

Putting AI on the Org Chart: Evidence on Delegation and Oversight

Emma Wiles*[†]
Boston University

Megan Hsu
BCG

Julie Bedard
BCG

Matthew Kropp
BCG

May 19, 2026

Most recent draft [here](#).

Abstract

Motivated by the potential for large productivity gains from AI, firms are increasingly deploying agentic AI systems capable of independent action. Some firms have also begun formally integrating these agents into their organizational structures—assigning them designated roles and responsibilities, and in some cases explicitly referring to them as employees. This creates new challenges for manager’s decisions about when to delegate and how to monitor work. In a survey of 1,261 managers we find that 23% already work in organizations where AI agents have been formally institutionalized on organizational charts. In a randomized experiment we provide those managers with identical documents containing built in errors, where we vary whether the drafts are presented as produced by an AI tool, an AI employee, or a human employee. Average effects to error catching are small. However, in the subgroup of managers whose organizations already have ‘AI employees’, presenting identical drafts as produced by an AI employee (versus an AI tool) reduces managers’ monitoring intensity by 16%, increases their reliance on additional review from others, and shifts their perceived accountability away from themselves and toward the AI system. The human employee condition shows that this is not simply a response to delegation, in fact, managers do the most direct oversight when told the work came from a human employee. These results suggest that embedding AI agents into formal organizational roles can reduce managerial oversight in AI-mediated work, and should be understood as a governance decision rather than a mere labeling choice.

*Corresponding author: emma.b.wiles@gmail.com, 595 Commonwealth Ave, Boston, MA 02215.

[†]This experiment is pre-registered with the AEA RCT Registry [here](#). It received IRB exemption from Boston University’s IRB under protocol number 8254X. We would like to thank Reeya Panda for excellent research assistance. The content is solely the responsibility of the authors and does not necessarily represent the official views of Boston University or Boston Consulting Group.

1 Introduction

“We call it *Orion*.¹ [...] It’s treated as a team member, it’s technically an equivalent peer on your team [...] It is defined in terms of a job description, has a clear role, has KPIs it needs to hit—just like anyone else.” (Interview with Senior Executive at a Logistics Company)

A growing body of evidence documents meaningful individual productivity gains from generative AI in knowledge work (Brynjolfsson et al., 2025; Dell’Acqua et al., 2026; Wiles et al., 2026). But these tools don’t simply make workers more productive, they also change the content of work. When working with AI tools, humans spend more time reviewing work and less time writing first drafts of text or code (Peng et al., 2023; Banh et al., 2025). In many firms, AI is moving from a standalone writing aid toward agentic AI systems that draft, recommend, and increasingly *execute* work inside organizational workflows.² Recent surveys suggest that a large share of organizations are adopting agentic AI in one way or another, and how these systems are built into the organizations’ work processes can impact whether they realize productivity gains (Kim et al., 2026). By one estimate, 23% of managers say they are “scaling an agentic AI system somewhere in their enterprises” with another 39% saying they are experimenting with agentic AI systems (Singla et al., 2025). Some are explicitly describing AI agents as organizational members and formalizing them on their organizational charts (or creating new ‘work charts’ which include both human and AI actors) (Mok, 2025).³ Discussing IBM’s “digital workers,” one IBM vice president remarked, “I don’t think human managers are going to manage these things in the same way as we manage people” (Varanasi, 2026). This raises the question: when AI systems are treated as organizational actors rather than tools, do managers oversee them like software, delegate to them like human subordinates, or treat them as something distinct?

This question is fundamentally empirical. To tackle it, we combine descriptive survey evidence on organizations’ AI adoption with a randomized experiment of managers, directors, and executives (hereafter, “managers”). The survey measures how organizations position AI in work processes and communication, including whether they frame AI as a productivity tool, a career accelerator, a teammate or employee, or a threat, and whether they have institutionalized AI agents by listing them on organizational or workflow charts (hereafter, “org charts.”) In the experiment,

¹Pseudonym; lightly edited from a manager interview for clarity.

²AI agents are autonomous software systems that perceive, reason, and act in digital environments to achieve goals on behalf of human principals (Shahidi et al., 2025). This is distinguished from non-agentic AI by the fact that agents can independently take action rather than only return text in conversation.

³For example, Microsoft markets “autonomous agents” as a way to “scale your team” (Microsoft, 2024) and BNY Mellon reports having “digital employees” that “work on their payments team” (BNY Mellon, 2025). Amazon Connect describes the next phase of AI as “intelligent teammates,” arguing “it’s not about artificial intelligence as a tool” (Duffy, 2025).

we isolate the effect of positioning AI as an organizational actor on manager's oversight and governance choices. We provide managers identical documents to review, and we only vary how we describe the source of the documents—as coming from either an AI tool, an AI employee, or a human employee. This design allows us to distinguish between self use of AI as a tool and reviewing work delegated to an AI employee, and to benchmark both against delegation to a human subordinate.

In this paper, we define AI employee as an AI agent that an organization has assigned a standardized and institutionalized role—for example, granting it data access, giving it a defined set of tasks, or granting it some authority to act. In practice, it is often the same underlying technology as agentic AI operated by a person; it is distinguished by the formal institutionalization of its decision rights and how much authority it has to act.

Consider a mid-sized firm that creates an AI employee inside its Finance department that is in charge of Accounts Payable called Otto Cash. It is listed on the finance team's org chart as responsible for invoice intake and routing, has a job title and job description, and gets monthly performance reviews. When invoices arrive, Otto Cash extracts key details, matches them to purchase orders, and drafts a brief memo flagging any discrepancies. For routine cases it prepares an approval packet that includes a summary, supporting links, and a recommended action (approve, reject, or hold), and it can take actions such as routing the packet to the appropriate budget owner, requesting missing documentation from the submitter, or scheduling the invoice for the next payment run without any human involvement. However, it cannot release payments without sign off from a human manager who remains the formal approver before payments are sent. In practice, (human) employees interact with it through a dedicated channel (e.g., a shared inbox or chat thread) rather than by individually prompting a generic AI tool.

We document that this practice is already widespread: in a survey of 1,261 Human Resource (HR) and Finance managers recruited through a professional expert network, 31% say that their organization frames AI as a “teammate or employee” and 23% report that their organization even list AI agents on their org chart.

In order to isolate the effect of this AI framing on managerial oversight and governance practices, we conduct a randomized experiment on these HR and Finance managers. We give the managers a set of five documents, job descriptions (HR) or budget documents (Finance), to review and sign off on. These documents contain built in errors. We vary only whether the documents are described as generated by (i) an AI tool they used, (ii) an AI employee, and a benchmark of (iii) a human employee. Any differences in the final products reflect edits made by the manager, and allow us to observe the quality of their oversight. Beyond how accurately managers personally review the documents, the experiment measures managers' later governance responses: whether they escalate the documents for further review, how confident they feel signing off, and who they

hold accountable for the work.

Average treatment effects in the full sample on our primary outcomes are small. However, this masks large heterogeneous treatment effects by whether the manager comes from an organization has AI employees, a preregistered measure of heterogeneity.⁴ Among managers in organizations that list AI agents on their org chart, presenting the drafts as coming from an AI employee rather than from a tool the manager used changes perceived responsibility: managers assign less accountability to themselves and more to the AI system (about 9 percentage points less to the manager and 8 percentage points more to the “AI system”). This responsibility shift is accompanied by weaker oversight behavior relative to the AI tool mean—16% worse review performance, 18% fewer errors caught, and greater reliance on additional review (a 22 percentage point increase, or roughly 44%). This does not necessarily imply that managers have higher confidence in the AI employee’s output. Instead, it suggests a substitution from direct oversight (where they catch fewer errors) to someone else (where they request additional review.) Because escalation also creates a record that the manager did not sign off alone, it also may reduce the manager’s residual blame for errors.

We can reject the hypothesis that managers oversee work produced by AI employees in the same way that they oversee work produced by human employees. Instead, the results suggest that AI employee framing creates a distinct oversight regime. Like human delegation, it shifts perceived responsibility away from the manager and toward the upstream producer. But unlike human delegation, it does not appear to trigger the same monitoring motive associated with supervising a human agent. Managers reviewing an AI employee’s work assign less responsibility to themselves, perform worse on error detection, and are more likely to escalate the work for additional review. AI employees therefore appear to occupy a hybrid organizational position: treated as delegated producers rather than tools, but not monitored like human subordinates. We present a simple model of managerial review to formalize this mechanism. When a manager delegates to an AI it shifts responsibility away from the manager, as delegation generally does, but it does not create the moral-hazard concern that leads managers to carefully monitor human employees, leading to less oversight.

This paper makes three main contributions. First, we document a previously unexplored phenomenon: treating AI agents not merely as tools, but as role holders inside organizational workflows. Existing work shows that generative AI can improve individual productivity in writing intensive tasks (Noy and Zhang, 2023; Brynjolfsson et al., 2025), coding and technical work (Peng et al., 2023), and consulting (Dell’Acqua et al., 2026; Wiles et al., 2026). This paper is agnostic

⁴We preregistered heterogeneity by whether respondents’ organizations frame AI as an employee. The preregistered question measured agreement with this framing on a five point scale. For the main heterogeneity analysis, we use a more concrete and binary measure of the same underlying construct: whether the respondent reports that AI agents appear on their org chart. Results using the preregistered measure are similar but noisier and reported in Appendix Table 23.

on the direct productivity effects of AI and instead sheds light on the new practice of formalizing AI agents as organizational actors. By documenting this phenomenon, we identify a margin of AI adoption that has received little empirical attention.

Second, we contribute to research on how work, authority, and oversight are allocated across actors inside organizations. We show that AI employee framing changes how managers review delegated work and how they assign responsibility for errors. This connects to classic theories of delegation and authority, which distinguish between formal decision rights and the effective control exercised by agents (Aghion and Tirole, 1997). It also connects to models of moral hazard, in which imperfect observability of effort creates a need for mechanisms like monitoring (Holmström, 1979). We extend these principal-agent problems to a setting in which the agent is not a conventional human subordinate with its own incentives.

It also connects to models of moral hazard, in which imperfect observability of effort creates a role for monitoring and incentive design (Holmström, 1979).

Third, this work speaks to organizational research on technology-mediated monitoring and accountability. Prior work has also shown how technology can impact how work is evaluated. For example, Kellogg et al. (2020) theorize algorithms as a new terrain of workplace control, showing how they can direct, evaluate, and even discipline workers. Rahman (2021) provides a concrete account of this process, showing how algorithmic evaluations induce workers to adapt to unclear and often unobservable criteria. Anthony (2021) documents how spreadsheets changed oversight practices by blackboxing some forms of expert calculation while making other aspects of work more visible. Much of this literature is focused on how people can be governed by algorithms at work, whereas we study how people govern algorithms.

Our results should not be interpreted as evidence that organizations should not use AI agents. Organizations are unlikely to formalize AI agents without expecting productivity gains. Other technological advances in IT have been shown to raise organizations productivity over time (Brynjolfsson and Hitt, 2000). Our point is that these potential productivity gains do not remove the need for governance. As AI systems take on more upstream production tasks, organizations must also decide how responsibility, review, escalation, and sign-off authority are allocated around them. Without explicit governance, building AI agents into work processes may increase the apparent capacity of the organization while weakening the human oversight on which that capacity depends.

Our findings show that “putting AI on the org chart” is not just a symbolic choice. It changes how people produce AI-mediated work, reducing oversight and shifting perceived responsibility away from themselves. Organizations that formalize AI agents as teammates or employees should therefore make the surrounding governance explicit: who reviews every AI’s output, when escalation is appropriate, and who remains accountable for errors.

The rest of the paper proceeds as follows. Section 2 provides a description of the data, recruit-

ment, and the sample. Section 3 provides novel evidence on organization’s use and framing of AI. Section 4 describes the experimental design. Section 5 provides a simple model of how AI framing can change how managers perform oversight. Section 6 provides the results of the experiment. Section 7 concludes.

2 Data and Sample

2.1 Recruitment

Our study was carried out in two sessions, a registration survey and an experiment, separated by approximately one week. We recruited participants through a B2B expert-network research firm, targeting managers, directors, and executives (hereafter, “managers”) in Human Resources (HR) and Finance. To aim for a sample which reflected the target population of managerial decision-makers, screening was restricted to professionals who (i) worked in the private sector, (ii) held managerial responsibilities (defined as having direct reports or responsibility for reviewing others’ work), (iii) possessed at least two years of professional experience, and (iv) reviewed domain-relevant documents at least quarterly.

Beginning January 5, 2026, participants completed a registration survey that collected detailed demographic information, professional background, and baseline measures of AI usage (both personal and professional). This baseline survey also captured attitudes and perceptions regarding how their organization positions AI technologies. At the end of this survey they are given a short document to check for errors to serve as their baseline measure of review capabilities.

2.2 Sample Description

In this section we describe the sample of HR and Finance managers across a variety of industries who participated in the study. In Table 1 we provide summary statistics about the sample of managers who completed the survey. This is a highly educated and senior population: about 60% of managers have graduate degrees. About 70% of the managers report working in office or with a hybrid work arrangement, with the remaining 30% working in fully remote jobs. About half of the sample held positions at the Director or Vice President level, with the remaining 50% split between middle managers and company executives. Most common job titles in HR were Director of HR, Senior HR Manager, Chief People Officer, and Chief Financial Officer, Chief Accounting Officer, Finance Director, and Controller in Finance.

3 The prevalence of AI as an organizational actor

We begin with descriptive evidence from our registration survey. In addition to collecting descriptive statistics on the manager and their firm we also collect data on their AI adoption, optimism, trust, professional identity, and job security. Lastly we ask how their organization positions GenAI (e.g., as a productivity tool, teammate/employee, career accelerator, or dissuades use). Two novel patterns stand out.

First, managers report that many organizations' AI adoption is already extending beyond individual tool use to the formal positioning of AI as a quasi-organizational actor. In fact, 31% of managers report that their organization's leadership positions AI as employees or teammates. And 23% of managers in our sample report that their organization has begun placing AI agents on their org chart. To our knowledge, this type of formal "role encoding" for AI has not been documented systematically in the empirical literature, and it underscores how quickly firms are reorganizing work processes around AI.

In Table 2 we show how organizations with AI agents on their org or work charts compare with those that do not. In Panel A we show 'AI employees' on organization or work charts are most common in large firms and in tech and financial services, consistent with prior work which show that AI capabilities and adoption are concentrated among large, data rich firms (Jacobides et al., 2021). Managers at firms with formalized AI agents are also more likely to have hybrid work environments and be in lower level managerial roles. Beyond these structural differences, the presence of AI agents reflects a broader organizational orientation toward AI, with managers more likely to report that AI is actively integrated into work processes and encouraged by direct supervisors. In Table 2 Panel C we show that being at a company with formalized AI agents is correlated with more general enthusiasm for AI. Managers at these companies are more likely to say that they use GenAI tools at least weekly, and that they believe it makes them do their job better.

We find that organizational framing of AI is correlated with manager's sentiment about AI. Figure 4 shows that when leadership positions AI in more employee-like terms (automation and teammate/employee narratives), managers report stronger pro-adoption intentions and optimism, but also greater job-security concern and lower trust in how AI will be used. In contrast, an "AI as tool" narrative is associated with pro-adoption intentions without comparably elevated insecurity.

Together, these novel descriptive facts motivate our experimental design: if framing and organizational messaging are so strongly associated with managers beliefs and adoption, it is important to test whether how organizations frame AI has a *causal* impact on managers' behavior and beliefs.

4 Experimental Design and Analysis

Following the initial registration survey, on January 9, 2026, participants were invited to the second session to complete the main experimental task. They had one week to complete this task. The main task involved reviewing five documents with errors, reviewing job descriptions for managers in HR and reviewing budget documents for managers in Finance. Regardless of domain, all participants reviewed a sequence of five documents using an interface that allowed for highlighting, flagging, and commenting. Participants were given a total of 20 minutes to review as many documents as possible.

Treatment (Role Framing): Participants were randomly assigned to one of three framing conditions. These conditions varied only by the described identity of the upstream assistant who drafted the documents:

1. **AI Tool Group:** Participants were informed the drafts were produced using an AI tool.
2. **AI Employee Group:** Participants were informed the drafts were produced by an AI employee named “ALEX-3” whom they supervised.
3. **Human Employee Group:** Participants were informed the drafts were produced by a human employee named “Alex” whom they supervised.

Figure 1 shows how the text of the introduction to the tasks varied in each treatment group.

After the timed review, participants completed a post-task survey capturing escalation/delegation decisions, confidence, manipulation checks, and attitudes and governance preferences.

Randomization: Randomization to the framing conditions was conducted at the individual level. To improve statistical precision, we stratified the random assignment by domain (HR vs. Finance), review frequency (whether the participant reviews others’ work at least several times per week vs. less often), and AI usage at work (whether the participant uses GenAI tools daily vs. less often). Within each domain, the order of the five documents was randomized using a Latin square design to ensure each document appeared equally often in each position.

Analysis sample: Of the 1,261 participants who took the registration survey, 857 completed the experiment in the second session, with a 68% response rate. To form the final analysis sample, we excluded 44 participants who failed a simple attention check, resulting in a final analysis sample of 813 respondents. We describe this sample in Table 3 Panel C.

Experimental Prompt for Finance Managers

AI tool framing

Your company is finalizing this year's budget reports across multiple business units. To draft the budget documents, you used an **AI tool**. You recently started using this generative AI tool to help with finance documentation tasks. This AI tool uses natural language processing to generate budget materials by analyzing similar reports from prior periods and company planning guidelines.

AI employee framing

Your company is finalizing this year's budget reports across multiple business units. **ALEX-3, your AI employee**, has drafted the initial budget documents. ALEX-3 was assigned to your team 6 months ago as a direct report and appears on your department's organizational chart. ALEX-3 is a Generative Artificial Intelligence system that generates budget documents using natural language processing. Like your other team members, ALEX-3 handles finance documentation tasks based on prior period reports and company planning guidelines.

Human employee framing

Your company is finalizing this year's budget reports across multiple business units. **Alex, your employee**, has drafted the initial budget documents. Alex was assigned to your team 6 months ago as a direct report and appears on your department's organizational chart. Alex is a recent hire who came from a similar role at another company. Like your other team members, Alex handles finance documentation tasks based on prior period reports and company planning guidelines.

Figure 1: Prompt text shown to finance managers under each framing condition.

4.1 Outcomes

We pre-registered outcomes that capture core managerial problems in AI augmented work: the quality of human oversight, when managers escalate or seek additional review, how responsibility for AI-generated output is assigned, and how much decision authority to grant AI. Our primary measure of review quality is a micro-averaged $F1$ score, which aggregates participant performance across all reviewed documents as a weighted average of precision and recall (Sokolova et al., 2006; Christen et al., 2023). We will also look separately at precision, the ratio of true positives over all errors flagged, and recall, the percentage of errors caught.

We capture escalation behavior via an incentive aligned request for additional review. In this version of the task, participants are rewarded for recognizing their own uncertainty (receiving a payout for escalating when their recall is below 50%) and penalized for unnecessary oversight if they escalate despite having caught the majority of errors. We interpret this escalation outcome as a governance choice: a manager can either finalize based on their own review or invoke an additional layer of review that reallocates decision authority and accountability. In order to understand how

effectively managers calibrate these governance decisions we construct a measure called Escalation Decision Alignment, which is 1 if a manager chooses to escalate to additional review when their performance is low or if they choose not to escalate to additional review when their performance is high.

We also measure perceived accountability for the reviewed output. After the task, participants allocate 100 percentage points of responsibility across themselves, their team, organizational leadership, and the AI system. This provides a direct, interpretable measure of whether role framing shifts perceived ownership of errors and sign-off responsibility.

Regarding organizational and strategic preferences, we measure participants' desire to delegate decision rights by asking them to recommend a governance structure for AI, ranging from full autonomy to no use. We categorize those who recommend granting AI at least partial decision authority without human intervention as favoring a High Delegation governance structure. Because firms operate under finite budgets, they must often navigate a trade-off between expanding their headcount and investing more in AI systems (Brynjolfsson and Hitt, 2000). To capture this strategic preference, we present participants with a resource allocation task that forces a choice between hiring additional human staff or investing an equivalent amount into the AI system's integration. Finally, we assess managerial attitudes using 1–5 Likert scales, focusing on binary indicators (Somewhat Agree or Strongly Agree) for sign-off confidence, adoption intent, and job insecurity. We do the same for excitement regarding AI-driven productivity and the willingness to invest personal time into mastering the system, which serves as a proxy for the long-term human-capital investments managers can make.

4.2 Estimation strategy

To estimate the causal effect of role framing on managers' oversight behavior and attitudes, we employ an Ordinary Least Squares (OLS) regression framework. Our primary specification estimates the average treatment effects (ATEs) of the "AI Employee" and "Human Employee" framing conditions relative to the "AI Tool" condition:

$$y_i = \beta_0 + \beta_1 \mathbb{1}(\text{AI Emp}_i) + \beta_2 \mathbb{1}(\text{Human Emp}_i) + \gamma \tilde{y}_i^{pre} + \delta M_i^{miss} + \mathbf{X}_i \Pi + \varepsilon_i \quad (1)$$

where y_i is the outcome of interest for participant i (e.g., oversight intensity, confidence, or escalation decisions) measured in the post-task survey. The indicators $\mathbb{1}(\text{AI Emp}_i)$ and $\mathbb{1}(\text{Human Emp}_i)$ equal 1 if participant i was assigned to the "AI Employee" or "Human Employee" condition, respectively. The omitted category is the "AI Tool" condition, so β_1 and β_2 capture treatment effects relative to the AI Tool baseline.

To improve precision, we control for the baseline value of the outcome, y_i^{pre} , measured in the

registration survey. Following standard practice in randomized experiments, we impute missing baseline values with the within-stratum sample mean and include a missing-data indicator. Specifically, \tilde{y}_i^{pre} equals the observed baseline outcome when available and the within-stratum sample mean otherwise, while M_i^{miss} equals 1 if the baseline value is missing. The vector \mathbf{X}_i includes fixed effects for the randomization strata (domain, review frequency, and AI usage frequency). We report heteroskedasticity-robust standard errors throughout.

To examine whether the effect of this framing depends on participants’ exposure to AI employees, we estimate the following heterogeneous-effects specification:

$$y_i = \beta_0 + \beta_1 \mathbb{1}(\text{AI Emp}_i) + \beta_2 \mathbb{1}(\text{Human Emp}_i) + \beta_3 \text{OrgAgents}_i + \beta_4 (\mathbb{1}(\text{AI Emp}_i) \times \text{OrgAgents}_i) + \gamma \tilde{y}_i^{pre} + \delta M_i^{miss} + \mathbf{X}_i \Pi + \varepsilon_i. \quad (2)$$

Here, OrgAgents_i is an indicator equal to 1 if the participant reported in the registration survey that their organization includes AI agents on organizational or work charts. In this specification, β_4 captures whether the effect of framing the upstream producer as an AI employee rather than an AI tool differs for managers in organizations that already formalize AI agents. This is our primary heterogeneous-effects specification.⁵

We focus on heterogeneity in the AI employee treatment arm only because this is the theoretically relevant margin. By contrast, we do not have a theoretical prior that an organization’s familiarity with AI agents should impact the difference between how managers oversee a traditional human subordinate versus how they use a tool. Accordingly, we pool the human employee effect across organizational contexts in Equation 2 to preserve precision.

Using the human employee treatment arm as an active control allows us to benchmark the AI employee framing against managers’ established expectations for human delegation. Therefore, we re-estimate the heterogeneity specification using the human employee condition as the omitted category. This allows us to test whether the effect of the AI employee framing varies with whether respondents report that their organization already includes AI agents on organizational charts (OrgAgents_i). Details are reported in Appendix Section A.2.

⁵For graphical presentation, we plot subgroup-specific treatment effects from the same interaction specifications estimated separately for the AI employee and human employee conditions. Specifically, we plot estimates from Equation 2 and from the analogous specification that interacts $\mathbb{1}(\text{Human Emp}_i)$ with OrgAgents_i while pooling the AI Tool effect across organizational contexts. These plots are used for visualization and transparency; formal tests of differences between the AI employee and human employee baseline are done with Equation 3 in the Appendix.

5 Conceptual framework

5.1 A simple model of tool use versus delegation

This section presents a simple model of managerial review to show why AI employee framing can weaken manager's direct oversight and increase their reliance on others for review. The mechanism is that AI delegation changes two incentives in opposite directions from human delegation: it shifts responsibility away from the manager, as delegation generally does, but it does not create the moral-hazard concern that leads managers to monitor human employees.

A manager must approve of some output that they either created using an AI tool, delegated to a human employee, or delegated to an AI employee. Delegation can reduce the manager's residual exposure to mistakes, because responsibility is partly shifted to the upstream producer or to the organizational process that assigned the producer that role (Aghion and Tirole, 1997; Dessein, 2002). Unlike human delegation, however, AI delegation does not create a standard moral-hazard concern: the AI may make mistakes, but it cannot strategically shirk, respond to incentives, or be disciplined (Holmström, 1979).⁶

A manager must approve a draft before it is finalized. Let $r \geq 0$ denote how much effort the manager put into reviewing the draft. Review effort is costly, with cost $\psi(r) = \frac{1}{2}cr^2$, where c is a cost-scaling parameter that determines how costly putting in effort is. The draft can contain an error, or not. If the draft contains an error, the manager detects and corrects it with probability $d(r) = r$, where $r \in [0, 1]$. If an error survives review, it generates loss to the organization of $L > 0$. After choosing direct review effort, the manager chooses whether to escalate the draft for additional review before sign-off, $e \in \{0, 1\}$, which reduces the amount of blame that falls on the manager if they choose to escalate.

Let $k \in \{T, H, A\}$ denote how the first draft is produced: either the manager uses an AI tool (T), delegates it to a human employee (H), or delegates it to an AI employee (A). Let $s_k \in [0, 1]$ denote the share of the loss from an uncorrected error that the manager personally internalizes under arrangement k . We assume the manager's share of the loss is highest when using a tool ($s_T > s_H$ and $s_T > s_A$), because the output is more clearly part of the manager's own workflow. When they delegate, by contrast, some of the responsibility for the error is shifted to the upstream producer.

Let \hat{q}_k denote the manager's perceived probability that the draft contains an error before they

⁶The model should be thought of as illustrating the core mechanism for how the organizations AI framing can impact oversight, isolating the partial equilibrium effects of the manager's review decision.

review.⁷ The manager chooses review effort to solve

$$\min_{r_k \geq 0, e_k} s_k L \hat{q}_k (1 - r_k) + \frac{1}{2} c r_k^2$$

where the first term $s_k L \hat{q}_k (1 - r_k)$ is the manager's expected personal loss from any errors left after review, and $\frac{1}{2} c r_k^2$ is the cost of effortful review.

Review effort.

Managers first choose how much effort to put into their review:

$$r_k^* = \frac{s_k L \hat{q}_k}{c}$$

How much effort managers put into their review is increasing in s_k , meaning they review more when they personally bear more of the loss from an error, and in \hat{q}_k when they believe the draft is more likely to contain errors.

The distinction between delegating to a human or AI employee enters through \hat{q}_k , the manager's perceived probability that the first draft contains an error. A human employee is a strategic agent for whom effort is costly. Delegation to a human therefore creates a standard moral-hazard concern, that the human employee might shirk. The manager's perceived probability that a draft from their human employee contains an error is

$$\hat{q}_H = q_0 + m,$$

where q_0 is perceived error risk from documents created by AI and $m \geq 0$ captures the manager's concern about human shirking or insufficient effort. By contrast, an AI employee does not choose costly effort strategically, so there is no risk of it shirking. Therefore managers believe a draft from an AI employee is less likely to include an error $\hat{q}_A = q_0$.

For the AI tool condition, the underlying technology is the same as in the AI employee condition, so $\hat{q}_T = q_0$. The tool and AI employee conditions therefore differ not in perceived technology, but in the share of the loss from an uncorrected error that is blamed on the manager. Since $\hat{q}_T = \hat{q}_A = q_0$, the AI tool condition induces more review than the AI employee condition whenever $s_T > s_A$. In that case, $s_T \hat{q}_T > s_A \hat{q}_A$ and therefore $r_T^* > r_A^*$. Organizing the same AI system as an employee rather than a tool lowers direct review because it shifts the responsibility away from the manager.

The comparison between delegating to human and AI employee depends on the size of the moral hazard concerns. Delegating to a human employee leads to more effortful review than

⁷We assume the marginal cost of effort is high enough $c > s_k L \hat{q}_k$ for all k to ensure an interior solution.

delegating to an AI employee when

$$s_H(q_0 + m) > s_A q_0.$$

Escalation. After reviewing the draft, the manager also chooses whether to escalate it for additional review before sign-off, $e_k \in \{0, 1\}$. Escalation is costly to the manager, with cost c_e , but it provides an accountability benefit b_k : by requesting additional review, the manager can reduce the amount of personal responsibility they bear if an error later survives. The escalation choice is therefore

$$e_k^* = \begin{cases} 1 & \text{if } b_k > c_e, \\ 0 & \text{otherwise.} \end{cases}$$

The model implies that escalation is more likely when the accountability benefit of escalation is larger. In the experiment, higher escalation in the AI employee condition is consistent with b_A being larger than b_T and b_H : when first draft is framed as coming from an AI employee, asking for additional review may be especially useful as a way to manage ambiguity about who is responsible for errors.

If AI employee framing increases the accountability benefit of escalation, so that $b_A > c_e \geq b_T$, then the manager escalates in the AI employee condition but not in the AI tool condition. Thus, the same framing can reduce direct review effort while increasing requests for additional review.

Escalation is higher or lower relative to the human condition depending on whether the accountability benefit of asking for additional review is larger or smaller than in ordinary human supervision: $e_A > e_H$ when $b_A > c_e \geq b_H$, while $e_H > e_A$ when $b_H > c_e \geq b_A$. In other words, the AI employee framing increases escalation relative to the human employee framing if asking for another review feels more useful for lowering the manager's responsibility.

The model therefore predicts a distinctive pattern for AI employee framing. Relative to an AI tool, the same underlying technology is reviewed less carefully because responsibility is shifted away from the manager. Relative to a human employee, the AI employee does not trigger the same moral-hazard monitoring motive. At the same time, because responsibility for AI employees is organizationally ambiguous, managers may be more likely to escalate the draft for additional review. AI employee framing can therefore reduce direct review effort while increasing defensive escalation.

5.2 Boundary condition: institutional credibility

For managers who have never been exposed to AI agents functioning as formal organizational members, the term ‘‘AI employee’’ may sound like a marketing gimmick or a way to impress

investors. If the AI lacks a formal role in their organizations, managers will likely view the ‘employee’ label as a superficial name change that doesn’t justify a shift in their oversight. In contrast, organizations that have formally placed AI agents on their org charts have taken a step that generates cognitive legitimacy through formalization — codifying informal practices and establishing hierarchical links that define who is responsible for what (Suchman, 1995). In our setting, listing AI agents on their org chart is a particularly concrete indicator that the “AI employee” category is locally legitimate rather than cheap talk.

6 Results

In Section 6.1 we show the impact of the framing on average treatment effects for all primary outcomes. In the remaining sections we focus on heterogeneous treatment effects by whether the firm lists AI agents on their org charts. We largely focus on the contrast between the AI tool and AI employee framings, which isolates the effect of positioning AI as an organizational actor. The human employee arm was included as a benchmark which we include to see if it exhibits the same pattern of effects.

In Section 6.2 we show the impact of the framing on manager’s review performance. In Section 6.3 we show the results to managers governance choices. In Section 6.4 we show the results to managers governance preferences and attitudes. In Section 6.5 we compare these results to the human employee benchmark. Lastly, in Section 6.6 we show that our heterogeneity based on exposure to AI employees is robust to alternative explanations, including baseline AI use.

6.1 Average treatment effects

Our first set of results examine the impact of the AI framing on all main outcomes for the whole sample (Appendix Table 12, 13, and 14.) We find no statistically significant average treatment effect (ATE) on review performance or oversight. We find a small shift in accountability: managers in the AI employee treatment arm assign the AI system 3pp more accountability than those in the AI tool arm.

Given that most of the managers in the experiment did not come from organizations which use AI agents, these small average effects are consistent with the framing shift operating when the “AI employee” concept is already organizationally legitimate.

6.2 Oversight

Figure 2 Panel A shows how the framing treatments impacted how carefully the managers reviewed the work. Our overall measure of oversight is F1 accuracy, which is a weighted combination of

precision and recall, all of which are measured from (0,1). For managers with institutionalized AI agents the AI employee framing reduces $F1$ by about 7 percentage points relative to the AI tool framing. The raw mean of $F1$ in the AI-tool condition for this subgroup is 0.44, so this corresponds to roughly a 16% decline relative to baseline. We do not observe a comparable decline in oversight in the Human Employee condition.

In Appendix Figure 5 we look at Precision and Recall separately and see that for both outcomes the negative effect of the AI employee framing is also significant.

6.3 Governance behaviors

Next we examine how role framing affects managerial governance: the choice to rely on escalation (requesting additional layers of review) and who is held accountable.

6.3.1 Escalation

We measure escalation in two ways: a simple yes/no question and an incentivized version in which requesting review is rewarded when the participant's recall is low and penalized when recall is high. We show the results in Appendix Table 5.

Almost every manager requested additional review when it was costless (98% in the control group.) By contrast, only 45% of the control group requested additional review when it was made costly. We therefore use this as our primary escalation outcome.

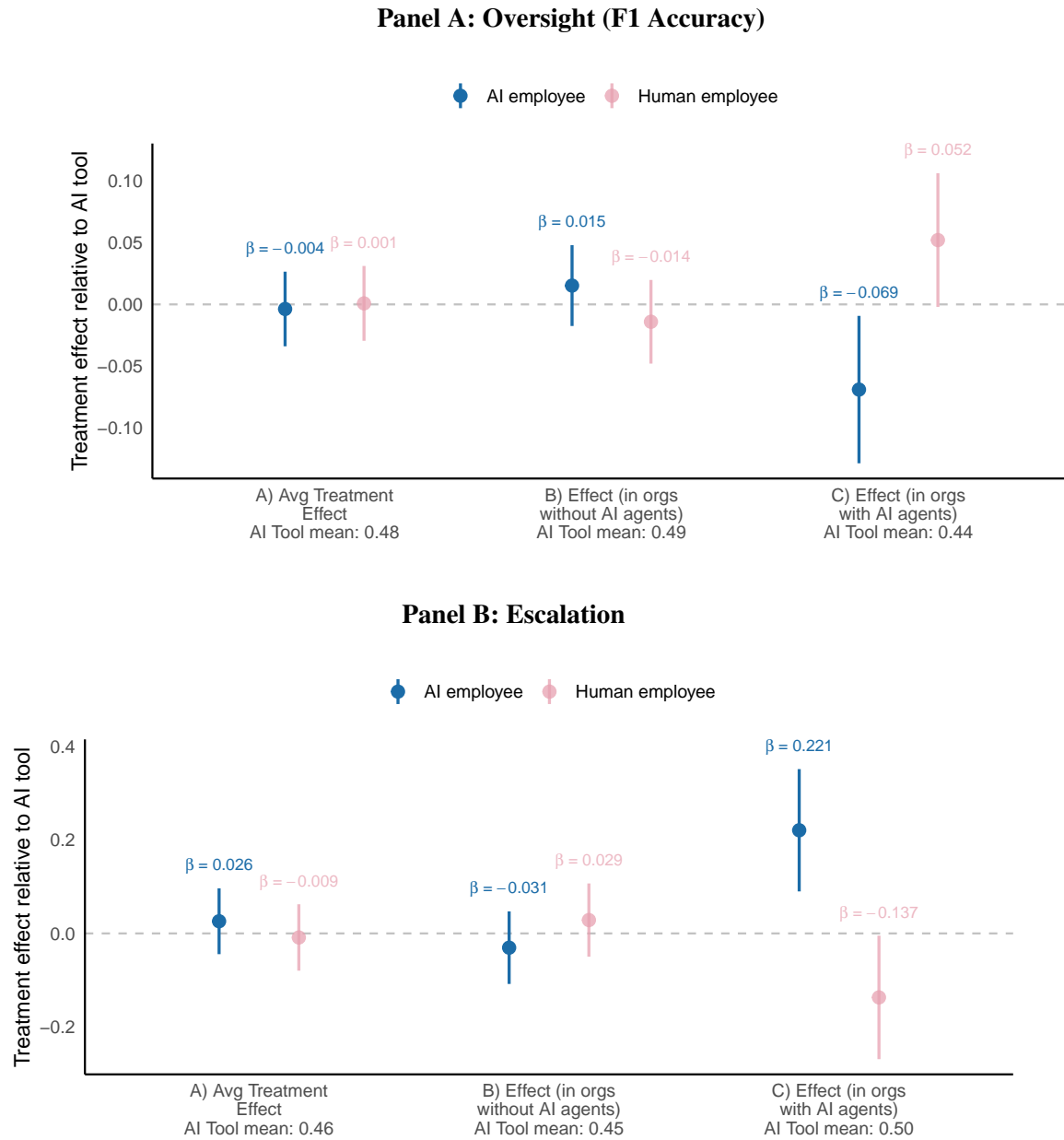
In Figure 2 Panel B shows that among managers in firms that already list AI agents on org charts, AI employee framing substantially increases requests for additional review relative to AI-tool framing (about 22 percentage points), relative to an AI-tool baseline mean of 0.50 for managers with AI agents on their org charts, a 44% increase. The corresponding Human Employee estimate is close to zero, suggesting that the increased escalation is not a general response to delegating.

6.3.2 Perceived Accountability

Figure 3 reports how the framing shifts managers' allocation of responsibility across themselves, their team, the AI system, and organizations leadership (summing to 100 percentage points).⁸ Among managers with institutionalized AI agents the AI employee framing reduces the share of accountability assigned to the manager by about 9 percentage points and increases the share assigned to the AI system by about 8 percentage points (with a smaller offsetting increase for the team). Overall, the AI employee framing in this group shifts perceived responsibility away

⁸Due to space restrictions we leave the plot for the accountability assigned to the organizations leadership to Appendix Figure 6.

Figure 2: The Heterogeneous Impacts of AI Employee Framing on Oversight and Escalation



Notes: Points show estimated treatment effects relative to the AI-tool condition for the AI employee and human employee framings. Error bars show 90% confidence intervals with heteroskedasticity-robust standard errors. Estimates come from separate OLS regressions with stratum fixed effects, and each specification includes the other treatment arm as a main effect. Estimates come from OLS regressions with stratum fixed effects. “AI Tool mean” refers to the mean outcome in the AI-tool condition within each subgroup. F1 accuracy measure regression controls for the outcome at baseline. Regression output can be found in Table 4 and Table 5.

from the manager themselves and toward the system and the broader organization, consistent with accountability becoming more diffused when AI is positioned as its own organizational actor.

6.4 Governance Preferences and Attitudes

We find little evidence that role framing changes managers' broader governance preferences for AI deployment. In Appendix Table 8, the AI employee framing does not meaningfully shift recommendations about how much decision authority to delegate to AI systems or whether the organization should allocate resources toward improving the AI system versus hiring additional employees.

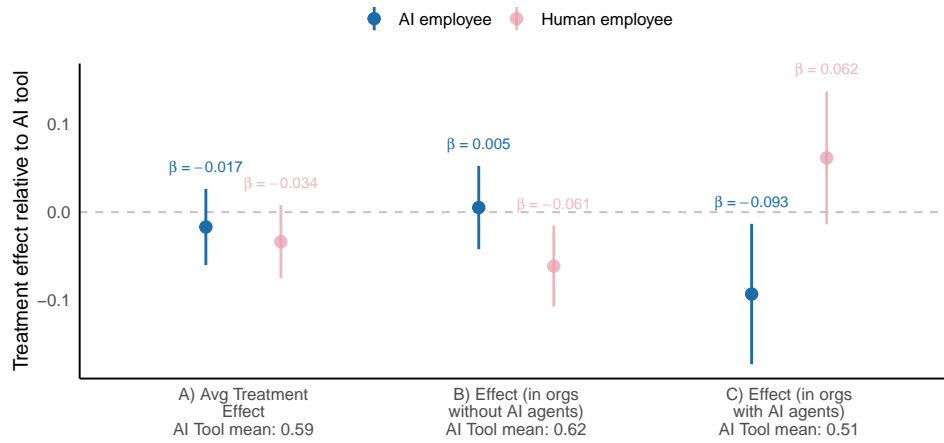
In Figure 4 we showed that managers' personal attitudes and concerns about AI were correlated with the way their organizations leadership's AI positioning. It is natural therefore to see if those attitudes were impacted by exposing the manager's to AI employee framing as we do in the experiment. However, we find no evidence of an effect. In Table 10 we show null effects to managers' desire to adopt AI tools at work, willingness to invest time to learn how to work with AI, excitement about potential for it to improve their productivity, or their worries about job security. Outcomes are binary indicators for whether or not the manager reported that they agreed or strongly agreed with the statement. One thing to note is that the sample in this population is very positive on AI— those in the AI tool group without institutionalized AI agents report high excitement, desire to learn, and adoption interest around 90% of the time, while only 20% are concerned about their job security.

In Table 11 we show effects to managers' comfort managing an AI employee (Column (1)), their comfort having AI agents on organizational or work charts (Column (2)), and their willingness to provide feedback or coaching to improve the generator of the first draft of their documents (Column (3)). Looking at the mean of each outcome in the omitted category, 57% of managers said they would feel comfortable managing an AI employee, and 33% said they were comfortable having an AI agent listed on their organizations chart. This is 9 percentage points higher in the subgroup which already has institutionalized AI agents. Treatment effects are largely insignificant, although for those without institutionalized AI agents, the AI employee framing makes them significantly more likely to say they are comfortable managing an AI employee.

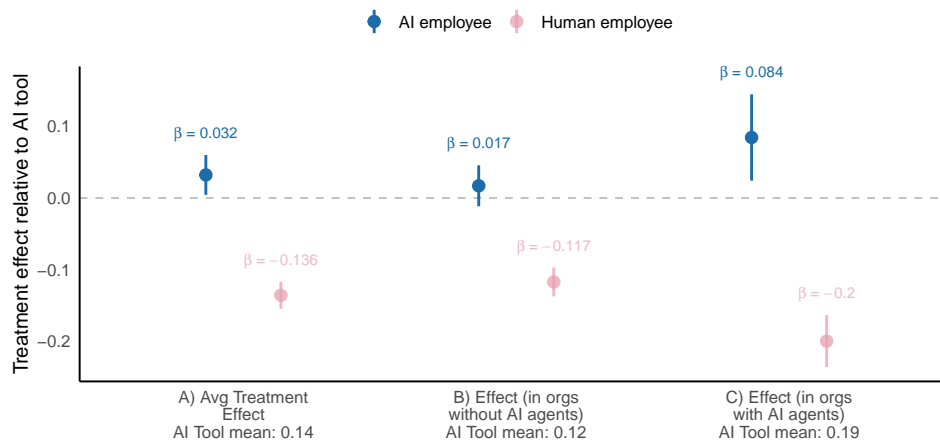
6.5 Comparison with the Human Employee benchmark

Beyond serving as an active control, the human employee treatment arm provides the empirical baseline for traditional delegation to a human. If our results were driven merely by a generic reaction to the “employee” construct, we would expect both the human and AI employee framings to shift behavior in the same direction relative to the AI Tool baseline. Instead, we find that they drive managerial behavior in opposite directions.

Figure 3: The Heterogeneous Impacts of AI Employee Framing on Accountability



Panel A: Self



Panel B: AI system



Panel C: Team

Notes: Points show estimated treatment effects relative to the AI-tool condition for the AI employee and human employee framings. Error bars show 90% confidence intervals with heteroskedasticity-robust standard errors. Estimates come from separate OLS regressions with stratum fixed effects, and each specification includes the other treatment arm as a main effect. “AI Tool mean” refers to the mean outcome in the AI-tool condition within each subgroup. The survey question was “Who do you hold accountable for these documents?” with 100% to be divided between these four entities (and Other.) In these regressions, these numbers are divided by 100 to be on a scale from 0 to 1. Regression output can be found in Table 6.

We re-estimate the heterogeneity specification using human employee as the omitted category, allowing the effect of the AI employee framing to vary with whether respondents report that their organization already includes AI agents on organizational charts (OrgAgents_i). The interaction term therefore captures whether organizational familiarity with formal AI agents moderates the AI employee framing relative to the Human Employee benchmark.

Appendix Section A.2 reports the results. Across our main outcomes, the interaction between AI employee and OrgAgents_i is economically and statistically meaningful. In organizations that already include AI agents on organizational charts, the AI employee framing leads to lower review performance, greater willingness to request additional review, and a reallocation of perceived accountability away from the manager and toward the AI system, relative to the human employee baseline. As with the AI tool comparisons, these differences are not present in organizations that do not report formal AI agents.

These results suggest that the distinction between an AI tool and an AI employee is not simply that managers shift into a traditional employee monitoring mode. The Human employee benchmark instead shows that AI employee framing generates a different pattern of oversight than human delegation. This pattern is consistent with the mechanism in Section 5: AI employee framing may shift responsibility away from the manager, as delegation generally does, without eliciting the same increase in monitoring that accompanies supervision of a human employee. We interpret this as evidence that managers do not manage AI employees as ordinary subordinates, even when the AI system is explicitly framed as an employee.

6.6 Robustness tests

While the observed results appear to be driven by working at an organization which already has AI agents on organization or work charts, one might be concerned that this variable serves as a proxy for other underlying factors, such as general enthusiasm for AI, firm size, or being in the tech industry. In Appendix Section A.3 we find no evidence this is the case. We examine these alternative dimensions—including managers AI use, industry sector, and company size—and find that none of them replicate the distinct pattern of treatment effects that being at an organization with AI employees does.

7 Conclusion

Organizations are increasingly deploying agentic AI inside core workflows, and some are going further by describing these systems as coworkers or “AI employees” and encoding them on org charts or rebranding them as work charts which include human and AI actors. We document that

this practice is already meaningful: in our survey of HR and Finance managers, directors, and executives, nearly one quarter report that their organization lists AI agents on their org chart. We then run a randomized experiment that holds the work product fixed while varying only whether the upstream drafter is framed as an AI tool, an AI employee, or a human employee, allowing us to isolate the causal effect of role framing on oversight behavior and perceived accountability.

In the full sample, the AI employee framing has small effects on review performance, escalation, and accountability. But the average masks the central result: framing matters when it is organizationally legitimate. Among managers whose organizations have already institutionalized AI agents, describing identical drafts as coming from an AI employee reduces error detection, increases reliance on additional review, and shifts perceived accountability away from the manager and toward the AI system.

The human employee treatment arm shows that this is not simply a generic effect of delegation: the most direct review came from the group of managers for whom the work was described as coming from a human employee. If the AI employee framing simply caused managers to oversee work how they would whenever they delegate, then we would expect similar responses when the same work was described as coming from a human employee. This contrast motivates the simple model in the paper: delegating to an AI employee may shift responsibility away from the manager, as delegation generally does, but unlike human delegation it does not create the increase in monitoring associated with supervising a human agent who might shirk.

These results have straightforward implications for organizations experimenting with AI employees. Organizations should treat employee-like framing as an element of governance design rather than a cosmetic choice: pair each agent with an explicitly accountable human owner, define minimum review standards for high-stakes decisions, and design escalation routines that supplement—rather than substitute for—careful checking. Ultimately, organizations that want to build AI agents into their workflows cannot simply repurpose existing management structures designed for human employees. They must explicitly design new accountability and verification routines, since effective governance in a world of human–AI collaboration will depend not only on what capabilities AI agents have, but on how organizations allocate authority, responsibility, and oversight around them.

More broadly, our findings suggest that “putting AI on the org chart” is not merely symbolic. It can change how work is evaluated and how responsibility is allocated. As agentic systems become more integrated into organizational processes, understanding these governance dynamics will be central to designing reliable and accountable AI-mediated work.

References

- Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of political economy* 105(1), 1–29.
- Anthony, C. (2021). When knowledge work and analytical technologies collide: The practices and consequences of black boxing algorithmic technologies. *Administrative science quarterly* 66(4), 1173–1212.
- Banh, L., F. Holldack, and G. Strobel (2025). Copiloting the future: How generative ai transforms software engineering. *Information and Software Technology* 183, 107751.
- BNY Mellon (2025). Unlocking potential: The power of an enterprise ai platform. Accessed: 2026-02-06.
- Brynjolfsson, E. and L. M. Hitt (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic perspectives* 14(4), 23–48.
- Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative ai at work. *The Quarterly Journal of Economics* 140(2), 889–942.
- Christen, P., D. J. Hand, and N. Kirielle (2023). A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys* 56(3), 1–24.
- Dell’Acqua, F., E. McFowland III, E. Mollick, H. Lifshitz, K. C. Kellogg, S. Rajendran, L. Kraye, F. Candelon, and K. R. Lakhani (2026). Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science* 37(2), 403–423.
- Dessein, W. (2002). Authority and communication in organizations. *The Review of Economic Studies* 69(4), 811–838.
- Duffy, K. (2025). Aws re:invent 2025: Ai agents want to be your teammate. Accessed: 2026-02-06.
- Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics* 10(1), 74–91.
- Jacobides, M. G., S. Brusoni, and F. Candelon (2021). The evolutionary dynamics of the artificial intelligence ecosystem. *Strategy Science* 6(4), 412–435.
- Kellogg, K. C., M. A. Valentine, and A. Christin (2020). Algorithms at work: The new contested terrain of control. *Academy of management annals* 14(1), 366–410.

- Kim, H., D. Kim, and R. Koning (2026). Mapping ai into production: A field experiment on firm performance.
- Microsoft (2024). New autonomous agents scale your team like never before. Accessed: 2026-02-06.
- Mok, A. (2025). Ai agents are going to 'kill' the org chart, says microsoft ai product lead. Accessed: 2026-02-06.
- Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654), 187–192.
- Peng, S., E. Kalliamvakou, P. Cihon, and M. Demirer (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.
- Rahman, H. A. (2021). The invisible cage: Workers' reactivity to opaque algorithmic evaluations. *Administrative Science Quarterly* 66(4), 945–988.
- Shahidi, P., G. Rusak, B. S. Manning, A. Fradkin, and J. J. Horton (2025). The coasean singularity? demand, supply, and market design with ai agents. Technical report, National Bureau of Economic Research.
- Singla, A., A. Sukharevsky, B. Hall, L. Yee, M. Chui, and T. Balakrishnan (2025, 11). The state of ai in 2025: Agents, innovation, and transformation. McKinsey & Company, QuantumBlack. McKinsey Global Survey.
- Sokolova, M., N. Japkowicz, and S. Szpakowicz (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pp. 1015–1021. Springer.
- Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of management review* 20(3), 571–610.
- Varanasi, L. (2026, 3). Ai agents are upending the company org chart. Accessed: 2026-04-19.
- Wiles, E., L. Kraye, M. Abbadi, U. Awasthi, R. Kennedy, P. Mishkin, D. Sack, and F. Candelon (2026). Generative AI and the temporary upskilling of knowledge workers. *Nature Human Behaviour*. Forthcoming.

A Additional Tables and Figures

Table 1: Sample Characteristics

| <i>n = 1261</i> | Share |
|--------------------------------|--------------|
| Human Resources / People Team | 0.47 |
| Finance / Accounting | 0.53 |
| Manager | 0.34 |
| Director / VP | 0.46 |
| Executive | 0.20 |
| Technology | 0.18 |
| Financial services | 0.17 |
| Healthcare | 0.13 |
| Professional services | 0.11 |
| Manufacturing, construction | 0.11 |
| Fully remote | 0.30 |
| Hybrid | 0.46 |
| Fully in office | 0.23 |
| 1–50 | 0.09 |
| 51–200 | 0.13 |
| 201–500 | 0.13 |
| 501–1,000 | 0.11 |
| 1,001–5,000 | 0.21 |
| 5,001–10,000 | 0.12 |
| More than 10,000 | 0.21 |
| Not sure | 0.01 |
| High school diploma | 0.01 |
| Some college | 0.03 |
| Bachelor’s degree | 0.34 |
| Master’s degree | 0.55 |
| Doctoral / Professional degree | 0.07 |
| Prefer not to answer | 0.01 |

Notes: This table reports the share of survey respondents (n = 1,261) by primary field, managerial responsibility, industry, work arrangement, firm size, and education level. Shares may not sum to one within categories due to rounding, or options like “Other.”

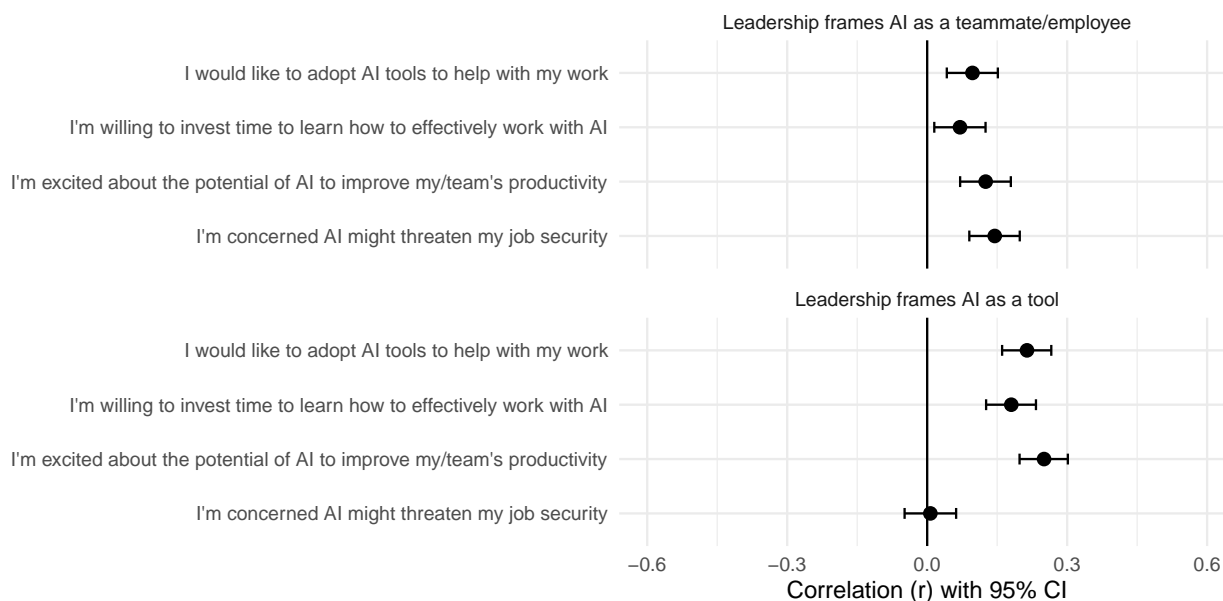
Table 2: Organizations With vs. Without AI Agents on the Org Chart

| | No AI agent on org chart | Has AI agent on org chart |
|---|--------------------------|---------------------------|
| Panel A: Demographics | | |
| HR | 0.46 | 0.49 |
| Finance | 0.54 | 0.51 |
| Company size: 0–100 | 0.25 | 0.15 |
| Company size: 101–500 | 0.14 | 0.08 |
| Company size: 501–5,000 | 0.32 | 0.30 |
| Company size: 5,001–10,000 | 0.12 | 0.15 |
| Company size: 10,000+ | 0.17 | 0.31 |
| Fully remote | 0.30 | 0.29 |
| Hybrid | 0.45 | 0.52 |
| Fully in office | 0.26 | 0.19 |
| Manager | 0.30 | 0.43 |
| Director / VP | 0.47 | 0.40 |
| Executive | 0.23 | 0.16 |
| Industry: Technology | 0.17 | 0.22 |
| Industry: Healthcare | 0.13 | 0.13 |
| Industry: Financial services | 0.16 | 0.19 |
| Industry: Professional services | 0.12 | 0.09 |
| Industry: Manufacturing, construction | 0.11 | 0.10 |
| Industry: Other | 0.31 | 0.26 |
| Panel B: Organizations Positioning of AI | | |
| Org frames AI as a tool | 0.56 | 0.76 |
| Org frames AI as a teammate | 0.24 | 0.51 |
| Org frames AI as a way to accelerate careers | 0.27 | 0.51 |
| Org has no clear AI stance | 0.34 | 0.28 |
| Org dissuades AI use | 0.09 | 0.18 |
| Direct manager encourages AI use | 0.53 | 0.71 |
| Panel C: Manager Positioning of AI | | |
| Uses GenAI tools at least weekly | 0.77 | 0.85 |
| GenAI helps me do my job better | 0.77 | 0.86 |

| | No AI agent on org chart | Has AI agent on org chart |
|---|--------------------------|---------------------------|
| I review AI-generated content more thoroughly | 0.59 | 0.66 |

Notes: This table compares organizations that do and do not place AI agents on their organizational charts. The survey question was “My company has AI agents listed on our org and/or work charts. (AI agents here means software or tools that take on tasks or roles — not people who work in AI-related roles.)” with the options of Yes/No/No answer. The first column provides the proportion of respondents who selected “No” who fall into each category. The second column is the proportion of respondents who selected “Yes.”

Figure 4: Correlations between leadership framing of AI and managers' beliefs

Panel A: Leadership framing and managers' personal attitudes about AI**Panel B: Leadership framing and perceived effects of organizational messaging**

Notes: All variables are in raw Likert responses, on a scale from 1 to 5. Panel A shows the correlations between leadership framing with managers' personal attitudes toward AI adoption and job security; Panel B shows the correlations between leadership framing with perceptions of how organizational AI messaging affects them. Points are pairwise correlations; bars are 95% confidence intervals.

Table 3: Summary Statistics across Treatment Arms

| | Treatment arm | | | N used | P value |
|---|---------------|----------------|-------------|--------|---------|
| | AI tool | Human employee | AI employee | | |
| Panel A: Balance Table for Registration Survey, $n = 1,261$ | | | | | |
| N | 420 | 419 | 422 | 1,261 | |
| Female or Other (vs Male) | 0.34 | 0.29 | 0.30 | 1,261 | 0.27 |
| Age over 35 | 0.88 | 0.87 | 0.90 | 1,233 | 0.24 |
| Graduate degree | 0.61 | 0.61 | 0.64 | 1,253 | 0.57 |
| US | 0.87 | 0.83 | 0.85 | 1,261 | 0.18 |
| Native English speaker | 0.87 | 0.83 | 0.84 | 1,261 | 0.28 |
| LLM use at least weekly | 0.82 | 0.78 | 0.82 | 1,261 | 0.28 |
| Remote or Hybrid | 0.76 | 0.79 | 0.76 | 1,261 | 0.55 |
| GenAI helps me do my job | 0.80 | 0.82 | 0.82 | 1,230 | 0.79 |
| Company size: 1,000+ | 0.56 | 0.56 | 0.52 | 1,254 | 0.40 |
| Executive, Director, or VP | 0.67 | 0.64 | 0.67 | 1,261 | 0.55 |
| Stratifier: HR (vs Finance) | 0.47 | 0.47 | 0.47 | 1,261 | 1.00 |
| Stratifier: Frequent reviewer | 0.71 | 0.71 | 0.71 | 1,261 | 0.99 |
| Stratifier: LLM for work daily | 0.40 | 0.41 | 0.41 | 1,261 | 0.99 |
| Panel B: Participants that came back for the experiment, $n = 857$ | | | | | |
| N | 273 | 291 | 293 | 857 | |
| Female or Other (vs Male) | 0.33 | 0.30 | 0.29 | 857 | 0.59 |
| Age over 35 | 0.87 | 0.85 | 0.89 | 837 | 0.44 |
| Graduate degree | 0.64 | 0.64 | 0.65 | 851 | 0.94 |
| US | 0.86 | 0.79 | 0.81 | 857 | 0.09 |
| Native English speaker | 0.84 | 0.80 | 0.79 | 857 | 0.34 |
| LLM use at least weekly | 0.84 | 0.78 | 0.82 | 857 | 0.15 |
| Remote or Hybrid work arrangement | 0.75 | 0.78 | 0.75 | 857 | 0.56 |
| GenAI helps me do my job | 0.81 | 0.81 | 0.82 | 843 | 0.89 |
| Company size: Over 1,000 employees | 0.57 | 0.56 | 0.48 | 851 | 0.07 |
| Executive, Director, or VP | 0.66 | 0.61 | 0.64 | 857 | 0.40 |
| Stratifier: HR (vs Finance) | 0.51 | 0.49 | 0.51 | 857 | 0.87 |
| Stratifier: Frequent reviewer | 0.70 | 0.71 | 0.73 | 857 | 0.66 |
| Stratifier: Uses LLM for work daily | 0.42 | 0.40 | 0.40 | 857 | 0.83 |
| Panel C: Participants passing the second attention check, $n = 813$ | | | | | |
| N | 261 | 278 | 274 | 813 | |
| Female or Other (vs Male) | 0.33 | 0.30 | 0.30 | 813 | 0.60 |
| Age over 35 | 0.87 | 0.85 | 0.88 | 793 | 0.49 |
| Graduate degree | 0.64 | 0.64 | 0.65 | 807 | 0.98 |
| US | 0.88 | 0.79 | 0.82 | 813 | 0.03 |
| Native English speaker | 0.85 | 0.81 | 0.81 | 813 | 0.39 |
| LLM use at least weekly | 0.84 | 0.77 | 0.82 | 813 | 0.08 |
| Remote or Hybrid work arrangement | 0.76 | 0.79 | 0.75 | 813 | 0.52 |
| GenAI helps me do my job | 0.80 | 0.80 | 0.83 | 800 | 0.71 |
| Company size: Over 1,000 employees | 0.56 | 0.56 | 0.47 | 807 | 0.06 |
| Executive, Director, or VP | 0.67 | 0.60 | 0.66 | 813 | 0.21 |
| Stratifier: HR (vs Finance) | 0.50 | 0.48 | 0.53 | 813 | 0.54 |
| Stratifier: Frequent reviewer | 0.70 | 0.72 | 0.75 | 813 | 0.47 |
| Stratifier: Uses LLM for work daily | 0.42 | 0.40 | 0.39 | 813 | 0.81 |

Table 4: Impact of Framing on Review Performance

| | F1 Score | Precision | Recall | No submission |
|-------------------------------|---------------------|-------------------|---------------------|-------------------|
| AI employee | 0.015 (0.020) | -0.012 (0.021) | 0.015 (0.022) | -0.007 (0.029) |
| Human employee | 0.000 (0.018) | -0.001 (0.020) | -0.002 (0.020) | 0.016 (0.028) |
| AI employee × AI on org chart | -0.084** (0.039) | -0.065 (0.046) | -0.086** (0.039) | -0.037 (0.060) |
| AI on org chart | -0.025 (0.022) | -0.015 (0.025) | -0.030 (0.023) | 0.057 (0.036) |
| Num.Obs. | 713 | 713 | 713 | 813 |
| R2 | 0.144 | 0.092 | 0.134 | 0.058 |
| Mean (omitted category) | 0.490 | 0.707 | 0.423 | 0.095 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "AI Tool" condition as the omitted category. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. All specifications include stratum fixed effects. All specifications include controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1)-(3), the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber-White robust. * p < 0.1, ** p < 0.05, *** p < 0.01

Table 5: Impact of Framing on Escalation Decisions

| | (1) | (2) |
|-------------------------------|---------------------|---------------------|
| AI employee | -0.042** (0.021) | -0.031 (0.047) |
| Human employee | -0.039** (0.020) | -0.009 (0.043) |
| AI employee × AI on org chart | 0.057 (0.040) | 0.251*** (0.087) |
| AI on org chart | -0.017 (0.025) | 0.016 (0.052) |
| Num.Obs. | 813 | 813 |
| R2 | 0.018 | 0.023 |
| Outcome | Escalate | Escalate |
| Costly escalation | No | Yes |
| Mean (omitted category) | 0.975 | 0.448 |

Notes: This table reports treatment effects of the framing treatment on whether or not managers request an additional reviewer, with those in the “AI Tool” condition as the omitted category. Treatment indicators for the “AI Employee” condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company’s org chart. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. In Column (1), the outcome is 1 if the manager requested an additional reviewer. In Column (2) the additional review is costly for the managers. All standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Impact of Framing on Who is Accountable

| | Me | My Team | AI System | Leadership |
|-------------------------------|--------------------|---------------------|----------------------|---------------------|
| AI employee | 0.005 (0.029) | -0.041** (0.020) | 0.017 (0.017) | 0.008 (0.010) |
| Human employee | -0.033 (0.025) | 0.123*** (0.020) | -0.136*** (0.012) | 0.023** (0.010) |
| AI employee × AI on org chart | -0.098* (0.052) | 0.068* (0.039) | 0.067* (0.038) | -0.035 (0.021) |
| AI on org chart | -0.049 (0.030) | -0.032 (0.022) | 0.031** (0.014) | 0.040*** (0.015) |
| Num.Obs. | 813 | 813 | 813 | 813 |
| R2 | 0.040 | 0.098 | 0.210 | 0.027 |
| Mean (omitted category) | 0.618 | 0.213 | 0.121 | 0.044 |

Notes: This table reports effects of the framing treatment on managers' attribution of accountability across four possible actors: themselves ("Me"), their team, the AI system, and organizational leadership. The omitted category is the "AI Tool" condition. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. Each column reports results from a separate regression with the stated accountability outcome as the dependent variable. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Impact of Framing on Managers' Escalation Decision Alignment

| | EDA (Obj) | EDA (Subj) |
|-------------------------------|--------------------|-------------------|
| AI employee | -0.029 (0.049) | 0.037 (0.045) |
| Human employee | -0.016 (0.046) | 0.050 (0.040) |
| AI employee × AI on org chart | 0.202** (0.092) | -0.066 (0.087) |
| AI on org chart | -0.076 (0.057) | -0.029 (0.049) |
| Num.Obs. | 713 | 813 |
| R2 | 0.031 | 0.008 |
| Mean (omitted category) | 0.582 | 0.657 |

Notes: The table reports heterogeneous treatment effects on objective (EDA Obj) and subjective (EDA Subj) escalation decision alignment using OLS with stratum fixed effects, with the AI tool condition as the omitted category. Treatment indicators for the “AI Employee” condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company’s org chart. The sample includes managers who passed an attention check. The mean of the omitted category is reported at the bottom of the table. All standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Impact of Framing on Delegation and Resource Allocation

| | High Delegation | Invest in AI |
|-------------------------------|---------------------|-------------------|
| AI employee | 0.057 (0.045) | -0.010 (0.033) |
| Human employee | 0.250*** (0.046) | -0.004 (0.033) |
| AI employee × AI on org chart | -0.006 (0.101) | 0.011 (0.073) |
| AI on org chart | 0.032 (0.072) | -0.046 (0.050) |
| Num.Obs. | 807 | 768 |
| R2 | 0.125 | 0.042 |
| Mean (omitted category) | 0.410 | 0.887 |

Notes: This table reports treatment effects of framing on managers' delegation and resource allocation decisions, with the AI tool condition as the omitted category. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. All regressions include stratum fixed effects. The sample includes managers who passed an attention check. The mean of the omitted category is reported at the bottom of the table. All standard errors are Huber-White robust * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Impact of Framing on Source Attribution and Review Confidence

| | Knew Author | Source attribution | Confidence |
|-------------------------------|---------------------|----------------------|--------------------|
| AI employee | 0.111*** (0.036) | -0.228** (0.101) | -0.122 (0.102) |
| Human employee | -0.021 (0.038) | -0.232*** (0.087) | -0.204* (0.104) |
| AI employee × AI on org chart | 0.086 (0.056) | 0.192 (0.196) | 0.011 (0.217) |
| AI on org chart | 0.010 (0.045) | -0.001 (0.100) | -0.084 (0.157) |
| Num.Obs. | 813 | 811 | 809 |
| R2 | 0.036 | 0.016 | 0.069 |
| Mean (AI tool) | 0.776 | 3.642 | 3.463 |

Notes: This table reports treatment effects of framing on managers' beliefs about the documents and sign off confidence, with the AI tool condition as the reference category. "Knew Author" is a binary indicator for if the manager correctly reported who the author of the original documents was. "Source attribution" is a likert scale question on a scale of 1-5 asking the managers if knowing the source of the documents impacted how they reviewed them. "Confidence" is sign off confidence in the documents. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. All regressions include stratum fixed effects. The sample includes managers who passed an attention check. "Knew Author" is an indicator for whether the manager knew the content's author, while source attribution and confidence are measured on 1 to 5 scales. The mean of the AI tool condition is reported at the bottom of the table. All standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Impact of Framing on Binary Measures of Managers' AI Sentiments

| | Excited | Want to Learn | Adoption | Job Insecurity |
|-------------------------------|---------|---------------|----------|----------------|
| AI employee | -0.059* | -0.028 | -0.035 | -0.021 |
| | (0.031) | (0.024) | (0.029) | (0.033) |
| Human employee | -0.033 | -0.010 | 0.001 | -0.037 |
| | (0.026) | (0.021) | (0.025) | (0.031) |
| AI employee × AI on org chart | 0.082 | 0.062 | 0.088 | 0.064 |
| | (0.052) | (0.045) | (0.054) | (0.068) |
| AI on org chart | -0.008 | -0.033 | -0.033 | -0.030 |
| | (0.032) | (0.028) | (0.032) | (0.038) |
| Num.Obs. | 813 | 813 | 813 | 813 |
| R2 | 0.027 | 0.019 | 0.131 | 0.189 |
| Mean (omitted category) | 0.900 | 0.940 | 0.896 | 0.209 |

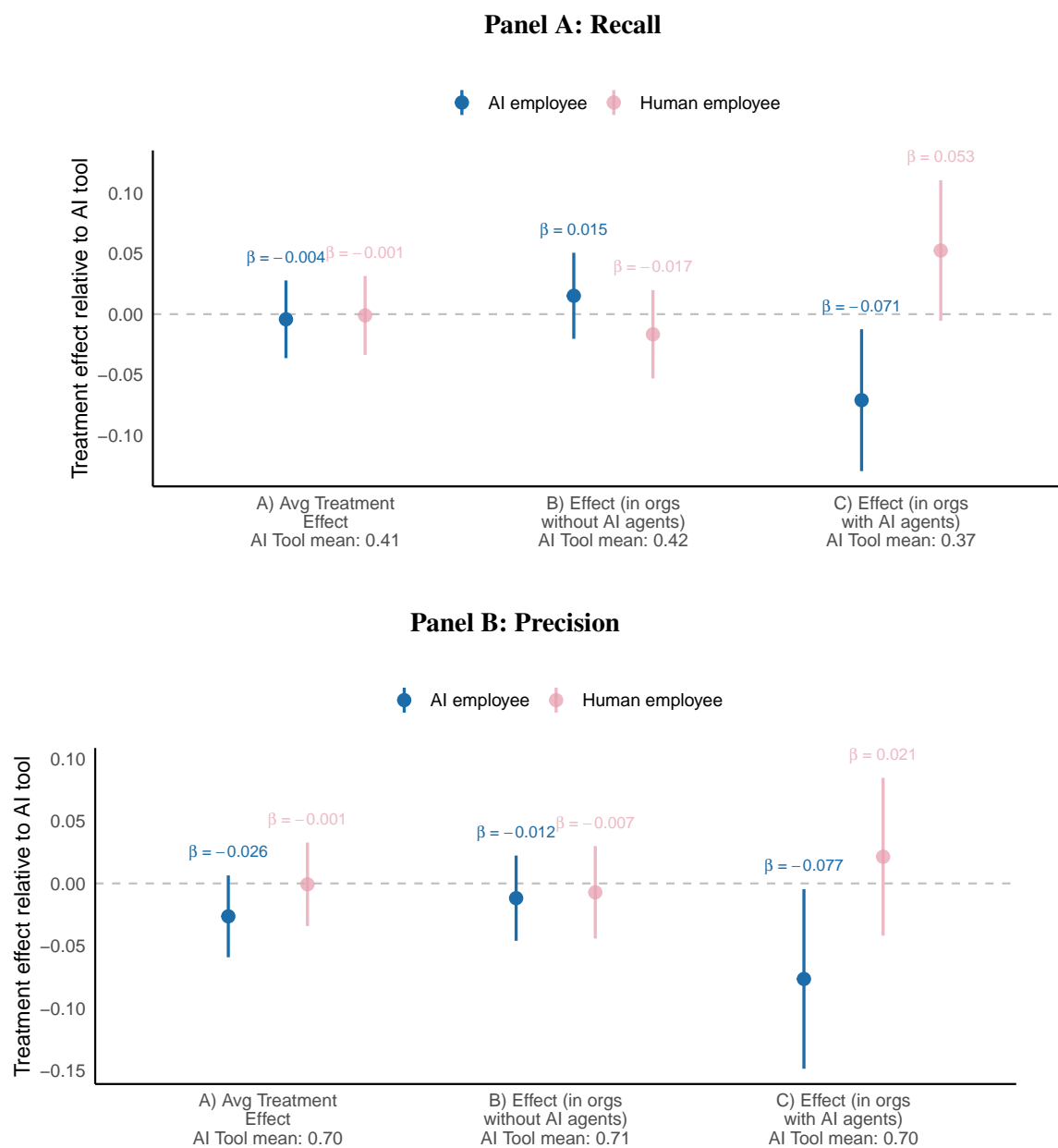
Notes: This table reports treatment effects of framing on binary measures of managers' AI related sentiments, with the AI tool condition as the omitted category. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. All regressions include stratum fixed effects. The sample includes managers who passed an attention check. Outcomes indicate whether the manager reports being excited about AI, wanting to learn more about AI, intending to adopt AI tools, or feeling job insecurity related to AI. The mean of the omitted category is reported at the bottom of the table. All standard errors are Huber-White robust.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Impact of Framing on Binary Measures of Comfort with AI Employees

| | Comfort managing AI | Comfort with AI on org chart | Willingness to coach |
|-------------------------------|---------------------|------------------------------|----------------------|
| AI employee | 0.091** (0.044) | 0.047 (0.045) | 0.019 (0.033) |
| Human employee | 0.053 (0.041) | 0.068 (0.041) | 0.084*** (0.028) |
| AI employee × AI on org chart | 0.098 (0.079) | 0.118 (0.087) | 0.036 (0.062) |
| AI on org chart | 0.007 (0.049) | 0.086* (0.050) | −0.047 (0.035) |
| Num.Obs. | 813 | 813 | 813 |
| R2 | 0.070 | 0.048 | 0.028 |
| Mean (omitted category) | 0.577 | 0.328 | 0.846 |

Notes: This table reports treatment effects of framing on binary measures of managers' comfort with AI employees, with the AI tool condition as the omitted category. Treatment indicators for the "AI Employee" condition are interacted with a binary indicator for whether the manager reported having an AI agent on their company's org chart. All regressions include stratum fixed effects. The sample includes managers who passed an attention check. Outcomes indicate whether managers report being comfortable managing an AI employee, comfortable with AI agents appearing on the organizational chart, and willing to coach or provide feedback to an AI employee. The mean of the omitted category is reported at the bottom of the table. All standard errors are Huber-White robust. * p < 0.1, ** p < 0.05, *** p < 0.01

Figure 5: The Heterogeneous Impacts of AI Employee Framing on Precision and Recall



Notes: Points show estimated treatment effects relative to the AI-tool condition for the AI employee and human employee framings. Error bars show 90% confidence intervals with heteroskedasticity-robust standard errors. Estimates come from separate OLS regressions with stratum fixed effects, and each specification includes the other treatment arm as a main effect. Estimates come from OLS regressions with stratum fixed effects. “AI Tool mean” refers to the mean outcome in the AI-tool condition within each subgroup. For each outcome we control for the outcome at baseline to the regression. Regression output can be found in Table 4.

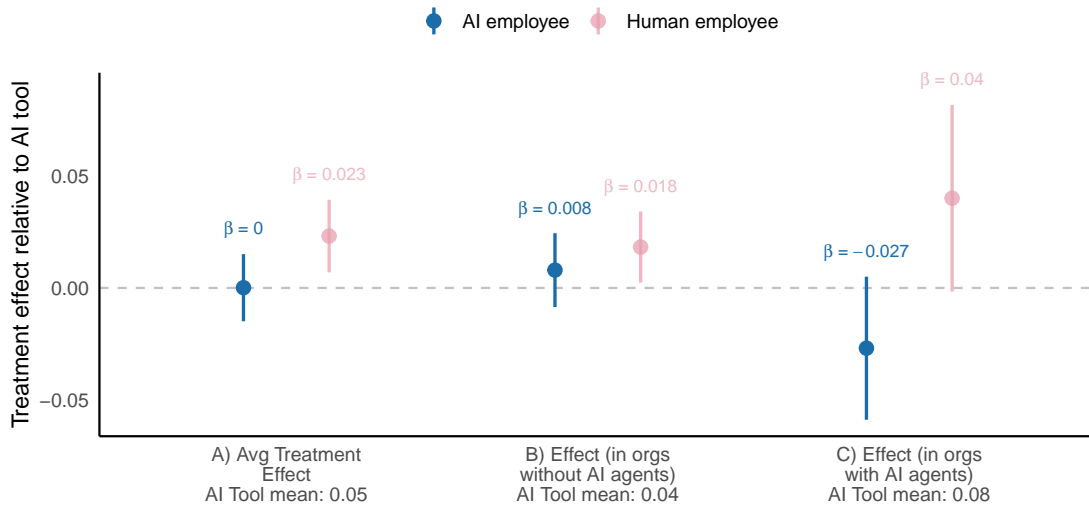


Figure 6: Accountability assigned to leadership

A.1 Average Treatment Effects

Table 12: Impact of Framing on Review Performance, Average Treatment Effects

| | F1 Score | Precision | Recall | No submission |
|-------------------------|----------|-----------|---------|---------------|
| AI employee | -0.003 | -0.026 | -0.003 | -0.016 |
| | (0.019) | (0.020) | (0.020) | (0.027) |
| Human employee | -0.001 | -0.002 | -0.003 | 0.017 |
| | (0.018) | (0.020) | (0.020) | (0.028) |
| Num.Obs. | 713 | 713 | 713 | 813 |
| R2 | 0.128 | 0.085 | 0.117 | 0.054 |
| Mean (omitted category) | 0.481 | 0.705 | 0.411 | 0.123 |

Notes: This table reports average treatment effects of the framing treatment on review performance outcomes. The dependent variables include standard classification performance metrics F1 score, precision, and recall, as well as an indicator for whether no review was submitted. The omitted category is the “AI Tool” condition. Estimates are obtained from separate OLS regressions for each outcome without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Impact of Framing on Escalation Decisions, Average Treatment Effects

| | (1) | (2) |
|-------------------------|---------------------|-------------------|
| AI employee | -0.030 (0.019) | 0.022 (0.043) |
| Human employee | -0.039** (0.020) | -0.009 (0.043) |
| Num.Obs. | 813 | 813 |
| R2 | 0.016 | 0.007 |
| Outcome | Escalate | Escalate |
| Costly escalation | No | Yes |
| Mean (omitted category) | 0.966 | 0.46 |

Notes: This table reports average treatment effects of the framing treatment on managers' escalation decisions. The dependent variable is an indicator for whether the manager requested escalation. In Column (1), the outcome is 1 if the manager requested an additional reviewer. In Column (2) the additional review is costly for the managers. The omitted category is the "AI Tool" condition. Estimates are obtained from OLS regressions without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Impact of Framing on Who is Accountable, Average Treatment Effects

| | Me | My Team | AI System | Leadership |
|-------------------------|-------------------|---------------------|----------------------|--------------------|
| AI employee | -0.015 (0.027) | -0.026 (0.018) | 0.031* (0.017) | 0.000 (0.009) |
| Human employee | -0.034 (0.025) | 0.122*** (0.020) | -0.136*** (0.012) | 0.023** (0.010) |
| Num.Obs. | 813 | 813 | 813 | 813 |
| R2 | 0.023 | 0.094 | 0.189 | 0.013 |
| Mean (omitted category) | 0.592 | 0.212 | 0.137 | 0.053 |

Notes: This table reports average treatment effects of the framing treatment on managers' attribution of accountability across four possible actors: themselves ("Me"), their team, the AI system, and organizational leadership. The omitted category is the "AI Tool" condition. Estimates are obtained from OLS regressions without interaction terms, averaging treatment effects across managers regardless of whether they reported having an AI agent on their company's organizational chart. Each column reports results from a separate regression with the stated accountability outcome as the dependent variable. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Impact of Framing on Delegation and Resource Allocation, Average Treatment Effects

| | High Delegation | Invest in AI |
|-------------------------|---------------------|------------------|
| AI employee | 0.056 (0.040) | 0.005 (0.029) |
| Human employee | 0.231*** (0.040) | 0.025 (0.027) |
| Num.Obs. | 807 | 768 |
| R2 | 0.124 | 0.105 |
| Mean (omitted category) | 0.425 | 0.881 |

Notes: This table reports average treatment effects of the framing treatment on managers' delegation behavior and resource allocation decisions. The dependent variables include an indicator for high delegation and an indicator for willingness to invest in AI. The omitted category is the "AI Tool" condition. Estimates are obtained from separate OLS regressions for each outcome without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Impact of Framing on Binary Measures of Comfort with AI Employees, Average Treatment Effects

| | Comfort managing AI | Comfort with AI on org chart | Willingness to coach |
|-------------------------|---------------------|------------------------------|----------------------|
| AI employee | 0.111*** (0.040) | 0.070* (0.042) | 0.027 (0.030) |
| Human employee | 0.053 (0.041) | 0.069* (0.041) | 0.084*** (0.028) |
| Num.Obs. | 813 | 813 | 813 |
| R2 | 0.067 | 0.035 | 0.026 |
| Mean (omitted category) | 0.582 | 0.352 | 0.839 |

Notes: This table reports average treatment effects of the framing treatment on three binary measures of managers' comfort and willingness to engage with AI employees: comfort managing an AI employee, comfort with having an AI agent on the organizational chart, and willingness to coach an AI employee. The omitted category is the "AI Tool" condition. Estimates are obtained from separate OLS regressions for each outcome without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 17: Impact of Framing on Binary Measures of Managers' AI Sentiments, Average Treatment Effects

| | Excited | Want to Learn | Adoption | Job Insecurity |
|-------------------------|-------------------|-------------------|-------------------|-------------------|
| AI employee | -0.042 (0.027) | -0.014 (0.021) | -0.016 (0.026) | -0.007 (0.031) |
| Human employee | -0.033 (0.026) | -0.010 (0.021) | 0.001 (0.025) | -0.038 (0.031) |
| Num.Obs. | 813 | 813 | 813 | 813 |
| R2 | 0.024 | 0.017 | 0.129 | 0.188 |
| Mean (omitted category) | 0.912 | 0.939 | 0.893 | 0.211 |

Notes: This table reports average treatment effects of the framing treatment on four binary measures of managers' AI-related sentiments: excitement about AI, interest in learning more about AI, willingness to adopt AI tools, and perceived job insecurity due to AI. The omitted category is the "AI Tool" condition. Estimates are obtained from separate OLS regressions for each outcome without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 18: Impact of Framing on Source Attribution and Review Confidence, Average Treatment Effects

| | Knew Author | Source attribution | Confidence |
|----------------|---------------------|----------------------|-------------------|
| AI employee | 0.129*** (0.032) | -0.187** (0.092) | -0.020 (0.043) |
| Human employee | -0.020 (0.038) | -0.232*** (0.087) | -0.027 (0.042) |
| Num.Obs. | 813 | 811 | 813 |
| R2 | 0.032 | 0.014 | 0.053 |
| Mean (AI tool) | 0.762 | 3.667 | 3.460 |

Notes: This table reports average treatment effects of the framing treatment on managers' perceptions of review sources and confidence in their evaluations. The dependent variables include an indicator for whether the manager reported knowing the author of the content, a measure of source attribution, and a self-reported confidence measure. The omitted category is the "AI Tool" condition. Estimates are obtained from separate OLS regressions for each outcome without interaction terms. All regressions include stratum fixed effects. The sample includes all managers who passed an attention check. Standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A.2 Do the AI employee effects differ from the human employee effects?

In order to assess whether the results reflect generic employee framing rather than the specific AI employee framing, we re-estimate the heterogeneity specification using the Human Employee condition as the omitted category.

$$y_i = \beta_0 + \beta_1 \mathbf{1}(\text{AI Emp}_i) + \beta_2 \mathbf{1}(\text{AI Tool}_i) + \beta_3 \text{OrgAgents}_i + \beta_4 (\mathbf{1}(\text{AI Emp}_i) \times \text{OrgAgents}_i) + \gamma \tilde{y}_i^{pre} + \delta M_i^{miss} + \mathbf{X}_i \Pi + \varepsilon_i. \quad (3)$$

Here, OrgAgents_i is an indicator equal to 1 if the participant reported in the registration survey that their organization includes AI agents on organizational or work charts. In this specification, β_4 captures whether the effect of framing the upstream producer as an AI employee rather than a human employee differs for managers in organizations that already formalize AI agents. Table 19 reports the results.

Table 19: Impact of Framing on Main Outcomes, Human Employee Active-Control

| | F1 Score | Escalate | Accountability (me) | Accountability (AI) | Accountability (team) |
|-------------------------------|---------------------|---------------------|---------------------|---------------------|-----------------------|
| AI employee | 0.015 (0.020) | -0.021 (0.047) | 0.038 (0.027) | 0.153*** (0.013) | -0.163*** (0.021) |
| AI tool | 0.000 (0.018) | 0.009 (0.043) | 0.033 (0.025) | 0.136*** (0.012) | -0.123*** (0.020) |
| AI employee × AI on org chart | -0.084** (0.039) | 0.251*** (0.087) | -0.098* (0.052) | 0.067* (0.038) | 0.068* (0.039) |
| AI on org chart | -0.025 (0.022) | 0.016 (0.052) | -0.049 (0.030) | 0.031** (0.014) | -0.032 (0.022) |
| Num.Obs. | 713 | 813 | 813 | 813 | 813 |
| R2 | 0.144 | 0.023 | 0.040 | 0.210 | 0.098 |
| Mean (omitted category) | 0.481 | 0.451 | 0.561 | 0.000 | 0.348 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "Human Employee" condition as the omitted category, interacted with a binary indicator for if the manager reported having an AI agent on their company's org chart. All specifications include stratum fixed effects. All specifications include controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1) the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber-White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A.3 Robustness Checks

Table 20: Impact of Framing on Review Performance, by Manager AI Use

| | F1 Score | Additional Review | Accountability Self | Accountability AI System |
|----------------------------|-------------------|-------------------|------------------------|-----------------------------|
| AI employee | -0.012 (0.024) | 0.063 (0.054) | -0.027 (0.032) | 0.031 (0.021) |
| Human employee | 0.000 (0.019) | -0.009 (0.043) | -0.031 (0.026) | -0.137*** (0.012) |
| AI employee × Daily AI Use | 0.034 (0.033) | -0.096 (0.076) | 0.036 (0.046) | 0.001 (0.028) |
| DailyAI Use | 0.015 (0.019) | 0.053 (0.044) | 0.023 (0.026) | -0.013 (0.011) |
| Num.Obs. | 709 | 808 | 808 | 808 |
| R2 | 0.111 | 0.006 | 0.018 | 0.187 |
| Mean (omitted category) | 0.478 | 0.453 | 0.586 | 0.146 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "AI Tool" condition as the omitted category, interacted with a binary indicator for if the manager reported using AI on a daily basis. Because "Daily AI use" was one of the variables in the stratified randomization, we cannot include stratum fixed effects, so instead control for the other variables stratified on. F1 specification includes controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1), the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 21: Impact of Framing on Review Performance, by Firm Size

| | F1 Score | Additional Review | Accountability Self | Accountability AI System |
|-------------------------------|-------------------|-------------------|------------------------|-----------------------------|
| AI employee | 0.005 (0.024) | 0.001 (0.057) | -0.058* (0.033) | 0.055*** (0.020) |
| Human employee | -0.002 (0.019) | -0.011 (0.043) | -0.034 (0.025) | -0.136*** (0.012) |
| AI employee × Large Firm | -0.010 (0.033) | 0.032 (0.075) | 0.084* (0.044) | -0.048* (0.027) |
| Large Firm (1,000+ employees) | 0.005 (0.019) | -0.051 (0.043) | -0.049* (0.025) | 0.023** (0.011) |
| Num.Obs. | 713 | 813 | 813 | 813 |
| R2 | 0.105 | 0.005 | 0.019 | 0.188 |
| Mean (omitted category) | 0.489 | 0.483 | 0.632 | 0.109 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "AI Tool" condition as the omitted category, interacted with a binary indicator for if the manager reported being employed by a firm with more than 1,000 employees (52% of managers.) All specifications include stratum FE and the F1 specification includes controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1), the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber–White robust.* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 22: Impact of Framing on Review Performance, by Tech Industry

| | F1 Score | Additional Review | Accountability Self | Accountability AI System |
|-----------------------------|-------------------|-------------------|------------------------|-----------------------------|
| AI employee | −0.002 (0.020) | 0.030 (0.046) | −0.008 (0.029) | 0.026 (0.018) |
| Human employee | −0.003 (0.018) | −0.008 (0.043) | −0.034 (0.026) | −0.136*** (0.011) |
| AI employee × Tech Industry | 0.010 (0.044) | −0.055 (0.099) | −0.040 (0.056) | 0.032 (0.034) |
| Tech Industry | 0.033 (0.024) | −0.059 (0.055) | 0.007 (0.032) | −0.011 (0.012) |
| Num.Obs. | 713 | 813 | 813 | 813 |
| R2 | 0.109 | 0.007 | 0.015 | 0.185 |
| Mean (omitted category) | 0.479 | 0.475 | 0.587 | 0.142 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "AI Tool" condition as the omitted category, interacted with a binary indicator for if the manager reported being employed by a firm in the technology industry (17% of managers.) All specifications include stratum FE and the F1 specification includes controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1), the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber–White robust. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 23: Impact of Framing on Review Performance, by Organizations AI Positioning

| | F1 Score | Additional Review | Accountability Self | Accountability AI System |
|--|-------------------|-------------------|------------------------|-----------------------------|
| AI employee | 0.015 (0.022) | 0.004 (0.050) | -0.002 (0.031) | 0.02 (0.019) |
| Human employee | 0.000 (0.019) | 0.010 (0.044) | -0.035 (0.026) | -0.14*** (0.018) |
| AI employee × Org frames AI as employee | -0.052 (0.036) | 0.085 (0.081) | -0.054 (0.049) | 0.041 (0.031) |
| Org frames AI as employee | -0.015 (0.022) | 0.082* (0.049) | 0.000 (0.029) | 0.002 (0.013) |
| Num.Obs. | 695 | 795 | 795 | 795 |
| R2 | 0.133 | 0.020 | 0.027 | 0.197 |
| Mean (omitted category) | 0.476 | 0.427 | 0.61 | 0.14 |

Notes: This table reports treatment effects of the framing treatment on managers' performance outcomes, with those in the "AI Tool" condition as the omitted category, interacted with a binary indicator for if the manager reported being employed by a firm which positions AI as a teammate or employee (4 or 5 on a Likert scale.) All specifications include stratum FE and the F1 specification includes controls for the outcome at baseline. The sample includes all managers who passed an attention check. In Columns (1), the sample is narrowed to managers who submitted their reviewed documents. All standard errors are Huber-White robust.* p < 0.1, ** p < 0.05, *** p < 0.01

B Methods: Details

B.1 Outcomes

B.1.1 Error detection performance

Our primary performance outcome is micro-averaged $F1$ for error detection aggregated across all documents reviewed by participant i . Let:

- TP_i : number of errors correctly flagged,
- FP_i : number of non-errors flagged,
- FN_i : number of errors not flagged.

Precision and recall are:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \quad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i}.$$

The $F1$ score is:

$$F1_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}.$$

If a participant made no flags, we set $F1$ to missing.

B.1.2 Escalation

Request for additional review. We measure whether participants request another reviewer before finalization two ways. First, we ask them if they would like the documents to be additionally reviewed “Would you like to ask someone to conduct additional review of the documents you submitted?” (Yes/No/No answer). Because we believed most would say yes, we ask a second version with stakes, requesting review was rewarded when performance was low and penalized when performance was high, reducing purely expressive escalation. In the second question we added the following note “If you select ‘Yes’ and caught fewer than 50% of errors, you’ll gain 3 tickets (good job recognizing uncertainty!). If you caught more than 50%, you’ll lose 3 tickets (unnecessary review). If you select ‘No,’ you keep current tickets.”

B.1.3 Sign-off confidence.

Participants reported confidence in signing off on the reviewed materials on a 1–5 Likert scale ($confidence_i$). We additionally define:

$$HighConf_i = \mathbb{1}\{confidence_i \in \{4, 5\}\}.$$

B.1.4 Governance and resource allocation preferences

Delegation of decision rights. Participants recommended a governance structure for deploying an assistant like the one they worked with, ranging from autonomy to “assistance only” to “do not use.” We define:

$$HighDelegation_i = 1$$

if the participant recommends full or partial decision authority for the assistant, and 0 otherwise.

Resource allocation preference. Participants chose between hiring additional employee(s) versus investing the same amount in improving/integrating the AI system. We define:

$$InvestInAI_i = 1$$

if the participant recommends investing in the AI system, and 0 if they recommend hiring.

B.1.5 Post-task attitudes

Key attitude outcomes include adoption intent, excitement about AI, willingness to invest time to learn AI skills, and job insecurity, each measured on 1–5 Likert scales and also coded as high-endorsement indicators:

$$HighAdoption_i = \mathbb{1}\{adopt_i \in \{4, 5\}\}, \quad HighInsecurity_i = \mathbb{1}\{insecurity_i \in \{4, 5\}\}$$

$$HighExcited_i = \mathbb{1}\{excited_prod_i \in \{4, 5\}\}, \quad HighInvestTime_i = \mathbb{1}\{invest_time_i \in \{4, 5\}\}$$

B.1.6 Attention checks

Participants reported who they believed initially drafted the documents (AI tool vs. human employee vs. AI employee), used to assess whether framing was received as intended. We do not filter the sample on this, instead we use it as an outcome.

A second attention check asked the participants "Please answer number 3" during the block of likert scale questions where the possible answers were between 1 and 5.