big data and cognitive computing

MDPI

Review

A Review of Human-AI Complementarities Across Multiple Dimensions of Organisational Complexity

Ganesh Sankaran [1,*], Marco A. Palomino [2,*] and Guido Siestrup [3]

[1] ZweiThreads Software LLP, Bengaluru 560048, India
[2] School of Natural and Computing Sciences, University of Aberdeen, Aberdeen AB24 3UE, UK
[3] Faculty of Business Information Systems, Hochschule Furtwangen University, 78120 Furtwangen, Germany; guido.siestrup@hs-furtwangen.de
* Correspondence: ganesh.sankaran@zweithreads.com (G.S.); marco.palomino@abdn.ac.uk (M.A.P.)

Abstract

The growing capabilities of artificial intelligence (AI) have not translated straightforwardly into organisational value. A persistent disconnect—the "last-mile problem"—arises from structural gaps between idealised AI tasks and real-world organisational contexts. Synthesising insights from organisational theory, cognitive science, and computer science, we have developed a five-dimensional diagnostic framework that maps the challenges of human-AI collaboration across Integration, Representation, Scale, Temporality, and Adequacy gaps. These gaps illuminate how socio-technical complexity, contextualised problem representations, interdependencies among agents, dynamic environments, and limitations in current AI reasoning collectively constrain full automation and demand human judgement. By reviewing the historical evolution of AI—from symbolic systems to machine learning, generative models, and emerging agentic approaches—we show that augmentation remains the dominant and most viable mode of use in complex environments. An illustrative system-dynamics example demonstrates how improvements in algorithmic performance do not automatically yield proportional system-level gains. Overall, our framework provides researchers with a conceptual lens and practitioners with a diagnostic tool for assessing complementarities and informing the design of human-AI collaborations. The framework is offered as a conceptual synthesis and diagnostic instrument rather than an empirically validated model.

Keywords: human-AI collaboration; organisational decision-making; complementarity; system dynamics; general-purpose technology; augmentation

Check for updates

Academic Editor: Shuping He

Received: 27 March 2026
Revised: 9 July 2026
Accepted: 14 July 2026
Published: 11 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The discourse surrounding AI and its role in organisational decision-making is often polarised between substitution-focused techno-optimism and dystopian fear. A more recent consensus, however, emphasises human–AI complementarity, recognising augmentation as the more realistic frame for the foreseeable future. This shift brings renewed attention to the challenge of evaluating outcomes in collaborative contexts, yet much of the literature still offers abstract principles without addressing the well-known 'last-mile problem'—the disconnect between AI's demonstrable capabilities and the difficulty of realising value in specific organisational settings [1,2]. This gap emerges when idealised capabilities encounter the complexity of real organisational systems. What remains missing is a systematic diagnostic framework for analysing where such gaps arise, alongside quantitative methods for evaluating proposed collaborative configurations within particular contexts.

This article is an integrative, narrative review. Rather than pursuing the exhaustive coverage of a systematic review, it synthesises conceptual and empirical contributions across organisational theory, cognitive science, and computer science to construct a diagnostic framework. Sources were identified by purposive, snowball search, starting from an anchor work that frames AI as a normal technology [2] and following citation chains across these literatures, with conceptual relevance to the last-mile problem taking priority over comprehensive enumeration. The scope is deliberately diagnostic: the review identifies the dimensions along which human-AI gaps arise, whereas quantitative evaluation of specific configurations falls outside its scope and is treated elsewhere [3].

Three questions guide the review. **RQ1:** Along which dimensions do gaps arise between an idealised AI task and a real-world organisational problem, and how does the strength of these gaps determine where human judgement remains necessary within collaborative human-AI configurations? **RQ2:** How have successive waves of AI, from symbolic to machine learning, generative, and agentic approaches, reshaped rather than removed the locus of human judgement in organisational decision-making? **RQ3:** Why do improvements in algorithmic performance not translate straightforwardly into proportional organisational value, and what does this imply for how human-AI configurations should be evaluated?

The contribution is integrative rather than foundational. Its novelty lies not in a new theory of human-AI collaboration but in synthesising previously separate literatures into a single diagnostic structure. Three features distinguish the framework from existing accounts. First, it moves beyond the binary automation-augmentation framing [4] towards a sliding scale of agency calibrated to gap strength. Second, whereas prior design frameworks address complementarity at a single level, whether task-level skill matching [5], team coordination [6,7], or organisational enablers [8], the five gaps span these levels within one structure. Third, each gap is linked to the decision phase at which human judgement is most salient, yielding a diagnostic rather than a prescriptive instrument. Existing sociotechnical and adoption accounts establish that context matters; this framework specifies along which dimensions, and with what consequence for the division of labour between humans and AI.

The framework identifies the fundamental, cumulative, and interdependent gaps between an idealised AI task and a real-world organisational problem. By synthesising insights from organisational theory, cognitive science, and computer science, we show that human judgement plays an essential role in bridging these gaps within collaborative human-AI configurations. Our framework offers researchers a theoretical lens and provides practitioners with a diagnostic tool for assessing gap strength across multiple dimensions. It is a conceptual synthesis and diagnostic instrument rather than an empirically validated model, and it has not been validated across multiple organisational contexts. However, translating these diagnostic insights into effective organisational practice requires quantitative evaluation methods tailored to specific cases, which fall outside the scope of this review.

The remainder of the review is structured as follows. First, we briefly trace the evolution of AI to frame the shifting nature of the challenge to human judgement, and we examine the modern productivity paradox to motivate the need for a deeper, systemic analysis. We then introduce and elaborate on the framework, detailing the gaps—Integration, Representation, Scale, Temporality, and Adequacy—that define the contemporary landscape of human–AI collaboration. It should be emphasised that the framework is diagnostic rather than prescriptive: it identifies dimensions along which gaps between idealised AI capabilities and organisational realities can arise. The strength of these gaps—shaped by organisational context, problem characteristics, and available complementary resources—

determines the degree of human involvement required in collaborative human-AI configurations. Where gaps are minimal, greater AI autonomy may be appropriate; where they are substantial, intensive human participation becomes necessary. This perspective rejects simplistic automation-augmentation dichotomies in favour of a sliding scale of human-AI agency, where the appropriate balance is determined by a situated assessment of gap strength across these dimensions.

2. Background

To situate the discussion, we now turn to the evolving nature of the AI challenge to human judgement. Despite successive technological waves, the productivity paradox persists, and the pressures on judgement have only intensified with newer forms of AI. To understand this trajectory, we begin with the foundations of symbolic AI before tracing how subsequent developments have reshaped the challenge.

2.1. Symbolic AI

A central goal of AI since its inception has been to create machines with human-like problem-solving skills and a capacity for self-improvement [9]. The pioneers of the field pursued this goal through a rationalist notion of intelligence, which posits rational deliberation as its primary ingredient [10]. This stands in contrast to the empiricist view, which holds that intelligence emerges from learning and perception [11].

The idea of thought as computation, a natural corollary of the rationalist view, was not novel; Leibniz, for instance, had analogised thought to arithmetic [12]. However, it was later developments in formal logic—most notably Frege's *Begriffsschrift* [13], which introduced modern quantification—building on algebraic logic (Boole) and work on relations and quantifiers (Peirce)—that provided the necessary conceptual tools to implement this ambitious vision on a digital computer [14].

This first wave of AI was built on the notion of digital computers as physical symbol systems [9,15,16]. The paradigm is underpinned by the knowledge-representation hypothesis [10,14], which posits that a system's intelligent behaviour results from manipulating a formal knowledge base [12]. The system's inferences are seen as manifestations of this internal knowledge, aligning with definitions of intelligence as goal-directed action [17]. While this approach yielded notable successes—such as the MYCIN medical diagnosis system and IBM's Deep Blue, which defeated Garry Kasparov in 1997 [14]—these examples also highlighted its fundamental constraints.

The primary constraint was the knowledge-acquisition bottleneck: the knowledge base had to be manually encoded by humans. This process was not only difficult to scale but also limited the system's capabilities to domains governed by articulable or explicit knowledge [18,19]. This limitation brought into sharp focus the challenge of tacit knowledge—the vast repository of intuitive understanding that humans possess but cannot easily articulate, famously captured in Polanyi's phrase, "we know more than we can tell" [20,21]. Minsky later framed this from an AI perspective, noting that 'easy things are hard [for AI]' [9].

In an organisational context, this limitation maps directly onto the distinction between routine and non-routine tasks [22]. Symbolic AI thus found its niche in automating a class of routine tasks that, while governed by explicit knowledge, required a degree of inference or problem-solving beyond simple computation. Non-routine tasks, however, depend on judgement—a form of tacit knowledge that Mintzberg called 'the word that the verbal intellect has given to the thought processes that it cannot articulate' [23].

2.2. The Machine Learning Paradigm

The symbolic approach, with its reliance on explicit rules, can be understood as an attempt to model the slow, conscious, and effortful cognition that Kahneman categorises as System 2 [24,25]. Proponents of this tradition mistakenly assumed that the fast, intuitive 'System 1' thinking underlying human proficiency in perceptual tasks—such as vision, speech, and natural language—would be simple to engineer [10]. This miscalculation revealed the paradigm's core limitation: its inability to transcend Polanyi's paradox by automating tasks for which the rules are subconscious or tacit. The resulting stagnation necessitated a new approach, which arrived with the ascendancy of machine learning (ML).

The ML paradigm inverts its symbolic predecessor; rather than beginning with hand-crafted knowledge, it is philosophically aligned with a Lockean tabula rasa [26], learning rules by finding patterns in data [27]. The subfield of deep learning—using artificial neural network architectures inspired by the brain [28]—became the paradigm's 'intellectual quasi-monopoly' [26]. Its success is largely the result of the growth in data and computational power [29], alongside scaling-law results [30].

In an organisational context, ML challenges the standard partnership by automating tasks whose instructions cannot be articulated but for which performance data exist [31]. By economising prediction [32], ML models have achieved superhuman performance in complex but convergent tasks with clear goals and stable rules—what Hogarth terms 'kind' learning environments [33].

However, the human role remains critical and shifts from rule specification to system design. The performance of ML models depends on human choices about architecture and data—inductive biases [34]. These models exhibit only local generalisation, failing outside their specific training domain [26], particularly in 'wicked' environments where feedback is misleading or the problem is novel [33]. Thus, the predictive paradigm excels at optimising convergent decisions but is ill-suited for divergent tasks that require innovation and synthesis, where alternatives are not given a priori [35,36]. This limitation—the dependence on labelled training data and the bottleneck it creates for scaling—motivated the development of generative approaches that could learn from unlabelled data.

2.3. Generative and Agentic AI

While deep-learning models trained under the supervised paradigm are optimised to produce point estimates—such as a numerical forecast or a class label—a transformer-based large language model learns the conditional probability distribution over possible continuations of a given context. A significant empirical finding is that although performance continues to grow with data, parameters, and compute [30], data wield a disproportionately large impact [37]. Supervised learning's dependence on labelled data is a bottleneck to scaling; the generative AI (GenAI) approach overcomes this by using next-token prediction as a self-supervised training task.

The most powerful large language models (LLMs)—those that generate text—best demonstrate the scaling potential of this approach, as eliminating the need for labelling allows much of the publicly available text on the Internet to be used for training [38]. Given the breadth of capabilities acquired this way, the term 'foundation models' denotes that they can be fine-tuned to perform a variety of downstream tasks. LLMs are relevant from a decision-making perspective because they exhibit state-of-the-art performance on benchmark tasks requiring language understanding and manipulation [36,39].

Although the underlying architecture of LLMs is a deep neural network, the key innovation in the transformer is self-attention, which enables contextually rich token representations by paying selective attention to other parts of the input sequence [40]. The combination of scale and attention yields a structured latent space that captures

statistical regularities informed by context [27,41], and this combination is responsible for the surprising ability of LLMs—trained on a narrow next-token prediction task—to demonstrate useful skills across domains, commensurate with the breadth of training data. Some researchers interpret this as emergence, where optimising for coherent text (e.g., lower perplexity) generates capabilities across a broad swathe of domains [42]. Yet Krakauer et al. caution that not all such cases are emergence in the strict scientific sense; they argue that scientific emergence requires a coarse-graining that yields a new explanatory level [43].

Despite scepticism about whether scaling improves the quality of representation in a veridical sense [44,45], optimising for fluency and context coherence yields skills relevant for decision-making not previously seen in AI. In particular, an LLM generates a probability distribution over continuations. By sampling from this distribution, LLMs can generate novel content and support decision-making tasks that utilise nascent abilities for abstraction, analogy, and reasoning across a wide variety of contexts.

Decision-making typically involves two stages: a divergent stage where options are generated, and a convergent stage where a decision is finalised. While the convergent stage emphasises optimisation and efficiency, the divergent stage requires creativity and problem-solving skills once presumed solely human—an assumption LLMs have started to challenge [36]. Concretely, novel problems require drawing on analogies that rely on finding abstractions to highlight similarities between situations [46]. Movement along an LLM's latent space—interpolating along lower-dimensional manifolds—can have effects akin to conceptual blending and analogy-making [39,47].

A further challenge to assumptions about human roles has been the augmentation of LLM capabilities with tool use, extending their reach from decisions to actions [48]. An LLM equipped this way can independently perform a well-defined task but is not yet suitable for complex workflows. A newer approach, agentic AI, aims at more realistic problems where multiple agents collaborate on task decomposition, communication, and coordination, adapt to feedback, and access persistent memory [49].

There is undeniably a rapid expansion in AI capabilities—most recently due to generative and agentic approaches—that poses a robust challenge to human judgement. Despite this, several recent studies highlight that the prevalent mode of AI use is augmentative rather than fully automative [50,51]. Mapping worker preferences alongside expert assessments of technological capabilities [52] also reveals a strong preference for augmentation, indicating a likely trajectory for new AI capabilities. This empirical reality aligns with the sliding scale perspective introduced earlier: as AI capabilities expand, complementarities shift and the specific loci of human contribution evolve, but gap strength across multiple dimensions continues to necessitate collaborative configurations rather than full automation in most complex organisational contexts.

The new AI capabilities, especially LLMs that can undertake divergent decision-making tasks, are creating new complementarities at the human–AI intersection (e.g., [53–55]). These emerging skills are often unrelated—or even negatively correlated—with traditional skills associated with the tasks [56]. For example, a broader view of decision-making that emphasises task interdependencies and instance-level variability highlights metacognitive skills important for delegation [57]. Thus, a picture emerges of redefinition and readjustment regarding the role of judgement rather than dispensing with it. This strengthens the consensus calling for a focus on the expanding scope of tasks humans and AI can accomplish together [41,58,59] instead of a narrower substitutive focus. Such delegation for novel task configurations requires a nuanced assessment of complementary strengths [5].

From a more technical perspective, LLMs that learn from language are focused on codifiable tasks, which Acemoğlu calls 'easy tasks' [58]. In contrast, 'hard tasks' relying

on tacit knowledge gained from experience require human judgement to varying degrees. Observed friction in the adoption of AI into workflows [51,60,61] serves as indirect evidence for the importance of tacit know-how beyond know-what captured in prediction targets [62]. Finally, alongside new opportunities, new forms of AI also present risks and amplify existing risks of ML, which learns correlations from large datasets but may not identify underlying causal mechanisms.

Del Giudice [45] argues that a model not based on causal knowledge can, in theory, improve prediction accuracy while degrading the accuracy of its representation of the underlying phenomenon. This evolution reveals that the challenge is not whether human involvement remains necessary—current evidence strongly suggests it does in most organisational contexts—but rather understanding the systematic sources of this necessity. The next section introduces a five-dimensional framework that identifies the specific types of gaps between idealised AI capabilities and organisational realities, enabling a more nuanced and context-sensitive assessment of gap strength and the appropriate intensity of human involvement in collaborative configurations.

2.4. The Modern Productivity Paradox

AI's promise as a general-purpose technology (GPT) is undeniable, yet scepticism persists regarding its transformative economic impact. Despite AI's rapid progress, labour-productivity growth has remained stagnant in many mature economies since slowing in the mid-2000s [19]. In the US, for instance, productivity growth fell from 1.73% between 1998 and 2007 to 0.53% between 2010 and 2019 [63]. This gap is reminiscent of the paradox observed during the ICT revolution, captured by Solow's quip that 'you can see the computer age everywhere but in the productivity statistics' [19].

While several explanations exist, the most influential factor appears to be implementation lags—the slow diffusion of benefits throughout the economy [19]. The promise of AI and the reality of slow productivity growth are not necessarily contradictory when viewed through this lens. Like previous GPTs, AI's diffusion follows an S-curve: a slow start due to costs and learning curves, followed by acceleration as competition and capabilities improve. This dynamic creates a significant performance gap between a few front-runner firms and a long tail of laggards that are slow to adopt or absorb the technologies.

2.5. From Implementation Lags to Systemic Gaps

The theory of implementation lags suggests that the full potential of a GPT like AI is unlocked through significant complementary changes [19,63,64]. Realising value requires more than deploying a new technology; it demands concurrent evolution across policies, infrastructure, co-innovations, and skills. For instance, organisations that extensively change their business processes when integrating AI are more likely to see significant benefits. This need for deep integration means AI cannot be viewed as a simple 'point solution.' Instead, it calls for a systems approach where humans and AI are interdependent parts of a socio-technical ensemble. Business value emerges not from the technology itself but from a firm's unique configuration of resources, developed through a long process of trial and error. Consequently, organisations must contend not with idealised problems from benchmarks but with firm-specific problem representations. This creates a disconnect between AI capability in the abstract and its contextualised, real-world application.

Implementation lags manifest along multiple dimensions, each representing a distinct source of friction between technological capability and organisational value creation. Embedding AI within existing organisational systems creates integration challenges, as the technology must be woven into complex socio-technical infrastructures rather than deployed in isolation. Translating benchmark performance to context-specific problems

generates representation challenges, where a task's properties in idealised settings differ substantially from its manifestation in organisational workflows. Coordinating multiple agents—both human and AI—in complex environments produces scale challenges, as interdependencies and feedback loops complicate efforts to optimise collective outcomes. Adapting to evolving conditions introduces temporal challenges, where models trained on historical data face dynamic environments that drift over time. Finally, ensuring AI capabilities adequately match problem requirements creates adequacy challenges, particularly when problems demand causal reasoning, common sense, or contextual understanding that current approaches struggle to provide.

Understanding this disconnect requires systematic analysis of these specific types of gaps. The framework that follows identifies five fundamental dimensions—Integration, Representation, Scale, Temporality, and Adequacy—along which such gaps manifest, enabling diagnostic assessment of gap strength and appropriate collaborative configurations. As illustrated by the historical analogy of electricity adoption, where full benefits were realised not by replacing steam engines but by reorganising factories around decentralised power (a shift from 'point solution' to 'system solution' elaborated above), the challenge is not merely technological substitution but systemic transformation.

3. A Framework for Mapping Human-AI Complementarities

The historical evolution of AI reveals an expanding frontier of automatable tasks, yet the translation of these capabilities into organisational value remains challenging. This suggests that a simple focus on technological milestones is insufficient. Instead, we propose a multi-dimensional framework that identifies the fundamental gaps between an idealised AI task and a real-world organisational problem. We demonstrate that bridging these gaps typically requires substantial human involvement within collaborative human-AI configurations, with the intensity of involvement determined by gap strength in specific organisational contexts. The framework is organised around five interdependent layers of complexity that separate an idealised task from its real-world counterpart: the Integration Gap, the Representation Gap, the Scale Gap, the Temporality Gap, and the Adequacy Gap. The subsections that follow therefore note only the involvement specific to each gap, whose intensity varies with gap strength along the sliding scale introduced above. Table 1 introduces the diagnostic boundaries of each gap, and Figure 1 illustrates them schematically.

Table 1. Diagnostic boundaries of the five gaps.

Gap (Transition)	Diagnostic Boundary (the Question It Asks)	Typical Symptoms	Representative Organisational Question	Primary Decision Phase
Integration (Isolated → Embedded)	Can a single AI application be embedded into existing socio-technical complements and routines?	System interdependence; complementarity tightness; routine embeddedness; infrastructure maturity; integration experience	How many and how deep are the routine, role, and control changes required to deploy and sustain this AI?	Intelligence

Table 1. *Cont.*

Gap (Transition)	Diagnostic Boundary (the Question It Asks)	Typical Symptoms	Representative Organisational Question	Primary Decision Phase
Representation (Intrinsic → Contextual)	Has the benchmark task been correctly framed and contextualised as the firm's actual problem?	Tacit-knowledge dependence; contextual richness; task-instance heterogeneity; mental-model diversity; problem structuredness	Do the chosen targets and problem framing faithfully capture the value the organisation cares about?	Design
Scale (Analytical → Systemic)	Does optimising the collective outcome require coordinating multiple interdependent agents?	Interdependence tightness; coordination complexity; agent heterogeneity; reasoning opacity; local-global tension	Are decisions largely local, or tightly coupled across units such that local optimisation risks global dysfunction?	Choice
Temporal (Static → Dynamic)	Is the environment stable, or subject to drift and performativity that invalidate historical learning?	Environmental volatility; drift susceptibility; performativity strength; adaptation requirements; change predictability	Is drift primarily covariate (retrainable) or concept and performative (requiring reformulation)?	Review
Adequacy (Technology-led → Problem-led)	Do current AI capabilities suffice for what the problem requires, as framed?	Causal-reasoning demands; data availability and quality; explanation requirements; novelty frequency; representation reliability	Do downstream decisions require causal or counterfactual claims beyond prediction?	Intelligence

Each gap is defined by a diagnostic boundary, manifests in characteristic symptoms (the factors that determine gap strength), and can be probed by a representative question. The gaps interact across phases; the final column marks the phase at which each gap's influence on the human-AI division of labour is most pronounced. Integration and Adequacy both surface at the Intelligence phase, reflecting that problem assessment evaluates organisational context and AI capability together.

Two boundaries warrant explicit clarification. The Representation and Adequacy gaps both concern mismatch, but along different dimensions: the Representation gap concerns whether the idealised task has been correctly framed and contextualised for the organisation, a problem-definition question in which mental models play a definitional role, whereas the Adequacy gap concerns whether current AI is sufficient for the problem so framed, a capability question turning on causal reasoning and the reliability of learned representations. The Integration and Scale gaps both involve coordination, but the Integration gap concerns embedding a single application within a firm's existing complements and routines, whereas the Scale gap concerns coordinating multiple interdependent agents to optimise a collective outcome. Appendix A sets out how the framework can be applied as a qualitative diagnostic.



Figure 1. Stylised illustration of the five gaps. Polygon sizes are conceptual, not empirical. Organisations could optionally adapt this as a qualitative diagnostic to track progress over time (see Appendix A).

3.1. The Integration Gap (Isolated → Embedded)

The Integration Gap addresses the transition from an isolated application to an embedded organisational system. A unifying theme here is AI as a general-purpose technology (GPT). While AI has long been considered a GPT [21], the versatility of generative AI has strengthened this claim [36]. A defining feature of GPTs is that realising their potential requires extensive adaptations to surrounding infrastructure, policies, and procedures; simple labour substitution yields limited benefits [1,63]. This necessitates viewing AI as situated within a complex socio-technical system where technical and social aspects are highly interdependent [65,66].

Three established organisational theories provide a coherent account of the resulting slow adoption trajectory. The theory of complementarity formalises how synergistic relationships between assets create value [64]. For established firms, the configuration of practices that generates these returns is a hard-won, tacit asset [64,67]. Given the sensitivity of this system, organisations rationally take a gradual approach when introducing a disruptive technology like AI. This perspective clarifies the 'implementation lag' seen with GPTs [21] and provides a convincing explanation for the productivity paradox [31,68].

The resource-based view (RBV) aligns with complementarity, explaining that competitive advantage comes from configurations of resources that are valuable, rare, inimitable, and non-substitutable [67,69]. Because a GPT like AI is broadly available [70], competitive advantage must derive from the unique, inimitable way an organisation integrates it into its system of complements to create a defensible 'resource-position barrier' [71].

The evolutionary view of routines provides a third lens, identifying organisational practices as the 'genes' or units of selection in a competitive environment [16,22]. These routines provide stability, and changing them is a slow, tentative process [48].

Together, these theories explain why organisations take a measured approach. To gain an advantage, a firm must integrate a widely available tool into its unique, hard-to-replicate internal systems [72]. As noted earlier, the electricity analogy applies here: benefits came from reorganising around the new technology, not from simple substitution [41,68]. Here the human contribution concentrates on change management, process re-engineering, and system integration.

3.2. The Representation Gap (Intrinsic → Contextual)

The Representation Gap—addressing the transition from a task's intrinsic properties to its contextual reality—explains why benchmark performance is a poor predictor of real-world value. A recent MIT Media Lab Project NANDA report [61] reports that approximately 95% of enterprise GenAI initiatives show no measurable profit-and-loss (P&L) impact, underscoring the difficulty of translating pilots into value; this headline figure has been widely reported (e.g., Fortune, August 2025). A socio-technical lens reveals why context is crucial. From this perspective, a task's essential properties are interactional, not inherent; its complexity emerges from interrelationships with other tasks, processes, and firm-specific complements—what Weaver [73] termed 'organised complexity' (see also [74]).

The cognitive mechanism behind this subjective framing is the mental model: a decision-maker's internal, simplified map of reality [75,76]. As boundedly rational actors, decision-makers use these models not to find optimal solutions but to satisfice [35,77]. By selecting what is relevant, the mental model plays a definitional role, fundamentally altering the nature of the task from its idealised benchmark form. For example, deep-learning models may be suitable for demand forecasting in functional products with abundant data, whereas innovative products with sparse data demand human-centred inputs [3]. The greater the reliance on such interactional properties, the greater the risk that benchmark performance will fail to translate. Abstraction is one practical mechanism behind this translation problem: the means by which a model preserves the latent structure of a task across superficially varying instances [9]. Such abstractions are not always faithful, since learned representations can fragment into spurious or entangled associations [44]. Abstraction is therefore not only a conceptual concern but a computational mechanism that shapes what an AI system retains, discards, and generalises.

3.2.1. Exploration, Exploitation, and New Modes of Collaboration

The contingent nature of organisational problem-solving implies a choice amongst multiple plausible mental models. The chosen representation is a high-leverage decision: it defines the problem's scope and delimits the space of possible actions. A complex problem may become tractable under a different representation, as when a mathematical derivation makes 'evident what was previously true but obscure' [16]. This creates a critical trade-off between searching for a better problem representation (exploration) and searching for better policies within that model (exploitation) [76,78].

The meta-decision to prioritise exploration is, in essence, a choice to engage in 'routine-changing' rather than 'routine-guided' behaviour [22]. Such a choice relies upon planning, counterfactual reasoning, and common sense, where humans retain clear superiority [9,14,79–81]. However, GenAI's ability to take on divergent tasks has added nuance [36]. The distinct cognitive styles of humans (deep inference from path-dependent learning) and AI (shallow inference from broad data patterns) present opportunities to design new collaborative modes that systematically manage the trade-off between search scope and depth [10,81]. Beyond conventional specialisation, an ensemble strategy—in

which humans and AI perform the same creative task—can improve outcomes through error cancellation and the synthesis of diverse perspectives [47,54].

3.2.2. The Challenge of Tacit Knowledge: A Case Study in Radiology

Even during the exploitation phase, where algorithms are presumed to excel, human judgement remains critical. Complex organisational tasks show high heterogeneity amongst task instances, requiring constant vigilance and meta-decisions about the appropriate mix of human and AI input [68,82]. Diagnostic radiology offers a vivid illustration. Despite early predictions that ML would replace radiologists [83], the demand for them has grown [41,84]. A field study of ML tool adoption in a US hospital highlights why: the task relies on 'situated, social, and tacit practices' not captured in training data [85]. This asymmetry—where experts rely on non-codifiable 'type II data'—creates the potential for genuine augmentation [54,62]. However, the opacity of ML tools and uncertainty about the 'ground truth' of training labels hinder the reconciliation of conflicting human–AI judgements [85]. The study found that meaningful collaboration only occurred when experts practised 'engaged augmentation,' investing significant effort in 'interrogative' activities to resolve uncertainties. This demonstrates that even in highly specialised, data-rich domains, tacit knowledge places a heavy burden on human judgement to unlock human–AI complementarity [5,80].

Here the human contribution concentrates on problem framing, mental-model selection, and managing the exploration-exploitation trade-off.

3.3. The Scale Gap (Analytical → Systemic)

The Scale Gap addresses the transition from an analytical task solvable by a single agent to a systemic problem requiring the coordinated action of multiple agents to optimise a collective outcome. Real-world problem-solving involves partitioning complex problems into sub-problems, solving them, and integrating partial solutions to achieve overall goals [86]. This systemic frame brings interdependencies to the fore. The decision space can be modelled as a performance landscape whose topology depends on interactions between decision variables. Drawing from evolutionary biology, 'ruggedness' denotes a landscape with many local optima as interdependencies increase [87,88].

Navigating such a rugged landscape, even for all-human teams, is fraught with coordination challenges, particularly under epistemic uncertainty. Organisational design research identifies two crucial, interdependent risks. The first is 'mutual confusion'—the inability to determine whether observed consequences stem from one's own actions or others'. The second is 'joint myopia'—the tendency to settle prematurely on a suboptimal consensus in regions where pay-offs are unambiguous [89]. These risks are interdependent: efforts to mitigate mutual confusion through information sharing can induce premature homogenisation of beliefs, causing joint myopia, and vice versa [90].

The introduction of AI as a novel learner can exacerbate these challenges. Because ML models learn with few explicit inductive biases and their reasoning is largely opaque, their exploratory nature can amplify mutual confusion, making causal attribution significantly more difficult [42,90]. Furthermore, ML's superior information-processing capacity and convergent logic can be susceptible to local rationality—optimising a narrow set of information to the detriment of the global goal [91,92].

The qualitatively different behaviour of AI [93] and the fact that functional adaptations in organisations are not guaranteed [66] highlight the need for deliberate design of human–AI systems. The challenge is not merely to solve sub-problems, but to orchestrate a collective intelligence that can learn partial mental models, account for interdependencies, and make trade-offs that may sacrifice local objectives for global ones.

Here the human contribution concentrates on systemic oversight, coordination design, and managing global-local trade-offs.

3.4. The Temporality Gap (Static → Dynamic)

The Temporality Gap addresses the transition from a static problem environment, typical of benchmarks, to a dynamic one. Organisational environments are in flux, subject to changes in customer preferences, competitive actions, and internal processes. This dynamism creates a challenge for AI models trained on historical data: a mismatch between the model's learning environment and the current target environment, a gap that can ultimately undermine performance, which can lead to degradation [33].

This mismatch often manifests as model drift, traceable to external and internal factors. External shifts (e.g., changing customer demographics) can invalidate core assumptions [94]. Internal changes (e.g., new processes) can alter downstream conditions. Technically, this drift falls into two categories: covariate drift, where the input distribution changes while the underlying rules remain valid, and concept drift, where the input–output relationships change [95,96]. While retraining can address covariate drift, concept drift challenges the problem formulation itself, often requiring human intervention to redefine the model.

A more profound temporal challenge arises from performativity, where the deployment of an AI changes the environment it seeks to model. Consider an AI that improves demand-forecast accuracy in a supply chain. This initial success is performative: greater accuracy reveals latent capacity constraints, leading to service failures. These failures trigger customer defection, which alters demand patterns. This creates feedback-induced concept drift, where success precipitates the very conditions that degrade future performance. Managing such reflexive dynamics requires strategic foresight and coordination—such as adapting capacity planning—that fall squarely in the domain of human judgement.

Ultimately, navigating the temporality gap requires a shift from 'set and forget' to continuous vigilance. Here the human contribution concentrates on monitoring for drift, diagnosing its sources, and managing performative effects.

3.5. The Adequacy Gap (Technology-Led → Problem-Led)

The Adequacy Gap addresses the transition from technology-led optimisation to problem-led sufficiency. A fundamental limitation is the ML paradigm's reliance on learning exclusively from data without incorporating prior knowledge [9,97]. While this approach has yielded impressive practical successes, it restricts adaptability to novel situations [98]. Deep neural networks, for instance, operate as 'black boxes,' with knowledge represented in complex, non-transparent statistical forms [19]. This opacity poses challenges for inference tasks where understanding the relationship between inputs and outputs is crucial, especially in high-stakes or regulated contexts.

LLMs and GenAI more generally continue this tradition of representing knowledge through shallow inference over a vast number of weakly correlated variables [10]. Due to this approach of bootstrapping knowledge entirely from raw data, LLMs—like traditional ML—are susceptible to spurious correlations [39]. For instance, in so-called large reasoning models, 'reasoning' is based on planning knowledge learned from text rather than first principles. Superficial alterations to planning problems, while retaining structure, can lead to sharp performance drops—a condition Kambhampati attributes to LLMs performing 'universal approximate retrieval' [99,100].

A notable instance of limitations in rule derivation is the 'reversal curse' [101], where an LLM trained only on 'A is B' fails to infer 'B is A.' A study examining underlying representations in GenAI-based models reports redundancy and fragmentation, including connections between unrelated ideas ('fractured entangled representations') [44]. In another

study testing an LLM's ability to infer a physical law from orbital data, the model failed to deduce the law despite accurate predictions [102]. This highlights the fallacy of equating prediction accuracy with explanatory depth [45].

The lack of causal knowledge in state-of-the-art AI means understanding is based on associations and cannot support interventions or counterfactual claims, which require reliable world models [79]. In decision-making settings, especially under sparse data and uncertainty, where current AI is vulnerable to flawed representations, human judgement must continue to play a vital role. Perhaps a manifestation of the unreliability of representations learned exclusively from data is the dynamic nature of AI's capabilities at the frontier that Dell'Acqua et al. characterise as 'jagged' [103]. In LLMs trained on most available Internet text, next-token prediction that emphasises fluency and coherence can contradict truthfulness. Moreover, fine-tuning often uses binary preference scores that may penalise acknowledging uncertainty [104]. These factors further emphasise the need for supervision and judgement in high-stakes settings.

Furthermore, because of the 'foundational' nature of LLMs, suitable for various downstream applications with fine-tuning, their limitations are inherited. The 'homogenisation' of models in the LLM paradigm, due to broad capabilities [42], differs from earlier deep-learning approaches and poses the risk of spreading flaws and biases uniformly, requiring human judgement as a counterbalance.

A related aspect of inheritance is the increasing use of model-generated data to train newer models. A problematic consequence—stemming from statistical approximation in sampling over token distributions—is the disappearance of tail events (the rare, low-frequency cases that lie in the tails of the data distribution) [105]. The implication for decisions is a decline in performance due to a lack of information about niche or under-represented contexts, which calls for vigilance even in situations currently suitable for automation and seemingly well understood.

Although agents, in principle, enable the expansion of automation to actions, this multiplies issues and risks associated with LLMs as perception and reasoning engines. Evaluations on realistic long-horizon benchmarks—e.g., WebArena (GPT-4 agent approximately 14% vs. humans approximately 78%), OSWorld (best agent approximately 12% vs. humans approximately 72%)—and enterprise scenarios (e.g., CRMArena-Pro) report brittle reasoning, susceptibility to input perturbations, weak long-horizon planning, and difficulty maintaining coherent goals over time.

Here the human contribution concentrates on causal reasoning, contextual interpretation, and managing capability-requirement mismatches.

3.6. Cross-Cutting Normative Arguments

Beyond limitations of current AI, scholars warn of normative considerations that call for reflective use, emphasising human agency and judgement. Several studies find that using LLMs for decision-making involving cognitive offloading can lead to atrophy of skills such as critical thinking [106,107]. This effect is not unique to LLMs; earlier transformative technologies also induced the 'irony of automation' [108], whereby mechanisation diminishes the very ability to delegate, which requires metacognition [106]. The radiology case discussed earlier is paradigmatic: traditional ML was often met with cognitive disengagement, whereas augmentative use depends on 'interrogative practices' that examine underlying knowledge claims [62].

However, the human-like communication skills of LLMs, expressly optimised for fluency, amplify the risk of disengagement. Even before GenAI, scholars [41,59,68] promoted a design approach that does not take tasks as given but reimagines them, recognising that the range of tasks humans and AI can perform together is broader. The increasing capabilities

strengthen this view. There is also a strong case for prioritising evaluations that assess joint performance rather than AI-only benchmarks [109].

In predictive AI, using labels as proxies for ground truth can create 'problem-formulation gaps' [2,96]. Next-token prediction in LLMs can worsen this because plausibility does not ensure truthfulness. The unavoidable gap between prediction and explanation introduces difficulties; combined with the 'foundational' nature of LLMs, this risks homogenisation towards approaches that mistake illusions of understanding for genuine causal understanding [110].

Narayanan and Kapoor advocate viewing AI as a 'normal technology.' Although they acknowledge its transformative potential, they argue that it remains a tool—flawed in ways that often defy intuition—and that this perspective highlights human agency in shaping outcomes and rejects technological determinism [2].

These normative considerations apply across the spectrum of human-AI collaborative configurations. Whether gap strength in a particular organisational context suggests greater AI autonomy or intensive human involvement, the risks of deskilling, cognitive disengagement, and overreliance on flawed representations persist. This underscores the importance of deliberate design choices that preserve human agency, critical thinking, and metacognitive capabilities regardless of the specific division of labour between humans and AI in a given setting.

3.7. Mapping the Five Gaps to the Decision Process

The five gaps developed in this paper can be organised according to where they most saliently affect human judgement within a stylised decision process. Figure 2 depicts a decision model adapted from Simon [111] and Mintzberg et al. [112], comprising four phases: Intelligence (problem assessment), Design (problem definition and alternative generation), Choice (evaluation and selection), and Review (action, outcomes, and feedback). The figure highlights where mental models—the decision-maker's internal, simplified maps of reality—play a definitional or generative role, particularly during problem assessment and framing within complex organisational settings.

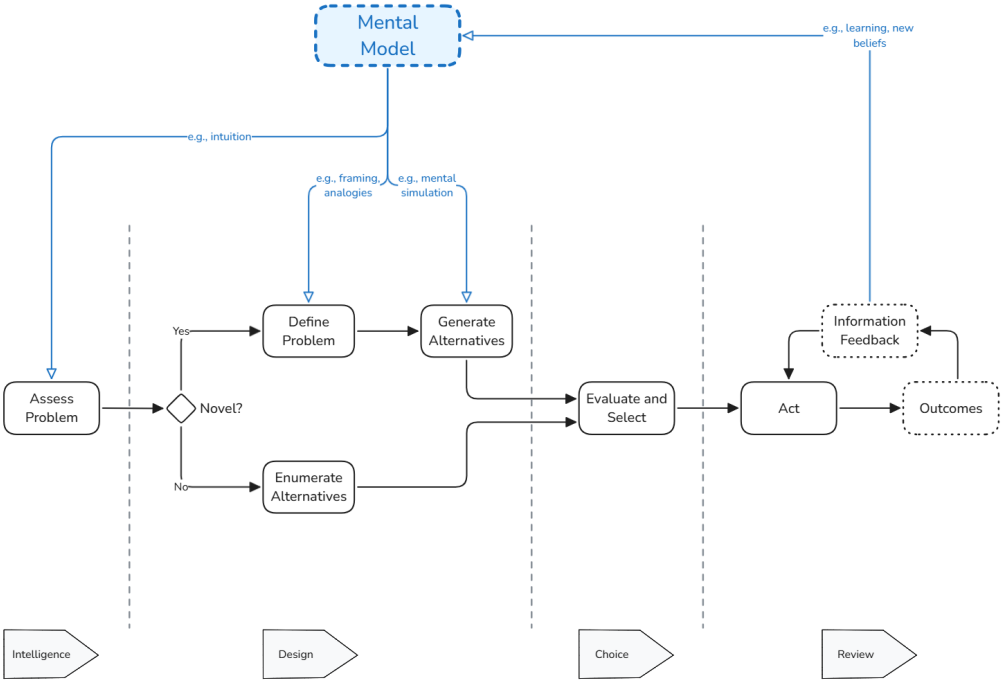


Figure 2. Decision Process Model with Mental Model Influence. Adapted from Simon [111] and Mintzberg et al. [112]. The model depicts four phases of organisational decision-making: Intelligence

(problem assessment), Design (problem definition and alternative generation), Choice (evaluation and selection), and Review (action, outcomes, and feedback). Blue arrows play a definitional or generative role; mental models inform all stages but are highlighted where their contribution is most substantial. Information feedback from outcomes updates mental models over time, reflecting organisational learning.

4. Towards Quantitative Evaluation: An Illustrative Example

The diagnostic framework developed above identifies the dimensions along which gaps arise; the mapping to the decision process locates these gaps within the phases where they most directly shape human involvement. Furthermore, depending on the level of decision-making at which an organisation seeks to implement AI—for example, strategic or tactical—the mapping to decision phases helps highlight the gaps most relevant to the challenge. In practice, assessing gap strength across these dimensions guides the calibration of the level of judgement needed at each decision phase, producing an initial hypothesis about the human-AI teaming configuration. A complementary question concerns how this hypothesis might be evaluated quantitatively—that is, whether a proposed configuration actually yields system-level performance improvements within a specific organisational context and under realistic operational conditions.

The existing literature offers qualitative design frameworks—organised around task-level skill complementarities [5], team-level coordination [6,7], outcome-oriented strategies [4,113], and organisational enablers [8]—that provide valuable guidance for generating hypotheses about effective teaming configurations given assessed gap strength. However, these frameworks do not enable quantitative evaluation of proposed configurations within organisation-specific contexts, accounting for unique problem framings and system-level interdependencies.

Addressing this last-mile challenge requires an approach capable of representing how algorithmic and human decision-making interact, not merely how each performs in isolation. Three requirements can be distilled: the model must be holistic, capturing interactions that generate endogenous explanations for system behaviour; causal, rendering the mental model of joint decision-making transparent; and quantitative, enabling the impact of interactions to be assessed in terms of outcome-oriented metrics. System dynamics, founded on the insight that behaviour is latent within the structure of a system [114], broadly satisfies these requirements. Its emphasis on correspondence with organisational reality rather than mathematical tractability through drastic simplification makes it well suited to modelling the dynamically complex interactions within human-AI sociotechnical systems.

Sankaran et al. [3] demonstrate one such approach using a bike-share system modelled on New York's Citi Bike, where customers rent and return bikes at docking stations. Recurring inventory imbalances—where stations experience stockouts even when the aggregate fleet is sufficient—are addressed through crowdsourced rebalancing, in which incentivised users help redistribute bikes. The complex interplay among planning actions, user behaviour in response to those actions, system inventory balance, and provider decisions to authorise incentives generates a rich set of feedback loops. This diversity of subproblems—demand forecasting, station clustering, causal inference of incentive effects, and the design of incentivisation rules—results in a blend of algorithmic and judgmental components within the organisation's decision structure.

A four-stage evaluation process—causal mapping, model formulation, partial-model testing, and integration and policy analysis—translates an initial augmentation hypothesis into a testable simulation. Partial-model testing verifies whether an ML intervention improves decision-making at the process level before system-wide integration. For instance, an LSTM-based demand forecasting model demonstrated a 26% accuracy improvement over exponential smoothing. However, whole-model testing—which integrates all subsys-

tems and activates all feedback processes—revealed that this local accuracy gain did not automatically translate to proportional system-level improvement. Performance across over 400 simulated policy scenarios proved highly sensitive to parameters governing risk perception, responsiveness to forecasts, and incentivisation timing—judgmental components that mediate between algorithmic outputs and system outcomes. The resulting performance landscape was rugged, with small parameter changes producing discontinuous shifts in outcomes, underscoring the importance of robust policies over narrowly optimal ones.

The methodology—the four-stage evaluation process and performance landscape mapping—transfers to other contexts where algorithmic and judgmental components require integration. The quantitative findings, however, are context-specific: calibration of judgmental and algorithmic content is necessarily firm-specific, and particular thresholds, delay values, and improvement percentages should not be extrapolated directly.

5. Conclusions

This paper has developed a five-dimensional framework for analysing the complications that arise when translating AI capabilities into organisational value. The Integration, Representation, Scale, Temporality, and Adequacy gaps collectively demonstrate why pure autonomy remains elusive in complex, real-world organisational contexts, despite rapid advances in generative and agentic AI. Each dimension reveals systematic challenges—from embedding AI within socio-technical systems to managing temporal dynamics and ensuring adequate problem representations—that typically require human involvement within collaborative configurations to navigate. The strength of these gaps, which varies substantially by organisational context, problem characteristics, and available complementary resources, determines the appropriate intensity of human participation along a sliding scale from greater AI autonomy to intensive human involvement.

While the framework is diagnostic, the illustrative example [3] demonstrates that the strategy advocated in this paper—conceptual framing of complementarities followed by precise quantification—can reveal dynamics invisible to either approach alone. The finding that a 26% improvement in forecast accuracy did not translate to proportional system-level gains underscores the importance of evaluating human-AI configurations holistically, within the specific decision structures and feedback processes of the focal organisation. Translating the diagnostic insights developed here into effective organisational practice entails quantitative evaluation methods that are necessarily specific to individual cases and therefore fall outside the scope of this review.

5.1. Methodological Recommendations

The five gaps are conceptual, and the framework that organises them is diagnostic: assessing gap strength yields an initial hypothesis about where judgement should reside, and hence about the human-AI teaming configuration, but it does not by itself establish whether that configuration delivers system-level value. The recommendation offered here therefore concerns the method by which such a hypothesis should be evaluated. As Section 4 observes, the existing qualitative design frameworks support the generation of a teaming hypothesis from an assessment of gap strength, yet they do not enable quantitative evaluation of that hypothesis within an organisation-specific context. Closing this methodological gap calls for an evaluation approach that is holistic, causal, and quantitative; system dynamics broadly satisfies these requirements.

Concretely, we recommend the four-stage evaluation process set out in Section 4—causal mapping, model formulation, partial-model testing, and integration and policy analysis—together with performance landscape mapping. Partial-model testing assesses an algorithmic intervention against process-level metrics before system-wide integration,

and so separates a local gain in predictive accuracy from a gain in decision quality. Whole-model testing, which activates all feedback processes, then establishes whether that gain survives integration: in the illustrative case, a 26% improvement in forecast accuracy did not translate proportionally into system-level performance. Performance landscape mapping across policy scenarios subsequently identifies the judgmental parameters—risk perception, responsiveness to forecasts, and incentivisation timing—to which outcomes are most sensitive, and favours robust policies over narrowly optimal ones.

These stages are procedural and should not be conflated with the decision phases of Section 3.7, which locate where within a decision human judgement is required. We offer the process as a methodological direction: its structure is not specific to the illustrative case, but empirical validation across multiple organisational contexts remains outstanding.

5.2. Limitations and Future Research

Several limitations should be noted. The framework is intended as a conceptual synthesis and diagnostic instrument rather than an empirically validated model. As an integrative, narrative review, source selection was purposive rather than exhaustive, and a different selection might weight the dimensions differently; the five-gap structure is offered as an analytically useful organisation of the field rather than an exhaustive taxonomy. The framework is diagnostic rather than prescriptive: it locates where judgement is required but does not by itself prescribe the optimal configuration. The illustrative example is a single case [3], and its quantitative findings, including particular thresholds, delays, and improvement percentages, are firm-specific and should not be extrapolated, although the four-stage method transfers. Finally, the qualitative diagnostic in Appendix A is indicative and would benefit from empirical validation.

Several directions follow from this framework. First, it should be extended from the supervised predictors considered here to generative and agentic AI, where multiple agents decompose tasks, communicate, and coordinate. The open question is whether the five gaps persist, and how their relative strength shifts, once the AI components themselves learn, adapt, and act; system dynamics, with its capacity to represent feedback and emergent behaviour, is a natural vehicle for the extension, and progress would be marked by a model of a multi-agent human-AI configuration whose system-level behaviour can be evaluated rather than assumed. Second, evaluation should move beyond parameter settings to the decision architecture itself, asking how alternative allocations of decision rights and roles between humans and AI reshape the performance landscape; here progress means showing that structural redesign, and not only parameter tuning, alters system outcomes. Third, the judgmental heuristics that stand in for human decision-makers should be validated against real behaviour, the question being whether these literature-grounded assumptions hold once decision-makers respond to AI-generated recommendations in deployment; progress would consist of calibrated judgmental components and evidence on how trust and reliance form in practice. A fourth, longer-horizon direction is to extend the approach across hierarchical levels, with human judgement acting as orchestrator, drawing on creativity, intuition, and abstraction to direct AI contributions, in line with the agentic paradigm now emerging in practice.

Author Contributions: G.S. (Ganesh Sankaran): Conceptualization; Investigation; Methodology; Resources; writing—original draft preparation; Validation; Visualization; M.A.P.: Conceptualization; Investigation; Methodology; Resources; writing—review and editing; Supervision; Validation; Project Administration; G.S. (Guido Siestrup): Conceptualization; Funding Acquisition; Methodology; Resources; writing—review and editing; Supervision; Validation; Project Administration. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the revision of this manuscript, the authors used Claude Opus 4.8 (Anthropic) to assist with revising and copy-editing text drafted by the authors, organising the point-by-point responses to reviewers, and auditing and standardising the reference list, including verification of bibliographic details. All conceptual contributions, including the framework and its arguments, are the authors' own. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: Author Ganesh Sankaran was employed by the company ZweiThreads Software LLP, India. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Applying the Framework: A Qualitative Diagnostic

The framework is applied by assessing the strength of each gap for the focal problem and reading off the implied division of labour. Table 1 gives, for each dimension, the symptoms that signal a strong gap and a representative diagnostic question; these serve directly as evaluation prompts. The assessment is a structured qualitative judgement rather than a precise measurement. Because the system of complements that generates value is organisation-specific, no single universal recipe applies; a Delphi-style procedure, in which decision owners and relevant domain experts assess each gap independently and then reconcile divergences, is well suited to producing a defensible reading. Where the gaps are weak, greater AI autonomy may be appropriate; where they are strong, intensive human participation becomes necessary, in line with the sliding-scale perspective developed above.

In practice, the diagnostic proceeds in four steps. First, the focal decision is delineated, together with the AI capability proposed to support it. Second, for each gap in turn, the diagnostic question in Table 1 is put to that decision, and the available evidence—the symptoms listed in Table 1, the data conditions, and the organisational context—is assembled. Third, each gap is assigned a qualitative strength on the sliding scale, with its evidential basis recorded alongside it; where several assessors are involved, the Delphi-style reconciliation described above is applied at this step. Fourth, the implied division of labour is read off gap by gap, identifying where human judgement must concentrate. The output is a per-gap qualitative reading: the strength labels are not converted into numeric scores, nor aggregated into a composite index, because the gaps are interdependent and their strength is organisation-specific.

To illustrate the procedure, Table A1 applies the diagnostic questions of Table 1 to the bike-share rebalancing case of Section 4.

To make the third step concrete, consider the Scale gap. The diagnostic question asks whether value depends on coordinating interdependent agents rather than on optimising the focal decision in isolation. The evidence in this case is that station inventories are coupled through user flows, that incentives act locally, and that improving a single station can therefore worsen system-wide balance, producing the rugged performance landscape observed. Because this coupling is both strong and pervasive, the gap is assessed as High rather than Moderate, and the division of labour that follows places systemic oversight with the human decision maker.

Read together, the strong Scale gap, the causal demands of the Adequacy gap, and the bounded temporal feedback place the bulk of human involvement in systemic oversight, causal interpretation, and the management of time-lagged effects, while the data-rich forecasting sub-task admits greater AI autonomy. This reading is consistent with the finding (Section 4) that a 26% improvement in forecast accuracy did not translate into a

proportional system-level gain: the judgmental components that mediate between forecast and outcome are precisely those the stronger gaps identify.

Table A1. Worked example: applying the diagnostic to the bike-share case of Section 4.

Gap	Strength	Basis in the Bike-Share Case	Where Human Judgement Concentrates
Integration	Moderate	Forecasting and causal-inference components must be embedded within the provider's existing crowdsourced-rebalancing operation (the Bike Angels programme); the forecast-accuracy gain delivered no process-level benefit until integrated—via a bias-correction layer—with the incentivisation logic.	Adapting incentivisation routines and decision interfaces.
Representation	Moderate	Demand is data-rich, so the forecasting sub-task is well posed; but framing the problem—de-censoring suppressed demand, clustering stations by demand symmetry, and selecting the service level (stockouts) as the value to capture—is judgement-laden, since factors beyond those the models consider may be relevant.	Problem framing, target selection, and goal setting.
Scale	High	Value depends on coordinating interdependent stations and users toward a system-wide inventory balance; because incentives are local, optimising a single station can be globally suboptimal, producing the rugged performance landscape observed.	Systemic oversight and coordination of incentivisation across the cluster.
Temporality	Moderate	Demand varies intraday and incentives feed back into local user behaviour over time, although they do not meaningfully influence overall demand; the drift is therefore real but bounded and local.	Monitoring availability and managing time-lagged incentive effects.
Adequacy	Moderate–High	Prediction suffices for forecasting, but estimating the effect of incentives requires Bayesian causal inference rather than prediction, and the system-level question is addressed through simulation rather than prediction alone.	Causal interpretation of incentive effects; matching capability to requirement.

The reading applies the diagnostic questions of Table 1 to the bike-share case. Strength labels are qualitative and per-gap; they are not scored or aggregated into an index.

References

1. Feigenbaum, J.; Gross, D.P. Organizational and Economic Obstacles to Automation: A Cautionary Tale from AT&T in the Twentieth Century. *Manag. Sci.* **2024**, *70*, 8520–8540. [CrossRef]
2. Narayanan, A.; Kapoor, S. *AI as Normal Technology*; Knight First Amendment Institute: New York, NY, USA, 2025.
3. Sankaran, G.; Palomino, M.A.; Knahl, M.; Siestrup, G. Towards a System Dynamics Framework for Human–Machine Learning Decisions: A Case Study of New York Citi Bike. *Appl. Sci.* **2024**, *14*, 10647. [CrossRef]
4. Raisch, S.; Krakowski, S. Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Acad. Manag. Rev.* **2021**, *46*, 192–210. [CrossRef]
5. Hemmer, P.; Schemmer, M.; Kühl, N.; Vössing, M.; Satzger, G. Complementarity in Human-AI Collaboration: Concept, Sources, and Evidence. *Eur. J. Inf. Syst.* **2025**, *34*, 979–1002. [CrossRef]
6. Shrestha, Y.R.; Ben-Menahem, S.M.; Von Krogh, G. Organizational Decision-Making Structures in the Age of Artificial Intelligence. *Calif. Manag. Rev.* **2019**, *61*, 66–83. [CrossRef]
7. Saenz, M.; Revilla, E.; Simón, C. *Designing AI Systems with Human-Machine Teams*; MIT Sloan Management Review; MIT Sloan School of Management: Cambridge, MA, USA, 2020.
8. Kolbjørnsrud, V. Designing the Intelligent Organization: Six Principles for Human-AI Collaboration. *Calif. Manag. Rev.* **2024**, *66*, 44–64.

9. Mitchell, M. Abstraction and Analogy-Making in Artificial Intelligence. *Ann. N. Y. Acad. Sci.* **2021**, *1505*, 79–101. [CrossRef] [PubMed]
10. Smith, B.C. *The Promise of Artificial Intelligence: Reckoning and Judgment*; MIT Press: Cambridge, MA, USA, 2019.
11. Summerfield, C. *These Strange New Minds: How AI Learned to Talk and What It Means*; Penguin Group: London, UK, 2025.
12. Levesque, H.J. *Common Sense, the Turing Test, and the Quest for Real AI*; MIT Press: Cambridge, MA, USA, 2017.
13. Frege, G. Begriffsschrift, A Formula Language, Modeled upon that of Arithmetic, for Pure Thought. In *From Frege to Gödel: A Source Book in Mathematical Logic*; Harvard University Press: Cambridge, MA, USA, 1879; Volume 1931, pp. 1–82.
14. Brachman, R.J.; Levesque, H.J. *Machines like Us: Toward AI with Common Sense*; MIT Press: Cambridge, MA, USA, 2023.
15. Newell, A.; Simon, H.A. Computer Science as Empirical Inquiry: Symbols and Search. *Commun. ACM* **1976**, *19*, 113–126. [CrossRef]
16. Simon, H.A. *The Sciences of the Artificial*, 3rd ed.; MIT Press: Cambridge, MA, USA, 1996.
17. Legg, S.; Hutter, M. Universal Intelligence: A Definition of Machine Intelligence. *Minds Mach.* **2007**, *17*, 391–444. [CrossRef]
18. Nonaka, I. The Knowledge-Creating Company. In *The Economic Impact of Knowledge*; Routledge: Oxfordshire, UK, 2009; pp. 175–187.
19. Brynjolfsson, E.; McAfee, A. The Business of Artificial Intelligence. *Harv. Bus. Rev.* **2017**, *7*, 1–2.
20. Polanyi, M. The tacit dimension. In *Knowledge in Organisations*; Routledge: Oxfordshire, UK, 2009; pp. 135–146.
21. Brynjolfsson, E.; Mitchell, T. What can Machine Learning Do? Workforce Implications. *Science* **2017**, *358*, 1530–1534. [CrossRef] [PubMed]
22. Nelson, R.R.; Winter, S.G. *An Evolutionary Theory of Economic Change*; Belknap Press of Harvard University Press: Cambridge, MA, USA, 1982.
23. Mintzberg, H. Planning on the Left Side and Managing on the Right. In *Leadership Perspectives*; Routledge: Oxfordshire, UK, 2017; pp. 413–426.
24. Kahneman, D. *Thinking, Fast and Slow*; Penguin: London, UK, 2011.
25. Bengio, Y. The Consciousness Prior. *arXiv* **2017**, arXiv:1709.08568.
26. Chollet, F. On the Measure of Intelligence. *arXiv* **2019**, arXiv:1911.01547.
27. Chollet, F. *Deep Learning with Python*; Manning Publications Company: Shelter Island, NY, USA, 2021.
28. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
29. Voosen, P. The Believers. *Chron. High. Educ.* **2015**, *61*. Available online: https://www.chronicle.com/article/the-believers-190147/ (accessed on 13 July 2026).
30. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling Laws for Neural Language Models. *arXiv* **2020**, arXiv:2001.08361.
31. Autor, D. *Polanyi's Paradox and the Shape of Employment Growth*; Working Paper 20485; National Bureau of Economic Research: Cambridge, MA, USA, 2014.
32. Agrawal, A.; Gans, J.; Goldfarb, A. What to Expect from Artificial Intelligence. *MIT Sloan Manag. Rev.* **2017**, *58*, 23–26.
33. Hogarth, R.M.; Lejarraga, T.; Soyer, E. The Two Settings of Kind and Wicked Learning Environments. *Curr. Dir. Psychol. Sci.* **2015**, *24*, 379–385. [CrossRef]
34. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*; Basic Books: New York, NY, USA, 2017.
35. Simon, H.A. *Administrative Behavior*; Simon and Schuster: New York, NY, USA, 2013.
36. Krakowski, S. Human-AI Agency in the Age of Generative AI. *Inf. Organ.* **2025**, *35*, 100560. [CrossRef]
37. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; Casas, D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training Compute-Optimal Large Language Models. *arXiv* **2022**, arXiv:2203.15556.
38. Raschka, S. *Build a Large Language Model (From Scratch)*; Manning Publications: Shelter Island, NY, USA, 2024.
39. Mitchell, M.; Krakauer, D.C. The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2215907120. [CrossRef] [PubMed]
40. Alammar, J.; Grootendorst, M. *Hands-On Large Language Models: Language Understanding and Generation*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2024.
41. Acemoglu, D.; Johnson, S. *Power and Progress: Our Thousand-Year Struggle over Technology and Prosperity*; Hachette UK: London, UK, 2023.
42. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the Opportunities and Risks of Foundation Models. *arXiv* **2021**, arXiv:2108.07258.
43. Krakauer, D.C.; Krakauer, J.W.; Mitchell, M. Large Language Models and Emergence: A Complex Systems Perspective. *arXiv* **2025**, arXiv:2506.11135.
44. Kumar, A.; Clune, J.; Lehman, J.; Stanley, K.O. Questioning Representational Optimism in Deep Learning: The Fractured Entangled Representation Hypothesis. *arXiv* **2025**, arXiv:2505.11581.

45. Del Giudice, M. The Prediction-Explanation Fallacy: A Pervasive Problem in Scientific Applications of Machine Learning. *Methodology* **2024**, *20*, 22–46. [CrossRef]
46. Lakoff, G.; Johnson, M. *Metaphors We Live By*; University of Chicago Press: Chicago, IL, USA, 2008.
47. He, V.F.; Shrestha, Y.R.; Puranam, P.; Miron-Spektor, E. Creative Together: Human-AI Ensembling for Idea Generation. *SSRN Electron. J.* **2025**. [CrossRef]
48. Murray, A.; Rhymer, J.; Sirmon, D.G. Humans and Technology: Forms of Conjoined Agency in Organizations. *Acad. Manag. Rev.* **2021**, *46*, 552–571. [CrossRef]
49. Sapkota, R.; Roumeliotis, K.I.; Karkee, M. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. *Inf. Fusion* **2025**, *126*, 103599. [CrossRef]
50. Handa, K.; Tamkin, A.; McCain, M.; Huang, S.; Durmus, E.; Heck, S.; Mueller, J.; Hong, J.; Ritchie, S.; Belonax, T.; et al. Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. *arXiv* **2025**, arXiv:2503.04761. [CrossRef]
51. Brynjolfsson, E.; Chandar, B.; Chen, R. *Canaries in the Coal Mine? Six Facts About the Recent Employment Effects of Artificial Intelligence*; Technical report; Stanford Digital Economy Lab: Stanford, CA, USA, 2025.
52. Shao, Y.; Zope, H.; Jiang, Y.; Pei, J.; Nguyen, D.; Brynjolfsson, E.; Yang, D. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce. *arXiv* **2025**, arXiv:2506.06576.
53. Krakowski, S.; Haftor, D.; Luger, J.; Pashkevich, N.; Raisch, S. Human-Centered Artificial Intelligence: A Field Experiment. *Manag. Sci.* **2026**, *72*, 57–72. [CrossRef]
54. Choudhary, V.; Marchetti, A.; Shrestha, Y.R.; Puranam, P. Human-AI Ensembles: When Can They Work? *J. Manag.* **2025**, *51*, 536–569.
55. Yu, G.; Kasumba, R.; Ho, C.J.; Yeoh, W. On the Utility of Accounting for Human Beliefs about AI Intention in Human-AI Collaboration. *arXiv* **2024**, arXiv:2406.06051.
56. Krakowski, S.; Luger, J.; Raisch, S. Artificial Intelligence and the Changing Sources of Competitive Advantage. *Strateg. Manag. J.* **2023**, *44*, 1425–1452.
57. Fügener, A.; Grahl, J.; Gupta, A.; Ketter, W. Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Inf. Syst. Res.* **2022**, *33*, 678–696. [CrossRef]
58. Acemoglu, D. The Simple Macroeconomics of AI. *Econ. Policy* **2025**, *40*, 13–58. [CrossRef]
59. Brynjolfsson, E. The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence. *Daedalus* **2022**, *151*, 272–287. [CrossRef]
60. Humlum, A.; Vestergaard, E. *Large Language Models, Small Labor Market Effects*; Working Paper 33777; National Bureau of Economic Research: Cambridge, MA, USA, 2025.
61. Challapally, A.; Pease, C.; Raskar, R.; Chari, P. *The GenAI Divide: State of AI in Business 2025*; Technical report; MIT NANDA: Cambridge, MA, USA, 2025.
62. Lebovitz, S.; Lifshitz-Assaf, H.; Levina, N. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity when Using AI for Medical Diagnosis. *Organ. Sci.* **2022**, *33*, 126–148. [CrossRef]
63. Manyika, J.; Spence, M. The Coming AI Economic Revolution: Can Artificial Intelligence Reverse the Productivity Slowdown? *Foreign Aff.* **2023**, *102*, 70.
64. Brynjolfsson, E.; Milgrom, P. Complementarity in Organizations. In *The Handbook of Organizational Economics*; Princeton University Press: Princeton, NJ, USA, 2013; pp. 11–55.
65. Pedreschi, D.; Pappalardo, L.; Baeza-Yates, R.; Barabasi, A.L.; Dignum, F.; Dignum, V.; Eliassi-Rad, T.; Giannotti, F.; Kertesz, J.; Knott, A.; et al. Social AI and the Challenges of the Human-AI Ecosystem. *arXiv* **2023**, arXiv:2306.13723.
66. Tsvetkova, M.; Yasseri, T.; Pescetelli, N.; Werner, T. A New Sociology of Humans and Machines. *Nat. Hum. Behav.* **2024**, *8*, 1864–1876. [CrossRef] [PubMed]
67. Peteraf, M.A. The Cornerstones of Competitive Advantage: A Resource-Based View. *Strateg. Manag. J.* **1993**, *14*, 179–191. [CrossRef]
68. Agrawal, A.; Gans, J.; Goldfarb, A. *Power and Prediction: The Disruptive Economics of Artificial Intelligence*; Harvard Business Press: Brighton, MA, USA, 2022.
69. Melville, N.; Kraemer, K.; Gurbaxani, V. Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Q.* **2004**, *28*, 283–322. [CrossRef]
70. Ben-Ishai, G.; Dean, J.; Manyika, J.; Porat, R.; Varian, H.; Walker, K. AI and the Opportunity for Shared Prosperity: Lessons from the History of Technology and the Economy. *arXiv* **2024**, arXiv:2401.09718.
71. Wernerfelt, B. A Resource-Based View of the Firm. *Strateg. Manag. J.* **1984**, *5*, 171–180. [CrossRef]
72. Brynjolfsson, E.; Rock, D.; Syverson, C. The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *Am. Econ. J. Macroecon.* **2021**, *13*, 333–372. [CrossRef]
73. Weaver, W. Science and Complexity. In *Facets of Systems Science*; Springer: Berlin, Germany, 1991; pp. 449–456.

74. Capra, F.; Luisi, P.L. *The Systems View of Life: A Unifying Vision*; Cambridge University Press: Cambridge, UK, 2014.
75. Sterman, J. *System Dynamics: Systems Thinking and Modeling for a Complex World*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2002.
76. Csaszar, F.A.; Levinthal, D.A. Mental Representation and the Discovery of New Strategies. *Strateg. Manag. J.* **2016**, *37*, 2031–2049. [CrossRef]
77. Gigerenzer, G. *How to Stay Smart in a Smart World: Why Human Intelligence Still Beats Algorithms*; MIT Press: Cambridge, MA, USA, 2022.
78. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
79. Pearl, J. Radical Empiricism and Machine Learning Research. *J. Causal Inference* **2021**, *9*, 78–82. [CrossRef]
80. Cao, S.; Jiang, W.; Wang, J.; Yang, B. From Man vs. Machine to Man + Machine: The Art and AI of Stock Analyses. *J. Financ. Econ.* **2024**, *160*, 103910. [CrossRef]
81. Raisch, S.; Fomina, K. Combining Human and Artificial Intelligence: Hybrid Problem-Solving in Organizations. *Acad. Manag. Rev.* **2025**, *50*, 441–464. [CrossRef]
82. Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; Mullainathan, S. The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. *arXiv* **2019**, arXiv:1903.12220.
83. Hinton, G. Geoff Hinton: On Radiology. Video, Creative Destruction Lab seminar "Machine Learning and the Market for Intelligence", Toronto, ON, Canada, 2016. Available online: https://youtu.be/2HMPRXstSvQ?si=dBMPHu6xmCpuokg7 (accessed on 13 July 2026).
84. Smith, G.; Funk, J. AI Has a Long Way to Go before Doctors Can Trust it with Your Life. *Quartz* **2021**. Available online: https://qz.com/2016153/ai-promised-to-revolutionize-radiology-but-so-far-its-failing/ (accessed on 13 July 2026).
85. Lebovitz, S.; Levina, N.; Lifshitz-Assaf, H. Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What. *MIS Q.* **2021**, *45*, 1501–1526. [CrossRef]
86. Morecroft, J.D.W. *Strategic Modelling and Business Dynamics: A Feedback Systems Approach*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
87. Kauffman, S.A. The Origins of Order: Self-Organization and Selection in Evolution. In *Spin Glasses and Biology*; World Scientific: Singapore, 1992; pp. 61–100.
88. Levinthal, D.A. Adaptation on Rugged Landscapes. *Manag. Sci.* **1997**, *43*, 934–950. [CrossRef]
89. Knudsen, T.; Srikanth, K. Coordinated Exploration: Organizing Search by Multiple Specialists to Overcome Mutual Confusion and Joint Myopia. *Adm. Sci. Q.* **2014**, *59*, 409–441. [CrossRef]
90. Sturm, T.; Gerlach, J.P.; Pumplun, L.; Mesbah, N.; Peters, F.; Tauchert, C.; Nan, N.; Buxmann, P. Coordinating Human and Machine Learning for Effective Organizational Learning. *MIS Q.* **2021**, *45*, 1581–1602. [CrossRef]
91. Glazer, R.; Steckel, J.H.; Winer, R.S. Locally Rational Decision Making: The Distracting Effect of Information on Managerial Performance. *Manag. Sci.* **1992**, *38*, 212–226. [CrossRef]
92. Gigerenzer, G.; Goldstein, D.G. Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychol. Rev.* **1996**, *103*, 650–669. [CrossRef] [PubMed]
93. Rahwan, I.; Cebrian, M.; Obradovich, N.; Bongard, J.; Bonnefon, J.F.; Breazeal, C.; Crandall, J.W.; Christakis, N.A.; Couzin, I.D.; Jackson, M.O.; et al. Machine Behaviour. *Nature* **2019**, *568*, 477–486. [CrossRef] [PubMed]
94. Melville, N.; Gurbaxani, V.; Kraemer, K. The Productivity Impact of Information Technology Across Competitive Regimes: The Role of Industry Concentration and Dynamism. *Decis. Support Syst.* **2007**, *43*, 229–242. [CrossRef]
95. Babic, B.; Cohen, I.G.; Evgeniou, T.; Gerke, S. When Machine Learning Goes Off the Rails. *Harv. Bus. Rev.* **2021**, 21–32. Available online: https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails (accessed on 13 July 2026).
96. Wang, A.; Kapoor, S.; Barocas, S.; Narayanan, A. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsible Comput.* **2024**, *1*, 1–45. [CrossRef]
97. Marcus, G. The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *arXiv* **2020**, arXiv:2002.06177.
98. Shestakofsky, B. Working Algorithms: Software Automation and the Future of Work. *Work Occup.* **2017**, *44*, 376–423. [CrossRef]
99. Kambhampati, S. Polanyi's Revenge and AI's New Romance with Tacit Knowledge. *Commun. ACM* **2021**, *64*, 31–32. [CrossRef]
100. Kambhampati, S. Can large language models reason and plan? *Ann. N. Y. Acad. Sci.* **2024**, *1534*, 15–18. [CrossRef] [PubMed]
101. Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A.C.; Korbak, T.; Evans, O. The Reversal Curse: LLMs Trained on "A is B" fail to learn "B is A". *arXiv* **2023**, arXiv:2309.12288.
102. Vafa, K.; Chang, P.G.; Rambachan, A.; Mullainathan, S. What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models. *arXiv* **2025**, arXiv:2507.06952.
103. Dell'Acqua, F.; McFowland, E., III; Mollick, E.R.; Lifshitz-Assaf, H.; Kellogg, K.; Rajendran, S.; Krayer, L.; Candelon, F.; Lakhani, K.R. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organ. Sci.* **2025**, *37*, 403–423. [CrossRef]
104. Kalai, A.T.; Nachum, O.; Vempala, S.S.; Zhang, E. Why Language Models Hallucinate. *arXiv* **2025**, arXiv:2509.04664.

105. Shumailov, I.; Shumaylov, Z.; Zhao, Y.; Gal, Y.; Papernot, N.; Anderson, R. The Curse of Recursion: Training on Generated Data Makes Models Forget. *arXiv* **2023**, arXiv:2305.17493.
106. Lee, H.P.; Sarkar, A.; Tankelevitch, L.; Drosos, I.; Rintel, S.; Banks, R.; Wilson, N. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 26 April–1 May 2025; pp. 1–22.
107. Kosmyna, N.; Hauptmann, E.; Yuan, Y.T.; Situ, J.; Liao, X.H.; Beresnitzky, A.V.; Braunstein, I.; Maes, P. Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task. *arXiv* **2025**, arXiv:2506.08872.
108. Bainbridge, L. Ironies of Automation. In *Analysis, Design and Evaluation of Man–Machine Systems*; Elsevier: Amsterdam, The Netherlands, 1983; pp. 129–135.
109. Haupt, A.; Brynjolfsson, E. Position: AI Should Not Be An Imitation Game: Centaur Evaluations. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge MA, USA, 2025; pp. 81526–81541.
110. Messeri, L.; Crockett, M.J. Artificial Intelligence and Illusions of Understanding in Scientific Research. *Nature* **2024**, *627*, 49–58. [CrossRef] [PubMed]
111. Simon, H.A. *The New Science of Management Decision*, rev. ed.; Prentice-Hall: Englewood Cliffs, NJ, USA, 1977.
112. Mintzberg, H.; Raisinghani, D.; Théorêt, A. The Structure of "Unstructured" Decision Processes. *Adm. Sci. Q.* **1976**, *21*, 246–275. [CrossRef]
113. Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; Weld, D.S. Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11405–11414. [CrossRef]
114. Meadows, D.H. *Thinking in Systems: A Primer*; Chelsea Green Publishing: London, UK, 2008.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.