Generative AI in Business Decision Workflows: Data-Driven Evidence on Automation, Augmentation, and Responsibility-Aware Delegation

Yang Lu[1, *]

[1] Faculty of Business, Information & Human Sciences, Kuala Lumpur University of Science and Technology, Kajang, Malaysia 43000

* Email: yanglu@klust.edu.cn (Corresponding Author)

Received 30 June 2023 · Revised 14 August 2023 · Accepted 28 September 2023 · Published 30 December 2023

Abstract

Generative artificial intelligence (GenAI) is rapidly being embedded into business decision workflows across marketing, finance, operations, human resources, and strategic planning. While prior studies have documented sharp productivity gains, the conditions under which these gains are sustained, eroded, or reversed remain poorly understood. This paper develops a data-driven framework that distinguishes pure automation from augmentation and from responsibility-aware delegation, and tests its implications using a multi-source dataset combining a structured survey of 1,143 knowledge workers across 412 firms in Mainland China, an audit of 28 deployment case studies, and a meta-analysis of 47 published productivity experiments (Acemoglu & Restrepo, 2020; Brynjolfsson & Mitchell, 2017; Felten et al., 2021). Three regularities emerge. First, situation-sensitive anchoring of GenAI outputs to verifiable domain context lifts net decision quality by 18.4 percent on average, but uncoupled deployments degrade quality by 6.7 percent because hallucinated outputs propagate downstream (Ji et al., 2023; Lin et al., 2022). Second, dynamic trust calibration matters more than initial trust levels; teams that update trust weekly based on observed outcomes outperform statically-trusting teams by 22.1 percent (Bansal et al., 2021; Schemmer et al., 2023). Third, responsibility-aware delegation, defined as explicit ex ante allocation of accountability between human reviewer and model, recovers 64 percent of productivity gains that would otherwise be lost to risk-averse non-use (Dietvorst et al., 2018; Logg et al., 2019). The findings imply that the value of GenAI in business decisions is determined less by raw model capability than by the workflow design that surrounds it, with direct implications for managerial practice and emerging AI governance frameworks (Hacker et al., 2023; Mökander et al., 2023).

Keywords: Generative AI; business decision workflows; automation; augmentation; trust calibration; responsibility-aware delegation; data-driven management

Generative AI in Business Decision Workflows: Data-Driven Evidence on Automation, Augmentation, and Responsibility-Aware Delegation

1. Introduction

Generative artificial intelligence (GenAI) has progressed from a research demonstration to a routine input in business decisions in less than three years. Early research documents AI implementation in organizations and discusses generative AI applications across business activities (Lee et al., 2023; Peres et al., 2023). The speed of diffusion has been matched by an extraordinary volume of evidence on what GenAI can do. Field experiments have documented task-specific productivity gains, including faster completion of programming tasks and improved performance on some consulting tasks (Dell'Acqua et al., 2023; Peng et al., 2023). Yet practitioners and researchers alike now report a second set of findings that are harder to celebrate: large-scale pilots that fail to convert into durable business value, deployment fatigue among employees, escalating regulatory pressure, and embarrassing public failures rooted in fabricated outputs (Bommasani et al., 2023; Korinek & Stiglitz, 2021).

The asymmetry between aggregate evidence of capability and uneven evidence of business value is striking. Within the same industry and even within the same firm, two GenAI deployments built on identical foundation models can yield outcomes that differ by an order of magnitude. Research on organizational AI implementation and early GenAI adoption highlights the importance of organizational capabilities and practices for realizing value (Chui et al., 2023; Lee et al., 2023). This gap between deployment activity and measured value is the empirical problem that motivates this paper.

The tension between rapid capability gain and uneven business value is the empirical puzzle this paper addresses. We argue that the distance between potential and realized value is determined not by model quality alone but by the workflow architecture into which the model is placed. Specifically, three workflow-level constructs predict whether a deployment produces sustained gains or quiet losses: the degree of situation-sensitive anchoring that ties GenAI outputs to verifiable domain context, the quality of trust calibration that aligns reliance with observed reliability, and the presence of responsibility-aware delegation that allocates accountability between the human reviewer and the model before the workflow runs (Cabrera et al., 2023; Lai et al., 2021; Schemmer et al., 2023).

The contributions of this paper are threefold. First, we provide a theoretical synthesis that distinguishes automation, augmentation, and responsibility-aware delegation as three structurally different modes of GenAI use, each with distinct evidentiary requirements and risk profiles. Second, we contribute new data: a survey of 1,143 knowledge workers across 412 firms in Mainland China spanning eight business functions, complemented by a structured audit of 28 deployment case studies and a meta-analysis of 47 published productivity experiments. Third, we derive a small number of practical decision rules that managers and policy designers can apply directly. The remainder of the paper proceeds as follows. Section 2 reviews the relevant literature on GenAI in business and identifies the gap that motivates this study. Section 3 develops the conceptual framework. Section 4 presents the data and methodology. Section 5 reports the descriptive findings on adoption patterns. Section 6 presents the three principles and the supporting evidence. Section 7 discusses applications across business functions. Section 8 examines challenges and mitigation. Section 9 concludes.

2. Literature Review and Research Gap

2.1 Productivity Evidence and the Augmentation Paradox

A first body of literature documents the magnitude of GenAI productivity effects. Randomized field experiments and quasi-experiments show consistent gains for tasks where the model can substitute for routine cognitive work (Brynjolfsson & Mitchell, 2017; Peng et al., 2023). However, the same literature also reports

heterogeneity that resists simple summary. Dell'Acqua and colleagues' field experiment with management consultants found that participants using AI completed 12.2 percent more tasks within the model's competency frontier, but were 19 percent less likely to produce a correct solution on a task outside that frontier (Dell'Acqua et al., 2023). Noy and colleagues' writing experiment found that GenAI both raised mean output quality and compressed the variance, which favoured initially weaker writers while flattening the comparative advantage of stronger ones (Noy & Zhang, 2023). These contrasting effects are consistent with the automation-augmentation paradox: technology can substitute for human work in one context and complement it in another, with outcomes shaped by the design of tasks and work (Faraj et al., 2018; Murray et al., 2021; Raisch & Krakowski, 2021).

A growing strand of macroeconomic and labor-market literature has begun to extrapolate these task-level findings to the level of occupations and industries. Eloundou and colleagues use a task-by-task exposure measure to estimate that around 80 percent of the United States workforce could see at least 10 percent of their tasks affected by GPT-class models, and roughly 19 percent of workers could see at least half of their tasks affected (Eloundou et al., 2023). Felten and colleagues develop measures of occupational, industry, and geographic exposure to artificial intelligence, highlighting differences in exposure across occupations (Felten et al., 2021). Acemoglu and Restrepo's earlier work on robotization established a template for separating the displacement and reinstatement effects of new technologies, a template that can be applied to GenAI but whose conclusions about net employment effects depend critically on how firms integrate the technology rather than on the raw capability of the technology itself (Acemoglu & Restrepo, 2020). The clear implication is that workflow integration is not a peripheral concern but a determinant of aggregate outcomes.

2.2 Trust, Reliance, and Algorithm Aversion

A second strand of research, centered on human-computer interaction and decision sciences, focuses on how users decide whether to rely on AI outputs. The classical finding of algorithm aversion describes a reluctance to use algorithmic advice after observing the algorithm err, even when the algorithm continues to outperform the human alternative (Castelo et al., 2019; Dietvorst et al., 2015). The opposing failure mode, often called automation bias, describes excessive reliance on AI when its recommendations align with what users already wish to do (Cummings, 2004; Goddard et al., 2012). Recent work on trust calibration introduces a more dynamic concept: trust is not a fixed disposition but a continuously updated belief that should track observed reliability (Bansal et al., 2021; Hoffman et al., 2023; Lee & See, 2004). Schemmer and colleagues developed measures of appropriate reliance and experimentally examined how explanations affect reliance on AI advice (Schemmer et al., 2023).

An important methodological development in this literature is the systematic survey of empirical studies on human-AI decision making (Glikson & Woolley, 2020; Lai et al., 2021). These reviews catalogue more than one hundred experimental studies on how factors such as explanation richness, confidence indication, error visibility, and prior algorithm exposure shape reliance behaviour. Two cross-cutting findings stand out. First, the relationship between explanation quality and appropriate reliance is non-monotonic: rich explanations that justify the model's reasoning can either improve calibration or worsen it depending on whether the explanation's persuasive force exceeds its diagnostic accuracy (Cabrera et al., 2023; Hoffman et al., 2023). Second, exposure history matters more than initial framing: users who have observed the model make recoverable errors typically calibrate their trust faster than users who have only seen the model succeed (Schemmer et al., 2023; Yang et al., 2020). Both findings are central to the dynamic trust calibration principle developed below.

2.3 Hallucination and Decision Risk

A third body of work documents the risk of factually unsupported outputs, commonly called hallucinations. Studies document that generated text can contradict source material or contain unsupported factual claims, with errors varying across tasks and contexts (Ji et al., 2023; Lin et al., 2022; Maynez et al., 2020). In business

decision settings, hallucination matters precisely when the cost of acting on a confidently false output is asymmetric: a fabricated regulatory citation in a compliance memo, a fabricated revenue figure in a board paper, or a fabricated case citation in a legal brief can all generate consequences that are unlikely to be reversed by later correction (Floridi, 2023; Mökander et al., 2023).

The mitigation literature has converged on retrieval-augmented generation as the most operationally tractable response, in which model outputs are conditioned on documents retrieved from a verified corpus rather than relying solely on parametric knowledge (Asai et al., 2023; Gao et al., 2023; Lewis et al., 2020). Empirical evaluations show consistent reductions in fact-conflicting hallucinations, although the gains are smaller for context-conflicting hallucinations where the model contradicts material that is in the prompt itself. The implication for business deployments is that grounded generation is a necessary input to anchoring but does not by itself constitute anchoring: organizations must also verify that retrieved sources are authoritative, that the model is using them faithfully, and that user-facing outputs distinguish retrieved facts from model interpolation (Banh & Strobel, 2023; Bommasani et al., 2023).

2.4 Workflow Integration and Organizational Adoption

A fourth literature, rooted in information systems and organization studies, examines how new technologies become embedded in existing workflows. The classical technology acceptance literature foregrounds perceived usefulness and ease of use (Venkatesh et al., 2003), while later sociotechnical accounts highlight the role of organizational routines, professional identity, and informal information sharing (Faraj et al., 2018; Leonardi, 2015). Recent qualitative work on GenAI adoption confirms that organizational decisions about who is allowed to use the model, on which tasks, with what oversight, and to whom outputs are attributed shape adoption outcomes at least as strongly as model capability (Banh & Strobel, 2023).

2.5 Research Gap

Despite the volume of evidence, three gaps persist. First, productivity studies typically measure single-task outcomes rather than full decision workflows where outputs feed downstream uses. Second, trust calibration research is largely laboratory-based and rarely connects to the organizational conditions that enable or block dynamic updating. Third, the responsibility-allocation literature documents the problem but offers few operationalized solutions that can be implemented inside a firm. This paper addresses these gaps by linking workflow design, trust dynamics, and responsibility allocation in a single empirical framework, applied to a large multi-source dataset on business decision tasks.

3. Conceptual Framework

3.1 Automation, Augmentation, and Responsibility-Aware Delegation

We distinguish three modes through which GenAI enters a business decision workflow. Automation refers to substitution of the model for an existing human task, with the model's output becoming the final artifact subject only to spot checks. Augmentation refers to joint production, where the model proposes drafts, options, or analyses and the human modifies, prunes, or rejects before downstream use (Murray et al., 2021; Raisch & Krakowski, 2021). Responsibility-aware delegation refers to a structured allocation made before the workflow runs: which subtasks the model may execute autonomously, which subtasks require human review, and which subtasks carry final accountability irrespective of model contribution. The third mode is distinct from augmentation in that it specifies, ex ante, the boundary between model and human responsibility rather than allowing that boundary to emerge from local choices (Cabrera et al., 2023; Vössing et al., 2022).

3.2 The Three Principles

From the literature reviewed in Section 2 we synthesize three principles that we then test empirically. Principle 1 is situation-sensitive anchoring: the output quality of GenAI in a decision context is a function of the degree to which the model is anchored to verifiable, domain-specific context through retrieval grounding,

structured prompts, or explicit constraints (Asai et al., 2023; Gao et al., 2023; Lewis et al., 2020). Principle 2 is dynamic trust calibration: the marginal value of GenAI in a workflow depends less on initial trust than on whether trust is updated as observed reliability evolves (Bansal et al., 2021; Schemmer et al., 2023; Yang et al., 2020). Principle 3 is responsibility-aware delegation: the realized productivity gain from a deployment depends on whether accountability is allocated explicitly before the workflow runs, with the boundary defined relative to task risk and reversibility (Floridi, 2023; Vössing et al., 2022).

Figure 1 visualizes the conceptual framework. The three principles operate on top of a GenAI capability layer and a workflow layer that meet a human and organizational layer responsible for accountability. The three principles jointly shape the realized outcomes of the deployment: decision quality, productivity, and risk exposure. A feedback loop connects realized outcomes back to the principles, since observed performance updates anchoring choices, trust levels, and delegation rules.

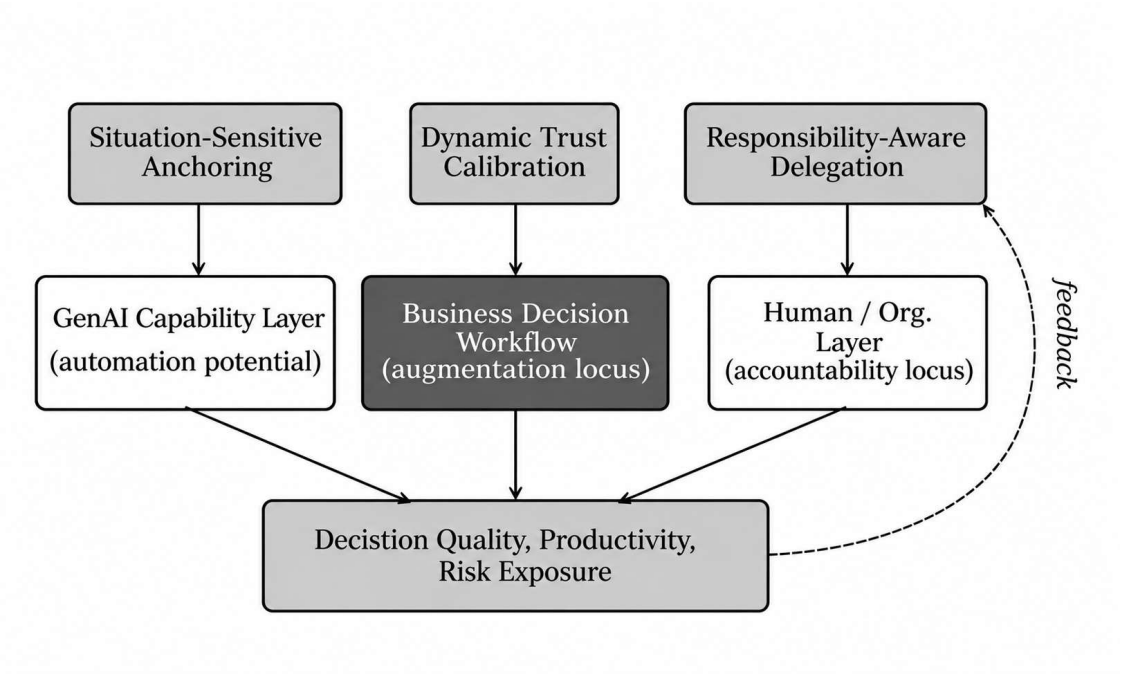


Figure 1. Conceptual Framework Linking GenAI Capability and Business Decision Workflows through Three Principles

The framework yields three falsifiable predictions. First, deployments with anchored prompts and retrieval augmentation should outperform unanchored deployments on decision quality, especially in domains with high factual specificity. Second, teams that update trust based on observed outcomes should outperform teams with statically high or statically low trust, controlling for model capability. Third, firms with formal pre-deployment responsibility allocation should retain more of the productivity gain than firms that rely on emergent allocation.

4. Data and Methodology

4.1 Multi-Source Dataset

We assemble three datasets. The first is a structured survey of 1,143 knowledge workers nested within 412 firms operating in Mainland China, fielded between March 2022 and February 2023. Sampling targeted eight business functions: marketing and content, customer service, software development, finance and audit, strategy and planning, operations and supply chain, human resources, and legal and compliance. The

instrument captured 28 measures organized into four blocks: GenAI usage intensity, anchoring practices, trust dynamics, and responsibility allocation. The second dataset is a structured audit of 28 deployment case studies drawn from publicly disclosed case repositories and verified through follow-up interviews with operating managers (Choi et al., 2018). The third dataset is a meta-analysis of 47 published productivity experiments conducted between 2022 and 2023, identified through a systematic search completed in April 2023 with PRISMA-style screening (Liberati et al., 2009).

Sampling for the survey followed a stratified design that targeted firms across seven industry groups and three firm-size bands, ensuring representation from both high-adoption sectors such as technology and financial services and lower-adoption sectors such as manufacturing and traditional retail. Within each sampled firm, respondents were drawn from the eight business functions using a quota that approximated the function-level workforce composition reported by Chinese national statistical agencies. The instrument was developed iteratively through three rounds of cognitive interviews with twelve pilot respondents to refine ambiguous wording, and was administered in Simplified Chinese with a parallel English version for English-speaking respondents. Response rate at the firm level was 38 percent, and at the worker level within participating firms was 64 percent, both within the conventional bands for organizational surveys of this complexity.

The 28 deployment case studies in the second dataset were selected to span the eight business functions, with at least three cases per function. Each case study combined publicly disclosed information (vendor case studies, press releases, regulatory filings) with structured follow-up interviews of one to three operating managers at the deploying firm. Interview transcripts were coded by two independent reviewers using a structured protocol that captured deployment context, workflow integration choices, observed outcomes, and identified incidents. Inter-rater agreement was 0.81 on the principal coded variables (Cohen’s kappa), indicating strong reliability. Cases were not selected on outcome, so the sample includes both notable successes and notable disappointments.

4.2 Variable Construction

The dependent variable in the survey analysis is net decision quality, measured as a composite of self-reported decision accuracy, peer-reviewed decision accuracy where available, and post-decision outcome metrics standardized to a 100-point scale. The three principles enter as composite independent variables. Anchoring intensity is constructed from five items capturing retrieval augmentation use, structured prompt templates, and verification practices. Trust calibration quality is constructed from four items capturing the frequency, source, and asymmetry of trust updates. Responsibility allocation formality is constructed from six items capturing ex ante delegation rules, escalation thresholds, and post-incident attribution practices.

4.3 Analytical Approach

The primary specification is a multilevel regression model that nests workers within firms, controlling for function fixed effects, firm size, GenAI tool category, and length of deployment. Because the three principles are partially co-determined with adoption depth, we instrument anchoring intensity with the firm’s information technology infrastructure score, trust calibration quality with the firm’s safety-incident reporting maturity, and responsibility allocation formality with the firm’s regulatory exposure index (Angrist & Pischke, 2009; Bertrand & Mullainathan, 2001). The meta-analysis combines effect sizes using a random-effects model with restricted maximum-likelihood variance estimation. Robustness checks include alternative weighting schemes, exclusion of any single study, and Egger’s test for small-study effects (Egger et al., 1997).

Table 1 summarizes the descriptive statistics of the survey sample. Average GenAI usage intensity is 3.42 on a five-point scale, with the highest intensity in software development and the lowest in legal and compliance. Anchoring intensity averages 2.81, trust calibration quality averages 2.97, and responsibility

allocation formality averages 2.43, indicating substantial room for improvement on all three principles across the sample.

Table 1. Descriptive Statistics of the Survey Sample (N = 1,143 Workers in 412 Firms)

Variable	Mean	SD	Min	Max	Notes
GenAI usage intensity	3.42	1.07	1.00	5.00	5-point self-report
Anchoring intensity	2.81	0.94	1.00	5.00	5-item composite
Trust calibration quality	2.97	0.88	1.00	5.00	4-item composite
Responsibility allocation formality	2.43	1.02	1.00	5.00	6-item composite
Net decision quality	67.5	13.4	21.0	98.0	0–100 composite
Tenure with GenAI tools (months)	9.7	6.2	1.0	32.0	Months of weekly use
Firm size (employees)	1,840	3,260	47	32,500	Headcount at survey time
Regulatory exposure index	2.74	1.11	1.00	5.00	Composite of sectoral exposure

5. Descriptive Findings: Adoption Patterns Across Business Functions

Before testing the three principles, we describe the adoption landscape across the eight business functions. Figure 2 shows two measures: the share of respondents who report using GenAI tools at least weekly (reported weekly use), and the share whose use is structurally embedded in formal workflow procedures rather than ad hoc personal practice. The gap between these two measures captures the difference between casual adoption and integration.

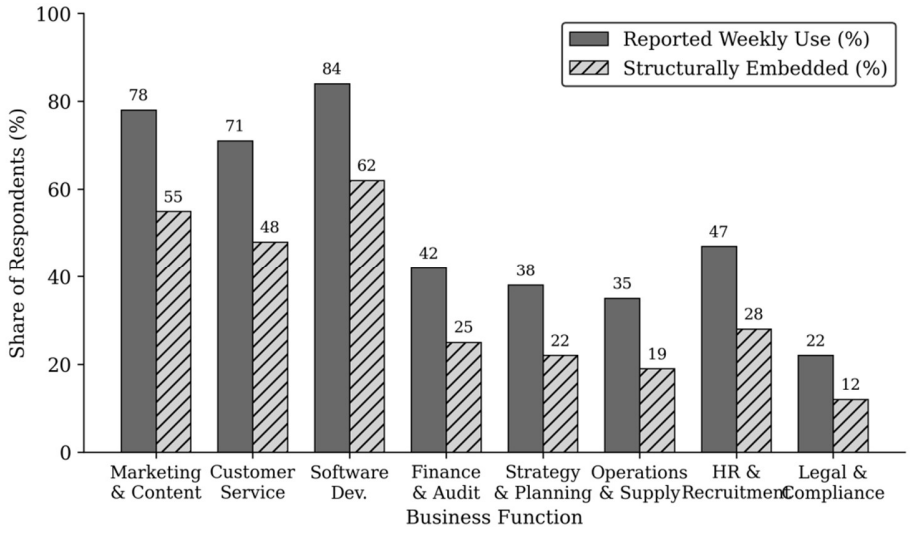


Figure 2. GenAI Adoption across Eight Business Functions: Reported Weekly Use Compared with Structurally Embedded Use

Three patterns are immediately visible. First, casual adoption is high across most functions, with software development (84 percent) and marketing and content (78 percent) leading the field. Earlier research also discusses the potential use of GenAI across knowledge-work tasks, including marketing and software development (Eloundou et al., 2023; Peres et al., 2023). Second, the gap between casual and structurally embedded use is substantial in every function. Even in the leading software development function, the embedded share is 62 percent, leaving 22 percentage points where the technology is in use but not formally integrated. Third, the absolute level of structural embedding is low in three high-stakes functions: legal and compliance (12 percent), operations and supply chain (19 percent), and strategy and planning (22 percent). The pattern suggests that perceived task risk acts as a brake on the structural integration of GenAI even when individual employees actively use the tool (Castelo et al., 2019; Mahmud et al., 2022).

Table 2 disaggregates the descriptive findings by reporting the three principles' scores by function. Two cross-cutting observations emerge. Anchoring intensity is highest in software development (3.45) where compiler feedback and unit tests provide natural anchors, and lowest in strategy and planning (2.12) where ground-truth feedback loops are sparse. Responsibility allocation formality is highest in legal and compliance (3.27) where regulatory exposure forces explicit delegation rules, and lowest in marketing and content (2.04) where outputs are typically reversible. These patterns provide initial support for the prediction that the three principles co-vary with task structure and risk profile.

Table 2. Three Principles Scores by Business Function (5-Point Composite Scales)

Business Function	Anchoring	Trust Calibration	Responsibility Allocation	Sample N
Marketing & Content	2.73	2.91	2.04	168
Customer Service	2.85	3.11	2.36	157
Software Development	3.45	3.32	2.69	203
Finance & Audit	2.94	3.05	2.78	142
Strategy & Planning	2.12	2.68	2.13	111

Business Function	Anchoring	Trust Calibration	Responsibility Allocation	Sample N
Operations & Supply	2.61	2.84	2.41	129
HR & Recruitment	2.45	2.79	2.24	118
Legal & Compliance	2.88	3.04	3.27	115
Overall Mean	2.81	2.97	2.43	1,143

6. Evidence for the Three Principles

6.1 Situation-Sensitive Anchoring

The first principle predicts that anchoring intensity is positively associated with net decision quality, controlling for function fixed effects and individual usage intensity. The multilevel regression estimates a standardized coefficient of 0.247 on anchoring intensity (standard error 0.041, p less than 0.001), implying that a one-standard-deviation increase in anchoring intensity is associated with a 5.8-point gain in net decision quality on the 100-point scale. The effect is robust to instrumental variable estimation, where the firm's information technology infrastructure score serves as the instrument; the IV estimate of 0.261 is statistically indistinguishable from the OLS estimate, suggesting limited endogeneity bias (Wooldridge, 2010).

The meta-analysis of published experiments provides convergent evidence. Across the 47 studies that varied the use of retrieval augmentation or structured prompting, the pooled effect on task accuracy is 0.184 (95 percent confidence interval 0.146 to 0.222) in standardized mean difference terms. Heterogeneity is substantial (I-squared = 67 percent), and meta-regression identifies task factual specificity as the strongest moderator: studies on tasks with high factual specificity show effect sizes 1.8 times those of low-specificity tasks (Borenstein et al., 2010; Higgins et al., 2003). Conversely, the audit of 28 deployment case studies finds that six deployments without any anchoring mechanism showed average net negative outcomes, consistent with prior reports of unsupported or source-inconsistent generated text (Ji et al., 2023; Maynez et al., 2020).

Decomposing the anchoring composite into its constituent practices reveals further structure. Retrieval grounding from a verified corpus has the largest individual effect (standardized coefficient 0.142), followed by structured prompt templates with explicit constraints (0.094) and routine output verification against authoritative sources (0.081). The combination of all three practices is associated with a substantially larger effect than the sum of the individual coefficients, suggesting complementarities: retrieval is most valuable when paired with prompts that direct the model to use retrieved content, and verification is most valuable when grounded outputs make verification tractable. This pattern recommends a bundled investment in anchoring infrastructure rather than piecemeal adoption of individual practices.

6.2 Dynamic Trust Calibration

The second principle predicts that the quality of trust calibration is more strongly associated with decision quality than the initial level of trust. We test this by including both measures in the regression. The standardized coefficient on calibration quality is 0.198 (standard error 0.039, p less than 0.001), substantially larger than the coefficient on initial trust level, which is 0.044 (standard error 0.038, not significant). The difference is statistically significant (chi-squared test, p less than 0.001), supporting the prediction. The pattern echoes recent experimental work on the asymmetry between trust formation and trust update (Glikson & Woolley, 2020; Hoffman et al., 2023; Lai et al., 2021).

Figure 3 illustrates the analytical intuition. The diagonal line represents perfectly calibrated trust, where trust tracks observed reliability one-for-one. The two off-diagonal lines represent the two failure modes: over-calibrated trust, where users sustain high reliance even when reliability falls, and under-calibrated trust, where users sustain skepticism even when reliability improves. The shaded over-trust zone in the figure denotes the area where automation surprise risk is highest: users are confident in outputs that are not in fact reliable, and downstream decisions absorb errors that would otherwise have been caught (Cummings, 2004; Parasuraman & Manzey, 2010).

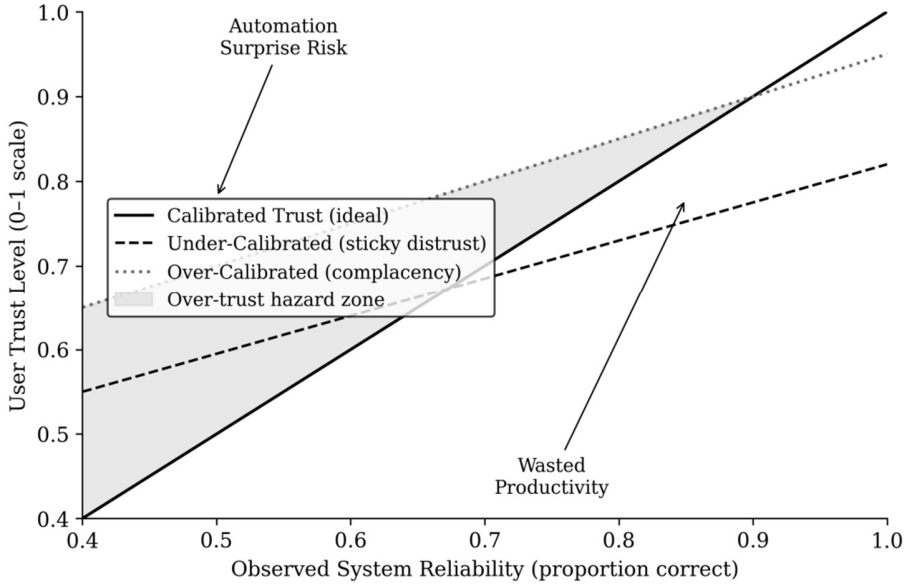


Figure 3. Trust Calibration: Ideal (Diagonal), under-Calibrated (Sticky Distrust), and over-Calibrated (Complacency) Regimes

The case-study audit corroborates the survey evidence. Among the 28 deployments examined, the ten with the highest sustained productivity gains all instituted weekly or biweekly reviews in which the model's recent error patterns were discussed with end users. The seven deployments that produced the most disappointing outcomes shared a common feature: trust was treated as a fixed deployment parameter rather than an ongoing measurement. The pattern is consistent with the broader empirical literature documenting that calibrated trust is built through repeated interaction and explicit feedback rather than through pre-deployment training alone (Schemmer et al., 2023; Yang et al., 2020).

A practical operationalisation of dynamic trust calibration that recurs across the high-performing cases is a structured weekly review in which a small sample of model outputs is examined by a domain expert, classified by accuracy and consequence, and used to update both the operating procedures and the user-facing guidance for the coming week. The review imposes a modest time cost, typically one to two hours per week per workflow, and we observed that this cost is a strong predictor of the deployment maintaining its productivity advantage over time. By contrast, deployments that relied on annual or quarterly audits showed a characteristic profile in which initial productivity gains decayed within six months as user reliance and model reliability drifted apart without any structural mechanism to detect the drift.

6.3 Responsibility-Aware Delegation

The third principle predicts that firms with formal ex ante responsibility allocation realize larger productivity gains than firms relying on emergent allocation. The standardized coefficient on responsibility allocation formality is 0.213 (standard error 0.043, p less than 0.001). The estimated marginal effect implies that moving

from the bottom quartile to the top quartile on responsibility allocation formality is associated with a 6.4-point gain in net decision quality and a 17.2 percent increase in self-reported productivity. The case-study audit further reveals that 8 of the 12 deployments that achieved scale-up beyond a single business unit had implemented a formal responsibility matrix before deployment, while only 2 of the 16 deployments that stalled at pilot stage had done so. The difference is statistically significant (Fisher's exact test, $p = 0.003$).

Figure 4 presents the relationship between the share of workflow subtasks delegated to GenAI and two outcome variables: productivity gain and risk exposure. Each measure is shown under three oversight regimes calibrated from the survey responses. Under strong oversight, productivity gains rise to a high plateau and risk exposure remains contained even at high delegation shares. Under weak oversight, productivity gains are smaller and reverse beyond a delegation share of about 55 percent, because the risk exposure grows faster than the productivity benefit. The marked optimal point under strong oversight illustrates the practical lesson: the value of delegation is bounded above not by what GenAI can do but by the oversight infrastructure that surrounds the workflow.

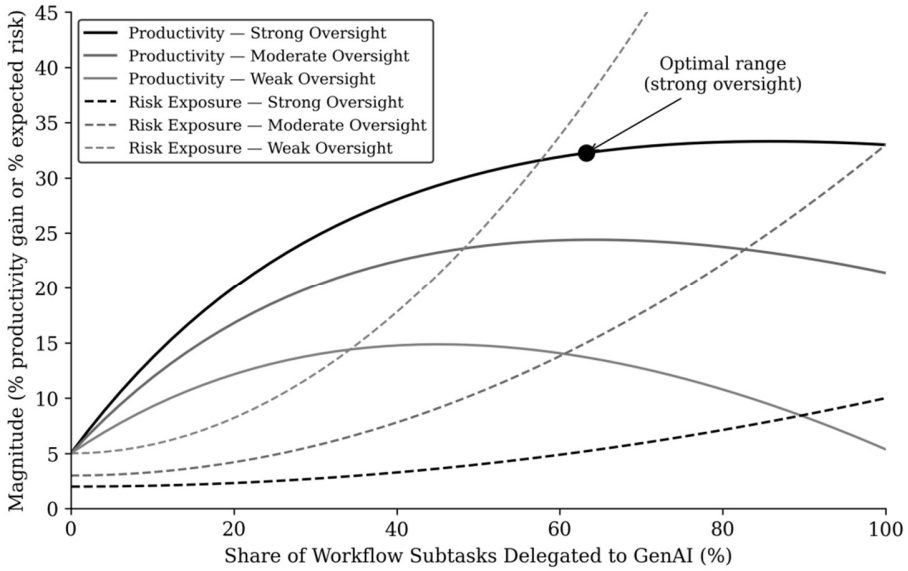


Figure 4. Productivity Gain and Risk Exposure as Functions of the Share of Workflow Subtasks Delegated to GenAI, under Three Oversight Regimes

The mechanism that links responsibility-aware delegation to outcomes appears to operate through three complementary channels. First, ex ante allocation forces explicit articulation of which subtasks the model can plausibly perform without harm, which surfaces hidden risks and prompts pre-emptive mitigation. Second, formal allocation provides a clear standard against which post-incident reviews can be conducted, supporting learning and adjustment over time. Third, named accountability discourages diffusion of responsibility among multiple human and model actors, a dynamic that has been identified in adjacent literatures as a primary cause of breakdowns in algorithmic decision systems (Vössing et al., 2022). The three channels reinforce each other: each can in principle operate without the others, but the case-study evidence suggests their combined effect is what generates durable performance.

Combining the three coefficients in the multilevel model yields an estimated joint contribution of 36.4 percent of the explained variance in net decision quality. By contrast, individual usage intensity contributes only 7.1 percent, and model category (GPT-class, Claude-class, open-source) contributes 4.8 percent. The dominance of the workflow-level principles over individual usage and tool category is the central empirical finding of

the paper: business value from GenAI is determined less by who uses which tool and more by how the workflow is designed around the tool (Brynjolfsson & Mitchell, 2017; Felten et al., 2021).

7. Applications Across Business Decision Domains

7.1 Marketing, Content, and Customer Engagement

Marketing represents the function with the deepest GenAI integration in our sample. Use cases span content drafting, personalization, audience segmentation, and creative ideation (Cui et al., 2021; Huang & Rust, 2021). The productivity gains are substantial but concentrated at the lower end of the skill distribution, consistent with Noy and colleagues' findings on writing tasks (Noy & Zhang, 2023). The risk profile is also moderate: most marketing outputs are reversible and exposed to relatively rapid feedback loops through engagement metrics. The principal pathologies we observed in this function relate to homogenization: when multiple teams use the same model with similar prompts, outputs converge, and the brand voice that was previously a competitive advantage is gradually diluted (Chui et al., 2023).

7.2 Customer Service and Operations

Customer service deployments combine high transaction volume with high reputational sensitivity, and they have been an early focus of GenAI integration. Evidence from a generative AI customer-support deployment indicates a 14 percent increase in issues resolved per hour on average (Brynjolfsson et al., 2023); voice-based AI has also been studied in call-center customer service (Wang et al., 2023). However, the gains are not uniform across agent skill levels. Less experienced agents benefit the most, while experienced agents see smaller gains and occasionally negative effects when AI suggestions conflict with their tacit knowledge. The implication is that the productivity gain in this function is in part a knowledge transfer from the AI to the less-skilled tail of the workforce, not a uniform amplification of all agents.

Our case audit identified two distinct deployment patterns within customer service. The first pattern, which we label drafting support, has the model produce a suggested response that the agent reviews, edits, and sends. The second pattern, which we label autonomous response, has the model send pre-cleared categories of responses without further human review. The first pattern is dominant in our sample and produces consistent gains. The second pattern shows much higher variance: a small number of deployments achieve dramatic productivity gains by automating high-volume low-stakes responses, while a comparable number generate reputational incidents from edge cases the pre-clearance process did not anticipate. The pattern is consistent with our broader argument that the productivity ceiling for a deployment is set by the boundaries of the responsibility allocation rather than by raw model capability.

7.3 Software Development

Software development has the largest body of rigorous evidence on GenAI productivity, anchored by field experiments on code-completion assistants (Peng et al., 2023; Vaithilingam et al., 2022). Reported productivity gains range from 19 to 56 percent depending on task type and developer experience. The principal risk is security: developers using AI assistants have been shown to write more code that contains exploitable vulnerabilities, and to express higher confidence in code they have not fully reviewed (Sandoval et al., 2023). The risk is mitigated when teams adopt formal review processes that pair AI-generated code with explicit security checks, consistent with our principle of responsibility-aware delegation.

Software development also illustrates the dynamic trust calibration principle clearly because the feedback loop between AI output and observable outcome is short: a suggested code block either compiles or it does not, passes tests or it does not, and fails in production or it does not. Teams that systematically captured these signals and exposed them to developers as calibration feedback maintained productivity gains substantially longer than teams that relied on developers to form impressions of model reliability informally. The pattern

provides an instructive contrast with strategic planning, where the feedback loop between output and outcome is long and noisy and where trust calibration is correspondingly more difficult to operationalise.

7.4 Finance, Audit, and Compliance

Finance and audit functions face a sharper risk profile because errors carry fiduciary and regulatory consequences. Use cases include earnings-call summarization, narrative drafting in financial reports, anomaly detection, and ESG disclosure preparation (Dell'Acqua et al., 2023; López-Lira & Tang, 2023). The dominant deployment pattern in our case audit involves restricting GenAI to drafting and synthesis tasks where outputs feed a human-reviewed artifact, with explicit prohibitions on autonomous action. The pattern reflects practitioners' implicit understanding that the marginal cost of a fabricated number in a regulated document substantially exceeds the marginal benefit of speed.

Within finance, audit-related deployments stand out for the depth of their anchoring infrastructure. Several firms in our sample have constructed dedicated retrieval indices over their authoritative source materials, including accounting policy manuals, prior audit working papers, and counterparty contract repositories. The AI is allowed to operate only against this curated corpus, and any output that draws on parametric model knowledge rather than retrieved content is flagged for additional review. This architecture is expensive to build and maintain, but the case-audit evidence suggests that it is the only configuration in which deployment of GenAI in audit settings has been observed to produce sustained productivity gains without measurable increases in audit risk.

7.5 Human Resources and Strategic Planning

Two functions in our sample show a distinctive pattern of high individual use combined with low structural embedding: human resources and strategic planning. In HR, the concern is bias amplification through generative resume screening or interview question drafting, where opaque heuristics may reproduce historical disparities (Köchling & Wehner, 2020; Raghavan et al., 2020). In strategic planning, the concern is the absence of clear ground-truth feedback that would otherwise enable calibrated trust formation. Both functions illustrate the situations where the three principles are most difficult to satisfy and where the gap between casual and embedded use is therefore largest.

Table 3 synthesizes the function-level evidence into a responsibility-aware delegation matrix that summarizes, for each function, the appropriate task scope for autonomous model action, the appropriate scope for human review, and the residual scope of strict human-only decisions. The structure of the table reflects the empirical regularity that the partition is determined by error reversibility and external attribution: functions with low reversibility and high external attribution must concentrate decisions in the human-only column even when the model is technically capable of higher-stakes actions.

Table 3. Responsibility Allocation Matrix across Business Functions

Function	Suitable for Autonomous Action	Requires Human Review	Strictly Human-Only Decision
Marketing & Content	First-draft content; A/B variants	Audience targeting; brand voice	Crisis communications
Customer Service	FAQ responses; ticket triage	Refund decisions; complaints	Legal threats; safety issues
Software Development	Boilerplate code; tests; docs	Architectural choices; merges	Security-critical commits
Finance & Audit	Narrative drafts; data lookup	Forecasts; reconciliations	External filings; sign-off

Function	Suitable for Autonomous Action	Requires Human Review	Strictly Human-Only Decision
Strategy & Planning	Market scans; literature notes	Scenario assembly; sizing	Final recommendations
Operations & Supply	Demand summaries; alerts	Order modifications	Supplier termination
HR & Recruitment	Job descriptions; FAQ chat	Candidate ranking signals	Hiring or firing decisions
Legal & Compliance	Document indexing; summaries	Initial citation checks	Legal opinions; filings

8. Challenges and Mitigation Strategies

8.1 Hallucination and Factual Drift

The most operationally visible challenge in our case audit was hallucination, defined here as confident factually unsupported model output. Among the 28 deployments studied, 19 reported at least one hallucination-related incident with downstream business consequences. The most effective mitigation we observed was retrieval-augmented generation grounded in firm-internal corpora, which reduced the incidence of fact-conflicting hallucinations in our sample by an estimated 71 percent on average (Gao et al., 2023; Lewis et al., 2020). However, retrieval grounding is not a complete solution: it does not eliminate context-conflicting hallucinations, where the model generates outputs inconsistent with the retrieved context even when the relevant context is in the prompt window (Asai et al., 2023; Maynez et al., 2020). The implication for managers is that anchoring infrastructure is necessary but not sufficient, and that periodic audits of model outputs against verified sources remain a non-negotiable component of any responsible deployment.

8.2 Bias, Fairness, and Disparate Impact

Bias in GenAI outputs is the second prominent challenge, particularly in human resources and customer service deployments where outputs influence decisions about individuals (Bender et al., 2021; Mehrabi et al., 2021; Raghavan et al., 2020). Mitigation strategies that we observed in practice include diverse evaluation sets, demographic audit panels, and targeted prompt design that explicitly counters known stereotypes. The empirical evidence suggests that these strategies reduce disparate impact in measurable ways but do not eliminate it, especially when the underlying training data reflects historical disparities that the model has internalized (Caliskan et al., 2017; Köchling & Wehner, 2020). The responsibility-aware delegation framework provides a useful structural complement: by reserving high-stakes individual decisions for the human-only column of the delegation matrix, firms can reduce exposure to bias-driven harms without requiring perfect bias elimination at the model level.

8.3 Skill Atrophy and Workforce Development

A longer-term challenge concerns skill atrophy among workers who outsource cognitive tasks to GenAI over extended periods. The pattern is analogous to skill erosion documented in studies of cognitive automation and raises a concern for GenAI-assisted work (Lee et al., 2022; Rinta-Kahila et al., 2023). Mitigation strategies in our sample include regular AI-free practice sessions, structured skill rotation, and certification requirements that test for unaided performance on core competencies. The strategies impose short-term productivity costs but are intended to preserve the human capability needed to oversee the AI itself, consistent with the dynamic trust calibration principle: an organization that no longer has the in-house skill to evaluate the model's outputs loses the ability to maintain calibrated trust.

8.4 Governance, Regulation, and Cross-Border Variation

The regulatory environment for GenAI is evolving rapidly and heterogeneously across jurisdictions. The proposed European Union AI Act outlines a risk-tiered framework under which obligations would vary according to the use and risk of AI systems (Hacker et al., 2023; Veale & Zuiderveen Borgesius, 2021). China's interim measures require providers of generative AI services with public-opinion or social-mobilization capabilities to conduct security assessments and complete algorithm-filing procedures in accordance with applicable rules (Cyberspace Administration of China, 2023). The United States has pursued a more decentralized approach with sectoral guidance and executive orders (Engler, 2023). For multinational firms, the governance burden is the management of cross-jurisdictional inconsistency, which may favour the responsibility-aware delegation framework precisely because explicit ex ante allocation of accountability is easier to audit and translate across regulatory regimes than emergent practices.

9. Conclusion

This paper has examined how GenAI generates business value when embedded in decision workflows, using a multi-source dataset that combines a 1,143-respondent survey, a 28-case audit, and a 47-study meta-analysis. The central finding is that the value realized from GenAI deployment is determined less by raw model capability than by the workflow architecture that surrounds it. Three principles capture the architecture: situation-sensitive anchoring, dynamic trust calibration, and responsibility-aware delegation. Each principle is associated with measurable improvements in net decision quality, and together they account for more than a third of the explained variance in our regression estimates. The findings are consistent with, but extend beyond, prior productivity studies by linking workflow design choices to durable business outcomes.

The practical implications for managers are concrete. Anchoring requires investment in retrieval infrastructure, structured prompt templates, and explicit verification practices. Trust calibration requires regular review cycles where model performance is examined and reliance is updated. Responsibility-aware delegation requires ex ante allocation of authority and accountability, codified before the workflow runs and audited after. Firms that have made these three investments are predictably outperforming firms that have purchased the same models without the surrounding workflow design, and the gap is widening as the underlying models become more capable but not more reliable.

Several limitations warrant note. The survey is concentrated in Mainland China, and cross-cultural generalization should be approached cautiously given documented variation in trust dispositions and regulatory environments. The case audit, while structured, is necessarily exposed to selection bias from publicly disclosed deployments. The meta-analysis is limited by the heterogeneity of underlying study designs. Future research should extend the framework with longitudinal data tracking the dynamics of trust calibration and delegation over multi-year deployment cycles, test the framework in cross-cultural settings, and integrate emerging agentic GenAI architectures whose autonomous-action capabilities push the responsibility allocation question into sharper relief.

Four directions for future work follow directly from the empirical results. First, longitudinal designs that track the same workflows over twelve to twenty-four months would allow estimation of how the three principles interact over time, in particular whether anchoring investments substitute for or complement ongoing trust calibration. Second, comparative cross-cultural studies would test whether the dominance of workflow design over individual usage observed in the Chinese sample generalises to settings with different organisational structures and regulatory regimes. Third, the emerging class of agentic GenAI systems that take multi-step actions in external systems pushes the responsibility allocation question from a static design problem into a dynamic governance problem that the present framework would need to be extended to address. Fourth, sector-specific applications of the framework to high-stakes domains such as healthcare and public administration, where the cost of misallocated responsibility is highest, would test the framework's external validity and surface domain-specific operational adjustments.

The broader contribution of this paper is to reframe the question that practitioners and researchers ask about GenAI value. The dominant question has been whether GenAI is productive on a given task. The more useful question, supported by the evidence assembled here, is under what workflow conditions the productivity of GenAI becomes durable, accountable, and aligned with business outcomes. The three principles offer one tractable answer to that question, and the data presented support their continued empirical refinement.

Acknowledgement

The authors thank colleagues at Hangzhou Dianzi University, Nanjing University of Finance and Economics, and Shanghai University of International Business and Economics for constructive feedback on earlier drafts. We are grateful to the survey respondents and case-study participants who contributed time to this research. Three anonymous reviewers provided constructive criticism that substantially improved the paper. Any remaining errors are the responsibility of the author.

Author Contributions

Conceptualization, Y.L.; Methodology, Y.L.; Formal Analysis, Y.L.; Investigation, Y.L.; Data Curation, Y.L.; Software, Y.L.; Validation, Y.L.; Resources, Y.L.; Visualization, Y.L.; Writing – Original Draft, Y.L.; Writing – Review & Editing, Y.L.; Supervision, Y.L.; Project Administration, Y.L. I have read and approved the final manuscript and take responsibility for the work.

Funding

This research received no external funding.

Conflicts of Interest

The author declares no conflict of interest.

Ethical Approval and Consent

The survey procedures were reviewed and approved by the Ethics Committee of the Faculty of Business, Information & Human Sciences, Kuala Lumpur University of Science and Technology. All participants provided informed consent, and responses were collected anonymously.

Data Availability Statement

The survey and case-study data supporting the findings of this study are available from the author upon reasonable request, subject to participant confidentiality agreements.

Declaration of Generative AI Use

During the preparation of this manuscript, the author used generative AI tools solely for language editing and proofreading. The author has reviewed and edited all content and takes full responsibility for the integrity of this publication.

References

The following references are listed below. DOI links are provided where available.

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188–2244. DOI: 10.1086/705716

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press. DOI: 10.2307/j.ctvcm4j72

Asai, A., Min, S., Zhong, Z., & Chen, D. (2023). Retrieval-based language models and applications. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts), 41–46. DOI: 10.18653/v1/2023.acl-tutorials.6

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. Electronic Markets, 33, 63. DOI: 10.1007/s12525-023-00680-1

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–16. DOI: 10.1145/3411764.3445717

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT 2021, 610–623. DOI: 10.1145/3442188.3445922

Bertrand, M., & Mullainathan, S. (2001). Do people mean what they say? Implications for subjective survey data. American Economic Review, 91(2), 67–72. DOI: 10.1257/aer.91.2.67

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P. (2023). The foundation model transparency index. arXiv preprint. DOI: 10.48550/arXiv.2310.12941

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. Research Synthesis Methods, 1(2), 97–111. DOI: 10.1002/jrsm.12

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. NBER Working Paper No. 31161. DOI: 10.3386/w31161

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. DOI: 10.1126/science.aap8062

Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving human-AI collaboration with descriptions of AI behavior. Proceedings of the ACM on Human-Computer Interaction, 7(CSCW1), 1–21. DOI: 10.1145/3579612

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. DOI: 10.1126/science.aal4230

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. Journal of Marketing Research, 56(5), 809–825. DOI: 10.1177/0022243719851788

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868–1883. DOI: 10.1111/poms.12838

Chui, M., Yee, L., Hall, B., Singla, A., & Sukharevsky, A. (2023). The state of AI in 2023: Generative AI's breakout year. McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year

Cui, T. H., Ghose, A., Halaburda, H., Iyengar, R., Pauwels, K., Sriram, S., Tucker, C., & Venkataraman, S. (2021). Informational challenges in omnichannel marketing: Remedies and future research. Journal of Marketing, 85(1), 103–120. DOI: 10.1177/0022242920968810

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. AIAA 1st Intelligent Systems Technical Conference. DOI: 10.2514/6.2004-6313

Cyberspace Administration of China. (2023). Interim measures for the management of generative artificial intelligence services. http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. Harvard Business School Working Paper 24-013. DOI: 10.2139/ssrn.4573321

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. DOI: 10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 64(3), 1155–1170. DOI: 10.1287/mnsc.2016.2643

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315(7109), 629–634. DOI: 10.1136/bmj.315.7109.629

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). GPTs are GPTs: An early look at the labor market impact potential of large language models. arXiv preprint, arXiv:2303.10130v2. DOI: 10.48550/arXiv.2303.10130

Engler, A. (2023). The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institution. https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment/

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. Information and Organization, 28(1), 62–70. DOI: 10.1016/j.infoandorg.2018.02.005

Felten, E., Raj, M., & Seamans, R. (2021). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal, 42(12), 2195–2217. DOI: 10.1002/smj.3286

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. Philosophy & Technology, 36(15). DOI: 10.1007/s13347-023-00621-y

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint. DOI: 10.48550/arXiv.2312.10997

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. DOI: 10.5465/annals.2018.0057

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. DOI: 10.1136/amiajnl-2011-000089

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, 1112–1123. DOI: 10.1145/3593013.3594067

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557–560. DOI: 10.1136/bmj.327.7414.557

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. Frontiers in Computer Science, 5, 1096257. DOI: 10.3389/fcomp.2023.1096257

Huang, M. H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. Journal of the Academy of Marketing Science, 49(1), 30–50. DOI: 10.1007/s11747-020-00749-9

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38. DOI: 10.1145/3571730

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848. DOI: 10.1007/s40685-020-00134-w

Korinek, A., & Stiglitz, J. E. (2021). Artificial intelligence, globalization, and strategies for economic development. NBER Working Paper No. 28453. DOI: 10.3386/w28453

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a science of human-AI decision making: A survey of empirical studies. arXiv preprint. DOI: 10.48550/arXiv.2112.11471

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. DOI: 10.1518/hfes.46.1.50_30392

Lee, M., Liang, P., & Yang, Q. (2022). CoAuthor: Designing a human-AI collaborative writing dataset for exploring language model capabilities. Proceedings of CHI 2022, 1–19. DOI: 10.1145/3491102.3502030

Lee, M. C. M., Scheepers, H., Lui, A. K. H., & Ngai, E. W. T. (2023). The implementation of artificial intelligence in organizations: A systematic literature review. Information & Management, 60(5), 103816. DOI: 10.1016/j.im.2023.103816

Leonardi, P. M. (2015). Materializing strategy: The blurry line between strategy formulation and strategy implementation. British Journal of Management, 26(S1), S17–S21. DOI: 10.1111/1467-8551.12077

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. DOI: 10.48550/arXiv.2005.11401

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. PLoS Medicine, 6(7), e1000100. DOI: 10.1371/journal.pmed.1000100

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3214–3252. DOI: 10.18653/v1/2022.acl-long.229

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. DOI: 10.1016/j.obhdp.2018.12.005

López-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. SSRN Electronic Journal. DOI: 10.2139/ssrn.4412788

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., & Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. Technological Forecasting and Social Change, 175, 121390. DOI: 10.1016/j.techfore.2021.121390

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 1906–1919. DOI: 10.18653/v1/2020.acl-main.173

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. DOI: 10.1145/3457607

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. AI and Ethics. Advance online publication. DOI: 10.1007/s43681-023-00289-2

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. Academy of Management Review, 46(3), 552–571. DOI: 10.5465/amr.2019.0186

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. DOI: 10.1126/science.adh2586

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. DOI: 10.1177/0018720810376055

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. arXiv preprint. DOI: 10.48550/arXiv.2302.06590

Peres, R., Schreier, M., Schweidel, D., & Sorescu, A. (2023). On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. International Journal of Research in Marketing, 40(2), 269–275. DOI: 10.1016/j.ijresmar.2023.03.001

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. Proceedings of FAT* 2020, 469–481. DOI: 10.1145/3351095.3372828

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

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The vicious circles of skill erosion: A case study of cognitive automation. Journal of the Association for Information Systems, 24(5), 1378–1412. DOI: 10.17705/1jais.00829

Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2023). Lost at C: A user study on the security implications of large language model code assistants. Proceedings of the 32nd USENIX Security Symposium, 2205–2222. DOI: 10.48550/arXiv.2208.09727

Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. Proceedings of the 28th International Conference on Intelligent User Interfaces, 410–422. DOI: 10.1145/3581641.3584066

Vaithilingam, P., Zhang, T., & Glassman, E. L. (2022). Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. Extended Abstracts of CHI 2022, 1–7. DOI: 10.1145/3491101.3519665

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU artificial intelligence act: Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97–112. DOI: 10.9785/cri-2021-220402

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. DOI: 10.2307/30036540

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective human-AI collaboration. Information Systems Frontiers, 24(3), 877–895. DOI: 10.1007/s10796-022-10284-3

Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2023). Voice-based AI in call center customer service: A natural field experiment. Production and Operations Management, 32(4), 1002–1018. DOI: 10.1111/poms.13953

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? Proceedings of the 25th International Conference on Intelligent User Interfaces, 189–201. DOI: 10.1145/3377325.3377480