract :

Generative artificial intelligence has entered the workplace faster than almost any technology before it, yet aggregate productivity has not risen in proportion. This thesis treats that disconnect as a management problem rather than a technological one. Its empirical starting point is the meta-analysis by Vaccaro et al. (2024), which found that across 106 experimental studies, human-AI combinations performed, on average, worse than the better of humans or AI alone. The thesis asks how organisations can coordinate task allocation, calibration and interface and process design to sustain human-AI collaboration.

The argument is structured by the automation-augmentation paradox of Raisch and Krakowski (2021), which holds that automation and augmentation are interdependent forces rather than a simple trade-off, so that sustainable productivity gains depend on managing the tension between them. From this lens the thesis derives three organisational levers: task allocation which maps work to the right actor along the jagged technological frontier (Dell'Acqua et al., 2023); calibration which concerns whether users hold accurate beliefs about AI reliability (Caplin et al., 2025); and interface and process design which configures systems to encourage appropriate reliance (Al-Refai et al., 2025). The three are integrated into a framework specified through six propositions.

The framework is tested through an abductive design (Dubois & Gadde, 2002) drawing on 28 semi-structured interviews across approximately 22 firms and 14 industries, analysed with the Gioia method (Gioia et al., 2013). The original three-department sampling frame did not hold once data collection began and recruitment was redirected toward deployment-side voices with cross-departmental visibility.

The framework survives confrontation with the evidence in a refined form: two propositions are reformulated, one is restated and a fourth dimension, the junior talent pipeline, emerges from the data. Two cross-cutting findings are named: human accompaniment as the binding condition under which the levers operate and pace-mismatch as the pressure exerted by rapid AI tool evolution. The thesis concludes that the tension is resolved not at the level of any single lever but across the four dimensions together, under the condition of human accompaniment.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm

Table of content

Introduction

Generative AI has entered the workplace at a pace that rivals the most transformative technologies of the past half-century, yet aggregate productivity statistics show no commensurate gain. This thesis examines that disconnect and locates it as a management problem rather than a technological one.

The topic emerged from a practical observation made during an internship in a firm undergoing its first wave of AI tool deployment: a striking gap between the public "honeymoon" discourse on AI productivity gains and the actual conditions under which AI was being used inside the firm. The public discourse did not match what professionals were doing with AI or what their managers were observing about adoption, calibration, and verification. The gap is analytically significant because firm-level decisions being made now, particularly around junior hiring, governance and tool procurement, will shape workforce architecture for the next decade. The thesis is an attempt to make the gap between AI rhetoric and AI practice visible at the level of organisational mechanism.

The meta-analysis by Vaccaro et al. (2024) establishes the empirical ground: across 106 experimental studies, human-AI combinations performed significantly worse on average than the best of either humans or AI alone. The literature reviewed in Chapters 1 and 2 shows that the under-performance is not random. It follows identifiable patterns related to the task's position on the jagged technological frontier (Dell'Acqua et al., 2023a), the user's calibration about AI reliability (Caplin et al., 2025), and the team-coordination costs that automation imposes on social work (Dell'Acqua et al., 2023b). Chapter 3 frames the management challenge through the theoretical lens of Raisch and Krakowski's (2021) automation-augmentation paradox: automation and augmentation are not a simple trade-off but interdependent forces and managing them together is what unlocks synergy.

Research question: *How can organisations coordinate task allocation, calibration, and interface and process design to sustain human-AI collaboration?*

The question is approached through an abductive research design (Dubois & Gadde, 2002). Chapters 4 through 6 develop the three organisational levers, task allocation, calibration training, and interface and process design, as the mechanisms through which firms navigate the paradox. Chapter 7 integrates the three into a framework testable through six theoretical propositions. Then chapter 8 specifies the methodology: 28 semi-structured interviews conducted across approximately 22 firms and 14 industries, analysed using the Gioia method (Gioia et al., 2013) across four iterative consolidation milestones. Next, chapter 9 reports the findings: the three-lever framework survives the confrontation with the data, but in a refined form. Two propositions are reformulated (P4 relocates the operative axis from seniority to disposition; P5 relocates the operative design mechanism from downstream warnings to upstream scoping), one is re-stated (P2 as no-governance baseline), and one additional aggregate dimension (Junior Pipeline) is added as an emergent contribution. Two transveral findings, human accompaniment as binding meta-condition and pace-mismatch as cross-cutting temporal pressure, are developed in the synthesis section. Chapter 10 draws out the theoretical, managerial and methodological implications.

The contribution this thesis seeks to make is twofold. Theoretically, it provides an empirical confrontation of the three-lever framework derived from Raisch and Krakowski (2021), Dell'Acqua et al. (2023a, 2023b), and Caplin et al. (2025), refining two of its propositions and adding one aggregate dimension the original framework did not anticipate. Practically, it documents an abductive design in action, with four iterative consolidation milestones and two methodologically transparent revisions visible in the audit trail. The most consequential difficulty encountered during the empirical phase was the abductive pivot itself: the initial sampling frame of three functional departments (Marketing, Legal, Sales) did not survive contact with the data. Early interviews surfaced that the paradox was most cleanly articulated by deployment-side voices with cross-departmental visibility, and recruitment was redirected accordingly. The pivot is documented in Chapter 8 §8.4 and is presented as a methodological flag rather than as a defect of the design.

The argument the thesis defends is empirical and specific. Sustainable productivity gains from AI depend on managing the tension between automation and augmentation. The tension is not resolved at the level of any single lever. It is resolved at the level of the four dimensions taken together (Task Allocation, Calibration, Interface and Process Design, Junior Pipeline), under the binding condition of human accompaniment against the cross-cutting pressure of rapid AI tool evolution. The under-performance documented by Vaccaro et al. (2024) reflects the prevalence of firms that have not yet made that integration. The upper-bound performance visible in the corpus reflects firms that have begun to.

Chapter 1: The Modern Productivity Paradox

1.1 The Scale of the Promise and the Persistence of the Paradox

Generative AI has entered the workplace at a pace that rivals the most transformative technologies of the past half-century. As Bick et al. (2024) document, "relative to each technology's first mass-market product launch, work adoption of generative AI has been as fast as the personal computer (PC), and overall adoption has been faster than either PCs or the internet." The scale and speed of this deployment make the question of aggregate productivity impact both urgent and surprisingly difficult to answer.

The difficulty is not new. Brynjolfsson et al. (2017) document what they call the modern productivity paradox: despite decades of investment in information technology, aggregate U.S. labor productivity growth declined by roughly half between the 1995–2004 period (2.8% annually) and the 2005–2016 period (1.3% annually). The most transformative technologies tend to produce the most pronounced measurement and implementation challenges and their macroeconomic benefits materialise slowly and unevenly. Generative AI intensifies rather than resolves this paradox and understanding why requires examining the four explanatory mechanisms Brynjolfsson and colleagues propose.

1.2 Four Explanations for the Productivity Paradox

Genuine economic disappointment. Acemoglu (2024) formalises this possibility in a task-based macroeconomic model, arguing that only roughly 4.6% of all tasks meet the conditions for economically significant AI automation given current capabilities and cost structures. This specific estimate may already understate current capabilities given the pace of model improvement, yet Acemoglu's underlying point holds: the share of tasks with cost-effective automation remains far narrower than aggregate enthusiasm suggests. The key distinction is between "easy-to-learn" tasks (where AI can synthesise reliable patterns from training data) and "hard-to-learn" tasks (where decision-making depends on contextual, tacit knowledge that resists codification).

Mismeasurement. Syverson (2010) documents persistent productivity dispersion across firms. This dispersion is consistent with the view that benefits of new technologies concentrate early in high-performing firms whose gains are partially absorbed in aggregate indices. Brynjolfsson et al. (2021) go further: the most valuable investments complementary to general purpose technologies (GPT), such as new business processes, organisational redesign, and human capital development, are intangible assets that do not appear in conventional national accounts. These investments create a productivity J-curve: measured productivity initially stagnates as firms invest in transformation, then rises as intangible capital accumulates. Adjusting for intangible capital tied to software, computer hardware and R&D yields total factor productivity levels 15.9% higher than official measures by 2017 (Brynjolfsson et al., 2021). Historical debates during France's post-war productivity movement (1944–1955) show that the challenge of measuring productivity during periods of technological change is not new (Boulat, 2006).

Concentrated and offset gains. Brynjolfsson et al. (2017) describe a "gold rush" dynamic in which AI-driven competition may produce net dissipative effects at the aggregate level even as individual winners prosper. The mechanism remains empirically contested.

Implementation lags and complementary investment. The fourth explanation, which Brynjolfsson et al. (2017) regard as the most empirically supported, is that gains from transformative technologies take time to materialise because they require complementary investments that cannot be accumulated quickly. This is particularly true of GPTs (Bresnahan & Trajtenberg, 1995): technologies with sufficiently broad applicability to reshape production across the entire economy, such as the steam engine, electrification and information technology. The more structurally disruptive a GPT's potential, the longer the adaptation period required and the more substantial the complementary investments in human capital, process redesign and institutional knowledge that must precede the productivity payoff. Generative AI exhibits the defining characteristics of a GPT (Brynjolfsson et al., 2017), and the implementation lag framework reads the current gap between AI's demonstrated capabilities and its aggregate productivity impact not as a verdict on AI's potential but as a measure of the organisational work that remains to be done.

1.3 Why Implementation Lags Are a Management Problem

The implementation lag explanation is simultaneously the most optimistic and the most demanding of the four. It predicts that productivity gains will eventually materialise but locates the bottleneck in the organisations deploying the technology, not in the technology itself. Bartel et al. (2007), in a plant-level analysis of IT adoption, document this in precise terms: plants that achieved genuine productivity gains simultaneously redesigned workflows, shifted business strategy toward customised production and raised operator skill requirements to include technical problem-solving the previous workflow did not demand. The productivity gain came from the co-evolution of technology deployment and human skill investment.

This co-evolution dynamic is the central management challenge this thesis examines. The aggregate productivity paradox is, at its core, a management paradox.

Chapter 2: The Micro-Foundations of AI-Driven Productivity

2.1 Heterogeneous Gains: Evidence from Customer Support

The most rigorous evidence on generative AI's impact in a real-world workplace comes from Brynjolfsson et al. (2025), who study the gradual introduction of a GPT-based conversational assistant among 5,172 customer-support agents at a Fortune 500 firm. The tool, fine-tuned on the firm's own high-performing agents' interactions, provided real-time suggestions and functioned as a continuously available repository of the firm's best-practice knowledge. Average productivity increased by 15%, driven by an 8.5% reduction in handling time and an improvement in issue resolution rates.

The aggregate figure hides a distributional story central to this thesis. Lower-skilled and less experienced workers saw productivity gains of approximately 30%; novices with two months of tenure performed at levels comparable to peers with over six months on the job, effectively compressing the experience curve. The mechanism is institutional knowledge democratisation. For the most experienced agents, the picture reversed: little to no productivity gain with small but statistically significant declines in output quality among top performers. A system trained on best practices may have little to offer workers who already embody them and may encourage cognitive shortcuts that displace superior independent judgement.

This pattern overturns the traditional narrative of skill-biased technical change, where technology systematically complements the most skilled workers (Bartel et al., 2007; Dixon et al., 2020). AI appears to act, at least in this first wave of deployment, as an equalising force that compresses the productivity distribution. AI assistance also improved the texture of work, with customers becoming measurably politer and escalations declining (Brynjolfsson et al., 2025), a spillover effect that recurs in the empirical interviews reported in Chapter 9.

2.2 The Jagged Frontier: Evidence from Knowledge Work

A critical question is whether AI's augmentation effects extend to the more complex, open-ended tasks that characterise knowledge work. Dell'Acqua et al. (2023a) address this directly in a field experiment with 758 strategy consultants at Boston Consulting Group.

The study's central theoretical contribution is the jagged technological frontier: AI's capabilities are uneven across tasks, with the boundary between what AI can and cannot do reliably being neither smooth nor predictable from the outside. For tasks within GPT-4's frontier (creative innovation, analytical synthesis, persuasive writing), consultants using AI improved output quality by 40% and completed work more than 25% faster, with below-average performers improving by 43% and above-average performers by 17%, a pattern consistent with the equalising dynamic in Brynjolfsson et al. (2025).

The results for tasks outside the frontier carry a different implication. When the task involved a form of quantitative reasoning for which GPT-4 had a systematic failure mode, consultants using AI were 19 percentage points more likely to submit a wrong answer than those working without it. The AI generated confident, well-structured, plausible-sounding output and highly trained professionals failed to detect the error. The danger is not simply that AI makes mistakes; it is that humans working with AI may progressively lose the situational awareness needed to detect those mistakes (Risko & Gilbert, 2016). The jagged frontier is asymmetric: inside it, AI is a powerful amplifier; outside it, it is a silent source of confident error.

2.3 Calibration as a Moderator

Caplin et al. (2025) introduce a finding complementary to the jagged frontier. Beyond raw task ability, they identify belief calibration, the alignment between an individual's confidence in their own judgement and their actual accuracy, as a decisive moderator of AI's benefits. Holding ability constant, workers with well-calibrated beliefs benefited approximately 20% more from AI assistance than their miscalibrated peers. The mechanism is that a calibrated worker knows when to defer to the AI and when to override it. Calibration is also trainable, even in short interventions (reviewed in Caplin et al., 2025), which makes it a specific organisational lever rather than a fixed individual trait. The full treatment of this finding sits in Chapter 5; what matters here is that calibration emerges as the binding individual-level variable behind AI's uneven distributional effects.

2.4 The Team Dimension and the Synergy Problem

Work is fundamentally social, and AI's impact on team dynamics introduces a distinct set of findings. Dell'Acqua et al. (2023b) study this in a laboratory experiment using a cooperative task: teams in which one human player was replaced by an automated co-worker were compared to a no-change control and to a traditional new-hire (human-replacement) condition. Even where the automated agent outperformed human players individually, its introduction into a human team degraded overall team performance. The mechanism was not algorithmic failure but the disruption of coordination routines, the tacit role understandings and shared expectations that human teams develop through repeated interaction. Low- and medium-skilled teams experienced the largest performance losses, while high-skilled teams were more able to absorb the automated co-worker into their routines. The effect was not driven by algorithm aversion but by an intrinsic preference for collaborating with other humans that automation cannot replicate.

Vaccaro et al. (2024), in a meta-analysis of 106 experimental studies, provide systematic confirmation. On average, human-AI combinations performed significantly worse than the best individual actor (g = -0.23), driven primarily by decision tasks. The most powerful moderator was relative individual performance: synergy emerged when the human alone outperformed the AI alone (g = 0.46); losses emerged when the AI alone outperformed the human (g = -0.54). The default assumption of complementarity is empirically unjustified. Vaccaro et al. (2024) functions as the empirical anchor for the management problem this thesis addresses.

2.5 The Hidden Cost: Motivation and Long-Term Skill Erosion

Even where AI collaboration produces immediate performance gains, a final body of evidence warns against treating those gains as unconditional. Wu et al. (2025), in four preregistered experiments with over 3,500 participants, document a dual effect: while AI consistently enhanced immediate task performance, these gains did not transfer to subsequent tasks performed independently. Self-Determination Theory provides the grounding: transitioning from AI-assisted to solo work is associated with reduced intrinsic motivation and increased feelings of boredom (Wu et al., 2025). Productivity metrics capturing only immediate task outputs systematically undercount the costs of AI collaboration, connecting to the cognitive offloading risk developed in Chapter 5 (Risko & Gilbert, 2016; Al-Refai et al., 2025).

The empirical regularities accumulated across §2.1-2.5 define the management problem that motivates Chapter 3: AI's benefits are heterogeneous, task-conditional, calibration-dependent, team-disruptive and motivationally costly. The question is no longer whether AI can improve productivity but how organisations can systematically structure the conditions under which it does.

Chapter 3: Beyond a Binary Choice — The Automation-Augmentation Paradox

3.1 The False Dichotomy That Structures the Debate

The evidence assembled in Chapter 2 leaves no doubt that AI's productivity impact is real but deeply conditional, mediated by task type, worker calibration, team dynamics and temporal horizon. Yet the dominant vocabulary of organisational AI strategy continues to frame the challenge as a binary choice: automate or augment. Automation implies that machines take over a human task entirely removing the human from the decision loop. Augmentation implies that humans collaborate closely with machines, each contributing complementary capabilities while the human retains ultimate authority (Raisch & Krakowski, 2021).

This framing appears in boardrooms, strategy reports, and policy documents. It is intuitively tractable and politically convenient: it gives managers a discrete strategic decision and a normative answer to reach for, since virtually all management advice of the past decade has recommended augmentation as the superior path. Raisch and Krakowski (2021) demonstrate that this framing is fundamentally misleading, not because the distinction is meaningless, but because the two are interdependent forces in dynamic tension. Choosing one does not neutralise the other; it intensifies the pressure that drives the organisation toward it.

3.2 The Anatomy of the Paradox

The paradox, in the precise theoretical sense used by Smith and Lewis (2011), requires two conditions: the two orientations must be contradictory, meaning they cannot both be fully satisfied simultaneously and they must be interdependent, meaning each generates the need for its opposite over time. Raisch and Krakowski (2021) demonstrate that automation and augmentation satisfy both conditions.

The contradiction operates at the task level. At any given moment, for any given task, an organisation must decide whether to keep the human inside or outside the decision loop. That choice is a question of who holds cognitive authority, the algorithm or the worker, and the two logics do not sit easily together. Automation follows the logic of algorithmic optimisation. Augmentation preserves the human capacity for contextual, value-led judgement. Raisch and Krakowski (2021), drawing on Lindebaum et al. (2020), frame this as formal rationality versus substantive rationality.

The interdependence operates across time and across tasks. Temporally, an organisation that automates a task today creates the conditions that will require augmentation tomorrow: the skills displaced by automation in one workflow cycle must be replaced with higher-order human capabilities in the next, as the frontier of automation advances and the remaining human tasks become more demanding. This dynamic echoes, at the firm level, the macroeconomic implementation lag documented by Brynjolfsson et al. (2017). The reverse dependence is equally important. Augmentation feeds automation: human oversight of AI outputs generates the corrections, examples and codified judgement that render previously “hard-to-learn” tasks automatable, each cycle of augmentation expands the automation frontier it operates within.

3.3 The Vicious Cycles of One-Sided Commitment

Organisations that fail to recognise this interdependence and commit to either pole tend to enter self-sabotage cycles. The automation-first organisation is drawn into a competitive race toward commodity work, in which rival firms match the same cost efficiencies and the differentiation that justified the investment disappears. More critically, sustained automation progressively erodes the human expertise required to adapt or override automated processes when they fail (Al-Refai et al., 2025; full treatment in Chapter 5). The cycle is empirically observable: Dell'Acqua et al.'s (2023b) cooperative-task experiment showed that even a technically superior automated agent degrades team performance by disrupting coordination routines (Chapter 2 §2.4).

The augmentation-first organisation faces a different self-defeating dynamic. Augmentation requires continuous, extensive human involvement in iterative learning cycles. Unlike automation, which functions within stable rule-governed domains, augmentation encounters human cognition in its full complexity. It cannot be standardised or reliably replicated, so augmentation initiatives fail frequently (Raisch & Krakowski, 2021). When they do, organisations face a perverse incentive: to justify prior investments, they may escalate commitment to further augmentation efforts, compounding the failure (Raisch & Krakowski, 2021). The Vaccaro et al. (2024) meta-analysis provides systematic empirical confirmation: across 106 experimental studies, human-AI combinations performed significantly worse than the best individual actor on average, with the negative effect concentrated in decision-making tasks (Chapter 2 §2.4). Augmentation does not reliably improve work without specific conditions and careful engineering.

3.4 The Paradox as an Organising Framework

What transforms this tension into an analytically productive framework is Raisch and Krakowski's (2021) three-stage path toward a virtuous cycle. The first stage is **acceptance**: recognising that the tension between automation and augmentation is not a problem to be resolved once but an organisational condition to be managed continuously, holding both orientations as simultaneously legitimate. The second stage is **differentiation**: deliberately identifying which tasks call for automation and which require augmentation. Raisch and Krakowski (2021) illustrate this with Symrise's master perfumers, who iterate between AI generating a landscape of fragrance formulations beyond human cognitive capacity and human experts applying holistic and aesthetic judgement to select the most promising options. The third stage is **integration**: designing the hand-off points between automated and augmented work so each mode supports the other, with humans retaining genuine accountability for the overall process. Accountability is therefore not an ethical preference but a structural condition for productive integration: when humans feel truly responsible, they are less likely to become complacent and more likely to engage in the reasoning that keeps expert judgement alive (Raisch & Krakowski, 2021, drawing on Larrick, 2004 and Skitka et al., 2000).

3.5 The Framework as a Map of the Evidence

The empirical regularities established in Chapter 2 map the paradox's facets at five levels: individual (Brynjolfsson et al.'s 2025 productivity-distribution compression), task (Dell'Acqua et al.'s 2023a jagged frontier), cognitive (Caplin et al.'s 2025 calibration as selfawareness problem), team (Dell'Acqua et al.'s 2023b Nintendo-style coordination disruption), and temporal (Wu et al.'s 2025 motivational costs manifesting on return to solo work).

The binary choice between automation and augmentation is empirically dangerous: organisations committing to one pole enter one of the two vicious cycles. The management challenge is not to choose but to navigate the dynamic relationship between them. The next three chapters develop three organisational levers: task allocation, calibration training, and interface and process design, as the primary mechanisms through which the paradox can be managed toward virtuous rather than vicious cycles.

Chapter 4: Lever 1 — Task Allocation as Frontier Navigation

4.1 The Allocation Problem Reframed

The automation-augmentation paradox developed in Chapter 3 resolves into a practical organisational challenge: for each task or sub-task in a workflow, who or what should perform it? On the surface, this sounds like a capability-matching exercise. In practice, it is a dynamic navigational problem in a landscape whose boundaries shift continuously, often in ways invisible to the workers operating within it. Effective task allocation requires a logic for where AI adds value, organisational practices for navigating the frontier in real time and institutional awareness that frontier navigation is itself a perishable capability.

4.2 Mapping the Frontier: The Logic of Task-Level AI Competence

Two complementary frameworks converge on the logic governing AI's task-level performance. Acemoglu's (2024) easy-to-learn versus hard-to-learn task distinction (Chapter 1 §1.2) does not map cleanly onto routine versus non-routine work: many ostensibly complex, high-status tasks are easy-to-learn for AI because their complexity is pattern-driven, while many tasks that appear routine are hard-to-learn because they depend on contextualised, experiential judgement that current AI cannot reliably replicate.

Dell'Acqua et al. (2023a) operationalise a related distinction empirically through the jagged frontier (Chapter 2 §2.2). One practical implication appears in Brynjolfsson et al.'s (2025) deployment: when the AI system had insufficient training data, it provided no suggestion, leaving the agent to respond independently. An AI that stays silent rather than confabulating can be read as a form of built-in frontier signalling that organisations can incorporate deliberately as discussed in Chapter 6.

Vaccaro et al.'s (2024) meta-analysis (Chapter 2 §2.4) provides the most useful allocation principle: the relative-performance heuristic. When the human alone outperforms the AI alone on a given task, the combined system generates genuine synergy and the opposite is true as well. The question a manager must ask is therefore not "can our AI perform this task?" but "does adding a human to the AI make the output better or worse?". Where human involvement adds genuine comparative advantage, through contextual judgement, ethical reasoning, relational intelligence or tacit domain knowledge, collaboration is productive. Where it does not, the evidence recommends either full automation or a redesign of the human's role toward oversight rather than co-production.

4.3 Navigating the Frontier in Practice: Centaur and Cyborg Models

Dell'Acqua et al. (2023a) identify two coherent strategies among consultants who managed outside-the-frontier tasks successfully. The **Centaur** model involves strategic division of labour: human workers switch between AI and human contributions at the sub-task level, maintaining clear cognitive ownership boundaries. The **Cyborg** model involves more granular integration: users intertwine their cognitive effort with the AI's at the sub-task level. This can produce extraordinary results inside the frontier but is more vulnerable to failure when the frontier has been crossed. In practice, the Centaur model may be more teachable and more robust to frontier uncertainty while Cyborg approaches may be better suited to creative contexts where the frontier is well-understood. The key design decision is which model to deploy for which category of task

4.4 The Opacity Problem: Who Knows Where the Frontier Is?

Three structural features of the frontier make it resistant to top-down mapping. First, **the frontier is not static**: AI capabilities evolve at a pace that surprises even AI researchers, and a task landscape mapped today may quickly become outdated, as the frontier is "expanding and changing" even to AI's own producers (Dell'Acqua et al., 2023a). Second, **the frontier is opaque to the user**: consultants who fell outside the frontier did not know they had done so because the AI's output was confident and superficially plausible (Dell'Acqua et al., 2023a). Organisations need structural safeguards (process checkpoints, domain expertise as external checks) because workers cannot rely on real-time introspection. Third, **the frontier is individual-contingent**: a task inside the frontier for a junior worker may fall outside it for a senior expert whose contextualised knowledge exceeds the AI's training data (Brynjolfsson et al., 2025). This connects to the calibration lever developed in Chapter 5 and to the co-evolutionary investment principle from Chapter 1 (Bartel et al., 2007).

4.5 The Training Deficit: The Long Shadow of Over-Delegation

When AI is deployed to handle tasks that would previously have been assigned to junior workers as structured learning opportunities, those workers lose the developmental experiences through which domain expertise is built (Dell'Acqua et al., 2023a). The frontier may be well-mapped today but if the next generation of senior practitioners never develops the foundational knowledge on which expert frontier navigation depends, the organisation's capacity for navigation will erode over time. This could produce training deficits with long-run consequences that short-run productivity metrics may not detect. The risk is partially offset by the accelerated skill acquisition in Brynjolfsson et al. (2025), but compressing the experience curve through AI assistance is not equivalent to accumulating deep expertise through direct practice and consequential error. The empirical extent of this training deficit, in its workforce-architecture form, is documented as the Junior Pipeline finding in Chapter 9.

4.6 Toward an Organisational Practice of Frontier Navigation

Four principles for effective organisational task allocation emerge. First, move from role-level to task-level analysis (Acemoglu, 2024). Second, apply the relative-performance heuristic as the primary allocation criterion (Vaccaro et al., 2024). Third, institutionalise frontier monitoring through regular retrospectives feeding a continuously updated map (Dell'Acqua et al., 2023a). Fourth, maintain deliberate human task ownership in frontier-adjacent domains: tasks just inside the frontier should be development assets, not only efficiency opportunities (Dell'Acqua et al., 2023a).

Task allocation is not a static organisational chart drawn up during an AI implementation project. It is a dynamic, continuously updated navigational practice that requires workers to know not just how to use AI but how to evaluate it. That evaluative capacity is the subject of the next chapter.

Chapter 5: Lever 2 — Calibration Training as a Meta-Skill

5.1 The Calibration Gap: Why Skill Alone Is Not Enough

Chapter 4 established that sustainable AI-assisted performance depends on matching the right actor to the right task. Yet even when task allocation is well-designed, a second and less visible barrier persists: workers' inability to accurately assess their own competence relative to the AI they are using. This failure of self-knowledge which the literature terms miscalibration, is one of the primary reasons why the productivity gains from AI remain so unevenly distributed across individuals (Caplin et al., 2025).

Caplin et al. (2025) provide the most rigorous treatment of this question to date. Using a controlled experiment, they decompose individual performance with AI into three components: ability (what a person actually knows), beliefs (what they think they know), and calibration (the alignment between the two). Their central finding is that, holding ability constant, workers with well-calibrated beliefs benefit substantially more from AI assistance: knowing how much you know is a significant complement to knowing itself. A worker who is well-calibrated knows when to defer to the AI and when to override it. Improving calibration is therefore not a question of teaching new skills but of teaching workers to know the limits of their existing skills.

5.2 Who Benefits and the Cost of Miscalibration

Caplin et al. (2025) identify a particularly consequential interaction: the workers who stand to gain the most from AI augmentation are those with low ability and high calibration. These individuals, aware of their own limitations, are precisely positioned to use AI as a genuine complement rather than as external validation. Workers with high ability but poor calibration gain the least as their confidence prevents them from adjusting their decision-making in response to AI signals.

This refines the pattern from Chapter 2. AI compresses the productivity distribution by benefiting lower-skilled workers more (Brynjolfsson et al., 2025) but compression is constrained by miscalibration. In a counterfactual where miscalibration is eliminated entirely, the reduction in performance inequality from AI adoption would be nearly twice as large as observed (Caplin et al., 2025). The implication is sharp: those who would benefit most from AI, workers with low ability, are typically also among the least calibrated. Hence, miscalibration disproportionately constrains the very gains AI could most clearly provide (Caplin et al., 2025). Miscalibration also offers a partial structural explanation for the Vaccaro et al. (2024) under-performance result: when workers cannot correctly assess when AI is likely to be superior to their own judgement, the resulting hybrid decision-making is noisier than either input alone.

5.3 Cognitive Offloading and Skill Erosion

The calibration problem is compounded by a temporal dynamic. Even workers who begin with reasonably well-calibrated beliefs can see that calibration deteriorate over time, as habitual AI use induces cognitive offloading (Risko & Gilbert, 2016). When workers routinely outsource judgement without actively verifying AI outputs, they progressively lose the feedback loops that keep their beliefs calibrated. The skill that underpins good calibration, the regular practice of forming and testing independent judgements, is precisely the skill that uncritical AI use threatens (Bastani et al., 2025; Al-Refai et al., 2025).

Al-Refai et al. (2025) frame this through the I-PACE model, arguing that when AI displaces human agency rather than scaffolding it, the result is skill degradation and overreliance. This connects to the broader cognitive offloading dynamic: routine reliance on AI tools can gradually weaken the cognitive skills most closely tied to judgement and evaluation (Risko & Gilbert, 2016). Bastani et al. (2025), in a mathematics education setting, found that students given unrestricted AI assistance showed learning gains during AI use but suffered performance degradation when the AI was removed. The generative AI had offloaded the cognitive work so completely that no durable skill transfer occurred, though students given guard-railed AI showed no such degradation suggesting the harm is a property of unguarded access rather than of generative AI itself. In organisational terms, this initiates a vicious cycle entirely consistent with the automation-augmentation paradox (Raisch & Krakowski, 2021): the short-term augmentation of task performance through AI can produce a longer-run erosion of the human capabilities that make augmentation valuable.

5.4 The Jagged Frontier's Calibration Demand

The frontier is not self-evident: even highly experienced professionals cannot easily identify whether a given task sits inside or outside AI's reliable capability space (Dell'Acqua et al., 2023a). Navigating it is therefore fundamentally a calibration task. Workers overconfident in AI's capabilities fail to apply corrective judgement at the boundary; workers underconfident fail to exploit genuine gains inside the frontier. Good frontier navigation requires exactly the metacognitive accuracy that Caplin et al. (2025) formalise as calibration. Calibration training is the individual-level complement to organisational-level frontier navigation.

5.5 Calibration as a Trainable Organisational Lever

A discouraging reading of this evidence would treat miscalibration as a fixed individual trait. Caplin et al. (2025) note that calibration, unlike raw ability or IQ, is a strong candidate for intervention. Training programmes have produced measurable effects in sessions lasting less than an hour (Gruetzemacher et al., 2024, as cited in Caplin et al., 2025), making calibration training a practically actionable organisational lever.

Three design principles emerge. Training should be **active** (workers practice forming independent judgements and receiving feedback, not being told that AI can be wrong), **domain-specific** (general AI literacy is insufficient because the frontier is task-specific), and **temporally sustained** (calibration erodes through cognitive offloading and must be maintained as ongoing practice).

Calibration training is, at its core, a mechanism for preserving the human skills that give augmentation its value. An organisation that invests in AI tooling without investing in calibration is progressively automating the judgement capacity of its workforce while calling the process augmentation. This confusion between the two poles of the paradox is precisely what Raisch and Krakowski (2021) identify as the source of vicious cycles.

5.6 Contextual Asymmetries Across Departments

A reasonable ex-ante expectation derived from this chapter is that calibration training should be department-specific: marketing risks accepting AI output that reads fluently but remains mediocre in substance, while legal and compliance settings may show seniority-asymmetric vulnerability. The empirical investigation in Chapter 9 §9.3 finds a different binding moderator (disposition over seniority), and Chapter 8 §8.4 documents the methodological pivot from departmental stratification to disposition-driven recruitment that the evidence forced.

Chapter 6: Lever 3 — Designing for Appropriate Trust

6.1 From Tool to Agent: A Shift in Interface Logic

The first two levers, task allocation and calibration training, operate primarily at the level of the individual worker or individual task. The third lever operates at a different level: the design of the organisational infrastructure through which human judgement and AI capability are made to interact. The third lever's complexity increases as AI systems take on more agentic behaviour.

A tool interface is passive: it responds to explicit human instructions and returns outputs the worker then evaluates. An agent interface involves a system that pursues goals, initiates actions, selects among strategies and operates across extended workflows with varying degrees of human oversight. The question is no longer how to help a worker use a capability more effectively but how to keep the human genuinely, not merely nominally, in control of a process the agent is largely executing. As Raisch and Krakowski (2021) observe, organisations tend initially to perceive automation and augmentation as a trade-off decision made at a specific point in time. This framing misses the process dynamics: automation in one task generates augmentation demands in adjacent tasks and the interaction between humans and agents becomes iterative and compounding. The interface is therefore the primary organisational mechanism through which the automation-augmentation tension is either managed productively or allowed to degrade into one of the vicious cycles described in Chapter 3.

6.2 Context Engineering as the Foundation

The quality of any AI agent's output depends in substantial part on the quality of the context it receives (Choi & Schwarcz, 2025). Context engineering is the deliberate organisational practice of structuring the informational environment in which an agent operates: the objectives and constraints encoded into the system, the institutional knowledge made available to it and the parameters defining the scope of its autonomy. It is not equivalent to prompt engineering which operates at the individual-user level; context engineering concerns the sustained, systematic design of the agent's operating environment, independent of any single interaction. Choi and Schwarcz (2025) demonstrate substantial performance variation across four prompting techniques (basic, chain-of-thought, few-shot, grounded) in a legal-analysis setting with grounded prompting consistently outperforming basic prompting. This supports the broader principle that the informational environment supplied to an AI is a performance-relevant design variable.

Brynjolfsson et al.'s (2025) customer-support deployment (Chapter 2 §2.1) illustrates thoughtful context engineering in practice: the AI assistant was trained on the firm's own high-performance interaction data and grounded in internal documentation rather than generic sources. Context engineering also connects to the jagged frontier (Chapter 4 §4.2): a well-engineered context narrows the agent's operational space to where it is known to perform reliably.

6.3 Compounding Corporate Knowledge

The second sub-lever addresses the dynamic dimension of the interface: the capacity of human-agent interactions to accumulate and compound institutional knowledge. Each interaction is a data point (decisions, corrections, overrides, ratifications) that can progressively refine the agent's contextual model. The interface becomes, over time, a mechanism for institutional learning.

Workers who actively engaged with AI suggestions in Brynjolfsson et al.'s (2025) study developed durable performance improvements that persisted even when the AI was unavailable in contrast to workers who frequently overrode suggestions. The dynamic operates in both directions: workers learn from the agent and the agent's value increases as its institutional context becomes richer. The risk of neglecting this feedback loop is illustrated by Dell'Acqua et al.'s (2023b) team-coordination findings (Chapter 2 §2.4): introducing an automated agent disrupts the tacit knowledge the team has developed because the AI cannot replicate that accumulated relational knowledge. Organisations that design their interfaces to capture and institutionalise AI-assisted knowledge convert a source of disruption into a sustained advantage.

6.4 Agent Proactivity and the Human-in-the-Loop Principle

The third sub-lever concerns a specific design choice: whether the agent surfaces its own uncertainty rather than proceeding silently on a best guess when context is missing or ambiguous.

Many current AI agent systems appear optimised for task completion rather than for accurately communicating output reliability. An agent that fills gaps silently produces outputs that appear authoritative regardless of underlying reliability. The human nominally remains in the loop but is functionally excluded from the decision point where judgement was most needed. Spielman and Le Blanc (2021) cite the Boeing 737 Max incidents as the canonical case: pilots whose reliance on automated systems had eroded their proactive engagement were unable to intervene appropriately when the system malfunctioned. The same dynamic operates at lower stakes in knowledge work where agents that do not surface ambiguity condition workers to treat AI outputs as reliable by default.

Agent proactivity is the design response to this risk. When the agent encounters a situation that falls outside its encoded context, that is ambiguous with respect to organisational norms or that involves a judgement call with high error costs, the appropriate response is to request human input rather than to generate a plausible-sounding output. This keeps the human genuinely in control by ensuring that the points of highest uncertainty are also the points of highest human engagement. Vaccaro et al. (2024) find that synergy emerges when humans and AI are assigned to the subtasks they handle best. Agent proactivity is the real-time analogue: the agent itself flags moments where reallocation to the human is warranted.

6.5 Bridge to the Integrative Framework

Task allocation, calibration training, and interface and process design each address the automation-augmentation paradox at a different level. They are conceptually separable but practically coupled: a well-allocated task assigned to a miscalibrated user under a poorly designed interface is unlikely to yield much improvement over working without AI at all. The next chapter integrates the three into a framework testable against organisational practice.

Chapter 7: An Integrative Framework

The three preceding chapters developed task allocation, calibration training and interface and process design as distinct organisational levers through which the automation-augmentation paradox can be managed. Each lever addresses the paradox at a different level: task allocation at the boundary between human and AI work, calibration at the user's judgement of AI output and interface and process design at the structural design of the tool and the workflow around it. Treated separately, each lever is empirically supported but analytically incomplete. This chapter integrates the three into a single framework that is testable against organisational practice.

The integrative framework rests on three claims about how the levers operate together. First, the levers are coupled, not independent. A calibrated user with a well-designed interface working on a poorly allocated task can only partially compensate for the misallocation. The same logic applies to every pairwise combination. The framework therefore predicts that organisational outcomes depend on the **coherence** of the three levers rather than on the strength of any single one. The vicious cycles Raisch and Krakowski (2021) describe are mechanisms by which an organisation's investment in one lever undermines the others; the virtuous cycle is the configuration in which the three reinforce each other.

Second, the levers operate on different time horizons. Task allocation is the fastest-moving lever: a manager can re-allocate a task in a single workflow iteration. Calibration training operates on a slower cycle: building the metacognitive skill that allows workers to evaluate AI output reliably takes time to develop and erodes if not maintained. Interface and process design operates on the slowest cycle: tool procurement, governance scaffolding and embedded skills change at the cadence of organisational change-management programmes. The framework therefore predicts that organisations attempting to manage the paradox through any single lever will misallocate effort across time horizons; effective management requires investment at all three cadences simultaneously.

Third, the framework yields six theoretical propositions, two per lever listed in Chapter 8 §8.2 and assessed against the empirical data in Chapter 9. The framework is presented here in its three-lever form for the methodology that follows. The empirical investigation in Chapter 9 reformulates two propositions where the data forces a relocation of the operative mechanism and surfaces one additional aggregate dimension that the original framework does not anticipate. That refinement is the contribution of the empirical chapter, not of this one.

Two limitations of the framework should be flagged. First, it operates primarily at the firm and individual level. It does not theorise macro-level forces (sovereignty, regulation, geopolitics, infrastructure cost) that the empirical investigation later surfaces as upstream determinants of organisational AI choices. Second, it assumes a stable temporal environment in which the levers can be designed deliberately. Chapter 9 §9.9 develops pace-mismatch as a cross-cutting temporal pressure that destabilises this assumption. Neither limitation invalidates the framework's central claim that the three levers must be managed together.

The methodology that follows operationalises this claim into testable form, specifying how the framework will be confronted with practice through 28 semi-structured interviews under the Gioia method with the abductive logic of Section 8.1 leaving room for refinement where the evidence warrants.

Chapter 8: Methodology

8.1 Research Objective and Epistemological Positioning

The preceding chapters constructed a theoretical framework organised around three organisational levers, namely task allocation, calibration training and interface and process design, through which firms may navigate the automation-augmentation paradox identified by Raisch and Krakowski (2021). The meta-analysis by Vaccaro et al. (2024) establishes that human-AI combinations frequently underperform the best of humans or AI alone and the experimental evidence reviewed in Chapters 4 through 6 suggests this failure follows identifiable patterns. What the literature does not yet provide is a detailed understanding of how these dynamics manifest in the daily routines of professionals who actually work with AI tools. The objective of this empirical investigation is to fill that gap.

The interviews pursue two related aims. The primary aim is to confront the three-lever framework with organisational practice. The secondary aim is to remain open to emergent themes that do not fit neatly into the existing framework allowing the data to extend or modify the theoretical structure where warranted. This dual orientation corresponds to an abductive research logic (Dubois & Gadde, 2002): the framework informs data collection while remaining provisional and subject to revision considering the empirical evidence. This positioning is coherent with the contingent nature of the paradox described by Raisch and Krakowski (2021).

8.2 Theoretical Propositions

To structure the dialogue between theory and data, six theoretical propositions have been derived from the literature. These are not hypotheses in the statistical sense but empirically grounded expectations that guide the interview protocol (Gioia et al., 2013). Table 8.1 maps each proposition to its parent lever (see Appendix A10. Interview guide).

Proposition	Lever	Statement
P1	Task Allocation	Professionals working in domains where AI capability varies sharply across seemingly similar tasks experience greater difficulty in deciding what to delegate to AI.

P2	Task Allocation	Organisations without explicit task-allocation guidelines rely on individual trial-and-error, producing inconsistent delegation patterns within the same team.
P3	Calibration	Professionals who can accurately judge when AI output is reliable benefit more from AI collaboration than those with higher domain expertise but poor calibration.
P4	Calibration	Junior professionals are more susceptible to over-reliance on AI output than senior professionals, moderated by the presence of organisational calibration mechanisms.
P5	Interface Design	The design of AI tools shapes verification behaviour: tools that present output as final answer encourage over-reliance; tools that present output as draft encourage verification.
P6	Interface Design	Organisations that integrate AI into structured workflows with human review checkpoints experience fewer quality failures than those where AI is used informally.

Table 1. *Six theoretical propositions, two per organisational lever.* The propositions operationalise the three-lever framework into testable claims that structure the empirical analysis.

The findings chapter assesses each proposition against the interview data. P4 and P5 were reformulated at consolidation_003 (see §8.7) and the reformulations are documented in Part B of Chapter 9..

8.3 Research Design

This study adopts a qualitative research design based on semi-structured interviews. The three levers operate through organisational routines, individual sensemaking and managerial decisions, phenomena that require depth of inquiry rather than breadth. The design is multi-case and iteratively-sampled, with the sampling logic itself evolving across four consolidation milestones as the analytical signal emerged, consistent with the abductive logic stated in Section 8.1.

8.4 Case Selection: Departmental and Industry Logic and the Abductive Pivot

Case selection follows the logic of theoretical sampling (Eisenhardt, 1989): participants are chosen because they represent contrasting configurations of the three levers. The initial sampling frame targeted three functional departments, Marketing and Content, Legal and Compliance and Sales, chosen because they represent contrasting positions on task creativity, regulatory exposure, and relational versus repetitive work. The expectation was that the same paradox would surface differently across these contexts.

As fieldwork progressed, the empirical signal located itself differently from the initial expectation. Early interviews surfaced that the automation-augmentation paradox is most cleanly articulated by deployment-side voices (Innovation Leads, IT Directors, AI implementation owners), while department-specific end-user perspectives accessed the paradox in fragmented ways. Following the abductive logic, recruitment evolved: subsequent batches added a fifth, cross-departmental "Insights" category. A second methodological reason supported the pivot: the data accumulated evidence that disposition and curiosity moderate AI use more reliably than seniority does (the P4 reformulation documented in Part

B of Chapter 9). Hence, deep stratification by tenure became less analytically necessary and recruitment effort was redirected toward cross-industry breadth. This evolution is consistent with established practice in iterative theoretical sampling (Eisenhardt, 1989; Charmaz, 2014). The trade-offs are acknowledged in Section 8.9.

8.5 Sample Design

The realised sample comprises 28 semi-structured interviews conducted across approximately 22 firms and 14 industries. The sample size exceeds the original target of 9 to 12 interviews stated at the protocol stage. Theoretical saturation took longer than originally anticipated and the corpus stabilised around interviews 25 to 28. The cross-firm and cross-industry breadth of the sample justifies a later saturation point than Guest et al. (2006) report for homogeneous samples. The realised distribution covers construction, technology and software, telecommunications, financial services, pharmaceutical, legal services, public-sector transport, insurance, tech research and policy, food retail, footwear retail, industrial chemicals, digital agencies, and water technology advisory. Seniority skews toward senior decision-makers (25 of 28) with one mid-senior interviewee, one individual contributor and one explicit junior. This distribution reflects the abductive pivot described above and its implications are addressed in Section 8.9. The full sample composition and recruitment-grid-prefix conventions are documented in A9. Methodology Detail.

8.6 Interview Protocol

Each interview follows a semi-structured format lasting approximately 45 minutes, conducted remotely via Microsoft Teams and recorded with written consent (Section 8.8). The guide comprises a shared core supplemented by department-specific questions (see Appendix A10. Interview guide) organised into four thematic blocks of roughly 10 minutes: context, task allocation (P1, P2), calibration (P3, P4), and interface and process design (P5, P6). A closing segment captures emergent themes. To avoid biasing responses, the specific theoretical constructs are not disclosed beforehand. A pilot interview was conducted before main data collection and three probes were added during fieldwork (champion network, single-failure reversion, AI-driven attrition). Their null replication results inform Chapter 9.

8.7 Data Analysis

Interviews are transcribed using Fireflies, followed by a manual review pass. Interviews are conducted in English, French or Spanish; excerpts used in the findings chapter are then translated.

The analysis follows the Gioia method (Gioia et al., 2013), proceeding in three stages: first-order codes kept close to the participant's wording; second-order themes applying theoretical interpretation; aggregate dimensions expected to align with the three levers but allowing for emergent dimensions consistent with the abductive logic. Coding is performed in a structured spreadsheet (column structure detailed in Appendix A9. Methodology Detail and see the excel for the detailed classification with roughly 500 statements identified across interviews)

The Gioia process is run iteratively across four consolidation milestones (consolidation_001 through consolidation_004), each occurring after a batch of five to seven additional interviews. Each consolidation reviews emerging patterns, identifies gaps for subsequent recruitment and locks decisions that subsequent batches must respect. This cadence is intrinsic to the abductive design and provides the audit trail through which the analytical structure matured. The detailed per-milestone log is preserved in Appendix A6. Consolidation Milestone Log.

Two methodologically transparent revisions emerged through this process. At consolidation_002, the "junior pipeline disruption" pattern was promoted from emergent to aggregate dimension because it had appeared in five or more firms across multiple industries with a coherent mechanism. At consolidation_003, "deskilling and cognitive offloading" was confirmed as a sub-theme inside calibration; "pace-mismatch and temporality" was located as a sub-theme of Proposition 6; and P4 and P5 were reformulated. The resulting data structure is presented in Chapter 9.

8.8 Ethical Considerations and Anonymisation Protocol

Along the interviews, several measures ensure ethical rigour. Participants receive a written informed consent form prior to each interview outlining the research purpose, voluntary nature of participation, recording, anonymisation and the right to withdraw. Consent was then confirmed at the beginning of each interview.

Anonymisation conventions are as follows. Interviewees are referenced exclusively by alphanumeric interview code (for example, `I-C-2`, `M-A-1`, see Appendix A9.3 Recruitment-grid prefix as artefact for more details). Firm names are anonymised by sector and size descriptors (for example, "a large pharmaceutical multinational", "a Belgian construction group"). Role titles are anonymised to role-family categories with sector context preserved where analytically necessary. Identifying details within direct quotations are redacted by square-bracketed elision or replaced with category descriptors. Cross-firm references between interviewees are preserved at the firm-letter level so that within-firm triangulation remains visible. Audio recordings are stored on encrypted local storage and will be deleted upon completion of the thesis. The use of AI tools in the preparation of this thesis is disclosed on the title page.

8.9 Validity, Reliability, and Limitations

Several measures enhance trustworthiness, following Lincoln and Guba (1986). **Credibility**: a consistent interview guide, a pilot interview and the iterative consolidation process with four milestones and locked decisions per milestone provide a transparent audit trail. **Transferability**: recruiting across approximately 22 firms and 14 industries captures a broader range of contexts than a single-case design would permit; findings are not statistically generalisable but illuminate mechanisms in comparable professional contexts. **Dependability**: the interview guide, coding spreadsheet, consolidation notes and final data structure are documented and available for review.

Four limitations should be noted. First, the sample concentrates on senior decision-makers with cross-departmental deployment visibility, a deliberate consequence of the abductive pivot (Section 8.4). Hence, the findings are stronger on firm-level paradox dynamics than on department-specific end-user practice. Second, recruiting across multiple organisations and industries introduces variability that cannot be fully ignored: differences may partly reflect firm-specific, industry-specific or national-cultural factors. Third, reliance on self-reported experience means accounts may be subject to social desirability bias. Triangulation against concrete examples mitigates but does not eliminate this. Fourth, two interviews (`L-G-1` and `M-B-1`) are paraphrase-only because of some network issues. There are marked [¶ paraphrase] at the point of use. Also, the trilingual nature of the corpus introduces a translation layer; original-language text is preserved in transcripts. These limitations circumscribe the claims that can be made but do not invalidate the findings.

Chapter 9: Findings and Discussion

9.1 Introduction

This chapter reports the findings from 28 semi-structured interviews conducted across approximately 22 firms and 14 industries between November 2025 and April 2026. The analysis follows the abductive logic stated in Chapter 8: the three-lever framework derived from the literature (Task Allocation, Calibration, Interface and Process Design) guided coding and interpretation, while the data was allowed to extend or modify that framework where the evidence warranted. Four iterative consolidation milestones (consolidation_001 through consolidation_004) progressively stabilised the analytical structure and produced two methodologically transparent revisions. First, "junior pipeline disruption", initially treated as an emergent theme, was promoted at consolidation_002 to an aggregate dimension of its own. Second, "human accompaniment" and "pace-mismatch", both surfacing in more than ten interviews, were locked respectively as a cross-cutting synthesis condition and as a sub-theme of Proposition 6. The data structure therefore comprises four aggregate dimensions plus a residual emergent bucket and supports a refined version of the three-lever framework rather than its straightforward confirmation.

The chapter is organised in three parts. Part A walks descriptively through the four aggregate dimensions and the residual emergent bucket. Part B confronts the framework with the data on a proposition-by-proposition basis. Part C synthesises the findings into a refined four-dimension framework, argues that human accompaniment functions as the binding meta-condition under which all four dimensions operate and treats pace-mismatch as the cross-cutting temporal pressure that shapes how firms manage the automation-augmentation paradox in practice. Throughout, interviewees are referenced by their anonymous code (for example, `I-M-2`, `M-A-1`) and presented by their substantive role rather than by the recruitment-grid prefix, in line with the anonymisation protocol described in Chapter 8.

Part A: Descriptive Findings

9.2 Task Allocation

Task Allocation, the first organisational lever, concerns how professionals and firms decide which tasks to delegate to AI and which to retain for humans. The data structure consolidates 117 statements under this dimension, distributed across 17 themes. Two themes dominate by cross-firm coverage. The first, a stakes-based delegation gradient, was articulated in 14 or more firms: AI is permitted on low-consequence work and excluded from the moments where the cost of an error is large. The second, the jagged frontier (Dell'Acqua et al., 2023a), was articulated in 15 or more firms: AI is reliably useful for retrieval, summarisation, and reformulation and unreliable for strategic novelty or long-horizon coherence.

A Corporate BIM Manager at a Belgian construction group articulated the stakes gradient bluntly: "for engineering calculations, we will not allow an AI to make the structural calculation of a building. Maybe it can help, but the final signature should be an engineer who checked and double-checked everything before we build [a flagship project]" (`I-M-2` S06). Stake-based exclusion was repeatedly framed as a structural feature of professional work rather than a transitional caution. The jagged-frontier pattern appears in two complementary forms. In its capability-mapping form, the Director of Estimating at the same construction group put it cleanly: "today if you ask AI to give a price for a type-A suspended ceiling, it can find the reference price. With measurement tools like BIM, it can extract quantities without problem. What it won't do is the big execution decisions, how to execute the works, where to put tower cranes, in what order to carry out the works in detail" (`I-M-3` S06). In its surprise form, interviewees describe AI failing on tasks they expected to be trivial. An Innovation Director at a telematics multinational described the boundary in terms of training-data availability: "if you ask Claude, Gemini, or Codex to add a table or diagram to your UI, it will do it without any issue. But if you're in our domain and you ask it to rewrite a fuel usage algorithm, it can't, because there's zero training data on that specific logic" (`M-K-1` S08)

Three other themes structure delegation decisions. **Sovereignty and intellectual property** (12+ firms): data sensitivity and jurisdictional control drive delegation choices upstream of any task-by-task consideration (`L-E-1` S04). **Regulatory or brand carve-out**: the EU AI Act excludes recruitment from full agentic deployment at one firm: "recruitment is seen as high risk; we need to have a human in the loop and it will not be a full agentic recruitment process" (`I-C-2` S05). **Retrieval-not-analysis as a safe delegation pattern** in legal and regulated work: AI surfaces references that the human then evaluates, never the analysis itself. A Senior Legal Counsel described the reading order as deliberately reversed: "before, you had to read enormous amounts of pages. Now it's reversed: give me all the articles referring to this risk, then I go directly to those. The risk is minimal because the tool is just referencing specific wording, it's not making analysis, just extracting references" (`L-M-4` S06).

Where firms did not have explicit rules, delegation was overwhelmingly governed by individual trial-and-error (16 statements, 7 firms; P2 baseline). A Deputy COO articulated the deliberately tool-agnostic posture: "we don't impose the tools, we say how things should be done, not which tool to use. People are free to choose their tools" (`L-M-1` S07). Even at a Digital Transformation Manager's firm with a formal sovereign tool, departmental variance was extreme: 95% adoption in IT, twice the token consumption per capita compared with the rest of the firm (`I-Q-2` S12). Nine statements across six firms describe deliberate moves beyond trial-and-error. A Head of Marketing articulated the sandbox-then-industrialise logic: "I call it wild west mode. Everyone is trying things in their own way. But there are practical targets: by year-end, I want the marketing team to have documented at least five best practice examples. If we prescribed what to do now, we'd miss out on the diversity of use cases people discover on their own" (`M-B-2` S12). A Risk Domain Manager described the structural precondition: "what matters for AI implementation is whether there is a defined process or not. In recruitment, there's a clear process; if you embed AI at one step of that process, every recruiter will use it at that point" (`M-C-1` S04).

Two themes the framework did not anticipate also shape delegation. **Function-asymmetric productivity gain** (4 statements, 3 firms): AI value materialises unevenly across functions of the same firm, with sales communication volume tripling while warehouse adoption is near zero (`M-K-1` S17). **Motivational or affective dimension** (3 statements, 3 firms): tasks are kept human or delegated on grounds of intrinsic satisfaction rather than capability. An IT Director at a footwear retailer articulated this as a deliberate principle: "I really want to preserve the pleasure of working. There are things AI can do, but I love doing them myself, I'll do them myself. And there are things I could delegate and I'll use AI instead because I'm not competent at them, or they're hyper-repetitive" (`M-S-1` S12). A psychosocial researcher framed it as an ethical question: "if AI can do a task well, but that task brings genuine pleasure, satisfaction, and motivation to a worker, should we give it to AI?" (`I-N-1` S24).

Two power-user cases (`S-I-1`, `S-J-1`) push the augmentation reading of Task Allocation to its boundary, with a sole-operator agency producing the output of a former 16-employee firm and a lead-generation startup productising AI as the technical engine of the business. These are read as boundary observations rather than median behaviour. More details about those cases in Annex A7. Power-User Case Summaries.

Practitioners do not navigate the jagged frontier through a unified rule but through a multi-criterion judgement combining capability, stakes, sovereignty, regulatory exposure, motivational fit and function-level economics. Where firms move beyond trial-and-error, they do so by mapping tasks formally not by issuing rules. Full supplementary quotes in Annex A1. Task Allocation: Supplementary Quotes

9.3 Calibration

Calibration, the second organisational lever, concerns whether professionals can accurately judge when AI output is reliable. The data structure consolidates 131 statements under this dimension, distributed across 20 themes. The dominant finding is that verification cost is the operative variable. Fourteen statements across eight firms describe verification work that erodes or even reverses the productivity gain from delegation. A Commercial Manager at a Latin American advisory firm framed the dilemma at the level of the individual decision: AI-and-then-review may be slower than skipping AI altogether (`S-L-2` S05). The telematics Innovation Director quantified the same problem at firm level: "since about a year and a half ago, more than 60% of written content in the company is written by AI. And from last month's fact-check run, only 34% of what was written was factually correct. But the fact-check itself is also done by AI, so we're not 100% sure how accurate the fact-check is either" (`M-K-1` S22). The most consequential illustration of miscalibrated trust came from the sole-operator agency founder: he fed multiple AIs the same prompt, framed from his own perspective and "almost all of them said write a letter, contest the decision, say you disagree. I followed that advice, and I regret it. The inspector who received my contestation letter called me and said: you didn't understand, the problem is not you, it's them. The AI had misunderstood. It put itself in defensive-solution mode because of how I framed the prompt" (`S-I-1` S04). A consequential letter was sent to a government ministry on bad AI advice.

Two patterns describe how practitioners attempt to keep verification cost down without abandoning verification itself. The first is multi-tool and source-citing triangulation. The editorial-agency founder described his verification ritual as "media literacy applied to AI": he asks the model for sources and rejects answers grounded in low-authority origins. The cautionary example he cites is a lawyer whose AI-generated legal conclusions "had used French jurisprudence instead of Belgian law" (`S-D-1` S06). The sole-operator agency founder pushed the triangulation pattern further by running the output of one model through a second model as a comprehension test: "the way it independently understands and reinterprets the sources and synthesizes them into a different format tells me whether a second AI has grasped what the first one said. That's cross-referencing sources" (`S-I-1` S08). The second pattern is stake-asymmetric verification: verification effort scales with downstream consequence not with tool. A finance contributor reports systematic rework as his default at work but a much lighter posture at home (`I-P-1` S07; S12).

A persistent challenge underneath these patterns is AI sycophancy: models that adjust toward user framing rather than testing it. The psychosocial researcher described this as "confirmation bias in AI. It wants to please you. That's very striking. Training on prompt quality is critically important because it clearly influences the result AI gives" (`I-N-1` S16). The sole-operator agency over-reliance incident already cited makes the sycophancy stakes concrete: a defensive-solution mode triggered by user-perspective framing propagated across multiple AIs before producing a real-world harm.

The most analytically significant finding under Calibration concerns the moderator of verification skill. The popular AI-and-experience expectation, adjacent to Dell'Acqua et al.'s (2023a) discussion of junior training pipelines (drawing on Beane, 2019), is that calibration accuracy diverges along seniority lines with junior professionals more susceptible to over-reliance. The data tells a more nuanced story. Eighteen statements across 11 firms converge on a different framing: curiosity, openness and prior personal exposure to AI moderate use more reliably than seniority does, an axis closer to Caplin et al.'s (2025) calibration construct than to the popular seniority framing. A Digital Transformation Manager articulated this most directly: "I wouldn't split junior-senior, I'd split by personality. Because there are curious juniors who'll learn faster than non-curious seniors. You'll have an inversion of roles. The catastrophic scenario: the un-curious junior. No experience, no willingness to try" (`I-Q-2` S27). An HR Technology lead at a food retailer reached the same conclusion from a different vantage: "it's not really about age. It's about openness to technology. I know people much younger than me who have never used ChatGPT and aren't interested. So it's less generational and more individual in my view" (`M-A-1` S09).

Two counter-cases preserve the original Proposition 4 claim on bounded grounds. The clearest is the only quantitative behavioural metric in the corpus: at the telematics multinational, where AI-generated pull requests are instrumented, "all junior engineers averaged 24 seconds for code approval. Average seniors: seven minutes and more. Seniors are genuinely reading the code line by line. Juniors check if it runs locally and approve" (`M-K-1` S12). A Luxembourg law firm provided a qualitative parallel (`L-G-1` S07 [¶ paraphrase]). The same Innovation Director also observed the inverse pattern in marketing, support, and project-management functions of the same firm where juniors outperform seniors with AI (`M-K-1` S18; cross-confirmed at `S-J-1` S09). The original Proposition 4 therefore holds only in narrow, high-stakes, expert domains where reading the output line by line is the operative quality gate. Outside those domains, disposition is the binding axis.

Three supporting clusters extend the calibration picture (full quote evidence in Appendix A2. Calibration: Supplementary Quotes). **Capability-building infrastructure** (15 statements, 9 firms): formal training programmes with the inflection point repeatedly placed at the shift from online to in-person delivery (`L-M-4` S03 paraphrased; `L-H-1` S05); prompt-craft framed as cognitive labour rather than effort-saving (`L-G-1` S14 [¶ paraphrase]). **Foundational deskilling and cognitive offloading** (13 statements, 7 firms): the industrial-group CEO articulating the binding concern as "don't lose the hand" (`M-L-1` S13); foundational deskilling observed in junior developers unable to debug AI-produced code without global view (`S-I-1` S09); a counter-pattern of comprehension-preservation discipline among professionals who deliberately under-delegate (`M-K-1` S15). **Heavy-user-as-most-sceptical** (3 statements, 3 firms): heaviest AI users in advisory and sole-operator contexts are also the most rigorous verifiers, suggesting the calibration risk is heavy AI use without the corresponding verification discipline not heavy AI use per se (`S-L-2` S07; `S-I-1` S05).

Two additional findings deserve flagging: **critical thinking as comparative advantage** (13 statements, 8 firms) framed as "value moves from answer-finding to question-asking" (`M-L-1` S04) and as the "judgment company" identity of an AI-core firm (`S-J-1` S02) and the **AI-as-onboarding metaphor** (`I-M-2` S08; `L-M-4` S08; `M-B-2` S09), which preserves human expertise authority while normalising AI use.

The Calibration dimension is the most evidence-dense in the corpus. Calibration emerges as a practiced, learned, multi-pattern discipline rather than a passive trait. Verification cost is real and persistent but bounded by triangulation, stake-asymmetry, and references. The naive seniority story does not hold: disposition does most of the work with junior over-acceptance retained only as a residual mechanism in narrow domains.

9.4 Interface and Process Design

Interface and Process Design, the third organisational lever, concerns how the tools and workflows around AI shape its use. The data structure consolidates 130 statements under this dimension across 21 themes. Two findings stand out by cross-firm coverage. The first, data sovereignty as the primary governance criterion, was articulated in 18 firms and is the strongest cross-industry pattern in the entire corpus. The second is the inadequacy of unstructured deployment: governance maturation lags adoption across most of the same firms.

Data sovereignty in this corpus is rarely about technical compliance alone. It is about jurisdictional control, intellectual property (IP) protection and increasingly geopolitical exposure. A finance contributor reported a planned strategic pivot from Copilot to Mistral driven by European sovereignty concerns: geopolitical sovereignty operates as an upstream determinant of toolchain choice (`I-P-1` S17). An HR Technology lead at a food retailer reported acquisition of an in-house AI company to enforce closed-loop sovereignty with GDPR as a second driver and sovereignty-by-architecture extended to vendor ownership (`M-A-1` S01). The Head of Marketing at a software firm captured the bidirectional logic of IP exposure: incoming exposure as third-party violation, outgoing exposure as own IP into the training set (`M-B-2` S02). The pattern is so consistent across the corpus that it functions less as a governance choice and more as a precondition for serious enterprise deployment.

Beneath this sovereignty floor, governance maturity diverges. Thirteen statements across nine firms describe governance that lags adoption. Shadow use precedes formal policy and rules emerge reactively. The founder of a sole-operator agency framed the dynamic crisply: prohibition produces shadow use and decentralised governance with traceability and education is the realistic alternative (`S-I-1` S12). A psychosocial researcher described the gap as systemic across firms she had studied: the governance gap is visible across rules, training, and managerial support (`I-N-1` S09).

Where firms move toward maturity, they do so through three families of mechanism. **Embedded governance** (12 statements, 5 firms): rules baked into tool ergonomics through skills or system-prompt-level configurations users do not even read, exemplified by "governance smuggled into reusable skills" at one tech firm (`I-O-1` S19) and by playbooks "default plus custom per business unit" at a construction multinational (`L-M-4` S02). **Mandatory human-in-the-loop at the expertise boundary** (11 statements, 7 firms): HITL as inviolable principle, anchored variously by regulator (GMP in pharma, `L-H-1` S02), corporate red line (telematics, `M-K-1` S11), or regulatory ceiling from the EU AI Act framed as "propose, not decide" (`M-C-1` S01). **Upstream scoping rather than downstream**

warning (9 statements, 7 firms): the design choice of "shifting calibration upstream into the model configuration, not just downstream into user verification" (`I-Q-1` S14), operationalised as an 85% confidence gate that declines low-confidence answers (`M-K-1` S25) and structured custom-GPTs with fixed contextual scaffolding (`M-S-1` S05).

A complementary finding is that surface-level UI warnings are widely acknowledged as inadequate. The Programme Manager stated the diagnosis plainly: "disclaimers acknowledged as legal protection, not effective behavioural intervention" (`I-Q-1` S13). The combined message of the upstream-scoping cluster and the disclaimer cluster is that Proposition 5 needs reformulation: it is not whether AI presents output as draft or as final answer that determines verification behaviour; it is whether the AI's scope, context, and confidence have been bounded upstream. This reformulation is developed in Part B.

Three additional governance families round out the picture. **Governance-by-restriction** (hard perimeter): 14 statements across nine firms describe scope-restricted tools and blocked external paths as the operative governance lever, exemplified by the finance contributor's firm where "external tools are blocked; only IT-Security-vetted internal tools are accessible" (`I-P-1` S11). **Behavioural instrumentation** (4 statements, 2 firms): per-user budget caps with cost and CO_2 visibility as a calibration nudge (`I-Q-1` S19), and behavioural ratios as proxy indicators of cognitive offloading (`I-Q-2` S33). **Codified scaffolding** (9 statements, 7 firms): written charters, AI Offices reporting to the executive committee and external regulatory alignment as the institutional anchor of governance.

A distinctive single-firm cluster at one food retailer (`M-A-1`) pulls together two participatory mechanisms the rest of the corpus did not surface: a co-authored knowledge base with user-rating loops paired with a deliberately "rough" in-house UX as a data-stays-inside trust signal (`M-A-1` S08; S13); and a stakeholder-asymmetric persuasion pattern where AI-first triage at the expert/lay boundary triggers resistance (`M-A-1` S07; S11).

A final theme under this dimension, locked at consolidation_003 as a sub-theme of Proposition 6, is pace-mismatch and temporality. The pace-mismatch finding has firm-level evidence at 10 or more firms but is structurally distinct from the rest of Interface Design because its mechanism is temporal rather than design-driven: tool evolution outruns approval cycles. The Head of Marketing at the software firm described the pressure as the binding governance friction (`M-B-2` S20). The IT Director at the footwear retailer made the cross-firm comparison concrete: a Thursday-to-Monday tool deployment cycle in a small firm, "about 3 months" at a comparable large industrial firm (`M-S-1` S20). Pace-mismatch is treated in Part C as the cross-cutting temporal pressure on the four-dimension framework rather than as a Proposition 6 sub-finding alone.

The data does not support the simple draft-versus-final-answer reading of Proposition 5 but a more demanding one: upstream scoping, recalibration checkpoints and structural perimeter dominate downstream warnings. Sovereignty operates as a near-universal precondition; governance maturity lags adoption almost everywhere; the firms that escape do so by encoding rules into reusable tool ergonomics, holding HITL as non-negotiable and binding the model upstream before it reaches the user. Full supplementary quotes in Annex A3. Interface and Process Design: Supplementary Quotes.

9.5 Junior Pipeline

The Junior Pipeline dimension was not in the original framework. It was promoted at consolidation_002 after the pattern appeared in five or more firms across multiple industries with a coherent underlying mechanism and was retained at consolidation_004 with cumulative coverage of 10 firms in eight industries. The data structure consolidates 15 statements as primary in this dimension with an additional eight cross-tagged statements primary in Calibration or Emergent.

The core mechanism is straightforward. Senior professionals who adopt AI substitute it for the work that juniors used to do. Junior hiring slows. The pipeline of professionals who will eventually become tomorrow's seniors thins. Within a 5-to-10-year horizon, the firms most aggressive about substitution face a continuity problem: AI cannot yet replace senior judgement and they will not have produced new seniors. A Managing Director at a software firm framed it as "the principal short-termism risk of AI adoption: structural feedback loop, AI displaces juniors, no pipeline to future seniors, expertise erosion in approximately 10 years" (`M-B-1` S03 [¶ paraphrase]).

Nine statements across seven firms describe this reduced-junior-hiring pattern in concrete present-tense terms. The Corporate BIM Manager described a legal AI tool displacing "two to three juniors" already (`I-M-2` S15). The telematics Innovation Director reported an explicit hiring stop on juniors alongside observed entry-level skill atrophy: "I know how to prompt" as a stand-in for know-how (`M-K-1` S13). The lead-generation startup founder reported the strongest articulation: a 13-to-3 headcount collapse paired with an explicit zero-junior hiring policy and a hiring criterion shift to "senior plus AI-native" (`S-J-1` S03; S04). The Commercial Manager at the advisory firm described AI-driven outbound (Apollo email sequences) replacing dedicated meeting-booking hires in a small sales team: "I got far more meetings than we had ever gotten by hiring people exclusively to book meetings" (`S-L-2` S01).

A counter-case is worth flagging because it occurs at the same firm as one of the strongest substitution cases. At the construction multinational, two senior voices articulate opposite postures within the same legal department, while a senior voice in a different department articulates deliberate junior recruitment maintenance "despite AI displacement potential, succession logic" (`I-M-3` S15). The Innovation Lead at the financial-market-infrastructure firm reframes the pattern more broadly as a pre-AI declining-junior-hiring trend with the AI-attribution still under discussion rather than enacted as zero-junior policy (`I-C-2` S14). The Junior Pipeline dimension therefore captures both enacted substitution and enacted preservation, with the same firm sometimes containing both.

Three narrower mechanisms extend the picture. **Mentoring and knowledge-transfer collapse**: senior preference for AI over junior collaboration as direct erosion of intergenerational transfer (`L-G-1` S11 [¶ paraphrase]). **Workforce-pyramid disruption at the offshore tier**: possible concentration on senior local labour if junior offshore work commoditises (`I-Q-2` S26). **Value-capture and forecasted reversal**: AI shifts billing from hourly to flat fee with no obvious revenue uplift (`L-G-1` S15 [¶ paraphrase]); a related shift to a senior-freelancer model is enacted at the lead-generation startup (`S-J-1` S04); the same founder forecasts a rehiring cycle of "quality collapse will force rehiring of seniors first, then eventually juniors" (`S-J-1` S05), echoed by the Risk Domain Manager's "digging our own graves" framing of early-adopter second-order resistance (`M-C-1` S06).

A nuance worth flagging is the department-conditioned reversal already discussed under Calibration: in marketing, support, and project-management functions, juniors outperform seniors with AI whereas in engineering, seniors retain the advantage (`M-K-1` S18; `S-J-1` S09). The Junior Pipeline dimension is therefore not monotonically negative for juniors. The binding variable is the domain-criticality of judgement: where reading the output line by line is the operative quality gate, seniors win; where speed and pattern-matching are the operative gate, juniors win.

Taken descriptively, the Junior Pipeline dimension is the most consequential emergent finding in the empirical investigation. The substitution decisions documented here, taken in the past 12 to 24 months, will reshape firms' competence base over the next 5 to 10 years. Extended evidence and Table A.1 in Annex A4. Junior Pipeline: Extended Evidence.

9.6 Emergent Themes

The Emergent bucket retains 113 statements distributed across 18 themes, plus 13 single-source observations. Two emergent themes have been promoted at earlier consolidations: human accompaniment as binding meta-condition (E1) and pace-mismatch as Proposition 6 sub-theme (E2), both treated in detail in Part C. Three additional themes deserve a descriptive pass before the synthesis.

The first is anxiety and declared-versus-actual dissonance (11 statements, 4 firms): surface adoption optimism masking underlying anxiety (`I-N-1` S15), silent resistance not surfaced because of organisational pressure (`I-Q-2` S11) and a senior observability gap where a CEO discovered pervasive AI use among his management committee "by raise of hands, he was the only non-user" (`M-L-1` S06).

The second is augmentation-versus-substitution rhetoric. Thirteen statements across eight firms record explicit corporate stances on whether AI replaces or augments. Five distinct firm-level positions emerge in the corpus, often coexisting across firms in the same industry. Table 9.1 maps them.

Position	Sector	Articulation	Anchor
Augmentation at equal headcount	Telecoms multinational	"At equal human capital, we should produce more with AI rather than keep a constant productivity level with fewer people"	I-Q-2 S06
Augmentation as official corporate stance	Food retailer	Explicit non-substitution declared as official corporate position; competitive parity plus augmentation framing	M-A-1 S02
Honest-uncertainty middle ground	Footwear retailer	Refusing to give false reassurance about future automation; demystification plus facilitator-led adoption	M-S-1 S13; S03
Substitution framed without euphemism	Financial-market-infrastructure firm	"Companies seek headcount reduction, framed without euphemism"	M-C-1 S05
Substitution enacted at firm identity	Lead-generation software startup	13-to-3 headcount collapse; zero-junior hiring policy; senior-freelancer model	S-J-1 S03; S04

Table 2. *How firms position themselves on the automation–augmentation axis.* Each articulation is drawn from an interview and anchored to its source segment, showing the range from augmentation at equal headcount to substitution built into the firm's operating model.

A sixth articulation, from an industrial-group CEO, frames humanness preservation as a corporate-strategy stance rather than a personal preference (`M-L-1` S11). The same paradox that Raisch and Krakowski (2021) theorise is therefore visible in the corpus as a structured set of firm-level postures rather than as a single normative direction.

The third theme is levelling and value retreat to advisory (8 statements, 3 firms): the Luxembourg founding partner directly echoes Chapter 2's finding on distribution compression (Brynjolfsson et al., 2025), with AI that "lifts low performers, drags down high performers" framed as an augmentation paradox (`L-G-1` S13 [¶ paraphrase]; see also long-horizon paradox at S12). Two single-source observations push the corpus to its boundary: AI-driven attrition (`M-L-1` S03) and inverted deskilling at the sole-operator agency (`S-I-1` S10), both flagged but not generalised. Capability discovery as ongoing organisational task also runs through the corpus (`I-Q-2` S24).

The Emergent bucket confirms that the corpus contains substantial signal the framework does not exhaust while the most consequential signals, human accompaniment and pace-mismatch, are folded back into the analytical structure in Part C. Extended evidence in Annex A5. Emergent Themes: Extended Evidence.

Part B: Confronting the Framework

Proposition	Status	Refinement and anchor evidence
P1 Task Allocation	Supported.	Multi-axis (capability, stakes, sovereignty, regulatory exposure, motivational fit, function-asymmetry), not capability-only. Long-horizon coherence added as structural ceiling on full delegation (I-O-2 S19; M-B-1 S02 [¶ paraphrase]). Surprise-form failures on tasks expected to be trivial extend the original frame (I-P-1 S04). See §9.2.
P2 Task Allocation	Supported, re-stated as no-governance baseline.	Trial-and-error confirmed as default in 16 statements / 7 firms (L-M-1 S07; I-Q-2 S12; M-F-1 S08). Mature firms substitute structured alternatives: top-down task mapping (I-C-2 S04), portfolio governance (I-Q-1 S08), sandbox-then-industrialise (M-B-2 S12), process-driven embedding (M-C-1 S04).

		Guidelines are not the operative remedy; mappings, sandboxes and embedded processes are. See §9.2.
P3 Calibration	Supported, enriched.	Calibration manifests as a multi-pattern competence: stake-asymmetric verification (I-P-1 S07; I-O-2 S04; S-L-2 S04), references-as-anchor (L-M-4 S06; S11), confidence-threshold gating (M-K-1 S25; S-J-1 S07). Enriched by single-failure reversion (M-S-1 S09) and operational-urgency suppression (L-H-1 S18). See §9.3.
P4 Calibration	Reformulated (locked at consolidation_003).	Calibration readiness is moderated by personal disposition (curiosity, prior personal/domestic reflex, role operationality) more reliably than by formal seniority, though junior over-trust on specific output classes remains a residual mechanism. Anchored at I-Q-2 S27, M-A-1 S09, I-C-2 S13, L-H-1 S14, M-K-1 S14. Residual mechanism anchored at M-K-1 S12 (quantified: 24s vs 7+ min). See §9.3 + prose below.
P5 Interface Design	Reformulated.	Upstream scoping (RAG, context-injection, confidence-threshold gating, in-house style grounding, process-visibility) shapes verification more reliably than downstream framing (draft labels, disclaimers, warnings). Anchored at I-Q-1 S14, M-K-1 S25, M-S-1 S05, S-I-1 S07, M-A-1 S13. Disclaimers acknowledged as inadequate at I-Q-1 S13. See §9.4 + prose below.
P6 Interface Design	Supported across all 28 interviews.	Sovereignty-by-design anchors upstream (18 firms: I-P-1 S02; M-K-1 S01; M-A-1 S01; L-H-1 S10; M-B-2 S02). Mandatory HITL at expertise boundary anchors downstream (11 firms: L-H-1 S02; M-K-1 S11; M-C-1 S01). Three governance families operate in between (embedded skills, behavioural instrumentation, codified scaffolding). Pace-mismatch and temporality, locked as P6 sub-theme at consolidation_003, are treated as cross-cutting in Part C. See §9.4.

Table 3. *Empirical status of the six propositions after confrontation with the data.* Each row pairs a verdict (supported, refined, restated, or reformulated) with its anchor evidence, documenting where the framework held and where it was revised.

Part B confronts the three-lever framework with the empirical data on a proposition-by-proposition basis. Table 3 summarises the verdict for each proposition; the prose below justifies the two reformulations (P4 and P5) where the data forced a relocation of the operative mechanism. The propositions that survive without reformulation (P1, P3, P6) and the proposition that survives re-stated (P2) carry their argument from the table cells.

Proposition 4 reformulation. The original wording predicts a junior-senior asymmetry in over-reliance, moderated by organisational mechanisms. Eighteen statements across 11 firms converge instead on a dispositional axis (anchored at `I-Q-2` S27; `M-A-1` S09; `I-C-2` S13; full set in Appendix A2. Calibration: Supplementary Quotes). The Digital Transformation Manager's framing is the cleanest in the corpus: "I wouldn't split junior-senior, I'd split by personality. The catastrophic scenario: the un-curious junior" (`I-Q-2` S27). Two counter-cases preserve the original mechanism as a residual: the only quantitative behavioural metric in the corpus (`M-K-1` S12, 24s vs 7+ min code-approval gap) and a Luxembourg law-firm qualitative parallel (`L-G-1` S07 [¶ paraphrase]). The reverse pattern, juniors outperforming seniors in marketing, support, and project-management functions, is cross-firm (`M-K-1` S18; `S-J-1` S09). The binding moderator is domain-criticality, not seniority. Proposition 4 in its original wording does not survive. In its reformulated wording (locked at consolidation_003) it is one of the strongest cross-firm findings in the corpus.

Proposition 5 reformulation. The original wording locates the design mechanism downstream: draft labels and warnings encourage verification. The data does not contradict the wording; it locates the binding mechanism elsewhere. The clearest articulation comes from the Programme Manager at the telecoms group: "shifting calibration upstream into the model configuration, not just downstream into user verification" (`I-Q-1` S14). The same manager states the downstream-only diagnosis directly: disclaimers are "legal protection, not effective behavioural intervention" (`I-Q-1` S13). The upstream alternative appears in operational form: an 85% confidence gate (`M-K-1` S25); structured custom-GPTs with fixed contextual scaffolding (`M-S-1` S05); visible chains-of-thought as trust mechanism (`S-I-1` S07); a deliberately "rough" in-house UX as data-stays-inside signal (`M-A-1` S13). What survives of the original distinction is the weaker claim that provenance labelling is a useful secondary mechanism. The reformulated proposition supplies one of the two main conclusions of the Interface Design dimension: design the constraint into the model, not into the warning.

None of the six propositions is contradicted in substance. P1, P3, P6 survive in original wording; P4 and P5 after reformulation; P2 in re-stated form. The Junior Pipeline addition (promoted at consolidation_002) gives the four-dimension framework Part C synthesises.

Part C: Synthesis

Part C integrates the four aggregate dimensions into a refined framework. Three claims are developed. First, the four-dimension framework is the analytical successor to the three-lever framework derived from the literature; the addition of the Junior Pipeline dimension is descriptive rather than theoretical, in the sense that the data made it visible. Second, human accompaniment functions as the binding meta-condition under which all four dimensions operate. Third, pace-mismatch is the cross-cutting temporal pressure on the framework.

9.7 The Refined Four-Dimension Framework

The three-lever framework derived from Raisch and Krakowski (2021), Dell'Acqua et al. (2023a, 2023b), Caplin et al. (2025), and Al-Refai et al. (2025) is a framework about the conditions of individual and team-level human-AI collaboration. Task Allocation, Calibration, and Interface and Process Design map onto the three sites of decision where the automation-augmentation paradox is actively negotiated: the boundary between human and AI work, the user's judgement of AI output and the structural design of the tool and the workflow around it.

The four-dimension framework that the data supports adds a fourth site: the firm's workforce architecture, particularly its junior pipeline. The addition is not a theoretical extension of Raisch and Krakowski (2021). It is an empirical observation that AI deployment, where it is enacted at scale, reshapes the workforce pyramid before the firm has resolved the calibration and process-design questions for its existing workforce. Substitution decisions are happening now, with senior interviewees describing already-reduced junior hiring (`I-M-2` S15; `L-M-4` S09; `M-K-1` S13; `S-J-1` S03; `L-H-1` S20; `S-L-2` S01), while the consequences are forecast over a 5 to 10-year horizon (`M-B-1` S03 [¶ paraphrase]).

The four dimensions are not independent. They are coupled through the augmentation-substitution choice mapped in Table 2. A firm that resolves the paradox in the augmentation direction tends to invest in Calibration, in upstream-scoping Interface design, and in preserving the Junior Pipeline. A firm that resolves it in the substitution direction tends to invest in Task Allocation that displaces juniors, in Interface design that scales the substitution, and may underinvest in Calibration because the burden moves to a smaller senior cadre. The data suggests that the four dimensions cohere into two firm-level strategic postures, with the paradox visible in the corpus as the five distinct positions of Table 2. The framework's value is in showing what choices each posture involves, not in prescribing which posture is correct. The current state of the corpus also suggests that few firms have made the choice explicit. The most common posture is one in which AI substitution is enacted at the task level while augmentation is articulated at the rhetorical level. The mismatch is not necessarily hypocritical, it reflects the fact that firms are still discovering what the four-dimension picture looks like in their own context.

9.8 Human Accompaniment as the Binding Meta-Condition

The strongest cross-cutting finding in the corpus is that human accompaniment, the active managerial work of supporting users through change, functions as the binding constraint on all four dimensions. Fourteen statements across seven firms describe this directly. The pattern recurs in the language of "accompaniment", "change management", "reception" and "user adoption" across firms in different industries and languages. A Programme Manager at the telecoms group articulated the diagnosis as "chronic under-investment in human accompaniment named as the most underestimated obstacle" (`I-Q-1` S25). An Innovation Lead at a financial-market-infrastructure firm reinforced the same point: change management identified as the binding implementation challenge with the tool-launches-itself fallacy explicitly named (`I-C-2` S16). The HR Technology lead at the food retailer named "accueil", the way users receive the tool, as "the determining variable in AI deployment success" (`M-A-1` S15). The IT Director at the footwear retailer described the per-person variation as the structural feature: change management as a per-person problem with large-firm budget unable to substitute for it (`M-S-1` S17; S19).

The mechanism through which accompaniment binds the four dimensions is straightforward: Task Allocation, Calibration, Interface design, and Junior Pipeline are all rules-as-written until a person enacts them. The gap between rule and practice is concrete in the data: 80% of users did not adopt a deployed AI scheduling tool despite formal rollout at one food retailer (`M-A-1` S04); at the construction multinational, three months of online training failed and in-person training in the fourth month was the inflection point (`L-M-4` S03); at the telecoms group, sustained behavioural change required a two-year embedded resource and reverted when the resource was withdrawn (`I-Q-2` S31; S32).

Accompaniment is treated as a meta-condition rather than a fifth dimension (locked at consolidation_003) because it does not have its own theoretical mechanism distinct from the four. It shapes whether each dimension's mechanism translates into practice. The same dispositional variation that determines Calibration outcomes determines accompaniment intensity; the same governance maturity that produces structured workflows determines whether accompaniment has organisational backing; the same firm-level substitution-or-augmentation stance that drives the Junior Pipeline dimension determines whether accompaniment is funded or hollowed.

The accompaniment finding has two broader implications. Methodologically, it explains why adoption-rate metrics without depth-of-integration measurement produce optimistic readings of the paradox: adoption rates measure tool penetration, sustained integration measures accompaniment success and they diverge in the corpus. Theoretically, the corpus suggests change management at the user-accompaniment level is the binding intra-firm variable that the macro productivity-paradox literature (Brynjolfsson et al., 2021; Acemoglu, 2024) names but does not microscope: whether the intangible complementarities they flag translate into firm-level productivity depends on per-user accompaniment. The corpus suggests it is the operative variable inside the paradox: where invested in, the four-dimension picture coheres; where not, the dimensions decouple and the paradox produces the Vaccaro et al. (2024) under-performance result.

9.9 Pace-Mismatch as the Cross-Cutting Temporal Pressure

The second cross-cutting finding is temporal: the pace of AI tool evolution outruns the pace of organisational absorption. Nine statements across seven firms describe this gap with strikingly consistent quantitative estimates: pace of model evolution outruns enterprise approval cycle which can create psychological fatigue from never-settled toolchain (`M-B-2` S20); AI strategy must be re-planned every approximately three months due to capability shifts (`I-O-1` S21); a 1 to 3-year window of competitive advantage from early adoption before commoditisation (`S-L-2` S09, corroborated at `S-J-1` S10).

The pace-mismatch finding interacts with each of the four dimensions. Under Task Allocation, it produces a moving boundary: what AI cannot do this quarter, it may be able to do the next and the firm's delegation rubric is therefore necessarily provisional. Under Calibration, it produces a moving competence: prompt-craft and verification heuristics must be relearned with each generation of model. Under Interface and Process Design, it produces a moving governance object: charters, perimeters and approval cycles drafted against today's tools become irrelevant against next year's tools. Under Junior Pipeline, it produces a moving prediction: the rate at which AI substitutes for juniors and the rate at which it might create new junior-friendly roles are both moving and firms making zero-junior decisions today are betting against a moving target.

The structural design responses observed in the corpus are consistent with the diagnosis. Modular swap-in architectures (`M-K-1` S10; `S-I-1` S03) are architectural hedges against tool obsolescence; sandbox-then-industrialise governance (`M-B-2` S12) and recalibration checkpoints (P5 cluster) are governance and calibration hedges against premature commitment and chained-AI drift; the lead-generation startup's "judgment company" framing (`S-J-1` S02) is an identity hedge in which the firm declares human judgement as its irreducible value-add and rebuilds its workforce around it.

Pace-mismatch behaves differently from the four dimensions and from accompaniment. It is not a site of decision. It is a structural property of the AI deployment environment that the firm must absorb. The corpus suggests that firms successfully managing the paradox treat pace-mismatch as a design constraint rather than as a transitional problem. Firms that treat pace-mismatch as transitional, waiting for tools to stabilise before formalising governance, instead accumulate shadow use and lose the optionality of structured deployment in keeping with the governance-lags-adoption pattern.

9.10 Implications for the Automation-Augmentation Paradox

The empirical investigation returns to the question with which the thesis began. The under-performance result documented by Vaccaro et al. (2024) is recognisable in the data: verification cost erodes productivity gain (C1 cluster, 14 statements / 8 firms); 80% non-adoption despite formal rollout (`M-A-1` S04); pace of model change makes long-horizon strategy obsolete (`I-O-1` S21); 60% of company writing is AI-generated with only 34% factually correct (`M-K-1` S22). The mechanism is not tool failure but coordination failure across the four dimensions with human accompaniment as the binding meta-condition and pace-mismatch as the cross-cutting temporal pressure.

Successful management is recognisable in firms that have made the four-dimension picture coherent at very different scales (a telecoms group, a footwear retailer, a sole-operator agency); Appendix A8. Firm-Level Patterns: Managing the Paradox Well vs Under-Performing (Table A.2) details the specific combinations observed. The corpus does not converge on a single best practice but on a structural reading: firms managing the paradox well treat the four dimensions as coupled, invest in accompaniment as the meta-condition and design for pace-mismatch as a permanent feature of the environment.

The argument of this thesis is therefore empirically supported in a specific form. Sustainable productivity gains from AI depend on managing the tension between automation and augmentation. The tension is not resolved at the level of any single lever but rather at the level of the four dimensions taken together, under the binding condition of human accompaniment, against the cross-cutting pressure of pace-mismatch. The under-performance documented by Vaccaro et al. (2024) reflects firms that have not yet made that integration. Implications for theory, practice and further research are developed in Chapter 10.

Chapter 10: Implications and Conclusion

This thesis began with the paradox documented in the meta-analysis by Vaccaro et al. (2024) and located it inside the automation-augmentation paradox of Raisch and Krakowski (2021), developing three organisational levers (task allocation, calibration, interface and process design) as the mechanisms through which firms negotiate the paradox. The empirical investigation confronted that framework with 28 interviews and refined it: two propositions reformulated, one re-stated, one aggregate dimension added (Junior Pipeline), and two cross-cutting findings (human accompaniment, pace-mismatch) locked. This chapter develops the theoretical, managerial and methodological implications, acknowledges the limitations, and identifies the questions further research should pursue.

10.1 Theoretical Contribution

The refined four-dimension framework is the theoretical contribution of this thesis. The addition of Junior Pipeline as a fourth dimension is descriptive, surfaced abductively from the data. The substantive contribution is the relocation of the operative mechanism in two of the three original levers. Under Calibration, the data forces a move from seniority to disposition as the binding moderator: curiosity, openness, and prior personal exposure to AI predict use quality more reliably than tenure does with junior over-trust retained only as a residual mechanism in narrow domain-critical contexts. Under Interface and Process Design, the data forces a move from downstream warnings to upstream scoping: confidence-threshold gating, RAG scoping, context-injection and process-visibility shape verification behaviour more reliably than draft labels or disclaimers. These reformulations relocate where in the workflow the design decision matters.

The two cross-cutting findings extend the contribution. Human accompaniment as binding meta-condition operationalises the user-accompaniment layer that the macro productivity-paradox literature (Brynjolfsson et al., 2021; Acemoglu, 2024) names but does not directly model. Pace-mismatch argues that the AI deployment environment is structurally non-steady-state, and firms successfully managing the paradox treat it as a design constraint rather than as a transitional problem. The thesis therefore extends Raisch and Krakowski's (2021) paradox formulation by grounding it in observable organisational practice.

10.2 Managerial Implications

Four implications follow for managers responsible for AI deployment.

First, the four dimensions are not separable. Deploying tools without process and without training, the most common posture in the corpus, produces precisely the under-performance documented by Vaccaro et al. (2024). A firm choosing aggressive substitution at the Task Allocation level should expect calibration burden to concentrate on a smaller senior cadre and Junior Pipeline consequences at a 5 to 10-year horizon; a firm choosing augmentation should expect to invest in calibration training as a sustained per-person practice. Making the substitution-vs-augmentation trade-off explicit is the first managerial recommendation.

Second, the operative remedy for unmanaged AI use is not behavioural rules. Mature firms substitute task mappings, sandboxes and process-driven embedding for trial-and-error (P2 re-statement). Managers attempting prohibition or behavioural prescription should expect shadow adoption rather than compliance. The operative interventions are upstream: scope-restricted tools, embedded skills in tool ergonomics, confidence-threshold gating and mandatory human-in-the-loop at the expertise boundary.

Third, calibration training is the most under-deployed managerial lever in the corpus. Existing training infrastructure typically goes unused for AI; online training repeatedly fails where in-person training is the inflection point and the prompt-craft skills that materialise AI productivity gain are the same skills that erode under uncritical AI use. Managers should treat calibration training as an active, domain-specific, temporally sustained practice rather than as a one-off rollout.

Fourth, Junior Pipeline decisions are being made now with consequences that will materialise over 5 to 10 years. The corpus documents reduced junior hiring in 10 firms across eight industries. Managers should make the substitution-or-augmentation choice explicit before further junior-hiring reductions lock in workforce-architecture consequences. The rehiring cycle one founder forecast (`S-J-1` S05) is a warning, not a prediction: firms aggressively reducing junior cadres now may face quality collapse and re-hiring at premium cost in 3 to 5 years.

10.3 Methodological Implications

The most consequential methodological implication is that adoption-rate metrics, even where carefully measured, may overstate the depth of real integration. Surface adoption optimism in the corpus repeatedly masked underlying anxiety, silent resistance and shadow use. Studies that measure tool penetration without measuring depth-of-integration produce optimistic readings of the paradox. A more rigorous follow-up would combine adoption-rate instrumentation with embedded ethnographic observation. The iterative-consolidation practice used in this thesis (four milestones with locked decisions per milestone) also demonstrates the value of explicit revision points in abductive qualitative research.

10.4 Limitations

Five limitations bound the claims made in this thesis. First, the sample concentrates on senior decision-makers and under-represents end-user practice. The findings are stronger on firm-level paradox dynamics than on department-specific end-user behaviour. Second, recruiting across multiple organisations and industries introduces variability that cannot be fully separated from firm-specific, industry-specific, or national-cultural factors. Third, reliance on self-reported experience exposes findings to social desirability bias, particularly around over-reliance; triangulation against concrete examples mitigated but did not eliminate this risk. Fourth, the corpus reflects a specific moment in the deployment of foundation-model AI in European and Belgian firms (late 2025 to early 2026). The pace-mismatch finding implies that findings tied to the current model generation (confidence-threshold gating, specific tool stacks) should be read as snapshots, while findings tied to more durable organisational mechanisms (sovereignty-by-design, accompaniment, dispositional moderator, junior-pipeline mechanism) are likely more stable. These limitations circumscribe but do not invalidate the four-dimension framework or the synthesis claims.

10.5 Further Research

Three avenues for further research follow directly from the findings and limitations. First, the empirical pivot left department-specific end-user mechanisms thinly documented. The dispositional moderator of calibration (P4 reformulation) would benefit from direct end-user ethnographic observation, ideally trying to capture calibration erosion through cognitive offloading (Risko & Gilbert, 2016; Al-Refai et al., 2025). Second, the Junior Pipeline finding rests on senior decision-maker forecasts over a 5 to 10-year horizon. A longitudinal study tracking junior hiring, mentoring and senior-cadre quality in firms that have enacted zero-junior policies would test the predicted rehiring cycle directly. Third, the pace-mismatch finding implies that organisational practice studies of AI deployment should be designed for replication: a multi-year study repeating the interview protocol every 18 to 24 months across the same firms would test whether the four-dimension framework, accompaniment and pace-mismatch hold across model generations.

10.6 Personal Assessment

What I take from these 28 interviews is that the substitution-versus-augmentation choice is the practically operative question of the current AI rollout, and most firms have not yet made it explicitly. The corpus shows substitution being enacted at the task level (reduced junior hiring, cost-anchored replacement of headcount, billing-model shifts) while augmentation is articulated at the rhetorical level (corporate stances on humanness, framing of "augmented humans at equal headcount", refusals to use the word "replace"). The mismatch is not always hypocritical: in many firms it appears to reflect genuine confusion about what each posture entails in operational terms. But the consequences of the mismatch fall on the workforce architecture, particularly on the junior cadre over a horizon longer than the rhetorical cycle. The most consequential managerial choice is to make the substitution-or-augmentation posture explicit before the operational decisions lock it in.

The most important shift in my own thinking during the empirical phase was the reformulation of Proposition 4. I began the research expecting a seniority story in which junior professionals over-trust AI and seniors push back at the boundary. The data forced a different reading: curious junior plus disengaged senior is a more common pattern than the textbook asymmetry suggests and disposition turns out to do more analytical work than tenure does. This is not a refinement of detail. It changes what calibration training should look like, who needs it and how organisations should screen for it. It also changes what the literature on age-cohort effects in AI adoption can and cannot tell managers.

The most underestimated finding, in my view, is the junior pipeline collapse. The substitution decisions being enacted now (zero-junior hiring policies, AI-as-junior-substitute, billing-model shifts that decouple productivity gain from monetisation) have consequences that will materialise over 5 to 10 years. Firms making these decisions in 2024 to 2026 are betting on a future they cannot yet evaluate. The rehiring cycle one founder forecast (`S-J-1` S05) reads less as a prediction and more as a warning: firms that aggressively reduce junior cadres now will face quality collapse and re-hiring at premium cost in 3 to 5 years. Whether they avoid that fate depends less on AI tool selection or governance documents than on how much they invest in human accompaniment, the binding meta-condition under which the four dimensions cohere.

My overall assessment of the current state of human-AI collaboration is that it is, on balance, slightly better than it would be without AI but operating far below the potential the technology offers. The upper-bound performance visible in the corpus, the firms that have made the four-dimension picture coherent at very different scales, shows that the gap between current practice and that potential is not a fixed feature of the technology. It is a feature of organisational choice. The four-dimension framework offered here is intended as a managerial diagnostic to help firms recognise where they currently stand and what choices remain available to them not as a prescription of a single correct posture.

10.7 Conclusion

This thesis sets out to ask how organisations can coordinate task allocation, calibration, and interface and process design to sustain human-AI collaboration. The literature provided a three-lever framework derived from Raisch and Krakowski (2021), Dell'Acqua et al. (2023a, 2023b), Caplin et al. (2025), and Al-Refai et al. (2025). The empirical investigation confronted that framework with 28 interviews and refined it: two propositions reformulated, one re-stated, a fourth aggregate dimension added, and two cross-cutting findings (human accompaniment, pace-mismatch) named. The argument the thesis defends is that sustainable productivity gains from AI depend on managing the tension between automation and augmentation. The tension is not resolved at the level of any single lever. It is resolved at the level of the four dimensions taken together under the binding condition of human accompaniment, against the cross-cutting pressure of pace-mismatch. The under-performance documented by Vaccaro et al. (2024) reflects the prevalence of firms that have not yet made that integration and the upper-bound performance visible in the corpus reflects firms that have begun to.

References

Acemoglu, D. (2024). The simple macroeconomics of AI (NBER Working Paper No. 32487). *National Bureau of Economic Research.* https://www.nber.org/papers/w32487

Al-Refai, O., Colunga Reyes, Y., Hammad, E., & Riley, C. (2025). Human-AI interactions: Cognitive behavioral, and emotional impacts. *IEEE Transactions on Technology and Society.* https://doi.org/10.48550/arXiv.2510.17753

Bartel, A. P., Ichniowski, C., & Shaw, K. (2007). How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills. *The Quarterly Journal of Economics, 122(4), 1721–1758.* https://doi.org/10.1162/qjec.2007.122.4.1721

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences.* https://doi.org/10.1073/pnas.2422633122

Bick, A., Blandin, A., & Deming, D. J. (2024). The rapid adoption of generative AI (Working Paper No. 2024-027). *Federal Reserve Bank of St. Louis.* https://doi.org/10.20955/wp.2024.027

Boulat, R. (2006). La productivité et sa mesure en France (1944-1955). *Histoire & Mesure, XXI(1), 79–110.* https://doi.org/10.4000/histoiremesure.1537

Bresnahan, T. F., & Trajtenberg, M. (1995). General purpose technologies "Engines of growth"? *Journal of Econometrics, 65(1), 83–108.* https://doi.org/10.1016/0304-4076(94)01598-T

Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics (NBER Working Paper No. 24001). *National Bureau of Economic Research.* https://doi.org/10.3386/w24001

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics, 13(1), 333–372.* https://doi.org/10.1257/mac.20180386

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics.* https://doi.org/10.3386/w31161

Caplin, A., Deming, D., Li, S., Martin, D., Marx, P., Weidmann, B., & Ye, K. J. (2025). The ABCs of who benefits from working with AI: Ability, beliefs, and calibration. *Management Science.* https://doi.org/10.1287/mnsc.2024.08994

Charmaz, K. (2014). Constructing Grounded Theory : A practical guide through qualitative analysis (2nd ed.). *SAGE Publications.* https://doi.org/10.7748/nr.13.4.84.s4

Choi, J. H., & Schwarcz, D. (2025). AI assistance in legal analysis: An empirical study. *Journal of Legal Education, 73(2), 384–419.* https://doi.org/10.2139/ssrn.4539836

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023a). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.4573321

Dell'Acqua, F., Kogut, B., & Perkowski, P. (2023b). Super Mario meets AI: Experimental effects of automation and skills on team performance and coordination. *The Review of Economics and Statistics, 1–47.* https://doi.org/10.1162/rest_a_01328

Dixon, J., Hong, B., & Wu, L. (2020). The employment consequences of robots: Firm-level evidence (Analytical Studies Branch Research Paper Series, No. 454). *Statistics Canada. Catalogue no. 11F0019M.* https://doi.org/10.2139/ssrn.3422581

Dubois, A., & Gadde, L.-E. (2002). Systematic combining: An abductive approach to case research. *Journal of Business Research, 55(7), 553–560.* https://doi.org/10.1016/S0148-2963(00)00195-8

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14(4), 532–550.* https://doi.org/10.2307/258557

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16(1), 15–31.* https://doi.org/10.1177/1094428112452151

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18(1), 59–82.* https://doi.org/10.1177/1525822X05279903

Lincoln, Y. S., & Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Directions for Program Evaluation, 1986(30), 73–84.* https://doi.org/10.1002/ev.1427

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review, 46(1), 192–210.* https://doi.org/10.5465/amr.2018.0072

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20(9), 676–688.* https://doi.org/10.1016/j.tics.2016.07.002

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies, 52(4), 701–717.* https://doi.org/10.1006/ijhc.1999.0349

Smith, W. K., & Lewis, M. W. (2011). Toward a theory of paradox: A dynamic equilibrium model of organizing. *Academy of Management Review, 36(2), 381–403.* https://doi.org/10.5465/AMR.2011.59330958

Spielman, Z., & Le Blanc, K. (2021). Boeing 737 MAX: Expectation of human capability in highly automated systems. In Advances in Human Factors in Robots, Drones and Unmanned Systems (Vol. 270, pp. 64–71). *Springer.* https://doi.org/10.1007/978-3-030-79997-7_8

Syverson, C. (2010). What determines productivity? (NBER Working Paper No. 15712). *National Bureau of Economic Research.* https://doi.org/10.3386/w15712

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour.* https://doi.org/10.1038/s41562-024-02024-1

Wu, S., Liu, Y., Ruan, M., Chen, S., & Xie, X.-Y. (2025). Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports.* https://doi.org/10.1038/s41598-025-98385-2

Appendix

This appendix contains supplementary evidence for Chapter 9. Quotes are organised by aggregate dimension (Task Allocation, Calibration, Interface and Process Design, Junior Pipeline, Emergent). The full quote bank and master coding file are in the attached excel. Anonymisation conventions follow Chapter 8 §8.8 throughout.

Contents

A1. Task Allocation: Supplementary Quotes

TA1 Stakes-based delegation gradient. A Head of Marketing at a software firm extended the engineering-stakes logic to product design: "we need to distinguish sharply between our own internal AI usage and AI within our products, because the processes our products support are much more sensitive and risk-critical than internal processes. If AI writes a bad text, I can fix it. But if AI takes a wrong decision on the shop floor and the user follows it blindly, the effect could be tremendous, expensive, dangerous. The closer you get to the shop floor, the less AI adoption you see. It's a trust problem" (M-B-2 S16).

TA2 Jagged frontier, surprise form. A finance contributor at a bank described a routine cross-document update that AI refused to perform meaningfully: "I had a document already written and another document with all the necessary figures. I just needed to update the Word document. I thought, that's good, it'll do it for me. But it wasn't capable. It highlighted everything I had to update in yellow. Like, yes, I know, I know I need to update it" (I-P-1 S04). The Global Head of QC at a pharmaceutical multinational described a different surprise form: AI was surprisingly strong on content discovery in a strategic-vision deliverable but surprisingly weak on narrative synthesis, requiring near-complete rework (L-H-1 S07).

TA3 Sovereignty / IP-tier-based delegation. A research-project manager: "the higher the sensitivity, whether data sensitivity, intellectual property, or sensitive sectors like defence, the less we'll use those tools. Even with a paid version where OpenAI commits not to train its algorithm on our data, we're not completely hacking-proof, and we won't put that in there" (L-E-1 S09). A Risk Domain Manager at a financial-market-infrastructure firm reported an in-house Copilot tuned to firm vocabulary as the operative reason for daily delegation (M-C-1 S03).

TA4a Trial-and-error baseline. The Head of Marketing at a software firm described shadow adoption running ahead of policy: "many people are already using AI, some even before it was officially approved, using personal subscriptions to ChatGPT or Mistral. Our company is full of engineers who love testing new things" (M-B-2 S15). A Head of Customer Care at a regional bus operator corroborated the shadow variant: "some agents will go to ChatGPT and ask it to generate a full response. Others use AI only for personal things" (M-F-1 S08).

TA4b Deliberate moves beyond trial-and-error. A GenAI Programme Manager at a telecoms group described a portfolio-management process: "business cases come in with an expected value. Steering and arbitration committees then evaluate which use cases get onto the roadmap based on ROI or on a complexity-versus-impact matrix, prioritising those that deliver the most value with manageable complexity" (I-Q-1 S08). An Innovation Lead at a financial-market-infrastructure firm described top-down task and skill mapping as a precondition for AI deployment, with the current bottom-up approach explicitly framed as transitional (I-C-2 S04).

TA5 Combinatorial-complexity tasks beyond human reach. The Global Head of QC described AI used precisely where combinatorial complexity exceeds practical human reach: "multiple parameters that are each within the acceptable limit can have a cumulative effect that pushes you out of spec. On your own, without AI to cross all those parameters, it would take an enormous amount of time. Too many parameters, too many possible and unimaginable combinations" (L-H-1 S04).

TA7a Time reallocation, cognitive workload shift. The Global Head of QC: "I could spend quite a lot of time organising workshops. Now I give three inputs to AI and everything's already been thought through. You spend much more time on the outcome, and making it as good as possible. You spend much more time on output quality" (L-H-1 S16). The IT Director at a footwear retailer described district-manager store audits structured by dictation-into-GPT: "before, after a full day, you're not mentally alert enough to remember everything from three store visits. Now, with this tool, note-taking happens on the fly" (M-S-1 S06).

TA10 Modular swap-in architecture. The telematics Innovation Director: "per-department AI workflow agents, one per department, each with its own data, prompts, and MCPs. They use the escalation bot, and the bot does the reasoning and asks questions to create a complete dossier. That dossier gets sent directly to legal" (M-K-1 S10). The sole-operator agency founder: "there's no perfect AI that can do everything 100% the way you want. You need to be able to switch: export the text as markdown, feed it into another AI, recontextualize with a different prompt to get the performance you need. No AI sits above all others" (S-I-1 S03).

TA11 Heuristic delegation rubrics. The Innovation Lead at the financial-market-infrastructure firm articulated the prediction-versus-judgment distinction: "AI is really good in making predictions but not super good or helpful in making judgments. There you should have a human in the loop at all times" (I-C-2 S06). An AI Adoption Lead at a tech firm articulated a two-criterion rule: "two types of tasks are good AI candidates. One: if the task is repetitive over time. The other: when the input context is small but the expected output is large" (I-O-1 S02).

TA12 Cost-anchored substitution. A CEO at an industrial group: "I was thinking of a recent issue about the tax status of French cross-border workers in Switzerland. AI didn't give a more precise answer than a lawyer would. It was just faster. I saved €5,000 by not calling a Swiss lawyer for that answer, and I had it in five minutes instead of waiting ten days" (M-L-1 S05).

A2. Calibration: Supplementary Quotes

C1 Verification cost erodes productivity gain. A Senior Legal Counsel: "AI is not yet 100% reliable. When you use the information it gives you, you have to scan for mistakes, that takes time. Any time we saw a mistake, we asked users to screenshot and send it to me so I could send it to LegalFly for them to improve the tool. That takes time away from actual work, you have to allocate time to spot errors. So users felt they were wasting their time" (L-M-4 S04). An AI Adoption Lead: "I've seen emails containing that kind of phrase ['If you need anything else, don't hesitate to ask me'], where you could tell the person copy-pasted the output without even checking. Always in internal deliverables, never external. But I have no doubt it will eventually happen externally" (I-O-1 S09).

C2 Multi-tool triangulation. A Deputy COO: "I often work in parallel, using Google as well. I re-verify things using Copilot, ChatGPT, or Google Gemini. When there's something I really want to verify, I go to Google" (L-M-1 S11).

C3 AI sycophancy. An AI Adoption Lead at a tech firm: "the way a query is formulated to AI will impact the response and the assertiveness of the response, and thus the degree of confidence, warranted or not, the person has in it. Some people say 'tell me why this is good' and AI immediately becomes a people-pleaser" (I-O-1 S13). The same lead reported a deliberate counter-pattern: "I'm hesitating on this topic, so maybe challenge me, interview me, ask me questions without knowing my answer, and after that interview give me your view on the question" (I-O-1 S14). The sole-operator agency founder: "today there's this fabulous, magical side to AI, and many people are dazzled. People are subjugated and stop thinking critically. 'If everyone talks about it, it must be true, so it must work.' The average person without technical, philosophical, or scientific background, without the habit of tracing where things come from, experiences AI the same way" (S-I-1 S11).

C4 Disposition over seniority. An Innovation Lead at the financial-market-infrastructure firm: "we work with some AI ambassadors. You see that those people are in most cases less tenured. On the other hand, we also have highly tenured people who are highly engaged and driving roadmap. Mostly younger people, less tenured people coming up with those solutions. Those are normally the people who also experiment a lot with it in their free time" (I-C-2 S13). A CEO at an industrial group framed his own late adoption: "personally, I haven't been using it for very long. I was somewhat resistant to the tool for ethical and familiarity reasons" (M-L-1 S01).

C5 Junior over-acceptance (residual P4). The telematics Innovation Director on instrumented data: "the lower the position of the person, the more easily they just accept what it says. The higher the position and background, the more they retry and control the answer. I see it in the data. My numbers from last month: all junior engineers averaged 24 seconds for code approval. Average seniors: seven minutes and more" (M-K-1 S12). A founding partner at a Luxembourg law firm observed a parallel pattern: a junior hire was much more fluent with AI tools than her older colleague but also more credulous, with calibration training flagged as necessary (L-G-1 S07 [¶ paraphrase]).

C6 Department-conditioned junior/senior inversion. The telematics Innovation Director: "our marketing manager is 45 and has one assistant who is 25. The 25-year-old does more work than the senior. For marketing, support, and project management, the younger people are faster. In engineering you still need critical thinking" (M-K-1 S18).

C8a Formal training as inflection. A Senior Legal Counsel: "three months online training failed. In-person fourth-month training was the inflection point for adoption" (L-M-4 S03 [paraphrased]). The Global Head of QC described mandatory training as a hard gate on tool access: completion is required before account provisioning (L-H-1 S05).

C8b Prompt-craft as competence. A founding partner at a Luxembourg law firm framed prompt formulation as cognitive labour, not effort-saving: the work moves upstream into the instruction (L-G-1 S14 [¶ paraphrase]). A CEO at an industrial group attributed an inaccurate AI output to his own prompting skill rather than to the tool, even from a self-described novice position (M-L-1 S02).

C15 Foundational deskilling. The CEO of an industrial group observed junior peers defaulting to ChatGPT for a sentence they could write themselves (M-L-1 S08), and articulated the binding concern as a red line: "don't lose the hand" (M-L-1 S13). The Digital Transformation Manager described the population-level dynamic as effectively irreversible (I-Q-2 S29). The sole-operator agency founder on junior developers: "loss of algorithmic and methodological reasoning, inability to debug AI-produced code without global view" (S-I-1 S09).

C17 Forced-exposure calibration ritual. The telematics Innovation Director: "forced-AI-only blitz cycles every approximately three weeks. Deliberate exposure to both AI strength and failure modes" (M-K-1 S07). The IT Director at a footwear retailer reported a parallel pedagogy: "deliberate critical-thinking training via deepfake and hallucination demonstrations" (M-S-1 S15).

C18 Heavy-user-as-most-sceptical. A Commercial Manager Latam at an advisory firm: the heaviest user of AI is also the most sceptical and most-reviewing, with verification reframed as the billable value-add in consulting (S-L-2 S07). The editorial-agency founder: AI outputs

framed as "complementary additions to pre-existing thinking, not as replacement", with systematic re-reading as the default (S-D-1 S05).
C20 AI-as-onboarding metaphor. The Corporate BIM Manager: AI is "framed organisationally as supervised junior; output requires verification" (I-M-2 S08). The Head of Marketing at a software firm: AI framed as new-hire metaphor with broad knowledge, missing context, and an expected error curve (M-B-2 S09).

A3. Interface and Process Design: Supplementary Quotes

P1 Governance lags adoption. The Head of Marketing at a software firm: "tool-silo problem, each AI sees only its own data domain; cross-tool integration via MCP and A2A in progress" (M-B-2 S18).
P2 Data sovereignty. The telematics Innovation Director: "we built our own MCP layer in-house to keep data localised, evidence-able, and out of vendor leak surface" (M-K-1 S01). A finance contributor: "strategic pivot from Copilot to Mistral driven by European sovereignty concerns" (I-P-1 S17). The Head of Marketing at a software firm: "incoming, third-party violation, and outgoing, own IP into training set" (M-B-2 S02). A founding partner at a Luxembourg law firm: "professional confidentiality obligation drives enterprise-tier procurement" (L-G-1 S10 [¶ paraphrase]). The Global Head of QC: "internal tools described as scope-limited and data-bounded by design, sovereignty enforced at the tool level" (L-H-1 S10). The sole-operator agency founder: "open-source self-hosted as preferred sovereignty pattern" (S-I-1 S13).
P3 Embedded skills. An AI Adoption Lead at a tech firm: "validation infrastructure deliberately expanded to catch AI errors" (I-O-1 S15). The editorial-agency founder: "brand-distinctiveness erosion at industry scale; high-grade prompt-craft as the defensive competence" (S-D-1 S10). The Senior Legal Counsel: "playbooks default plus custom per business unit as the firm's primary AI-deployment configuration unit; Senior Legal Counsel as playbook-customisation owner" (L-M-4 S02).
P4 Mandatory HITL. The Global Head of QC: "codified mandatory human-in-the-loop for safety-critical investigations, motivated by observed AI errors and patient safety stakes" (L-H-1 S02). A GenAI Programme Manager at a telecoms group: "human-in-the-loop is a stated, scope-binding principle, including for future agentic systems" (I-Q-1 S09).
P6 Behavioural instrumentation. A GenAI Programme Manager: "per-user budget cap with cost and CO_2 visibility, used to nudge model selection and curb overuse" (I-Q-1 S19). The Digital Transformation Manager: "planned use of behavioural ratios (planning/implementation, bug rate) as proxy indicators of cognitive engagement" (I-Q-2 S33). The telematics Innovation Director: "AI policy as living document, monthly revisions; signed acknowledgement creates legal liability frame" (M-K-1 S04).
P7 Codified scaffolding. A GenAI Programme Manager: "codified governance, charter, pledge, dedicated AI Office reporting to ExCo" (I-Q-1 S10). The Head of Marketing at a software firm: "just-released firm-wide AI policy anchored in EU AI Act and country-specific transpositions" (M-B-2 S04). An HR Technology lead at a food retailer: "codified governance scaffolding, audits, guidelines, feedback duties, security monitoring" (M-A-1 S12).
P11 Recalibration / confidence-threshold gating. The telematics Innovation Director: "explicit confidence-threshold gate (85%+) baked into the MCP. AI declines low-confidence answers" (M-K-1 S25). A GenAI Programme Manager: "dual-front design, upstream scoping (RAG, instructions) plus user-controlled creativity dial" (I-Q-1 S14). The IT Director at the footwear retailer: "structured custom-GPTs with fixed contextual scaffolding as the firm's primary AI-deployment mechanism" (M-S-1 S05). The Head of Marketing at a software firm: "context-tailored agents as the next governance layer, upstream scoping over downstream warning" (M-B-2 S11).
P12 Governance-by-restriction / hard perimeter. A finance contributor: "external tools are blocked; only IT-Security-vetted internal tools are accessible" (I-P-1 S11). The Senior Legal Counsel: "built-in anonymisation as embedded privacy protection; no additional behavioural

guidelines beyond tool-level safeguards" (L-M-4 S12). The Global Head of QC: "internal tools scope-limited and data-bounded by design" (L-H-1 S10).

P15 Output structure shapes net productivity. The Global Head of QC: "interaction style and presentation surface (clarifying questions, formulation) directly govern trust and continued use" (L-H-1 S15). An HR Technology lead at a food retailer: "opaque AI output drives mistrust; explainability shapes adoption" (M-A-1 S06).

P17 Distinctive trust-building cluster (single firm). An HR Technology lead at a food retailer reported two unusual mechanisms together: a co-authored knowledge base with user-rating loops as a participatory trust mechanism (M-A-1 S08), and a deliberately "rough" in-house UX as a "data-stays-inside" trust signal, trading aesthetic polish for trust legitimation (M-A-1 S13). The same firm also reported a stakeholder-asymmetric persuasion pattern: the same tool persuasive at management layer (via dashboard) but not at operational layer (target users), with AI-first triage at the expert/lay boundary triggering adoption resistance (M-A-1 S07; S11).

P21 Champion network. A GenAI Programme Manager: "structured network of internal champions, with differentiated role mandates" (I-Q-1 S05). The telematics Innovation Director: "voluntary peer learning forum scaled from 5 to 140 attendees over time; no top-down OKR mandate" (M-K-1 S06).

A4. Junior Pipeline: Extended Evidence

The Junior Pipeline dimension was promoted at consolidation_002 after the pattern appeared in five or more firms across multiple industries with a coherent underlying mechanism. Table A.1 summarises the cumulative cross-firm evidence as of consolidation_004 (10 firms, eight industries).

Anchor	Sector	Substantive pattern
I-M-2 S15	Construction	Legal AI tool already reduces junior hiring needs in legal department
L-M-4 S09	Construction	"LegalFly approximates a junior associate; firm needs fewer juniors as a result"
M-K-1 S13	Telematics	Explicit hiring stop on juniors; observed entry-level skill atrophy ("I know how to prompt")
S-J-1 S03; S04	Software	13-to-3 headcount collapse; zero-junior hiring policy; senior-freelancer model
L-H-1 S20	Pharmaceutical	Predicted continuity problem: late-adopting seniors substitute AI for juniors → reduced hiring → expertise erosion
M-B-1 S03 [¶]	Software	Structural feedback loop framed as the principal short-termism risk of AI adoption (~10y horizon)
S-L-2 S01	Water-tech advisory	AI-driven outbound replaced dedicated meeting-booking hires
I-Q-2 S26	Telecoms	Workforce-pyramid disruption: structural threat to junior-executing offshoring model

L-G-1 S15 [¶]	Legal services	Billing-model shift from hourly to flat fee with no obvious revenue uplift
L-G-1 S11 [¶]	Legal services	Senior preference for AI over junior collaboration → mentoring collapse

Table A.1. Reduced-junior-hiring evidence across the corpus

Within-firm contradiction at one construction multinational (Firm M). The construction multinational Firm M contains three coded voices on Junior Pipeline that articulate different postures. I-M-2 S15 (Corporate BIM Manager) reports LegalFly already displacing 2 to 3 juniors in the legal department. L-M-4 S09 (Senior Legal Counsel in the same legal department) corroborates the displacement: LegalFly approximates a junior associate, so the firm needs fewer juniors. I-M-3 S15 (Director of Estimating, a different department in the same firm) articulates deliberate junior recruitment maintenance despite AI displacement potential, on succession-logic grounds. The same firm therefore contains both enacted substitution (legal) and enacted preservation (estimating). This is the strongest within-firm triangulation in the corpus and is the empirical basis for the Junior Pipeline finding's framing as a strategic choice rather than as a deterministic consequence of AI adoption.
Pre-AI declining trend (counter-framing). An Innovation Lead at a financial-market-infrastructure firm reframes the pattern more broadly: the firm's junior hiring had been declining prior to AI adoption, with the AI-attribution still under discussion rather than enacted as a zero-junior policy (I-C-2 S14). The Junior Pipeline finding therefore captures both AI-driven substitution and pre-existing workforce-architecture trends that AI is now accelerating.
Predicted rehiring cycle (forecast). The lead-generation startup founder forecast a second-order rehiring cycle: "quality collapse will force rehiring of seniors first, then eventually juniors", framed as a societal-versus-firm responsibility distinction (S-J-1 S05). The Risk Domain Manager at the financial-market-infrastructure firm articulated a related second-order resistance: early adopters realising they are most exposed, with the resulting backlash framed as "digging our own graves" (M-C-1 S06).

A5. Emergent Themes: Extended Evidence

E1 Human accompaniment as binding meta-condition (cumulative). The pattern appears in 14 statements across 7 firms. Beyond the anchor quotes in Chapter 9 §9.15, additional supporting evidence includes: I-Q-2 S35 (vertical-versus-horizontal structural tension as binding architectural problem of AI integration); I-R-1 S03 (WIIFM as precondition for adoption); I-Q-2 S36 (predicted structural collision between organic role evolution and rigid corporate structure).
E2 Pace-mismatch and temporality (cumulative). Beyond the anchors in Chapter 9 §9.16: I-Q-2 S09 (adoption pressure framed as competitive, market-level, not internal-managerial); M-F-1 S11 (organisational time-horizon dramatically out of phase with AI evolution rate in public sector); M-S-1 S16 (refusal to commit to forward-looking task allocation rules in light of pace-mismatch).
E4 Anxiety / declared-vs-actual dissonance. The psychosocial researcher: "perceived gains conditional on absence of pressure, current honeymoon period" (I-N-1 S15); "minority job-security anxiety; majority positive overall" (I-N-1 S03). The Digital Transformation Manager: "silent resistance, non-believers do not surface their resistance because of organisational and societal pressure" (I-Q-2 S11); "surface 'no time' excuse masks underlying helplessness in the face of AI" (I-Q-2 S30).
E8 Augmentation-versus-substitution rhetoric (extended). Table 9.1 in the main text maps five positions. Additional articulations in the corpus include I-C-2 S15 ("task-not-job as the operative substitution unit"; 40-50% task-automation forecast without full job elimination), I-

R-1 S15 (AI value reframed as augmentation, quality/depth, rather than time-saving), and I-Q-1 S04 (managed urgency framing: adoption sold as personal-survival, not just productivity). The five core positions remain the analytical anchor.
E11 Levelling and value retreat to advisory. The Luxembourg founding partner: "future legal role projected as legal-technical hybrid; human judgement preserved in advisory work" (`L-G-1` S18 [¶ paraphrase]). An AI Adoption Lead at a tech firm: "AI compresses organisational layers, pulls developers back toward client contact" (`I-O-2` S11). A former CTO at a software firm: "human accompaniment reframed as the surviving value proposition against AI substitution" (`I-O-2` S22).
Single-source observations. Two boundary observations are flagged in Chapter 9 §9.6 and warrant slightly fuller treatment here:

- **AI-driven attrition (`M-L-1` S03).** A graphic designer at the industrial group left the firm in part because of AI-related friction: their refusal to use AI tools combined with the firm's ecological-mission ethos created a values conflict the firm could not reconcile. This is the only documented case of AI-driven attrition in the corpus; further investigation is flagged in Chapter 10.
- **Inverted deskilling (`S-I-1` S10).** The sole-operator agency founder reported that his extreme AI fluency had degraded his ability to collaborate with humans on delegation and coordination tasks. This is the only documented case of AI fluency producing a human-collaboration deficit; the case is read as a boundary observation about the upper end of the AI-fluency spectrum, not a generalisable mechanism.

A6. Consolidation Milestone Log

The four consolidation milestones documented the iterative shaping of the analytical structure. The detailed logs are preserved in the project repository; what follows is a high-level summary of the decisions locked at each milestone.
Consolidation_001 (after L-M-1). First-batch baseline established. Initial three-lever framework confirmed as analytically tractable. No structural revisions. Flagged: weak coverage on Sales-as-primary-role and Junior-as-primary-role.
Consolidation_002 (after L-H-1, end of batch 2). Two structural revisions. First, "junior pipeline disruption" pattern, which had appeared in five or more firms across multiple industries with a coherent mechanism (late-adopting seniors substituting AI for juniors → reduced junior hiring → mentoring and knowledge-transfer collapse), was promoted from emergent theme to aggregate dimension. Second, "deskilling and cognitive offloading" was confirmed as a sub-theme inside the Calibration dimension rather than as a standalone aggregate dimension, because its mechanism operates through verification habit rather than through delegation choice.
Consolidation_003 (after M-S-1, end of batch 3). Two structural revisions. First, "pace-mismatch and temporality" was located as a sub-theme of Proposition 6 (Interface and Process Design), because its mechanism operates through governance design under tool-evolution pressure. Second, "human accompaniment" was named as the binding meta-condition for the Chapter 9 synthesis section, not as a fifth aggregate dimension, because it does not have a theoretical mechanism distinct from the four dimensions; it shapes whether each dimension's mechanism translates into practice. Proposition 4 was reformulated to reflect the dispositional moderator finding. Proposition 5 was reformulated to reflect the upstream-scoping mechanism.
Consolidation_004 (after L-M-4, end of batch 4 - final). No structural revisions; closure milestone. Theoretical saturation reached for the four aggregate dimensions: each new interview extends existing themes rather than introducing new structural patterns. Single-source emergent themes retained as illustrative observations. Probes from consolidation_003 yielded null replication. M-K-1 S12 remained the only quantitative behavioural metric in the corpus, flagged as a methodological limitation.

A7. Power-User Case Summaries

Three interviews sit at the upper bound of AI fluency in the corpus and are read as boundary observations:
S-I-1 Sole-operator digital agency. The founder operates a digital agency built entirely around AI orchestration, with no employees. He reports producing output equivalent to his prior 16-employee firm (S-I-1 S01), uses a deliberate multi-AI orchestration architecture (S-I-1 S03), and has codified a post-incident verification procedure after an over-reliance incident sent a wrong letter to a government ministry (S-I-1 S04; S05). The boundary observation is inverted deskilling: extreme AI fluency degrading his own human-collaboration capacity (S-I-1 S10).
S-J-1 Lead-generation software startup. The founder runs a firm whose product is itself an AI system: semantic interpretation of company descriptions is the technical engine of the business (S-J-1 S01). The firm enacted a 13-to-3 headcount collapse and a zero-junior hiring policy in the past 18 months (S-J-1 S03; S04). The firm-identity framing is "judgment company": distrust by default, with human added value located in judgement rather than in execution (S-J-1 S02; S12). The founder forecasts a 1 to 3-year window of competitive advantage from early AI adoption (S-J-1 S10).
M-K-1 Telematics multinational. The Innovation Director leads parallel AI deployments across the firm. The firm has instrumented AI-generated pull requests, producing the only quantitative behavioural metric in the corpus: 24-second junior code-approval times versus 7-plus-minute senior times (M-K-1 S12). The firm has built its own MCP layer in-house with an 85% confidence-threshold gate (M-K-1 S01; S25). The Innovation Director also documented a forced-AI-only blitz cycle every three weeks as a deliberate calibration ritual (M-K-1 S07). The case sits between the median behaviour and the sole-operator boundary: a corporate context with power-user fluency at the deployment-lead level.
These three cases are deliberately retained to surface patterns that emerge only at the extreme of AI integration. Where they appear in the findings chapter, they are presented as boundary indicators rather than as median behaviour.

A8. Firm-Level Patterns: Managing the Paradox Well vs Under-Performing

Chapter 9 §9.17 argues that firms managing the paradox well share three features (the four dimensions are coupled, accompaniment is invested in, pace-mismatch is treated as a design constraint), and that firms under-performing share the opposite. Table A.2 details the specific combinations observed in the corpus.

Company	Task Allocation	Calibration	Interface and Process Design	Junior Pipeline	Accompaniment	Pace-mismatch response
Managing well: Telecoms multinational (I-Q-1 + I-Q-2)	Portfolio governance with quick-wins / structural bets (I-Q-1 S08); RAG-scoped HR chatbot to first-line tier (I-Q-1 S03)	Mandatory critical-thinking training with accountability framing (I-Q-1 S17, S18); recalibration checkpoints between AI stages (I-Q-2 S05); forced exposure to AI failure modes	In-house sovereign tool with one-way data flow (I-Q-1 S01); behavioral instrumentation via per-user budget cap with cost/CO2 visibility (I-Q-1 S19); written charter + AI Office reporting to ExCo (I-Q-1 S10); HITL stated as a priori design principle (I-Q-1 S09); structured champion network with differentiated mandates (I-Q-1 S05)	Workforce-pyramid disruption acknowledged; offshore junior model flagged as structurally threatened (I-Q-2 S26); not yet enacted as zero-junior policy	2-year embedded resource produced movement, removed produced full reversion (I-Q-2 S31; S32); chronic under-investment in accompaniment named as the most underestimated obstacle (I-Q-1 S25)	Sandbox-then-industrialise (I-Q-2 S13); adoption framed as moving target, each capability wave restarts diffusion (I-Q-1 S07)
Managing well: Footwear retailer (SME) (M-S-1)	Explicit four-criterion delegation rubric (repetitive, low-value, time-saving, disliked) (M-S-1 S11); structured custom-GPTs with fixed contextual scaffolding as primary deployment unit (M-S-1 S05)	Forced-exposure calibration ritual via deepfake / hallucination demonstrations on ongoing cadence (M-S-1 S15); single-failure-reversion documented as adoption risk (M-S-1 S09)	Modular swap-in architecture across LLM and image-gen backends, designed for moving-target tool environment (M-S-1 S10); structured GPT-level governance (S05)	N/A at this scale (30-store SME)	Demystification -led adoption with facilitator (M-S-1 S03); change management as per-person problem invariant to firm size (M-S-1 S17; S19); honest-uncertainty stance about future substitution (M-S-1 S13)	Thursday-to-Monday tool deployment cycle vs ~3 months at large firm baseline (M-S-1 S20); modular architecture as explicit hedge against tool obsolescence (S10); refusal to commit to forward-looking allocation rules (S16)
Managing well (boundary): Sole-operator agency (S-I-1)	Deliberate multi-AI orchestration with task-specific specialisation (S-I-1 S03); modular swap-in across LLMs and image-gen	Post-incident codified verification procedure (AI-prep + expert-final-verification scoped by user expertise) (S-I-1 S05); cross-AI verification as comprehension test (S08); heavy-user-as-most-sceptical pattern enacted	Open-source self-hosted as preferred sovereignty stance (S-I-1 S13); process-visibility via reading the AI's chain-of-thought (S07); decentralised governance with traceability proposed as alternative to prohibition (S12)	N/A (sole operator). Reports foundational deskilling in junior developers observed externally (S-I-1 S09)	N/A at this scale	Modular tool stack as primary architectural hedge against obsolescence (S-I-1 S03)

Boundary observation flagged						Inverted deskilling: extreme AI fluency erodes the operator's ability to collaborate with humans (S-I-1 S10). Single-source; reported as a warning that the upper-bound power-user posture has its own failure mode.
Under-performing: Default posture (composite)	Trial-and-error baseline; deliberately tool-agnostic posture (L-M-1 S07); extreme departmental variance even where a formal sovereign tool exists (I-Q-2 S12); shadow adoption pre-dating governance (M-B-2 S15)	Verification cost erodes productivity gain (C1 cluster, 14 statements / 8 firms); copy-paste of unedited AI output documented (I-O-1 S09); large-scale erosion of factual accuracy under heavy AI authorship (60% AI-written, 34% factual) (M-K-1 S22)	Governance lags adoption (P1 cluster, 13 statements / 9 firms); systemic gap across rules, training, and managerial support (I-N-1 S09); shadow AI undisclosed accumulates risk (I-N-1 S18); disclaimers acknowledged as legal protection only (I-Q-1 S13)	Reduced junior hiring enacted in present tense (I-M-2 S15; L-M-4 S09; M-K-1 S13; S-J-1 S03); substitution by default rather than explicit choice	Mandatory training infrastructure exists for other topics but unused for AI (I-P-1 S01); 80% non-adoption rate even after launch (M-A-1 S04); tool-velocity overruns absorption capacity (M-A-1 S05)	Waiting for tools to stabilise before formalising governance; approval cycles outrun by capability shifts (M-B-2 S20; I-O-1 S21); accumulating shadow use as a consequence

Table A.2. Detailed combinations of the four dimensions, accompaniment, and pace-mismatch response across firm-level patterns

The contrast the table makes visible is structural rather than evaluative. Firms classified as managing well in this corpus are not necessarily achieving net positive productivity gains from AI relative to a no-AI baseline; what they are doing is investing coherently across the four dimensions and the two cross-cutting conditions, with a posture (augmentation, substitution, or honest middle ground) that has been made explicit. Firms classified as under-performing are not necessarily failing as businesses; what they are doing is accumulating coordination costs across the four dimensions that will materialise as quality, retention, and continuity problems on a horizon longer than the current AI rollout cycle.

The three "managing well" cases operate at radically different scales (a multinational telecoms group, a 30-store footwear retailer SME, and a sole-operator agency). The cross-scale convergence on the same structural features (coupled dimensions, accompaniment investment, pace-mismatch as design constraint) is the empirical basis for the Chapter 9 §9.10 claim that the framework operates as a managerial diagnostic rather than as a scale-bound prescription.

A9. Methodology Detail

- A9.1 Full sample composition by code

Code	Sector	Substantive role	Seniority
I-M-2	Construction multinational	Corporate BIM Manager	Senior
I-M-3	Construction multinational	Director of Estimating	Senior
L-M-1	Construction multinational	Deputy COO	Senior
L-M-4	Construction multinational	Senior Legal Counsel	Senior
I-O-1	Software firm	AI Adoption Lead	Senior
I-O-2	Software firm	Former CTO/COO	Senior
I-O-3	Software firm	Software engineer	IC
M-B-1	Software firm	Managing Director	Senior
M-B-2	Software firm	Head of Marketing / AI strategy	Senior
M-K-1	Telematics multinational	Innovation/IT Director	Senior
S-J-1	Lead-generation startup	Founder / CEO	Senior
I-Q-1	Telecoms group	GenAI Programme Manager	Senior
I-Q-2	Telecoms group	Digital Transformation Manager	Senior
I-P-1	Bank	Finance contributor	Mid-senior
M-C-1	Financial-market-infrastructure firm	Risk Domain Manager	Senior
L-H-1	Pharmaceutical multinational	Global Head QC	Senior
L-G-1	Luxembourg law firm	Founding partner	Senior
M-F-1	Regional bus operator (public sector)	Head of Customer Care	Senior
I-R-1	Insurance / prevention	HR Change-Mgmt Director	Senior
L-E-1	Tech research / policy	Research-project manager	Junior
I-N-1	Tech research / policy	Psychosocial researcher	Senior
M-A-1	Food retailer	HR Technology lead	Senior
M-S-1	Footwear retailer	IT Director + AI Facilitator	Senior
M-L-1	Industrial group	CEO / Group Chairman	Senior
S-D-1	Editorial agency	Founder / Editorial Director	Senior
S-I-1	Sole-operator agency	Founder / Project Architect	Senior
S-L-2	Water-tech advisory	Commercial Manager Latam	Senior
I-C-2	Financial-market-infrastructure firm	Innovation Lead P&O	Senior

Total: 28 interviews across ~22 firms and 14 industries. Within-firm triangulation is preserved at the firm-letter level (for example, I-M-2, I-M-3, L-M-1, L-M-4 all sit within the same construction multinational).

- A9.2 Coding spreadsheet column structure

Coding is performed in a structured spreadsheet (master_coding.csv) with one row per text segment. Columns: interview code; recruitment-grid prefix; substantive role; first-order code; second-order theme; aggregate dimension; supporting proposition (P1-P6 or none); confidence level (high / medium / low); cross-department note (when the segment describes behaviour in a department different from the interviewee's own label); analytical notes. The structure enables systematic filtering and comparison across interviews and feeds the iterative consolidation milestones documented in §A6.

- A9.3 Recruitment-grid prefix as artefact

Each interview is identified by a code of the form X-A-N, where X is the recruitment-grid prefix (M for Marketing, L for Legal, S for Sales, I for Insights), A is a firm-letter identifier, and N is the within-firm interviewee sequence. The prefix reflects the initial sampling category at the moment of contact, not the substantive role at the time of interview. As recruitment evolved toward deployment-side voices, the substantive role frequently diverged from the prefix: in the final sample, 9 of 17 L-prefix and M-prefix files show a substantive role that does not match the prefix label. Each interview's substantive role is recorded in its file metadata (see Table above). The prefix is preserved in the interview code for audit-trail traceability.

- A9.4 Source caveats

Two interviews (L-G-1 and M-B-1) are paraphrase-only. For L-G-1, the audio transcript was lost for the substantive portion; only the closing fragment survives. For M-B-1, the interview was conducted under conditions that did not permit a transcript. Statements drawn from these interviews are based on translated researcher notes and are marked [¶ paraphrase] at the point of use in Chapter 9; they are not used for verbatim quotation.

A10. Interview guide

Bloc 0 : Introduction (2 min)
Bonjour et merci d'avoir accepté cet entretien. Je m'appelle Maxime Kohnen, je suis étudiant en master à l'UCLouvain. Mon mémoire porte sur la façon dont les professionnels utilisent l'IA dans leur travail au quotidien. Il n'y a pas de bonnes ou de mauvaises réponses, ce qui m'intéresse c'est votre expérience concrète. L'entretien est anonyme et dure environ 45 minutes. Avec votre accord, je vais enregistrer notre conversation pour la retranscrire ensuite. Est-ce que c'est ok pour vous ?
Bloc 1 : Contexte et profil (8 min)
Objectif : comprendre le rôle, l'ancienneté et le niveau d'usage de l'IA. Permet aussi de mettre la personne à l'aise.
Q1. Pouvez-vous me décrire votre rôle actuel et vos responsabilités principales ?
Q2. Depuis combien de temps occupez-vous ce poste, et depuis combien de temps travaillez-vous dans ce domaine en général ?
Q3. Quels outils d'IA utilisez-vous dans votre travail ? À quelle fréquence ?
→ Relance : Vous les utilisez tous les jours, quelques fois par semaine, de temps en temps ?
Q4. Comment avez-vous commencé à les utiliser ? C'était une initiative personnelle, une décision de l'équipe, ou une directive de l'entreprise ?
Bloc 2 : Répartition des tâches entre humain et IA (~10 min)
Objectif : explorer comment la personne décide ce qu'elle délègue à l'IA et ce qu'elle garde.
Proposition 1 : Les professionnels dans des domaines où les capacités de l'IA varient fortement d'une tâche à l'autre éprouvent plus de difficultés à décider quoi déléguer.

Proposition 2 : En l’absence de directives organisationnelles, la délégation repose sur le tâtonnement individuel, ce qui crée des pratiques incohérentes au sein d’une même équipe.
Q5. Concrètement, pour quelles tâches utilisez-vous l’IA, et pour lesquelles préférez-vous ne pas l’utiliser ?
→ Relance : Qu’est-ce qui fait qu’une tâche vous semble « adaptable à l’IA » ou pas ?
Q6. Est-ce que la répartition entre ce que vous faites vous-même et ce que vous confiez à l’IA a évolué depuis que vous avez commencé à l’utiliser ?
→ Relance : Pouvez-vous me donner un exemple concret de quelque chose que vous déléguez maintenant mais que vous ne déléguiez pas au début, ou l’ inverse ?
Q7. Est-ce qu’il vous est arrivé de confier une tâche à l’IA et de réaliser ensuite que le résultat n’était pas fiable ? Que s’est-il passé ?
Q8. Dans votre équipe, est-ce que tout le monde utilise l’IA de la même manière, ou est-ce que chacun a un peu sa propre approche ?
→ Relance : Est-ce qu’il y a des règles ou des guidelines dans votre organisation sur ce qu’on peut ou ne peut pas faire avec l’ IA ?
Bloc 3 : Évaluation et confiance envers les résultats de l’IA (~10 min)
Objectif : explorer comment la personne juge la qualité des outputs IA, et si cette capacité diffère selon l’expérience.
Proposition 3 : Les professionnels qui jugent correctement quand l’IA est fiable et quand elle ne l’est pas en tirent davantage de bénéfices que ceux qui ont une expertise supérieure mais une mauvaise calibration.
Proposition 4 : Les juniors sont plus susceptibles de sur-confiance envers l’IA, mais cette asymétrie est modérée par les mécanismes organisationnels (processus de relecture, feedback, formation).
Q9. Quand vous recevez un résultat généré par l’IA, comment décidez-vous si vous pouvez lui faire confiance ou pas ?
→ Relance : C’est un réflexe automatique ou vous vérifiez systématiquement ?
Q10. Est-ce qu’il vous est arrivé de faire confiance à un résultat de l’IA qui s’est révélé incorrect ? Comment l’avez-vous découvert ?
Q11. Et à l’inverse, avez-vous déjà rejeté un résultat de l’IA alors qu’il était en fait correct ?
Q12. Est-ce que votre capacité à évaluer les résultats de l’IA a changé avec le temps ? Comment ?
Q13. Si vous pensez à un collègue plus junior ou plus senior que vous, est-ce que vous observez des différences dans la façon dont ils utilisent ou font confiance à l’IA ?
→ Relance : Avez-vous déjà vu quelqu’un dans votre équipe accepter un résultat de l’ IA sans le vérifier, et ça a posé problème ?
Bloc 4 : Outils, workflows et organisation (~10 min)
Objectif : comprendre comment l’intégration de l’IA dans les processus et les outils influence le comportement de vérification.
Proposition 5 : Les outils qui présentent l’output IA comme une réponse finale encouragent la sur-confiance, tandis que ceux qui le présentent comme un brouillon encouragent la vérification.
Proposition 6 : Les organisations qui intègrent l’IA dans des workflows structurés avec des points de contrôle humain ont moins de problèmes de qualité que celles où l’IA est utilisée de manière informelle.
Q14. Comment l’IA est-elle intégrée dans votre flux de travail concrètement ? C’est un outil à part que vous ouvrez séparément, ou c’est intégré dans vos outils habituels ?
Q15. Est-ce que la façon dont l’outil présente ses résultats influence votre manière de les utiliser ? Par exemple, est-ce que vous traitez différemment un texte généré complètement par l’IA par rapport à une suggestion que vous devez compléter ?
Q16. Est-ce que votre organisation a mis en place des processus spécifiques autour de l’utilisation de l’IA ? Par exemple, des étapes de validation, des relectures, des restrictions ?

→ Relance : Si non : est-ce que vous pensez que ce genre de cadre serait utile, ou est-ce que la flexibilité actuelle vous convient ?
Q17. Est-ce que vous avez observé des changements dans votre façon de travailler depuis que l'IA est arrivée ? Des compétences que vous exercez moins, ou des habitudes qui ont changé ?
→ Relance : Est-ce qu'il y a des choses que vous faisiez manuellement avant et que vous avez arrêté de faire ? Est-ce que ça vous inquiète ou pas du tout ?
Bloc 5 : Clôture (~5 min)
Q18. Si vous deviez identifier le plus grand défi dans la collaboration avec l'IA dans votre métier, ce serait quoi ?
Q19. Est-ce qu'il y a un aspect de votre expérience avec l'IA dont on n'a pas parlé et qui vous semble important ?
Merci beaucoup pour votre temps et vos réponses. C'était très enrichissant. Si vous avez des questions par la suite, n'hésitez pas à me contacter.