When Corporate AI Adoption Backfires *

Joanne Juan Chen — *Boston University*

Brandon Yueyang Han — *University of Maryland*

March 15, 2026

[Click for the Latest Version]

Abstract

Firms are increasingly adopting predictive artificial intelligence (AI) to improve decision-making by combining advanced data analysis with managerial judgment. While AI provides more precise information to support managerial decision-making, its adoption can nevertheless reduce shareholder profits and the aggregate welfare, when both managerial and AI-provided information on project quality are highly precise and managerial bias is uncertain ex ante. In such cases, investment misallocation worsens, with overinvestment in low-quality projects and underinvestment in high-quality ones, even after the manager's compensation contract has been re-optimized. We propose remedies through information design and contracting approaches. Our analysis underscores the importance of governance reforms to guide AI adoption beyond simply enhancing AI's precision.

Key words: *AI-augmented Decision-making, Uncertain Bias, Optimal Contracting, Information Granularity, Information Design*

JEL codes: *G30, G34, D82, D86*

*We are grateful to Briana Chang, Yeon-Koo Che, Lin William Cong, Amil Dasgupta, Jerome Detemple, Daniel Ferreira, Naveen Gondhi, Alex Xi He, Zhiguo He, Yunzhi Hu, Shaowei Ke, John Kuong, Pete Kyle, Dan Luo, Andrew Newman, Dino Palazzo, Xuyuanda Qi, Jian Sun, Andrea Vedolin, Vish Viswanathan, Lucy White, Yizhou Xiao, Hao Xing, Liyan Yang, Hongda Zhong, Junyuan Zou, and all seminar and conference participants at Tsinghua Theory and Finance Workshop, CICF, SFS Cavalcade North America, Finance Theory Group Meeting in Hong Kong, Boston University Economics Micro Theory Workshop, and Boston University Questrom Brownbag for helpful comments and suggestions. All errors remain our own. Joanne Juan Chen: Department of Finance, Questrom School of Business, Boston University; Email: joanchen@bu.edu. Brandon Yueyang Han: University of Maryland Robert H. Smith School of Business; Email: yhan1@umd.edu.

1 Introduction

Firms across a wide range of industries are rapidly adopting artificial intelligence (AI) to support high-stakes investment and operational decisions. Human resources departments use machine-learning tools to screen job candidates, venture capital firms deploy algorithms to assess and rank startups, and companies increasingly rely on predictive tools to evaluate capital projects, acquisitions, and strategic initiatives. In these settings, AI has become central to information processing and managerial decision-making.

Much of the existing discussion of AI adoption focuses on advances in the technology itself. Efforts are devoted to improving model accuracy, expanding data inputs, and refining prediction algorithms,[1] often with the implicit assumption that more precise AI-generated information naturally leads to better decisions. Yet, for many high-stakes decisions, AI outputs are ultimately interpreted and acted upon by human managers, whose judgment, incentives, and discretion critically shape outcomes.[2] Understanding how AI-generated information is integrated into managerial decision-making, and how this integration interacts with managerial incentives, is therefore critical for assessing the value of corporate AI adoption. Despite its practical importance, this organizational dimension of AI integration remains relatively underexplored.

In this paper, we demonstrate that corporate AI adoption can backfire: despite improving information accuracy, it may reduce shareholder profits and even aggregate welfare. Absent appropriate organizational design and governance mechanisms, non-adoption may be optimal, even when the technology itself is highly reliable. Moreover, we characterize sufficient conditions under which corporate AI adoption backfires: (i) the manager's estimate of project quality without AI is highly precise; (ii) the AI's estimate of project quality is highly precise; and (iii) the manager may derive personal benefits from project approval, with the presence of such bias uncertain ex

[1] For instance, the annual Artificial Intelligence Index Report published by Stanford University tracks AI model performance, accessibility, and affordability over time (Maslej et al., 2025).

[2] A recent McKinsey & Company report shows that firms' difficulties in extracting value from AI primarily reflect organizational and incentive misalignment rather than technological limitations (Mayer et al., 2025).

ante. We further propose a set of remedies that address these challenges and enable effective AI integration.

Our core insight is that AI changes not only the accuracy of information but also its granularity. Human managers typically rely on coarse, categorical assessments when actions are discrete, focusing on a limited set of salient factors (e.g., Gabaix, 2014). Predictive AI, by contrast, generates finely differentiated estimates with greater dispersion (e.g., Fuster et al., 2022; Chen et al., 2024).[3] While such granularity is valuable in frictionless environments, it can be detrimental when managerial objectives diverge from those of shareholders. More granular AI estimates increase the prevalence of borderline decisions, where managers can act on personal bias while facing limited downside risk. The firm's optimal response of tightening incentives to curb these distortions entails additional costs, which in some cases outweigh the efficiency gains from improved information.

To formalize this mechanism, we develop a model of AI-augmented decision-making in which shareholders hire a manager to make an investment decision on a risky project. A project succeeds with a probability that depends on two components: a viability dimension, capturing soft, qualitative factors that are difficult for AI to assess, and a technical quality dimension, which is amenable to data-driven evaluation. The manager observes noisy signals about both dimensions and chooses whether to approve the project. He may also enjoy a private benefit from approval, such as personal preference, relationship-based considerations, or noncontractible perks. Crucially, this bias is uncertain ex ante and is privately observed by the manager.

Absent AI, the manager's information about technical quality is coarse, leading to a discrete distribution of posterior success probabilities. The principal designs an optimal contract that trades off investment efficiency against incentive costs. In this case, improving the manager's information precision always increases shareholder profits, even when managerial bias is uncertain, because more informative signals polarize posterior beliefs, relax incentive constraints, and reduce compensation costs. The introduction of AI fundamentally alters this logic. AI provides

[3] In Chen et al. (2024), the authors develop a novel method that utilizes recent advances in large language models (LLMs) to recover overlooked information in categorical variables.

a more precise estimate of technical quality that strictly Blackwell-dominates the manager's signal. However, once AI information is incorporated into decision-making, the distribution of perceived success probabilities becomes smoother and more concentrated around the investment threshold. As a result, a larger set of realizations emerges in which the manager's decision is sensitive to personal bias. The principal can mitigate these distortions by tightening incentives, but doing so requires higher expected compensation.

Our main result shows that when AI precision is sufficiently high and managerial bias is sufficiently uncertain, the increase in total agency costs induced by AI exceeds the efficiency gains from more accurate information. In this case, AI adoption backfires: shareholders are strictly worse off than in a world without AI. Moreover, because part of the loss reflects real misallocation, rather than pure transfers to the manager, aggregate welfare can also decline. These welfare losses arise even though AI improves information in the Blackwell sense and even though contracts are optimally adjusted after AI adoption.

Importantly, this backfiring mechanism does not arise in two benchmark cases. First, when there is no agency conflict, AI adoption always improves efficiency, as standard information theory would predict. Second, when managerial bias is deterministic and commonly known, the principal can fully offset it through a simple contract, and more precise information again weakly increases profits. It is the interaction between refined AI information and uncertainty about managerial bias that generates the inefficiency. This distinction helps reconcile our results with existing theories of information and incentives while highlighting a new channel through which advanced technologies can undermine firm performance.

After identifying the conditions under which AI adoption can backfire, we examine how firms can redesign information structures and incentive contracts to better integrate AI insights into managerial decision-making. We propose two remedies. The first is an information design approach in which the firm intentionally coarsens AI-generated information before communicating it to the manager. Instead of disclosing finely graded estimates, the firm groups AI predictions into a small number of categories aligned with decision-relevant thresholds. This coarsening reduces the prevalence of intermediate states where managerial bias distorts investment decisions,

while retaining most information pertinent to the investment decision. Because AI systems can credibly commit to a pre-specified disclosure rule, such information design is especially practical in algorithmic environments. The second remedy is an AI-contingent contract that links managerial compensation to AI estimates. By tailoring incentives to AI information, such contracts reduce investment distortions and lower expected agency costs. Each remedy provides a practical mechanism for integrating AI into organizations and resolves the incentive distortions underlying backfiring.

Related Literature This paper connects the classic literature on moral hazard and optimal contracting with the emerging literature on corporate AI adoption.

First, this paper builds on the classic literature on moral hazard, observability, and optimal contracting, pioneered by Mirrlees (1976), Holmström (1978, 1979), Holmström and Tirole (1997), among others. The novelty of this work lies in its focus on a setting with uncertain managerial bias, in which the manager's ability to derive private benefits from project approval is unknown ex ante. This environment is natural in organizational settings but remains underexplored in the literature.

We also add to the expanding body of literature on AI adoption and its implications. Much of the existing research focuses on the macroeconomic effects of AI, examining its impact on economic growth, employment, income inequality, and financial stability. Comunale and Manera (2024) provide a comprehensive review.

Research on corporate AI adoption and its implications is an emerging field. Empirical studies closely related to our work include Hoffman et al. (2018), Lyonnet and Stern (2022), Agarwal et al. (2023), Angelova et al. (2023), Dell'Acqua et al. (2023), Babina et al. (2024), and McElheran et al. (2025), among others. Hoffman et al. (2018) find that managers frequently overrule technology recommendations because they are biased or mistaken, leading to worse hires for the firm. Angelova et al. (2023) find that a majority of the judges underperform the algorithm when they make a discretionary override. In particular, McElheran et al. (2025) show that industrial AI adoption can initially backfire, reducing productivity and profits due

to organizational adjustment frictions, underscoring the importance of organizational design in determining the value of corporate AI adoption.

A growing theoretical literature studies AI adoption in organizational and decision-making environments, addressing a range of mechanisms that differ from the focus of this paper. For example, Weitzner (2024) analyzes algorithm aversion arising from asymmetric information and reputational concerns faced by human decision-makers. Zhong (2025) develops a framework for integrating humans and technologies in multilayered decision-making, showing that optimal authority depends on a decision maker's ability to correct errors relative to introducing new ones, with AI often most effective as the final decision maker. Huang and Ouyang (2025) examines the limitations of AI advisors in processing soft information, comparing unbiased but rigid AI advice with biased yet more interpretive human judgment in financial advising contexts. Ferreira and Li (2026) studies how AI adoption by CEOs alters information sharing with corporate boards, showing that AI can substitute for board advice, weaken monitoring incentives, and reduce CEO pay. In their setting, AI's skill-augmenting capability enhances firm value as it increases CEO productivity and alleviates moral hazard. Our paper focuses on the distinction between human and AI information processing and examines how human-AI integration may exacerbate incentive problems, distorting investment decisions, and destroying firm value and welfare, even when AI signals are highly precise.

The remainder of the paper is organized as follows. Section 2 presents the model setup. Section 3 characterizes optimal contracts and examines managerial decisions both before and after AI adoption. Section 4 analyzes the mechanism and the conditions under which AI adoption backfires. Section 5 introduces two remedies to address the challenges of AI integration. Section 6 concludes.

2 The Model

2.1 Model Setup

Shareholders (the principal / she) hire a manager (the agent / he) to evaluate and implement an investment project that requires cost c and yields payoff $y \in \{0, 1\}$. The project succeeds with probability p, which is not directly observed. The manager obtains information about p and chooses an action $a \in \{A, R\}$ based on the obtained information, where A denotes approval and R rejection. If the firm adopts predictive artificial intelligence, additional information about p becomes available for decision-making.

The Success Probability p:

The project's success probability p summarizes all information that can be potentially acquired about the payoff y.[4] This probability is determined by two layers: project viability ν and technical quality q. A project can succeed only if it is viable, and conditional on viability its success probability is determined by its technical quality:

$$p = \begin{cases} q, & \text{if } \nu = 1, \\ 0, & \text{if } \nu = 0. \end{cases} \tag{1}$$

The first layer ν captures project viability based on soft information, such as behavioral, organizational, and contextual factors that are difficult to observe or verify through formal data. For example, an entrepreneur may lack leadership, exhibit erratic behavior, or operate within a dysfunctional team, or a loan applicant may divert funds toward unrelated high-risk activities. Such factors can render a project incapable of success even when its hard metrics appear strong. Moreover, hard information itself may be strategically manipulated through selective reporting,

[4]Project outcomes involve inherent randomness, so some components of the payoff cannot be learned in advance. The variable p therefore represents the learnable component of project quality. Conditional on p, no further information can improve inference about y.

falsified metrics, or other forms of misrepresentation. In these cases, the project is not viable and $\nu = 0$. We assume that a fraction λ_ν of projects in the population are viable and $\nu = 1$.[5]

Conditional on being viable, the second layer q captures the project's technical quality. It reflects hard information such as quantifiable performance metrics and features accessible to data-driven evaluation. To allow for rich heterogeneity across projects, q is modeled as a continuous random variable uniformly distributed on $[0, 1]$, i.e., $q \sim U[0, 1]$.

Human Manager Information Processing:

Humans have limited time and cognitive capacity to process all relevant information, and therefore rely on a simplified, sparse model of the world that attends only to variables of first-order importance, as formalized in Gabaix (2014).[6] In the same spirit, we assume that the human manager focuses on a subset of factors summarized by X_q to infer project quality q and by X_ν to infer the viability indicator ν.

The signal X_q is a binary variable taking values in $\{H, L\}$, with $\mathbb{P}(X_q = H) = \theta_q$. This categorical formalization reflects the tendency of human decision makers to coarse-grain information when actions are discrete.[7] For simplicity, we abstract from the full specification of the joint distribution of q and X_q and instead characterize only the manager's posterior expectation $\mathbb{E}[q \mid X_q]$, which is relevant for his decision-making. This posterior expectation is denoted by

[5] In an earlier version of the paper, soft information was modeled as a continuous variable with an additive effect on the success probability. That formulation yields more complicated expressions but leads to the same economic insights.

[6] Humans cannot process all available information due to time constraints and cognitive limitations. This observation is extensively documented in the psychology literature (e.g., Pashler (1999) and Nobre and Kastner (2014)), and is the underlying idea behind the rational inattention literature in economics and finance (e.g., Sims (2003), Van Nieuwerburgh and Veldkamp (2010), and Gabaix (2014)). Sims (2003), Van Nieuwerburgh and Veldkamp (2010), and Gabaix (2014) adopt different approaches in modeling attention constraints but share similar spirits.

[7] Yang (2020) show that when information acquisition is endogenous, the optimal information structure is binary for binary decisions, as finer distinctions raise information costs without enhancing decision quality. In this paper, we follow the spirits of Yang (2020). To keep the model parsimonious, we take a reduced-form approach and treat X_q as exogenous. Endogenizing the manager's information structure would not change the mechanism of the model and would lead to qualitatively similar results.

$\widehat{q}_M$:

$$\widehat{q}_M \equiv \mathbb{E}[q|X_q] = \begin{cases} \zeta_H, & \text{when } X_q = H, \\ \zeta_L, & \text{when } X_q = L. \end{cases} \tag{2}$$

The gap $(\zeta_H - \zeta_L)$ measures how much the manager's posterior assessment of q shifts across signal realizations and thus captures the manager's information precision about project quality.

Similarly, the signal X_ν on project viability also takes values in $\{H, L\}$, with $\mathbb{P}(X_\nu = H) = \theta_\nu$. The manager's posterior expectation is denoted by $\widehat{\nu}$:

$$\widehat{\nu} \equiv \mathbb{E}[\nu|X_\nu] = \begin{cases} \eta_H, & \text{when } X_\nu = H, \\ \eta_L, & \text{when } X_\nu = L. \end{cases} \tag{3}$$

The gap $(\eta_H - \eta_L)$ measures the manager's information precision about project viability.

AI Information Processing:

In contrast to a human manager, artificial intelligence has difficulty capturing the soft-information-based viability, ν. However, AI excels at processing large volumes of data and can accommodate high-dimensional inputs $(X_{q1}, X_{q2}, \ldots, X_{qn})$ that would overwhelm a human decision-maker, allowing it to form a highly accurate estimate of the project's technical quality, q. Accordingly, we assume that the AI's inference about q is more accurate than that of a human manager and strictly subsumes his signal X_q. Specifically, AI's estimate of q is denoted by

$$\widehat{q}_{AI} \equiv \mathbb{E}[q|X_{q1}, X_{q2}, \ldots, X_{qn}] = \mathbb{E}[q|X_{q1}, X_{q2}, \ldots, X_{qn}, X_q],$$

which is more informative than the manager's estimate, $\widehat{q}_M$.

AI's estimate $\widehat{q}_{AI}$ is generally close to the true project quality q, reflecting the ability of modern AI technologies to process high-dimensional data. Nevertheless, the estimate may not be precise. We use the maximal integrated difference between the cumulative distribution functions of q and $\widehat{q}_{AI}$ to measure the inaccuracy of AI's estimate, denoted by $\epsilon(\widehat{q}_{AI})$:

$$\epsilon(\widehat{q}_{AI}) = \sup_{x \in [0,1]} \int_0^x [F_q(t) - F_{\widehat{q}_{AI}}(t)]dt. \tag{4}$$

$\epsilon(\widehat{q}_{AI})$ decreases as estimates become more informative, as will be analyzed in Section 2.2.

The Contract and Managerial Incentives:

Both the principal and the manager are risk neutral, and the manager is protected by limited liability. Upon hiring the manager, the principal offers a contract specifying compensation contingent on project outcomes, (w_S, w_R, w_F), where w_S is the wage if the project succeeds, w_F is the wage if it fails, and w_R is the wage if the manager rejects the project.

In addition, the manager may have a personal bias towards certain projects and enjoy a private benefit b from approval. This private benefit may arise from noncontractible perks associated with particular projects (e.g., travel or entertainment), informal side benefits, or other preference-based motives. It is uncertain *ex ante* and exists with probability φ. The manager learns its realization only after taking office and acquiring sufficient information about the project. We refer to it as the "**uncertain managerial bias**" and denote it by B. The variable B takes values in $\{0, b\}$, with $\mathbb{P}(B = b) = \varphi$.

Since B captures the manager's personal bias, it is assumed to be independent of the project's success probability p and its signals. In particular, a manager may have idiosyncratic reasons to favor certain proposals or applicants that are unrelated to their underlying quality or likelihood of success.[8] The realization of B is the manager's private information. We assume that the manager acts in the principal's favor when he is indifferent.

Assumptions on Parameter Ranges:

We assume that an unfavorable signal on viability, $X_\nu = L$, renders the project unprofitable, underscoring the pivotal role of viability in the decision. Absent AI, profitability further requires a favorable managerial signal on project quality, $X_q = H$, reflecting the scarcity of high-quality projects. The two signals are therefore non-redundant. The parameter restrictions below ensure these conditions hold:

$$\max\left\{\eta_L, \eta_H \zeta_L\right\} < c < \eta_H \zeta_H \,, \tag{5}$$

[8]For example, a loan officer might favor an applicant from the same cultural background or ethnic group, which has no clear or systematic effect on the probability of repayment.

In addition, we restrict the magnitude of the manager's bias for the reasons outlined below. We focus on environments in which, without AI, the manager approves profitable projects ($\mu_M > c$) and rejects unprofitable ones ($\mu_M < c$). We further assume that the managerial bias is sufficiently small so that retaining the manager remains optimal for the principal. The following bounds guarantee these outcomes:

$$b<\theta_\nu\left(1-\max\left\{\frac{\eta_L}{\eta_H},\frac{\zeta_L}{\zeta_H}\right\}\right)\min\Big\{(1-\varphi)\theta_q(\eta_H\zeta_H-c),\ \varphi(1-\theta_q)\big(c-\max\{\eta_L,\eta_H\zeta_L\}\big),\theta_q(\eta_H\zeta_H-c)-\frac{(\lambda_\nu-c)^2}{2\lambda_\nu\theta_\nu}\Big\}. \tag{6}$$

2.2 Preliminary Analyses

The manager forms expectations about the project's success probability based on the available information. Let $\mu = \mathbb{E}[p|\mathcal{I}]$ denote this expectation, where $\mathcal{I} = \{(X_\nu, X_q, B)\}$ in the absence of AI, and $\mathcal{I} = \{(X_\nu, \widehat{q}_{AI}, X_q, B)\}$ when the manager has access to the refined information provided by the predictive AI. We denote the corresponding expectations by μ_M and μ_{AM}, respectively.

Informativeness of Signals:

Our characterization and comparison of signal informativeness are consistent with Blackwell (1953). Formally, a signal X is Blackwell-more-informative than X' if X' is statistically equivalent to a garbled, noisy version of X, generated by a stochastic mapping from X.

Among the manager's possible signals about the project quality q, a less Blackwell-informative signal corresponds to a lower $\zeta_H - \zeta_L$. Consider a garbled signal X'_q that flips $X_q = L$ to H with probability $\tau\theta_q$ and reports it correctly otherwise, and flips $X_q = H$ to L with probability $\tau(1-\theta_q)$ and reports it correctly otherwise. The marginal distribution of X'_q remains identical to that of X_q. The difference in conditional expectation for the garbled signal is given by

$$\zeta'_H-\zeta'_L = \big[(1-\tau(1-\theta_q))\zeta_H+\tau(1-\theta_q)\zeta_L\big]-\big[\tau\theta_q\zeta_H+(1-\tau\theta_q)\zeta_L\big] < (1-\tau)(\zeta_H-\zeta_L),$$

so that the posterior gap is shrunk by a factor of $1-\tau$.

For the AI, the inaccuracy measure $\epsilon(\widehat{q}_{AI})$ respects the Blackwell order. Consider two estimates, $\widehat{q}_{AI}$ and $\widehat{q}'_{AI}$, which are the posterior means of q based on two different sets of signals, with $\widehat{q}_{AI}$ derived from the more informative set. The posterior mean tracks the true value of q more closely with more informative signals, whereas with less precise signals, it updates little from the prior mean. Consequently, the distribution of $\widehat{q}_{AI}$ across signal realizations is more dispersed than that of $\widehat{q}'_{AI}$. Formally, the distribution of $\widehat{q}_{AI}$ is a mean-preserving spread of that of $\widehat{q}'_{AI}$. From the integrated CDF condition, for all $x \in [0,1]$, we have

$$\int_{-\infty}^{x} F_{\widehat{q}_{AI}}(t)dt \geqslant \int_{-\infty}^{x} F_{\widehat{q}'_{AI}}(t)dt\,. \tag{7}$$

It follows from the definition in (4) that $\epsilon(\widehat{q}_{AI}) \leqslant \epsilon(\widehat{q}'_{AI})$.

Principal's Problem:

The principal's problem is to design the optimal contract (w_S, w_R, w_F) that maximizes her expected payoff given her information set. The principal's payoff is $1 - w_S - c$ if the project succeeds, $-w_F - c$ if the project fails, and $-w_R$ if the project is rejected by the manager. Thus, the principal's problem can be expressed as

$$\max_{(w_S,w_R,w_F)} \mathbb{E}\big[\big\{p(1-w_S-c) + (1-p)(-w_F-c)\big\}\mathbf{1}_{\{a=A\}} - w_R\mathbf{1}_{\{a=R\}}\big]\,, \tag{8}$$

where $\mathbf{1}_{\{a=A\}}$ and $\mathbf{1}_{\{a=R\}}$ are indicator functions denoting the manager's actions – approval and rejection, respectively. Since $\mathbf{1}_{\{a=A\}}$ and $\mathbf{1}_{\{a=R\}}$ are both $\mathcal{I}$-measurable and $\mu = \mathbb{E}[p|\mathcal{I}]$, the problem can be rewritten using the law of iterated expectations as follows:

$$\max_{(w_S,w_R,w_F)} \mathbb{E}\big[\big\{\mu(1-w_S-c) + (1-\mu)(-w_F-c)\big\}\mathbf{1}_{\{a=A\}} - w_R\mathbf{1}_{\{a=R\}}\big]\,. \tag{9}$$

First-step Analysis of the Contract:

The principal sets $w_S \geqslant w_F$ and $w_R \geqslant w_F$ to incentivize the manager to approve projects with high success probabilities and reject those with low success probabilities. For any contract triplet (w_S, w_R, w_F) where $w_F > 0$, an equivalent contract $(w_S - w_F, w_R - w_F, 0)$ achieves the same incentive at a reduced cost. Therefore, in the optimal contract, w_F should be set to 0, and the optimal contract takes the form $(w_S, w_R, 0)$.

Manager's Decision Rule:

Given the contract, the manager approves the project if approval yields a weakly higher expected payoff for him than rejection, with indifference resolved in favor of the principal:

$$\begin{cases} \mu w_S + (1-\mu) w_F + b & > & w_R, & \text{if } B = b, \\ \mu w_S + (1-\mu) w_F & \geqslant & w_R, & \text{if } B = 0. \end{cases} \tag{10}$$

He rejects the project if the reverse inequalities hold.

Figure 1 plots the manager's expected payoff from approval or rejection as a function of μ in these two cases. For simplicity, we set $w_F = 0$, following the previous analysis. Hence, the manager's expected payoff from approving the project is $w_S\mu + b$ when the private benefit is present, and $w_S\mu$ otherwise. His expected payoff from rejecting the project is w_R.

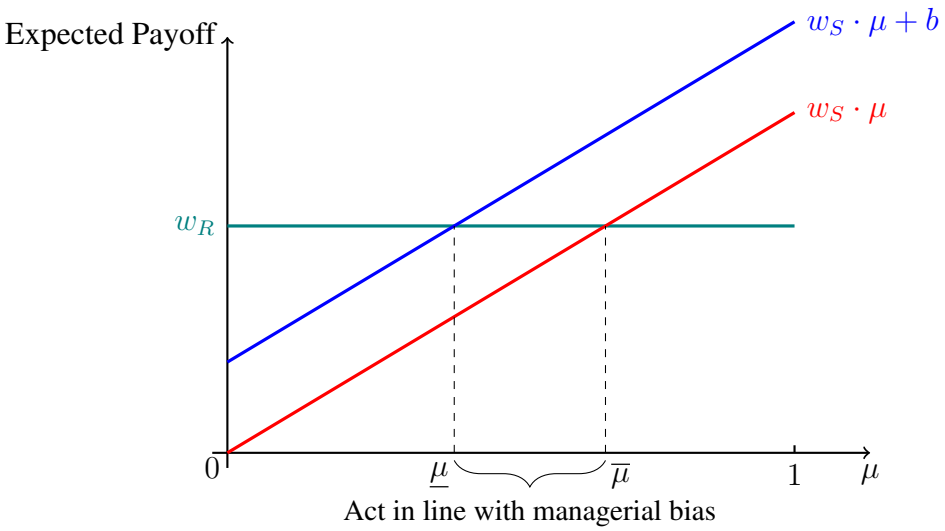

Figure 1: Manager's Expected Payoffs under Information Set $\mathcal{I}$

From Figure 1, we observe that there exists a range of μ for which the manager approves the project when the private benefit is present and rejects it otherwise. We refer to this behavior as "**acting in line with managerial bias**". The two thresholds in this range correspond to the points where the manager's expected payoff from rejection intersects with the expected payoff from approval, with and without private benefit, respectively.

Let $\underline{\mu}$ and $\overline{\mu}$ denote these two thresholds. Then,

$$\underline{\mu} = \frac{w_R - b}{w_S}, \quad \overline{\mu} = \frac{w_R}{w_S}. \tag{11}$$

The manager acts in line with his personal preference when μ lies in the interval $(\underline{\mu}, \overline{\mu})$. Therefore, the manager's decision rule as a function of his posterior belief about the project's success probability, μ, can be illustrated by the following diagram over the interval $[0, 1]$.

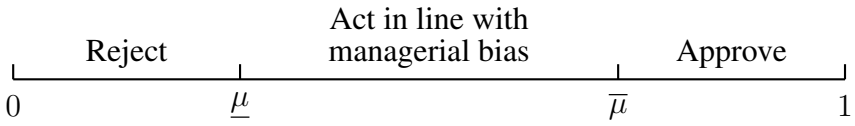

Figure 2: Manager's Decision Rule as a Function of μ

3 Optimal Contracts Before and After AI Adoption

In this section, we characterize the optimal contracts and analyze managerial decisions and expected payoffs. Section 3.1 describes the equilibrium without AI. Section 3.2 examines how AI affects the manager's assessment of the project and, in turn, alters managerial decisions, leading to an equilibrium in which the principal trades off managerial incentive costs against losses from suboptimal investment decisions.

3.1 Optimal Contract Before AI Adoption

According to previous analyses, the manager makes predictions on p based on X_q and X_ν in the absence of AI. The manager's expectation in this case, denoted by μ_M, is given by:

$$\mu_M = \mathbb{E}[p|X_\nu, X_q] = \eta_i \zeta_j, \quad \text{for } X_\nu = i \text{ and } X_q = j, \text{ where } i, j \in \{H, L\}. \tag{12}$$

These four realizations of μ_M each occur with probabilities of $\theta_\nu\theta_q$, $\theta_\nu(1-\theta_q)$, $(1-\theta_\nu)\theta_q$ and $(1-\theta_\nu)(1-\theta_q)$, respectively.[9]

[9]To be more precise, $\mu_M = \mathbb{E}[p|\mathcal{I}] = \mathbb{E}[p|X_\nu, X_q, B]$ in this scenario. We exclude private benefit B from the conditioning information in this equation because B is independent of both the success probability p and the signals, X_ν and X_q.

The principal chooses w_S and w_R, which in turn determine the manager's decision thresholds, $\underline{\mu}$ and $\overline{\mu}$. The principal would like to set both thresholds close to the investment cost c, so that profitable projects are approved and loss-making projects are rejected. However, placing both thresholds near c entails high incentive costs, as both w_S and w_R increase with the distance between the thresholds, $\overline{\mu}-\underline{\mu}$:

$$w_S = \frac{b}{\overline{\mu}-\underline{\mu}}\,, \quad w_R = \frac{\overline{\mu}b}{\overline{\mu}-\underline{\mu}}\,, \tag{13}$$

as implied by Equations (11). Consequently, the principal optimally trades off investment efficiency and incentive costs and maintains a positive distance between the decision thresholds.

In equilibrium, the principal chooses the wages such that $\overline{\mu} = \eta_H\zeta_H$ and $\underline{\mu} = \max\{\eta_H\zeta_L, \eta_L\zeta_H\}$. Setting $\overline{\mu}$ above $\eta_H\zeta_H$ would forgo profitable investment opportunities, while setting $\underline{\mu}$ below $\max\{\eta_H\zeta_L, \eta_L\zeta_H\}$ would induce the acceptance of unprofitable projects. This equilibrium is characterized in Proposition 1 below.

Proposition 1 *In the absence of AI,*

1.) the optimal contract is

$$w_S = \frac{b}{\eta_H\zeta_H - \max\{\eta_H\zeta_L, \eta_L\zeta_H\}}, \quad w_R = \max\left\{\frac{\zeta_H b}{\zeta_H-\zeta_L}, \frac{\eta_H b}{\eta_H-\eta_L}\right\}, \quad w_F = 0\,;$$

2.) the manager's action rule is

to approve the project if $(X_\nu, X_q) = (H, H)$, and to reject the project otherwise;

3.) the principal's expected profit, denoted by $\pi_{B,M}$, is given by

$$\pi_{B,M} = \mathbb{E}\big[(\mu_M(1-w_S)-c)\mathbf{1}_{\{a=A\}} - w_R\mathbf{1}_{\{a=R\}}\big] = \theta_\nu\theta_q(\eta_H\zeta_H-c) - \max\left\{\frac{\zeta_H b}{\zeta_H-\zeta_L}, \frac{\eta_H b}{\eta_H-\eta_L}\right\}.$$

In this equilibrium, the manager approves the project if and only if he receives two good signals, which occurs with probability of $\theta_\nu\theta_q$. In this case, the manager's estimate of the project's success probability, μ_M, equals $\eta_H\zeta_H$. Although the manager may occasionally obtain private benefits, these do not affect his decision-making: in equilibrium, the manager's actions are independent of his personal bias.

The principal's expected profit is denoted by $\pi_{B,M}$, where the subscript B captures the uncertain presence of private benefits, while M indicates that project information is acquired by

the manager. The principal's expected profit equals the expected project payoff $\theta_\nu\theta_q(\eta_H\zeta_H - c)$, minus the expected wage paid to the manager $\max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}$. Result 1 below demonstrates how it varies with the manager's information precision in equilibrium.

Result 1 *The principal's expected profit is strictly increasing in the precision of the manager's information on project quality, $\zeta_H - \zeta_L$.*

This result follows directly from differentiating the principal's expected profit, $\pi_{B,M}$, in Proposition 1. The underlying intuition is as follows. Higher information precision allows the manager to more accurately distinguish between high-quality and low-quality projects. In addition, as $\zeta_H - \zeta_L$ increases, the manager's posterior belief μ_M becomes more polarized, which weakly relaxes the incentive constraint and thereby reduces incentive costs. As a result, the principal's expected profit increases when the manager has access to more precise information on project quality q.

Result 2 *The manager's expected payoff,*

$$\max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\} + \theta_\nu\theta_q\phi b\,,$$

weakly decreases in the precision of the manager's information on project quality $\zeta_H - \zeta_L$.

The manager's expected payoff consists of his expected wage and private benefit, given by $\max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}$ and $\theta_\nu\theta_q\phi b$ respectively. The expected wage depends on the precision of the information component that binds the incentive constraint. Specifically, if $\eta_H\zeta_L > \eta_L\zeta_H$, it decreases with $\zeta_H - \zeta_L$; otherwise, it decreases with $\eta_H - \eta_L$. The expected private benefit is unaffected by either $\zeta_H - \zeta_L$ or $\eta_H - \eta_L$. This analysis demonstrates that the manager's total information rent declines with the precision of his information. This finding runs counter to the conventional wisdom in the contracting literature, which typically suggests that agents with more informative private signals extract greater rents due to their informational advantage.

3.2 AI-augmented Decision-making

AI can analyze large volumes of data and handle numerous variables related to the project's inherent quality and technical soundness, allowing it to distinguish subtle differences in project quality that the manager would treat as indistinguishable. Consequently, the AI's estimate on the project quality, $\widehat{q}_{AI}$, is more granular, capturing substantially richer underlying heterogeneity than the manager's coarser assessment $\widehat{q}_M$.

With AI's estimate on the project quality, $\widehat{q}_{AI}$, the manager's expectation of p is[10]

$$\mu_{AM} = \mathbb{E}[p|X_\nu, \widehat{q}_{AI}, X_q] = \begin{cases} \eta_H\, \widehat{q}_{AI}, & \text{if } X_\nu = H\,, \\ \eta_L\, \widehat{q}_{AI}, & \text{if } X_\nu = L\,. \end{cases} \tag{14}$$

Let $F(\mu_{AM})$ denote the cumulative distribution function (CDF) of μ_{AM} and $f(\mu_{AM})$ its density when it exists. For expositional convenience, we adopt a slight abuse of notation by using $F(\cdot)$ to denote all CDFs of the manager's expectation μ in different cases, with the understanding that the specific distribution is determined by context.[11]

The principal's problem is to design a contract that maximizes the project's expected payoff net of managerial compensation, taking into consideration the AI-augmented decision-making process of the manager. Because of the one-to-one mapping between the wage pair (w_S, w_R) and the two thresholds $(\overline{\mu}, \underline{\mu})$ as shown in Equations (13), the problem can be reformulated as a maximization over $(\overline{\mu}, \underline{\mu})$. Specifically, it can be rewritten as follows, with the expectation expressed in integral form:

$$\max_{(\overline{\mu},\, \underline{\mu})} \int_{\overline{\mu}}^{1} \big[\mu_{AM}(1-w_S) - c\big] dF(\mu_{AM}) + \int_{\underline{\mu}}^{\overline{\mu}} \big[\varphi\big(\mu_{AM}(1-w_S) - c\big) + (1-\varphi)(-w_R)\big] dF(\mu_{AM}) + \int_{0}^{\underline{\mu}} (-w_R) dF(\mu_{AM})\,. \tag{15}$$

The principal's expected profit takes this form because the manager approves the project with certainty if $\mu_{AM} \geqslant \overline{\mu}$, approves it with probability φ if $\mu_{AM} \in (\underline{\mu}, \overline{\mu})$, and rejects it with

[10] Similar to the analysis of μ_M in Equation (12), we exclude B from the conditioning information because it is independent of ν, q, X_ν, and X_q.

[11] For example, $F(\mu_M)$ and $F(\mu_{AM})$ refer to the CDFs of μ_M and μ_{AM}, respectively.

certainty if $\mu_{AM} \leqslant \underline{\mu}$, as analyzed in Section 2.2. This expression can be further decomposed into the following three components:

$$\underbrace{\int_c^1 (\mu_{AM} - c) dF(\mu_{AM})}_{\text{No-agency-conflict Surplus}} - \underbrace{\left[\varphi \int_{\underline{\mu}}^{c} (c - \mu_{AM}) dF(\mu_{AM}) + (1-\varphi) \int_c^{\overline{\mu}} (\mu_{AM} - c) dF(\mu_{AM})\right]}_{\text{Surplus Loss from Suboptimal Investment Decisions}}$$

$$- \underbrace{\left[\left(\int_{\overline{\mu}}^1 \mu_{AM} dF(\mu_{AM}) + \varphi \int_{\underline{\mu}}^{\overline{\mu}} \mu_{AM} dF(\mu_{AM})\right) w_S + \left(\varphi F(\underline{\mu}) + (1-\varphi) F(\overline{\mu})\right) w_R\right]}_{\text{Managerial Compensation}} . \quad (16)$$

The first component represents the project's expected surplus absent agency conflict ($\varphi = 0$), where the manager makes the investment decision on behalf of the principal. In this case, a project is approved if and only if $\mu_{AM} > c$. The second component captures the expected loss of project surplus arising from suboptimal investment decisions. When the manager may be biased ($\varphi > 0$), his decision rule differs from the no-agency-conflict benchmark. Specifically, when the bias is present ($B = b$), the manager approves projects with $\mu_{AM} \in (\underline{\mu}, c)$ that should be rejected; conversely, when it is not ($B = 0$), the manager rejects projects with $\mu_{AM} \in (c, \overline{\mu})$ that should be approved. Each scenario occurs with positive probability, reducing project surplus relative to the benchmark. The third component represents managerial compensation, reflecting the cost of incentivizing the manager so as to increase the likelihood that he makes good decisions.

The principal aims to limit project surplus loss from suboptimal investment decisions, which requires tightening the decision thresholds by increasing $\underline{\mu}$ and decreasing $\overline{\mu}$. However, tightening the thresholds is costly, since managerial compensation rises with $\underline{\mu}$ and falls with $\overline{\mu}$, as shown in Equation (13). In equilibrium, the principal selects $\underline{\mu}$ and $\overline{\mu}$ to minimize the sum of managerial compensation and project surplus loss from suboptimal decisions, thereby maximizing Expression (16). Proposition 2 characterizes the results from this optimization.

Proposition 2 *With AI in place,*

1.) the optimal contract is

$$w_S = \frac{b}{\overline{\mu}^* - \underline{\mu}^*}, \quad w_R = \frac{\overline{\mu}^* b}{\overline{\mu}^* - \underline{\mu}^*}, \quad w_F = 0;$$

where $(\overline{\mu}^*, \underline{\mu}^*)$ *maximizes the principal's expected profit in Expression* (15)*. In particular, when* μ_{AM} *is continuously distributed,* $\overline{\mu}^*$ *and* $\underline{\mu}^*$ *are implicitly determined by the following system of equations:*

$$\begin{cases} (1-\varphi)(\overline{\mu}^*-c)f(\overline{\mu}^*)=\dfrac{b}{(\overline{\mu}^*-\underline{\mu}^*)^2}\left[\displaystyle\int_{\overline{\mu}^*}^{1}\mu_{AM}dF(\mu_{AM})+\varphi\int_{\underline{\mu}^*}^{\overline{\mu}^*}\mu_{AM}dF(\mu_{AM})+\varphi\underline{\mu}^*F(\underline{\mu}^*)+(1-\varphi)\underline{\mu}^*F(\overline{\mu}^*)\right] \\ \varphi(c-b-\underline{\mu}^*)f(\underline{\mu}^*)=\dfrac{b}{(\overline{\mu}^*-\underline{\mu}^*)^2}\left[\displaystyle\int_{\overline{\mu}^*}^{1}\mu_{AM}dF(\mu_{AM})+\varphi\int_{\underline{\mu}^*}^{\overline{\mu}^*}\mu_{AM}dF(\mu_{AM})+\varphi\overline{\mu}^*F(\underline{\mu}^*)+(1-\varphi)\overline{\mu}^*F(\overline{\mu}^*)\right]; \end{cases}$$

2.) *the manager's action rule is to approve the project with certainty if* $\mu_{AM} \geqslant \overline{\mu}^*$*, to approve it when private benefits exist if* $\mu_{AM} \in (\underline{\mu}^*, \overline{\mu}^*)$*, and to reject it with certainty if* $\mu_{AM} \leqslant \underline{\mu}^*$*;*

3.) *the principal's expected payoff* $\pi_{B,AM}$ *is given by Expression (15) with* $(\overline{\mu}, \underline{\mu}) = (\overline{\mu}^*, \underline{\mu}^*)$*.*

In this equilibrium, the manager acts in line with his personal bias when the expected probability of success, μ_{AM}, falls within the range $(\underline{\mu}^*, \overline{\mu}^*)$. The principal tolerates these potentially suboptimal decisions because narrowing the gap between these decision thresholds would impose excessively high incentive costs. Thus, suboptimal investment decisions occur with positive probability in equilibrium.

Result 1 shows that in the absence of AI, the principal's expected profit is strictly increasing in the precision of the manager's estimate on project quality. Result 3 below examines how the precision of AI's estimate of project quality affects the principal's expected profit.

Result 3 *The principal's expected profit does not necessarily increase and may instead diminish as AI's estimate of the project quality* $\widehat{q}_{AI}$ *becomes more precise.*

Because the AI's estimate of project quality subsumes that of the manager, its information precision Blackwell dominates the manager's. However, greater precision of the AI's information on project quality need not improve the principal's expected profit, unlike the manager's signal.

The main reason for these two contrasting results is that greater precision of the manager's signal induces a more polarized distribution of posterior estimates, which relaxes the incentive constraint. By contrast, a more precise AI estimate is more granular and captures richer heterogeneity in project quality, thereby tightening the incentive constraint, even though more precise

information always increases the no-agency-conflict surplus. Therefore, the effect of AI adoption on the principal's profit is ambiguous. This observation is central to our analysis and underlies the possibility that AI adoption may backfire, a mechanism we examine in detail in the following section.

4 When AI Adoption Backfires

AI improves the precision of success probability estimates by providing a more accurate estimate of the project quality. Specifically, the AI-augmented estimate, μ_{AM}, provides a more precise assessment of the success probability p than the traditional manager-based estimate μ_M. This improved accuracy in assessing the likelihood of project success may translate into better-informed and more efficient decisions.

For instance, when the signals are mixed, such as when $X_\nu = H$ and $X_q = L$, the manager rejects the investment based on information available to him, which is an optimal decision. However, this may no longer be optimal if the firm adopts AI, which uncovers additional information and insights. AI may assess the project quality more favorably after identifying additional factors and underlying statistical patterns. If AI's estimate of the quality, $\widehat{q}_{AI}$, is higher than the threshold c/η_H, the project should be accepted rather than rejected. Conversely, when both X_ν and X_q are equal to H, the manager perceives the project as promising. However, if AI uncovers other relevant factors indicating that the project quality is below the investment threshold, the project should be rejected despite the manager's positive assessment initially.

In an ideal world without agency conflicts, where managers' interests are fully aligned with those of shareholders, AI adoption always improves decision-making efficiency. By providing more precise information about a project's likelihood of success, AI enables managers to better distinguish between worthwhile and unworthy investments. This result follows directly from Blackwell's theorem, which implies that more informative signals lead to better decisions in the absence of frictions.

Importantly, even when agency conflicts are present, AI adoption continues to deliver unambiguous benefits as long as the manager's bias is non-stochastic and known to the principal. In this case, the principal can fully account for the manager's bias when designing the optimal contract, so AI-enhanced information improves alignment without introducing new distortions and therefore does not backfire. The following lemma formally establishes results in these two benchmark cases.

Proposition 3 ***(Deterministic Bias with General Information Structure)***
Consider a manager with a commonly known deterministic bias, $B \equiv b$, and an arbitrary information set $\mathcal{I}$. $\forall\, b \geqslant 0$,

1.) *the optimal contract is $(w_S, w_R, w_F) = (0, b, 0)$;*

2.) *the manager's action rule is*
to approve the project if $\mu = \mathbb{E}[p|\mathcal{I}] \in [c - b, 1]$, and to reject the project otherwise;

3.) *the principal's expected profit is*
$$\int_0^1 \max\{\mu - c, -b\} dF(\mu)\,;$$

4.) *the principal's expected profit from the project is strictly higher when the manager has access to more information. Formally, for any information set $\mathcal{I}' \supseteq \mathcal{I}$,*
$$\int_0^1 \max\{\mu' - c, -b\} dF(\mu') \geqslant \int_0^1 \max\{\mu - c, -b\} dF(\mu)\,, \quad \textit{where } \mu' = \mathbb{E}[p|\mathcal{I}'].$$

When the manager has a commonly known deterministic bias of magnitude b, the principal can fully align incentives by offering a contract with a rejection wage of $w_R = b$, which compensates the manager for the forgone private benefit in the event of rejection. Under this contract, the manager approves the project if and only if $\mu \geqslant c - b$. Because the managerial bias is perfectly offset and no additional distortions are introduced, more precise information directly translates into more efficient outcomes by Blackwell's theorem.

The no-agency-conflict benchmark, in which $b = 0$, is a special case of the above analysis. Let $\pi_{0,AM}$ and $\pi_{0,M}$ denote principal's expected profit in this benchmark, with and without AI, respectively. The difference between the two is referred to as the "**efficiency gain from**

information", which is always nonnegative:

$$\pi_{0,AM} - \pi_{0,M} = \int_c^1 (\mu_{AM} - c) dF(\mu_{AM}) - \int_c^1 (\mu_M - c) dF(\mu_M) \geqslant 0\,.$$

Similarly, for a deterministic managerial bias at level b, the principal's expected profit with AI is weakly higher than that without AI, that is, $\pi_{b,AM} \geqslant \pi_{b,M}$.

However, when the managerial bias is uncertain, the above result may no longer hold. In this case, additional information from AI does not guarantee higher expected profits for the principal. This contrasts with Result 1, which shows that more precise information from the manager consistently improves the principal's profit, even in the presence of uncertain managerial bias.

The possibility that AI-augmented managerial decisions may backfire in corporate settings is the central focus of our paper. The intuition stems from the interaction between the uncertain managerial bias and the distinct information structures with and without AI. The subsequent two paragraphs elaborate on these two aspects sequentially.

When managerial bias is uncertain, there exists an interval $(\underline{\mu}, \overline{\mu})$ within which the manager's decisions are driven entirely by his bias, as illustrated in Figures 1 and 2. Within this interval, the manager approves a project when the private benefit is present, which occurs with probability φ, and rejects it otherwise, with probability $1 - \varphi$, regardless of the project's likelihood of success. Notably, this interval does not exist when the managerial bias is deterministic.

The role of the information structure is illustrated by contrasting the distributions of posterior estimates with and without AI. Without AI, the manager focuses on a limited number of salient factors, leading to a multi-peaked distribution of the posterior estimate of the project's success probability, μ_M. By contrast, AI processes a broad set of variables and large volumes of data, producing a smoother and less spiky distribution for the AI-enhanced estimate, μ_{AM}. As shown in Figure 3, this shift from the red arrows to the blue shaded areas increases the likelihood that μ_{AM} falls within the interval $(\underline{\mu}, \overline{\mu})$ in which the manager's action is driven by personal bias, thereby raising the probability of suboptimal investment decisions, making it possible for AI adoption to backfire.

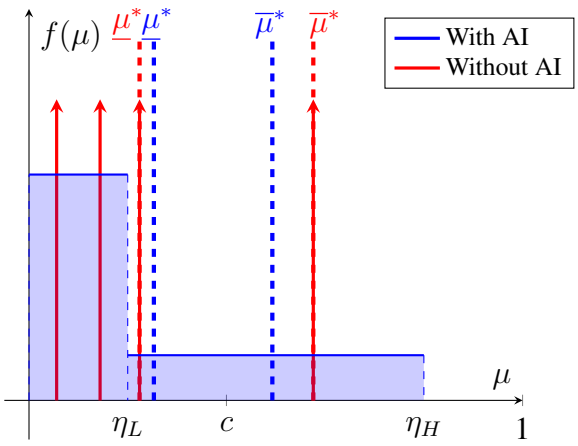

Figure 3: Distribution of Posterior Expectations and the Decision Thresholds

The figure depicts the distributions of managers' posterior expectations without AI, denoted by μ_M, and with AI, denoted by μ_{AM}. The distribution of μ_M is discrete and represented by four red arrows, whereas the distribution of μ_{AM} is illustrated by the blue shaded areas. The dashed vertical lines at $\underline{\mu}^*$ and $\overline{\mu}^*$ represent the optimal decision thresholds. For this illustration, the AI technology is assumed to infer project quality without error $\widehat{q}_{AI} = q$. Parameter values are $\varphi = 0.5$, $b = 0.005$, $c = 0.4$, $\lambda_\nu = 0.5$, $\theta_\nu = 0.5$, $\theta_q = 0.5$, $\eta_H = 0.8$, $\eta_L = 0.2$, $\zeta_H = 0.72$, and $\zeta_L = 0.28$.

In equilibrium, the principal sets the thresholds $\overline{\mu}^*$ and $\underline{\mu}^*$ by trading off the loss from suboptimal investment against the incentive cost required to mitigate this loss. A higher probability mass within this interval increases the incidence of biased, suboptimal decisions. The principal responds by shifting the thresholds inward to reduce such decisions, as shown in Figure 3, but this adjustment entails higher incentive costs.

To quantify this trade-off, we define the total agency cost as the sum of the "Surplus Loss from Suboptimal Investment Decisions" and "Managerial Compensation". The total agency cost with AI corresponds to the gap between the principal's expected profits with AI in the presence and absence of agency conflicts, $\pi_{0,AM} - \pi_{B,AM}$. When this total agency cost is sufficiently large, exceeding the sum of the total agency cost without AI $\pi_{0,M} - \pi_{B,M}$, and the efficiency gain $\pi_{0,AM} - \pi_{0,M}$, AI adoption backfires, leading to a lower expected profit than without AI:

$$\pi_{B,AM} - \pi_{B,M} = \underbrace{\left(\pi_{0,AM} - \pi_{0,M}\right)}_{\text{Efficiency Gain from AI}} + \underbrace{\left(\pi_{0,M} - \pi_{B,M}\right)}_{\text{Total Agency Cost without AI}} - \underbrace{\left(\pi_{0,AM} - \pi_{B,AM}\right)}_{\text{Total Agency Cost with AI}} < 0\,.$$

Figure 4 illustrates how the principal's profits with and without AI vary with the probability that the private benefit is present, φ. In line with Proposition 3, AI adoption always increases

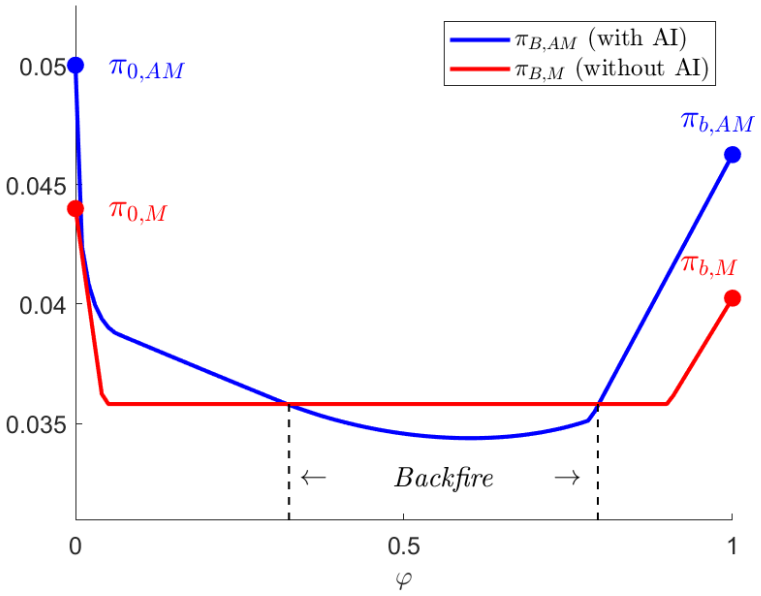

Figure 4: Uncertain Bias and AI Adoption Backfire

The figure shows the principal's profit as a function of φ, with AI represented by the blue curve and without AI by the red curve. The no-agency conflict benchmarks, $\pi_{0,AM}$ and $\pi_{0,M}$, along with the deterministic managerial bias cases, $\pi_{b,AM}$ and $\pi_{b,M}$, are highlighted using blue and red dots. AI adoption reduces the principal's profits whenever the blue curve lies below the red. Parameter values are $b = 0.005$, $c = 0.4$, $\lambda_\nu = 0.5$, $\theta_\nu = 0.5$, $\theta_q = 0.5$, $\eta_H = 0.8$, $\eta_L = 0.2$, $\zeta_H = 0.72$, and $\zeta_L = 0.28$.

the principal's profit in the absence of agency conflicts ($\varphi = 0$) or when managerial bias is deterministic ($\varphi = 1$). This result also extends to cases where φ is close to 0, in which the principal effectively ignores the low-probability managerial bias, and to cases where φ is close to 1, in which the principal can implement a contract similar to the $\varphi = 1$ case and safely treat the rare instances in which the private benefit is absent as negligible. However, for intermediate values of φ, where managerial bias is highly uncertain and not negligible, the increase in total agency costs induced by AI adoption can outweigh the associated efficiency gains, causing AI adoption to backfire.

Proposition 4 below characterizes sufficient conditions for corporate AI adoption to backfire.

Proposition 4 ***(Sufficient Conditions for AI Adoption to Backfire)***

A sufficient condition for AI adoption to reduce shareholders' expected profits is that both the manager's and the AI's estimates of project quality are sufficiently precise, and that the

managerial bias occurs with an intermediate level of probability. Specifically,

$$\begin{cases} \zeta_H - \zeta_L \geqslant \zeta^*, \\ \epsilon(\widehat{q}_{AI}) \leqslant \epsilon^*(\zeta_H - \zeta_L), \\ \varphi \in \left[\underline{\varphi}^*\big(\zeta_H - \zeta_L, \epsilon(\widehat{q}_{AI})\big), \overline{\varphi}^*\big(\zeta_H - \zeta_L, \epsilon(\widehat{q}_{AI})\big) \right], \end{cases}$$

where ζ^* is the solution to

$$\theta_\nu\theta_q\eta_H(1-\theta_q)\zeta^* - \frac{b}{2\zeta^*} = \frac{\theta_\nu(\eta_H-c)^2}{2\eta_H} - \theta_\nu\theta_q\left(\frac{\eta_H}{2}-c\right) - \frac{3}{2}\left(\frac{\theta_\nu c^2 b^2}{4\eta_H}\right)^{\frac{1}{3}} + \left(\frac{3}{2}-\theta_q\right)b\,, \quad (17)$$

ϵ^* is the solution to

$$\left[\left(\frac{\theta_\nu c^2 b^2}{4\eta_H}\right)^{\frac{1}{3}} + b\right] 2\eta_H\sqrt{\epsilon^*} = \theta_\nu\theta_q(\eta_H\zeta_H - c) - \frac{\theta_\nu(\eta_H-c)^2}{2\eta_H} + \frac{3}{2}\left(\frac{\theta_\nu c^2 b^2}{4\eta_H}\right)^{\frac{1}{3}} - \left(\frac{\zeta_H}{\zeta_H-\zeta_L}+\frac{1}{2}\right)b\,, \quad (18)$$

and $\underline{\varphi}^*, \overline{\varphi}^*$ are the two solutions to the following equation for φ:

$$\left(\frac{3}{2} - 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)\left(\frac{\theta_\nu c^2 b^2\varphi(1-\varphi)}{\eta_H}\right)^{\frac{1}{3}} - \left(\frac{\zeta_H}{\zeta_H-\zeta_L} + \varphi + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)b = \frac{\theta_\nu(\eta_H-c)^2}{2\eta_H} - \theta_\nu\theta_q(\eta_H\zeta_H - c). \quad (19)$$

As demonstrated in Proposition 4, backfiring occurs when both managerial and AI-provided information on project quality are highly precise, and managerial bias is uncertain ex-ante. The intuition regarding the role of precise managerial information is straightforward: when the manager already possesses highly accurate information, he is likely to take the correct action without AI assistance. Consequently, AI-augmented decision-making has a limited scope to improve outcomes.

Notably, high accuracy of the AI estimate contributes to backfiring. It is helpful to revisit Result 3, which shows that the principal's expected profit does not necessarily increase, and can even decline, as the AI's estimate $\widehat{q}_{AI}$ becomes more precise. The mechanism stems from the interaction between AI information and managerial bias: refinement in the AI-provided signal partitions projects into finer categories, producing more posterior realizations in the intermediate range where managerial bias influences decisions. Consequently, total agency cost rises relative to the case with less precise AI information.

The increase in total agency cost outweighs efficiency gains from more precise information when the AI inaccuracy $\epsilon(\widehat{q}_{AI})$ falls below the threshold $\epsilon^*(\zeta_H - \zeta_L)$, and in particular, when the AI perfectly reveals project quality and $\epsilon(\widehat{q}_{AI}) = 0$. This suggests that simply introducing more accurate AI does not automatically improve firm performance. Realizing the full potential of AI requires a careful calibration of managerial incentives and information flows, a topic we discuss in detail in Section 5.

In some cases, AI adoption could even reduce aggregate welfare. Although, in the analysis above, a decline in the principal's profits is often accompanied by an increase in managerial compensation, this increase does not always offset the principal's loss. Instead, surplus losses arising from suboptimal investment decisions may exceed the efficiency gains from AI information, in which case aggregate welfare declines. Proposition 5 characterizes the conditions under which corporate AI adoption leads to welfare loss.

Proposition 5 ***(AI Adoption and Welfare Loss)***

With more stringent conditions compared to Proposition 4, the aggregate welfare of the principal and the manager decreases with AI adoption. Specifically, there exists $\zeta^{**} > \zeta^*$, $\epsilon^{**}(\cdot) < \epsilon^*(\cdot)$, *and* $\left[\underline{\varphi}^{**}(\cdot), \overline{\varphi}^{**}(\cdot)\right] \subset \left[\underline{\varphi}^{*}(\cdot), \overline{\varphi}^{*}(\cdot)\right]$ *such that when*

$$\begin{cases} \zeta_H - \zeta_L \geqslant \zeta^{**}, \\ \epsilon(\widehat{q}_{AI}) \leqslant \epsilon^{**}(\zeta_H - \zeta_L), \\ \varphi \in \left[\underline{\varphi}^{**}\big(\zeta_H - \zeta_L, \epsilon(\widehat{q}_{AI})\big), \overline{\varphi}^{**}\big(\zeta_H - \zeta_L, \epsilon(\widehat{q}_{AI})\big)\right], \end{cases}$$

the sum principal's expected profit and manager's expected payoff is lower with AI

$$\varphi \int_{\underline{u}^*}^{\overline{u}^*} (\mu_{AM} - c + b) dF(\mu_{AM}) + \int_{\overline{u}^*}^{1} (\mu_{AM} - c + \varphi b) dF(\mu_{AM}) < \theta_\nu \theta_q (\eta_H \zeta_H - c + \varphi b)\,. \quad (20)$$

Proposition 5 demonstrates that when both the manager's and AI's estimates of project quality are remarkably accurate, corporate AI adoption can paradoxically reduce aggregate welfare. In this case, the efficiency gains from AI information are limited, while distortions in investment decisions driven by managerial bias remain significant, leading to a decline in the expected surplus from investment.

This decline reflects a clear misallocation of resources: high-quality projects can be underfunded, while lower-quality projects may receive excessive investment. The misallocation not only reduces the principal's profit but also erodes aggregate efficiency, creating a deadweight loss in which only part of the surplus is captured by managerial compensation.

5 Information and Contract Design for AI Integration

In this section, we propose and discuss potential remedies to the AI adoption challenges identified above, focusing on an information design that integrates governance considerations and an AI-contingent contract that adjusts managerial compensation in response to AI signals.

5.1 Information Design: Coarsened AI Signals

By producing finely differentiated assessments of project quality, AI can create situations in which managers can act on personal bias while facing only limited risk from project outcomes. To address this challenge, we propose an information design approach that strategically coarsens the AI-provided signals. By aggregating the information into broad categories, the firm can restrict the scope for bias-driven decisions. This design aligns managerial actions with project quality, fosters more efficient investment decisions, and achieves these improvements without incurring excessive incentive costs.

Information design is particularly feasible and relevant in the context of AI. Unlike human information senders, whose disclosures may be influenced by shifting incentives, an AI algorithm can be programmed to commit credibly to a pre-specified information structure. This allows the firm to implement signal designs reliably, ensuring that managerial decisions are guided by the intended informational structure. Consequently, interventions such as structuring signals around decision-relevant thresholds are readily implementable with AI, whereas comparable designs are often difficult to enforce with human information sources.

Formally, the principal observes the manager's project signal X_q and receives an AI-generated estimate of project quality, $\widehat{q}_{AI}$. Rather than fully disclosing $\widehat{q}_{AI}$, the principal communicates coarsened information by indicating which element of a partition the pair $(X_q, \widehat{q}_{AI})$ belongs to. Each partition groups together realizations that the principal treats as equivalent for communication purposes, effectively "bunching" values of $\widehat{q}_{AI}$ while leaving X_q unchanged. Given a partition element, the manager forms a posterior expectation of project quality based on the coarsened information.

In equilibrium, the mapping from $\widehat{q}_{AI}$ to the manager's action is monotone: for a given X_q, higher AI estimates lead to weakly more favorable decisions. Intuitively, if a lower-quality project were approved while a higher-quality project were rejected, the principal could swap their partition assignments to raise expected profits without violating the manager's incentive constraints.

The formal characterization of the optimal partition, the associated manager's decision rule, and the equilibrium contract are presented in Proposition 6.

Proposition 6 *The principal structures AI information for the manager as follows:*

1.) The principal partitions $(X_q, \widehat{q}_{AI}) \in \{H, L\} \times U[0,1]$ *into six regions,*

$$Q_1 = \{H\} \times [0, \underline{\widehat{q}}_H], \qquad Q_2 = \{H\} \times (\underline{\widehat{q}}_H, \overline{\widehat{q}}_H), \qquad Q_3 = \{H\} \times [\overline{\widehat{q}}_H, 1],$$
$$Q_4 = \{L\} \times [0, \underline{\widehat{q}}_L], \qquad Q_5 = \{L\} \times (\underline{\widehat{q}}_L, \overline{\widehat{q}}_L), \qquad Q_6 = \{L\} \times [\overline{\widehat{q}}_L, 1].$$

Upon observing $(X_q, \widehat{q}_{AI})$*, the principal communicates the region containing the realization. Thresholds* $\underline{\widehat{q}}_H$*,* $\overline{\widehat{q}}_H$*,* $\underline{\widehat{q}}_L$ *and* $\overline{\widehat{q}}_L$ *are set to maximize the principal's expected profit.*

Let $\widetilde{q}_j = \mathbb{E}[q \mid (X_q, \widehat{q}_{AI}) \in Q_j], j \in \{1, \ldots, 6\}$ *denote the manager's posterior expectation of project quality; in equilibrium, this partition satisfies* $\widetilde{q}_4 = \widetilde{q}_1$*, and* $\widetilde{q}_6 = \widetilde{q}_3$*.*

2.) The optimal contract is

$$w_S = \frac{b}{\eta_H(\widetilde{q}_3 - \widetilde{q}_1)}, \quad w_R = \frac{\widetilde{q}_3 b}{\widetilde{q}_3 - \widetilde{q}_1}, \quad w_F = 0.$$

3.) The manager's action rule is

if $X_\nu = H$, rejects with certainty upon observing Q_1 or Q_4, acts in line with managerial bias upon observing Q_2 or Q_5, and approves with certainty upon observing Q_3 or Q_6;

if $X_\nu = L$, rejects with certainty.

4.) The principal's expected profit is

$$\theta_\nu \eta_H \left[\varphi \mathbb{P}(Q_2)\widetilde{q}_2 + \mathbb{P}(Q_3)\widetilde{q}_3 + \varphi \mathbb{P}(Q_5)\widetilde{q}_5 + \mathbb{P}(Q_6)\widetilde{q}_6\right](1-w_S) - \theta_\nu \left[\varphi \mathbb{P}(Q_2 \cup Q_5) + \mathbb{P}(Q_3 \cup Q_6)\right](c - w_R) - w_R$$

The principal's optimal information design groups together, for each value of X_q, all projects that the manager would unambiguously approve in one partition and all projects that she would unambiguously reject in another. By grouping projects that elicit the same managerial action, the principal reduces the incentive costs associated with projects whose approval would otherwise hinge on borderline judgments. The thresholds for each partition are chosen so that the incentive constraints bind, aligning the manager's posterior expectations with the contract's decision thresholds and maximizing the principal's expected profit. Figure 5 provides a stylized illustration of this partition.

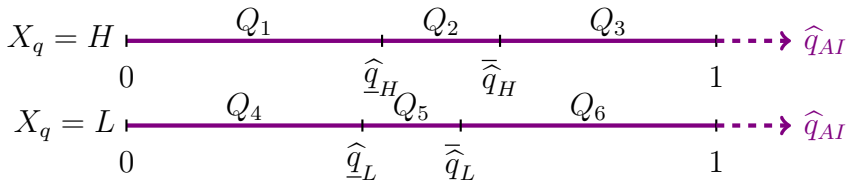

Figure 5: Illustration of the Partition of $(X_q, \widehat{q}_{AI})$

The figure depicts a stylized partition of $\{H, L\} \times [0, 1]$ into six regions $Q_1, \ldots, Q_6$ defined by thresholds $\underline{\widehat{q}}_H$, $\overline{\widehat{q}}_H$, $\underline{\widehat{q}}_L$ and $\overline{\widehat{q}}_L$. Coarsened disclosure of the AI estimate induces pooling of posterior beliefs.

Result 4 compares the principal's profit under this structure with the benchmark outcomes in Propositions 1 and 2.

Result 4 *Under the optimal information partition, the principal's profit surpasses both the no-AI benchmark and the full-AI-information scenario.*

Because the coarsened information structure is chosen optimally, it outperforms alternative approaches: revealing only the manager's private signal X_q (reverting to the no-AI benchmark)

or fully disclosing $\widehat{q}_{AI}$ (providing complete AI information). By carefully balancing incentives with information provision, this design improves project selection and enables the principal to make more productive use of AI insights.

5.2 Contract Design: AI-contingent Contracts

An alternative approach provides the manager with full AI-generated information while tying their payoff directly to the AI reports. Because the manager's posterior expectation depends on the AI estimate, the principal can tailor incentives to these updated beliefs. This contract reduces investment distortions and preserves the efficiency gains from precise AI information.

As discussed in Section 4, backfiring can occur when the manager's expectation μ_{AM} falls within a moderate range, leaving room for suboptimal investment decisions driven by managerial bias. Under the optimal contract described in Proposition 2, the manager's decision thresholds, $\underline{\mu}^*$ and $\overline{\mu}^*$, are determined by $F(\mu_{AM})$, the distribution of the manager's expectations, and remain unchanged regardless of $\widehat{q}_{AI}$, the information provided by AI.

Because the AI-provided information is public and verifiable, the principal can condition on it to obtain a more precise distribution, $F(\mu_{AM}|\widehat{q}_{AI})$. The principal optimizes the following objective function in place of Equation (15):

$$\int_c^1 (\mu_{AM}-c)dF(\mu_{AM}|\widehat{q}_{AI}) - \left[\varphi \int_{\underline{\mu}}^c (c-\mu_{AM})dF(\mu_{AM}|\widehat{q}_{AI}) + (1-\varphi)\int_c^{\overline{\mu}} (\mu_{AM}-c)dF(\mu_{AM}|\widehat{q}_{AI})\right]$$
$$- \left[\left(\int_{\overline{\mu}}^1 \mu_{AM}dF(\mu_{AM}|\widehat{q}_{AI}) + \varphi\int_{\underline{\mu}}^{\overline{\mu}} \mu_{AM}dF(\mu_{AM}|\widehat{q}_{AI})\right) w_S + \left[\varphi F(\overline{\mu}|\widehat{q}_{AI}) + (1-\varphi)F(\underline{\mu}|\widehat{q}_{AI})\right] w_R\right]. \tag{21}$$

Similar to Section 3.2, the principal seeks to minimize the total cost of suboptimal investment decisions and managerial compensation, now conditional on a given value of $\widehat{q}_{AI}$. The optimal contract selects $\underline{\mu}$ and $\overline{\mu}$ so that the distribution $F(\mu_{AM}|\widehat{q}_{AI})$ assigns minimal probability to the interval $(\underline{\mu}, \overline{\mu})$, reducing potential losses from suboptimal decisions, while keeping the interval wide to control incentive costs. We therefore propose the following AI-contingent contract, which links wages to $\widehat{q}_{AI}$, to address AI adoption challenges.

Proposition 7 *When the wage is contingent on AI's report,*

1.) The optimal contract is

$$\big(w_S, w_R, w_F\big) = \begin{cases} \left(\dfrac{k}{\widehat{q}_{AI}}\,, \eta_H k\,, 0\right), & \widehat{q}_{AI} \in \left[\dfrac{c}{\eta_H} + \dfrac{\eta_H k - (1-\theta_\nu\varphi)b}{\eta_H\theta_\nu(1-\varphi)}\,, 1\right] \\ \big(0, b, 0\big), & \widehat{q}_{AI} \in \left[0\,, \dfrac{c}{\eta_H} + \dfrac{\eta_H k - (1-\theta_\nu\varphi)b}{\eta_H\theta_\nu(1-\varphi)}\right), \end{cases}$$

where $k \equiv \dfrac{b}{\eta_H - \eta_L}$.

2.) The manager's action rule is:

if $X_\nu = H$*, approve the project with certainty if* $\mu_{AM} \geqslant c + \frac{\eta_H k - (1-\theta_\nu\varphi)b}{\theta_\nu(1-\varphi)}$*, only approve projects with positive bias if* $\mu_{AM} \in (c-b, c + \frac{\eta_H k - (1-\theta_\nu\varphi)b}{\theta_\nu(1-\varphi)})$*, and reject it with certainty if* $\mu_{AM} \leqslant c - b$*; if* $X_\nu = L$*, reject the project with certainty.*

The equilibrium features two wage schemes. For high $\widehat{q}_{AI}$, the principal sets $\overline{\mu} = \eta_H \widehat{q}_{AI}$ and $\underline{\mu} = \eta_L \widehat{q}_{AI}$ to balance investment efficiency with incentive costs. Substituting into Equation (13) yields the wage scheme described in the proposition. When $\widehat{q}_{AI}$ is low, the expected surplus from the project is minimal even if $X_2 = H$, so the principal sets $w_S = 0$ and offers a rejection wage $w_R = b$ to offset any private benefits from approval.

Within the first wage scheme, the rejection wage w_R remains constant, while the success wage w_S is higher for projects with a lower AI estimate $\widehat{q}_{AI}$. This pattern may seem surprising: one might expect the principal to offer lower rewards for projects assessed by AI to have a lower likelihood of success, anticipating greater concern about managers approving weak projects for private benefits. The economic mechanism underlying this result points in the opposite direction. When the AI estimate is lower, the project's success probability responds less to the manager's judgment, weakening the impact of managerial discretion on outcomes. Conditional on choosing to incentivize the manager, the principal must therefore rely on stronger monetary incentives to discipline behavior, which leads to a higher success wage w_S.

The AI-contingent contract proposed in Proposition 7 reduces losses from suboptimal investment decisions relative to the non-contingent contract in Proposition 2. Moreover, it effectively

addresses corporate AI adoption challenges by ensuring that, for any level of AI information, the principal's expected profit exceeds the no-AI benchmark, as established in Result 5.

Result 5 *Under the optimal AI-contingent contract, the principal's profit exceeds the no-AI benchmark characterized in Proposition 1 for any AI information that refines the manager's signal* X_q.

In summary, the AI-contingent contract enhances the efficient use of AI information by reducing misallocation from suboptimal investment decisions and managerial compensation. This result underscores AI-contingent contracting as an effective organizational response to agency conflicts with AI-assisted decision-making.

6 Conclusion

This paper examines the organizational consequences of adopting predictive AI for managerial decision-making. In particular, we show that the value of AI adoption depends critically on how AI-generated information interacts with managerial incentives. In a setting where managers make high-stakes investment decisions and may derive uncertain private benefits from project approval, AI adoption can backfire: despite refining information and Blackwell-dominating human signals, it can reduce shareholder profits and even aggregate welfare. The key mechanism is that AI changes not only the accuracy but also the granularity of information, increasing the prevalence of borderline decisions in which managerial discretion and bias distort investment choices. When managerial bias is sufficiently uncertain, the resulting increase in agency costs can outweigh the efficiency gains from more precise information, leading to misallocation even under optimally adjusted contracts. We propose two remedies to restore the gains from AI adoption: coarsening AI-generated information to limit bias-driven decisions, or using AI-contingent contracts which link managerial compensation to AI estimates to tighten incentives.

Several directions for future research may warrant further exploration. One possibility is to examine how information design and AI-contingent contracting might be combined in practice,

allowing firms to jointly tailor disclosure and incentives. Another is to consider alternative organizational responses, such as reallocating decision authority between humans and AI, or long-term contracts and dynamic incentive schemes. More broadly, extending the analysis to settings with multiple projects or team-based decisions may provide further insight into how firms integrate AI into organizational decision-making.

References

Acemoglu, D. (2021). Harms of AI. Working paper, National Bureau of Economic Research.

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working paper, National Bureau of Economic Research.

Angelova, V., Dobbie, W. S., and Yang, C. (2023). Algorithmic recommendations and human discretion. Working paper, National Bureau of Economic Research.

Babina, T., Fedyk, A., He, A., and Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151:103745.

Bergemann, D. and Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, pages 265–272.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Chen, Y., Fang, H., Zhao, Y., and Zhao, Z. (2024). Recovering overlooked information in categorical variables with llms: An application to labor market mismatch. Working paper, National Bureau of Economic Research.

Comunale, M. and Manera, A. (2024). The economic impacts and the regulation of AI: A review of the academic literature and policy actions.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Ferreira, D. and Li, J. (2026). Artificial intelligence in the boardroom. *Available at SSRN 5409542*.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710.

Hoffman, M., Kahn, L. B., and Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Holmström, B. (1978). *On Incentives and Control in Organizations.* Stanford University.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, pages 74–91.

Holmström, B. and Tirole, J. (1997). Financial intermediation, loanable funds, and the real sector. *the Quarterly Journal of economics*, 112(3):663–691.

Huang, J. and Ouyang, S. (2025). Soft information, hard decisions: Ai advising. *Available at SSRN 5518278*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lyonnet, V. and Stern, L. H. (2022). Venture capital (mis) allocation in the age of AI. *Fisher College of Business Working Paper*, (2022-03):002.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., et al. (2025). Artificial intelligence index report 2025. *arXiv preprint arXiv:2504.07139*.

Mayer, H., Yee, L., Chui, M., and Roberts, R. (2025). Superagency in the workplace. *McKinsey & Company report. January*.

McElheran, K., Yang, M.-J., Kroff, Z., and Brynjolfsson, E. (2025). The rise of industrial ai in america: Microfoundations of the productivity j-curve (s). *Working Paper*.

Mirrlees, J. A. (1976). The optimal structure of incentives and authority within an organization. *The Bell Journal of Economics*, pages 105–131.

Nobre, K. and Kastner, S. (2014). *The Oxford handbook of attention*. Oxford Library of Psychology.

Pashler, H. (1999). *The Psychology of Attention*. MIT Press.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.

Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *The Review of Economic Studies*, 77(2):779–805.

Weitzner, G. (2024). Reputational algorithm aversion. *arXiv preprint arXiv:2402.15418*.

Yang, M. (2020). Optimality of debt under flexible information acquisition. *The Review of Economic Studies*, 87(1):487–536.

Zhong, H. (2025). Optimal integration: Human, machine, and generative ai. *Management Science*.

Appendix

Proof of Proposition 1

When $\overline{\mu} = \eta_H\zeta_H$ and $\underline{\mu} = \max\{\eta_H\zeta_L, \eta_L\zeta_H\}$, the manager approves the project only if $(X_\nu, X_q) = (H, H)$ and rejects the project in all other cases. The manager's wages, w_S for approval and w_R for rejection, are determined by Equation (13) and are expressed as follows:

$$w_S = \frac{b}{\eta_H\zeta_H - \max\{\eta_H\zeta_L, \eta_L\zeta_H\}}, \quad w_R = \frac{\eta_H\zeta_H b}{\eta_H\zeta_H - \max\{\eta_H\zeta_L, \eta_L\zeta_H\}} = \max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}. \tag{A1}$$

The principal's expected profit is given by:

$$\begin{aligned}\mathbb{E}\left[(\mu_M(1-w_S)-c)\mathbf{1}_{\{a=A\}} - w_R\mathbf{1}_{\{a=R\}}\right] =& \theta_\nu\theta_q(\eta_H\zeta_H - c) - \theta_\nu\theta_q\eta_H\zeta_H w_S - (1-\theta_\nu\theta_q)\, w_R \\ =& \theta_\nu\theta_q(\eta_H\zeta_H - c) - \max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}.\end{aligned} \tag{A2}$$

Next, we show that no other contract can generate a higher level of profit.

Consider a contract in which either $w_S > 0$ and $\overline{\mu} > \eta_H\zeta_H$, or $w_S = 0$ and $w_R \geqslant b$. In this case, the manager approves projects with $(X_\nu, X_q) = (H, H)$ only in the presence of managerial bias, that is, with probability no greater than φ. From assumption (6), the principal's expected payoff from the project under this contract is bounded above by:

$$\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] \leqslant \varphi \cdot \theta_\nu\theta_q(\eta_H\zeta_H - c) < \theta_\nu\theta_q(\eta_H\zeta_H - c) - \max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}.$$

Similarly, for a contract in which either $w_S > 0$ and $\underline{\mu} < \eta_H\zeta_L$, or $w_S = 0$ and $w_R < b$, the manager approves projects with $(X_\nu, X_q) = (H, L)$ with probability at least φ. The principal's expected payoff is bounded above by:

$$\begin{aligned}\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] &\leqslant \theta_\nu\theta_q(\eta_H\zeta - c) - \varphi\theta_\nu(1-\theta_q)(c - \eta_H\zeta_L) \\ &< \theta_\nu\theta_q(\eta_H\zeta_H - c) - \max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}.\end{aligned}$$

Under a contract where $w_S > 0$ and $\underline{\mu} < \eta_L\zeta_H$, the manager approves projects with $(X_\nu, X_q) = (H, L)$ with probability at least φ. As in the analysis above, this contract cannot generate a higher profit than that in Equation (A2).

For all remaining contracts, which satisfy $w_S > 0$, $\overline{\mu} \leqslant \eta_H\zeta_H$, and $\underline{\mu} \geqslant \max\{\eta_H\zeta_L, \eta_L\zeta_H\}$, the principal's expected profit is bounded above by:

$$\theta_\nu\theta_q(\eta_H\zeta - c) - \frac{\overline{\mu}b}{\overline{\mu}-\underline{\mu}} \leqslant \theta_\nu\theta_q(\eta_H\zeta_H - c) - \max\left\{\frac{\zeta_H b}{\zeta_H-\zeta_L}, \frac{\eta_H b}{\eta_H-\eta_L}\right\},$$

with equality holding if and only if $\overline{\mu} = \eta_H\zeta_H$ and $\underline{\mu} = \max\{\eta_H\zeta_L, \eta_L\zeta_H\}$. ■

Proof of Result 1

By the law of iterated expectations,

$$\mathbb{E}[\mathbb{E}[q \mid X_q]] = \mathbb{E}[q] = \tfrac{1}{2},$$

which implies

$$\theta_q\zeta_H + (1-\theta_q)\zeta_L = \tfrac{1}{2} \quad \Rightarrow \quad \zeta_H = \tfrac{1}{2} + (1-\theta_q)(\zeta_H - \zeta_L).$$

Substituting into the principal's expected profit yields

$$\begin{aligned}\Pi &= \theta_\nu\theta_q(\eta_H\zeta_H - c) - \max\left\{\frac{\zeta_H b}{\zeta_H - \zeta_L}, \frac{\eta_H b}{\eta_H - \eta_L}\right\} \\ &= \theta_\nu\theta_q\big(\eta_H\big[\tfrac{1}{2} + (1-\theta_q)(\zeta_H - \zeta_L)\big] - c\big) - \max\left\{(1-\theta_q)b + \frac{b}{2(\zeta_H - \zeta_L)}, \frac{\eta_H b}{\eta_H - \eta_L}\right\}.\end{aligned}$$

The first term is increasing in $\zeta_H - \zeta_L$, while the variable component of the incentive term is decreasing and the alternative is constant. Therefore, Π is strictly increasing in $\zeta_H - \zeta_L$. ■

Proof of Result 2

The manager's expected payoff is

$$\max\left\{\frac{\zeta_H b}{\zeta_H-\zeta_L}, \frac{\eta_H b}{\eta_H-\eta_L}\right\} + \theta_\nu\theta_q\phi b,$$

Since the second term is constant in $\zeta_H - \zeta_L$, monotonicity depends only on the maximum term.

Note that

$$\frac{\zeta_H b}{\zeta_H - \zeta_L} \leqslant \frac{\eta_H b}{\eta_H - \eta_L} \iff \eta_H \zeta_L \leqslant \eta_L \zeta_H .$$

Hence, if $\eta_H \zeta_L > \eta_L \zeta_H$, the maximum is attained by $\frac{\zeta_H b}{\zeta_H - \zeta_L}$, which is decreasing in $\zeta_H - \zeta_L$ (from Result 1). Otherwise, the maximum is attained by $\frac{\eta_H b}{\eta_H - \eta_L}$ and is constant in $\zeta_H - \zeta_L$.

Therefore, the manager's expected payoff is weakly decreasing in $\zeta_H - \zeta_L$.

Proof of Proposition 2

The derivatives of w_S and w_R with respect to $\overline{\mu}$ and $\underline{\mu}$ are expressed as follows:

$$\frac{dw_S}{d\overline{\mu}} = -\frac{b}{(\overline{\mu}-\underline{\mu})^2}, \quad \frac{dw_R}{d\overline{\mu}} = -\frac{\underline{\mu} b}{(\overline{\mu}-\underline{\mu})^2}, \quad \frac{dw_S}{d\underline{\mu}} = \frac{b}{(\overline{\mu}-\underline{\mu})^2}, \quad \frac{dw_R}{d\underline{\mu}} = \frac{\overline{\mu} b}{(\overline{\mu}-\underline{\mu})^2} . \tag{A3}$$

Now, differentiating the objective function (16) with respect to $\overline{\mu}$:

$$\begin{aligned} &-(1-\varphi)(\overline{\mu}-c)f(\overline{\mu}) - \left(\int_{\overline{\mu}}^{1} \mu_{AM} dF(\mu_{AM}) + \varphi \int_{\underline{\mu}}^{\overline{\mu}} \mu_{AM} dF(\mu_{AM}) \right) \frac{dw_S}{d\overline{\mu}} \\ &- \left(\varphi F(\underline{\mu}) + (1-\varphi) F(\overline{\mu})\right) \frac{dw_R}{d\overline{\mu}} - (1-\varphi)\overline{\mu} f(\overline{\mu}) w_S + (1-\varphi) f(\overline{\mu}) w_R = 0 . \end{aligned}$$

Using (A3) and the condition $w_R = \overline{\mu} w_S$, and then evaluating at $\overline{\mu} = \overline{\mu}^*$, $\underline{\mu} = \underline{\mu}^*$, we obtain:

$$(1-\varphi)(\overline{\mu}^*-c)f(\overline{\mu}^*) - \frac{b}{(\overline{\mu}^*-\underline{\mu}^*)^2} \left[\int_{\overline{\mu}^*}^{1} \mu_{AM} dF(\mu_{AM}) + \varphi \int_{\underline{\mu}^*}^{\overline{\mu}^*} \mu_{AM} dF(\mu_{AM}) + \varphi \underline{\mu}^* F(\underline{\mu}^*) + (1-\varphi)\underline{\mu}^* F(\overline{\mu}^*) \right] = 0. \tag{A4}$$

Differentiating the objective function with respect to $\underline{\mu}$:

$$\begin{aligned} &-\varphi(\underline{\mu}-c)f(\underline{\mu}) - \left(\int_{\overline{\mu}}^{1} \mu_{AM} dF(\mu_{AM}) + \frac{1}{2} \int_{\underline{\mu}}^{\overline{\mu}} \mu_{AM} dF(\mu_{AM}) \right) \frac{dw_S}{d\underline{\mu}} \\ &- \left(\varphi F(\underline{\mu}) + (1-\varphi) F(\overline{\mu})\right) \frac{dw_R}{d\underline{\mu}} - \varphi \underline{\mu} f(\overline{\mu}) w_S + \varphi f(\overline{\mu}) w_R = 0 . \end{aligned}$$

Using (A3) and the condition $w_R = \underline{\mu} w_S + b$, and then evaluating at $\overline{\mu} = \overline{\mu}^*$, $\underline{\mu} = \underline{\mu}^*$, we obtain:

$$\varphi(c-b-\underline{\mu}^*)f(\underline{\mu}^*) - \frac{b}{(\overline{\mu}^*-\underline{\mu}^*)^2} \left[\int_{\overline{\mu}^*}^{1} \mu_{AM} dF(\mu_{AM}) + \varphi \int_{\underline{\mu}^*}^{\overline{\mu}^*} \mu_{AM} dF(\mu_{AM}) + \varphi \overline{\mu}^* F(\underline{\mu}^*) + (1-\varphi)\overline{\mu}^* F(\overline{\mu}^*) \right] = 0. \tag{A5}$$

Thus, the equilibrium thresholds $\overline{\mu}^*$ and $\underline{\mu}^*$ satisfy the conditions in (A4) and (A5), and are implicitly defined by this system of equations. ■

Proof of Result 3

Consider an example where the AI's estimate of the project quality, $\widehat{q}_{AI}$, takes values $q_1 < q_2 < \cdots < q_{2m}$ with equal probabilities, and the optimal contract based on this information satisfies

$$\underline{\mu}^* = \eta_H q_{2n}, \quad \overline{\mu}^* = \eta_H q_{2l-1}.$$

Now, suppose we define an alternative information structure that aggregates adjacent AI estimates and communicates the averages

$$\frac{q_1 + q_2}{2}, \quad \frac{q_3 + q_4}{2}, \quad \cdots, \quad \frac{q_{2m-1} + q_{2m}}{2}.$$

This coarser information yields the same approval and rejection decisions but reduces the manager's expected wage, as the thresholds are relaxed to

$$\underline{\mu}^* = \eta_H \frac{q_{2n-1} + q_{2n}}{2}, \quad \overline{\mu}^* = \eta_H \frac{q_{2l-1} + q_{2l}}{2}.$$

Hence, the principal's expected profit does not necessarily increase with more precise AI estimates of $\widehat{q}_{AI}$ and may in fact decline.

Proof of Proposition 3

We now demonstrate that the optimal contract takes the form $(w_S, w_R, w_F) = (0, b, 0)$ when the manager has a fixed, non-stochastic bias $B \equiv b$ from project approval.

Given the manager's expectation μ, the principal's payoff is given by $\mu(1-w_S) - c$ if the project is approved, and $-w_R$ if it is rejected. Under the contract $(0, b, 0)$, the manager is indifferent between approving and rejecting the project in all states. As a result, the manager defers to the principal's preferences and approves the project if and only if the expected payoff

from approval, $\mu - c$, exceeds the payoff from rejection, $-b$. The principal's resulting payoff is therefore $\max\{\mu - c, -b\}$.

Now consider alternative contracts. If $w_R \geqslant b$, the principal's payoff given the manager's expectation μ is bounded by

$$\max\{\mu(1-w_S) - c, -w_R\} \leqslant \max\{\mu - c, -b\}\,. \tag{A6}$$

If $w_R < b$, the manager's expected payoff from project approval $\mu w_S + b$ always exceeds that from rejection w_R, and therefore, the manager approves all projects. In this case, the principal's expected payoff is bounded above by

$$\int_0^1 (\mu - c) dF(\mu) \leqslant \int_0^1 \max\{\mu - c, -b\} dF(\mu)\,. \tag{A7}$$

Combining (A6) and (A7), we conclude that the optimal contract is $(w_S, w_R, w_F) = (0, b, 0)$.

We now turn to a comparison of the principal's expected payoffs when the manager has access to different information. From the law of iterated expectations, we have:

$$\mu = \mathbb{E}[p|\mathcal{I}] = \mathbb{E}\big[\mathbb{E}[p|\mathcal{I}']|\mathcal{I}\big] = \mathbb{E}[\mu'|\mathcal{I}]\,. \tag{A8}$$

The distribution of μ' is obtained by redistributing the probability of μ in a manner that preserves the mean while increasing dispersion. Formally, $F(\mu')$ is a mean-preserving spread of $F(\mu)$.

Since the function $\max\{\mu - c, -b\}$ is convex in μ, from Jensen's inequality, it follows that the principal's expected payoff increases with the precision of manager's information.

$$\int_0^1 \max\{\mu' - c, -b\} dF(\mu') \geqslant \int_0^1 \max\{\mu - c, -b\} dF(\mu)\,. \tag{A9}$$

Substituting in $\mu' = \mu_{AM}$ and $\mu = \mu_M$, we obtain $\pi_{b,AM} \geqslant \pi_{b,M}$. ■

Proof of Proposition 4

Rather than focusing on the optimal thresholds $\underline{\mu}^*$ and $\overline{\mu}^*$, we consider arbitrary $\underline{\mu}$ and $\overline{\mu}$ and derive a lower bound on the total agency cost that holds for $\underline{\mu} \leqslant c \leqslant \overline{\mu}$.

From assumption (4), the CDF of AI estimate of the project quality $\widehat{q}_{AI}$ satisfies

$$F_{\widehat{q}_{AI}}\left(\frac{\overline{\mu}}{\eta_H}\right) - F_{\widehat{q}_{AI}}\left(\frac{\underline{\mu}}{\eta_H}\right) \geqslant \frac{\overline{\mu}-\underline{\mu}}{\eta_H} - 2\sqrt{\epsilon(\widehat{q}_{AI})}\,. \tag{A10}$$

Applying (14), the probability that $\mu_{AM} \in [\underline{\mu}, \overline{\mu}]$ is bounded below by $\theta_\nu\left(\frac{\overline{\mu}-\underline{\mu}}{\eta_H} - 2\sqrt{\epsilon(\widehat{q}_{AI})}\right)$.

Consequently, when the manager forms expectation μ_{AM} with AI assistance, the "Surplus Loss from Suboptimal Investment Decisions" expression in equation (16) admits the following lower bound:

$$\begin{aligned}
&\varphi\int_{\underline{\mu}}^{c}(c-\mu_{AM})dF(\mu_{AM}) + (1-\varphi)\int_{c}^{\overline{\mu}}(\mu_{AM}-c)dF(\mu_{AM}) \\
=&\varphi\int_{\underline{\mu}}^{c}[F(\mu_{AM})-F(\underline{\mu})]d\mu_{AM} + (1-\varphi)\int_{c}^{\overline{\mu}}[F(\overline{\mu})-F(\mu_{AM})]d\mu_{AM} \\
\geqslant&\varphi\int_{\underline{\mu}}^{c}\theta_\nu \max\left\{\frac{\mu_{AM}-\underline{\mu}}{\eta_H}-2\sqrt{\epsilon(\widehat{q}_{AI})},0\right\}d\mu_{AM} + (1-\varphi)\int_{c}^{\overline{\mu}}\theta_\nu \max\left\{\frac{\overline{\mu}-\mu_{AM}}{\eta_H}-2\sqrt{\epsilon(\widehat{q}_{AI})},0\right\}d\mu_{AM} \\
\geqslant&\frac{\theta_\nu\varphi(1-\varphi)}{2\eta_H}\left(\overline{\mu}-\underline{\mu}-4\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2\,.
\end{aligned} \tag{A11}$$

The "Managerial Compensation" expression in (16) is bounded below by:

$$\begin{aligned}
&\left[\int_{\overline{\mu}}^{1}\mu_{AM}dF(\mu_{AM}) + \varphi\int_{\underline{\mu}}^{\overline{\mu}}\mu_{AM}dF(\mu_{AM})\right]w_S + \left[\varphi F(\underline{\mu}) + (1-\varphi)F(\overline{\mu})\right]w_R \\
>&\left[\int_{\overline{\mu}}^{1}\overline{\mu}dF(\mu_{AM}) + \varphi\int_{\underline{\mu}}^{\overline{\mu}}\underline{\mu}dF(\mu_{AM})\right]w_S + \left[\varphi F(\underline{\mu}) + (1-\varphi)F(\overline{\mu})\right]\overline{\mu}w_S \\
>&\int_{0}^{1}\left(\varphi\underline{\mu} + (1-\varphi)\overline{\mu}\right)dF(\mu_{AM})\cdot w_S \\
=&\left(\overline{\mu} - \varphi(\overline{\mu}-\underline{\mu})\right)\cdot\frac{b}{\overline{\mu}-\underline{\mu}} \\
>&\left(\frac{c}{\overline{\mu}-\underline{\mu}} - \varphi\right)b\,.
\end{aligned} \tag{A12}$$

Combining (A11) with (A12), we obtain a lower bound for $\pi_{0,AM} - \pi_{B,AM}$, the total agency cost from "Surplus Loss from Suboptimal Investment Decisions"and "Managerial compensation":

$$\begin{aligned}
\pi_{0,AM} - \pi_{B,AM} >&\frac{\theta_\nu\varphi(1-\varphi)}{2\eta_H}\left(\overline{\mu}-\underline{\mu}-4\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 + \left(\frac{c}{\overline{\mu}-\underline{\mu}} - \varphi\right)b \\
\geqslant&\left(\frac{3}{2} - 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)\left(\frac{\theta_\nu\varphi(1-\varphi)}{\eta_H}\right)^{\frac{1}{3}}(cb)^{2/3} - (\varphi + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})})b\,.
\end{aligned}$$

Thus, the increment in cost compared to the case without AI is bounded below by:

$$\begin{aligned}&\left(\pi_{0,AM}-\pi_{B,AM}\right)-\left(\pi_{0,M}-\pi_{B,M}\right)\\&>\left(\frac{3}{2}-2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)\left(\frac{\theta_\nu\varphi(1-\varphi)}{\eta_H}\right)^{\frac{1}{3}}(cb)^{2/3}-\left(\frac{\zeta_H}{\zeta_H-\zeta_L}+\varphi+2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)b\,.\end{aligned}\tag{A13}$$

The maximum of the above expression with respect to $\epsilon(\widehat{q}_{AI})$ and φ is bounded below by

$$\frac{3}{2}\left(\frac{\theta_\nu}{4\eta_H}\right)^{\frac{1}{3}}(cb)^{2/3}-\left(\frac{\zeta_H}{\zeta_H-\zeta_L}+\frac{1}{2}\right)b\,,\tag{A14}$$

which corresponds to the right-hand side of (A13) when $\epsilon(\widehat{q}_{AI})=0$ and $\varphi=\frac{1}{2}$.

AI adoption backfires when the efficiency gain from information $\pi_{0,AM}-\pi_{0,M}$ is insufficient to offset the increase in costs compared to managerial compensation without AI. By Proposition 3, $\pi_{0,AM}$ is bounded above by the expected profit under full information about project quality,

$$\pi_{0,AM}\leqslant\int_{c/\eta_H}^{1}\theta_\nu(\eta_H q-c)dq=\frac{\theta_\nu(\eta_H-c)^2}{2\eta_H}\,.\tag{A15}$$

Thus the upper bound of the efficiency gain is given by

$$\pi_{0,AM}-\pi_{0,M}\leqslant\frac{\theta_\nu(\eta_H-c)^2}{2\eta_H}-\theta_\nu\theta_q(\eta_H\zeta_H-c)\,,\tag{A16}$$

where the no-agency-conflict profit with the manager's information is $\pi_{0,M}=\theta_\nu\theta_q(\eta_H\zeta_H-c)$.

The threshold ζ^* is determined by equation (17), which equates the right-hand-sides of (A14) and (A16). Accordingly, when $\zeta\geqslant\zeta^*$, there exists a range of AI inaccuracy $\epsilon(\widehat{q}_{AI})$ and bias probability φ for which AI adoption results in net losses for the principal. The threshold is ϵ^* similarly determined by equation (18). The thresholds $\underline{\varphi}^*$ and $\overline{\varphi}^*$ are the two solutions in $(0,1)$ to Equation (19), which equates the right-hand sides of (A13) and (A16). These solutions exist whenever $\zeta_H-\zeta_L\geqslant\zeta^*$ and $\epsilon(\widehat{q}_{AI})\leqslant\epsilon^*(\zeta_H-\zeta_L)$. Hence, the conditions of Proposition 4 suffice to make the principal's expected profit under AI-augmented decision-making lower than the no-AI benchmark. ■

Proof of Proposition 5

The efficiency gain is bounded above by (A16).

Evaluated at the optimal thresholds $\underline{\mu}^*$ and $\overline{\mu}^*$, equation (A11) provides a lower bound on the "Surplus Loss from Suboptimal Investment Decisions" expression:

$$\frac{\theta_\nu \varphi(1-\varphi)}{2\eta_H}\left(\overline{\mu}^* - \underline{\mu}^* - 4\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 . \tag{A17}$$

Aggregate welfare therefore decreases whenever the efficiency gain fails to compensate for this surplus loss. To establish this proposition, it is sufficient to derive a lower bound on $\overline{\mu}^* - \underline{\mu}^*$. We next establish an upper bound on total compensation costs for arbitrary $\underline{\mu}$ and $\overline{\mu}$, and use this bound to demonstrate why a small separation between the decision thresholds is suboptimal.

From assumption (4), the CDF of AI estimate of the project quality $\widehat{q}_{AI}$ satisfies

$$F_{\widehat{q}_{AI}}\left(\frac{\overline{\mu}}{\eta_H}\right) - F_{\widehat{q}_{AI}}\left(\frac{\underline{\mu}}{\eta_H}\right) \geqslant \frac{\overline{\mu}-\underline{\mu}}{\eta_H} + 2\sqrt{\epsilon(\widehat{q}_{AI})}\,. \tag{A18}$$

Applying (14), the probability that $\mu_{AM} \in [\underline{\mu}, \overline{\mu}]$ is bounded above by $\theta_\nu\left(\frac{\overline{\mu}-\underline{\mu}}{\eta_H} + 2\sqrt{\epsilon(\widehat{q}_{AI})}\right)$.

Consequently, when the manager forms expectation μ_{AM} with AI assistance, the "Surplus Loss from Suboptimal Investment Decisions" expression in equation (16) admits the following upper bound:

$$\begin{aligned}
&\varphi\int_{\underline{\mu}}^{c}(c-\mu_{AM})dF(\mu_{AM}) + (1-\varphi)\int_{c}^{\overline{\mu}}(\mu_{AM}-c)dF(\mu_{AM}) \\
=&\varphi\int_{\underline{\mu}}^{c}[F(\mu_{AM})-F(\underline{\mu})]d\mu_{AM} + (1-\varphi)\int_{c}^{\overline{\mu}}[F(\overline{\mu})-F(\mu_{AM})]d\mu_{AM} \\
\leqslant&\varphi\int_{\underline{\mu}}^{c}\theta_\nu \max\left\{\frac{\mu_{AM}-\underline{\mu}}{\eta_H}+2\sqrt{\epsilon(\widehat{q}_{AI})},0\right\}d\mu_{AM} + (1-\varphi)\int_{c}^{\overline{\mu}}\theta_\nu \max\left\{\frac{\overline{\mu}-\mu_{AM}}{\eta_H}+2\sqrt{\epsilon(\widehat{q}_{AI})},0\right\}d\mu_{AM} \\
=&\frac{\theta_\nu\varphi}{2\eta_H}\left(c-\underline{\mu}+2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 + \frac{\theta_\nu(1-\varphi)}{2\eta_H}\left(\overline{\mu}-c+2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 .
\end{aligned} \tag{A19}$$

The "Managerial Compensation" expression in (16) is bounded above by:

$$\left[\int_{\overline{\mu}}^{1}\mu_{AM}dF(\mu_{AM}) + \varphi\int_{\underline{\mu}}^{\overline{\mu}}\mu_{AM}dF(\mu_{AM})\right]w_S + \left[\varphi F(\underline{\mu}) + (1-\varphi)F(\overline{\mu})\right]w_R < w_s = \frac{b}{\overline{\mu}-\underline{\mu}}\,. \tag{A20}$$

Combining (A19) with (A20), we obtain an upper bound for $\pi_{0,AM} - \pi_{B,AM}$, the total agency cost from "Surplus Loss from Suboptimal Investment Decisions"and "Managerial compensation":

$$\pi_{0,AM} - \pi_{B,AM} < \frac{\theta_\nu \varphi}{2\eta_H}\left(c - \underline{\mu} + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 + \frac{\theta_\nu(1-\varphi)}{2\eta_H}\left(\overline{\mu} - c + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right)^2 + \frac{b}{\overline{\mu}-\underline{\mu}} \tag{A21}$$

The following thresholds

$$\underline{\mu} = c - \left(\frac{\eta_H b}{2\theta_\nu}\right)^{\frac{1}{3}}, \qquad \overline{\mu} = c + \left(\frac{\eta_H b}{2\theta_\nu}\right)^{\frac{1}{3}} \tag{A22}$$

are admissible, and as a result

$$\pi_{0,AM} - \pi_{B,AM} < \frac{\theta_\nu}{2\eta_H}\left[\left(\frac{\eta_H b}{2\theta_\nu}\right)^{\frac{1}{3}} + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right]^2 + \left(\frac{\theta_\nu b^2}{4\eta_H}\right)^{\frac{1}{3}}. \tag{A23}$$

The "Managerial Compensation" expression for the optimal thresholds is lower than this upper bound

$$\left(\frac{c}{\overline{\mu}^* - \underline{\mu}^*} - \varphi\right) b < \frac{\theta_\nu}{2\eta_H}\left[\left(\frac{\eta_H b}{2\theta_\nu}\right)^{\frac{1}{3}} + 2\eta_H\sqrt{\epsilon(\widehat{q}_{AI})}\right]^2 + \left(\frac{\theta_\nu b^2}{4\eta_H}\right)^{\frac{1}{3}}. \tag{A24}$$

This inequality in turn establishes a lower bound on the gap between the optimal thresholds,$\overline{\mu}^* - \underline{\mu}^*$. The rest follows from the Proof of Proposition 4. ■

Proof of Proposition 6

We proceed in several steps.

Step 1: Threshold Monotonicity.

Let $\widetilde{q}_j = \mathbb{E}[q \mid (X_q, \widehat{q}_{AI}) \in Q_j$ denote the manager's posterior expectation given the coarsened signal.

When the manager receives a favorable viablity signal $X_\nu = H$, his conditional expectation of the project success probability is given by $\mu_{AM} = \eta_H \widehat{q}_{AI}$. Given decisions thresholds $\underline{\mu}, \overline{\mu}$ from the contract, the manager rejects if $\widetilde{q}_j \leqslant \underline{\mu}/\eta_H$, approves if $\widetilde{q}_j \geqslant \overline{\mu}/\eta_H$, and act in line with managerial bias otherwise.

Now we establish that conditional on $X_\nu = H$, the manager's action must be monotone in $\widehat{q}_{AI}$: if a project with estimate $\widehat{q}'$ is rejected with certainty, all projects with $\widehat{q}_{AI} < \widehat{q}'_{AI}$ must also be rejected; similarly, if a project is approved with certainty, all higher $\widehat{q}_{AI}$ must be approved. Otherwise, the principal could the partition assignments of $\widehat{q}_{AI}$ and $\widehat{q}'_{AI}$ to raise expected profits without violating the manager's incentive constraints.

Hence, there exists thresholds $\underline{\widehat{q}}_H, \overline{\widehat{q}}_H$ for $X_q = H$ and $\underline{\widehat{q}}_L, \overline{\widehat{q}}_L$ for $X_q = L$ such that the manager rejects if $\widehat{q}_{AI} \leqslant \underline{\widehat{q}}_k$, approves if $\widehat{q}_{AI} \geqslant \overline{\widehat{q}}_k$, and otherwise act in line with his bias.

Step 2: Optimal Partitioning.

All projects with $X_q = H$ and $\widehat{q}_{AI} \in [0, \underline{\widehat{q}}_H]$ should be grouped within the same partition, as this lowers the threshold $\underline{\mu}$ and reduces the associated managerial compensation cost. Intuitively, mixing low-quality projects with medium-quality ones would weaken the manager's tendency to approve projects due to managerial bias. Similarly, all projects with $X_q = L$ and $\widehat{q}_{AI} \in [0, \underline{\widehat{q}}_L]$ should be assigned to a single partition.

The same logic applies to the partitions for high-quality projects: $X_q = H$, $\widehat{q}_{AI} \in [\overline{\widehat{q}}_H, 1]$ and $X_q = L$, $\widehat{q}_{AI} \in [\overline{\widehat{q}}_L, 1]$.

In summary, the principal groups all low-quality projects into one partition and all high-quality projects into one partition:

$$Q_1 = \{H\} \times [0, \underline{\widehat{q}}_H], \quad Q_4 = \{L\} \times [0, \underline{\widehat{q}}_L],$$

$$Q_3 = \{H\} \times [\overline{\widehat{q}}_H, 1], \quad Q_6 = \{L\} \times [\overline{\widehat{q}}_L, 1].$$

Medium-quality projects are placed in the remaining partitions:

$$Q_2 = \{H\} \times (\underline{\widehat{q}}_H, \overline{\widehat{q}}_H), \quad Q_5 = \{L\} \times (\underline{\widehat{q}}_L, \overline{\widehat{q}}_L).$$

In equilibrium, the incentive constraints are binding for the extreme partitions, so $\widetilde{q}_4 = \widetilde{q}_1$ and $\widetilde{q}_6 = \widetilde{q}_3$.

Step 3: Manager's Action Rule.

Given the upper bound for managerial bias (6), it's never optimal for the principal to set $\underline{\mu} < \eta_L$. As a result, the manager rejects all projects if he observes a low viablity signal $X_\nu = L$.

Translating the partition into the manager's decisions in Step 1, we obtain the following action rule: if $X_\nu = H$, the manager rejects with certainty upon observing Q_1 or Q_4, acts according to her managerial bias upon observing Q_2 or Q_5, and approves with certainty upon observing Q_3 or Q_6; if $X_\nu = L$, he rejects all projects with certainty.

Step 4: Optimal Contract.

The principal sets the wages to bind the incentive constraints for the extreme partitions:

$$w_S = \frac{b}{\eta_H(\widetilde{q}_3 - \widetilde{q}_1)}, \quad w_R = \frac{\widetilde{q}_3 b}{\widetilde{q}_3 - \widetilde{q}_1}, \quad w_F = 0. \tag{A25}$$

This ensures the manager's expected payoff aligns with the threshold behavior while minimizing the incentive costs.

Step 5: Principal's Expected Profit.

The manager approves the project with probability

$$\theta_\nu \Big[\varphi \mathbb{P}(Q_2) + \mathbb{P}(Q_3) + \varphi \mathbb{P}(Q_5) + \mathbb{P}(Q_6)\Big], \tag{A26}$$

and rejects it with the complementary probability.

Conditional on approval, the project yields a successful outcome with probability $\eta_H \widetilde{q}_j$ when $(X_q, \widehat{q}_{AI}) \in Q_j$. Therefore, the overall probability that the project is both invested and succeeds is

$$\theta_\nu \eta_H \Big[\varphi \mathbb{P}(Q_2)\widetilde{q}_2 + \mathbb{P}(Q_3)\widetilde{q}_3 + \varphi \mathbb{P}(Q_5)\widetilde{q}_5 + \mathbb{P}(Q_6)\widetilde{q}_6\Big]. \tag{A27}$$

Consequently, the principal's expected profit is given by:

$$\begin{aligned} \pi_{\text{Partition}} = {} & \theta_\nu \eta_H \Big[\varphi \mathbb{P}(Q_2)\widetilde{q}_2 + \mathbb{P}(Q_3)\widetilde{q}_3 + \varphi \mathbb{P}(Q_5)\widetilde{q}_5 + \mathbb{P}(Q_6)\widetilde{q}_6\Big](1 - w_S) \\ & - \theta_\nu \Big[\varphi \mathbb{P}(Q_2 \cup Q_5) + \mathbb{P}(Q_3 \cup Q_6)\Big](c - w_R) - w_R \,. \end{aligned} \tag{A28}$$

■

Proof of Result 4

Because the principal can choose either the no-AI benchmark (revealing no AI information) or the full-AI-information benchmark (revealing all AI information) as feasible information partitions, the profit under the optimal information partition weakly dominates profits under both benchmarks. We now show that the optimal partition is in fact strictly superior.

Consider the partition defined by

$$\widehat{\underline{q}}_H = \widehat{\overline{q}}_H = \zeta_L \quad \text{and} \quad \widehat{\underline{q}}_L = \widehat{\overline{q}}_L = \zeta_H.$$

This partition separates the lowest-quality projects among those with $X_q = H$ and the highest-quality projects among those with $X_q = L$. Relative to the no-AI benchmark, it rejects low-quality projects that the manager would otherwise approve and accepts high-quality projects that the manager would otherwise reject, thereby strictly increasing expected project surplus.

By construction, $\widetilde{q}_1 < \sup Q_1 = \zeta_L$ and $\widetilde{q}_6 > \inf Q_6 = \zeta_H$. Moreover,

$$\widetilde{q}_3 > \mathbb{E}[q \mid X_q = H] = \zeta_H \quad \text{and} \quad \widetilde{q}_4 < \mathbb{E}[q \mid X_q = L] = \zeta_L.$$

These inequalities imply that the principal can strictly reduce the manager's wage payments while maintaining incentive compatibility.

In summary, the no-AI benchmark is strictly dominated by the partition constructed above and therefore cannot be optimal.

For the full-AI-information benchmark, consider the partition

$$\widehat{\underline{q}}_L = \widehat{\underline{q}}_H = \frac{\underline{\mu}^*}{\eta_H} \quad \text{and} \quad \widehat{\overline{q}}_L = \widehat{\overline{q}}_H = \frac{\overline{\mu}^*}{\eta_H}.$$

Here $\underline{\mu}^*$ and $\overline{\mu}^*$ are the optimal thresholds under full AI information. This partition replicates exactly the project approval and rejection decisions of the full-AI-information benchmark.

However, it strictly reduces the manager's wage payments, since

$$\widetilde{q}_1, \widetilde{q}_4 < \sup Q_1 = \sup Q_4 = \frac{\underline{\mu}^*}{\eta_H} \quad \text{and} \quad \widetilde{q}_3, \widetilde{q}_6 > \inf Q_3 = \inf Q_6 = \frac{\overline{\mu}^*}{\eta_H}.$$

Therefore, the full-AI-information benchmark is strictly dominated by the partition constructed above and cannot be optimal. ■

Proof of Proposition 7

For a given value of $\widehat{q}_{AI}$, the manager's expectation μ_{AM} takes the value of $\eta_H \widehat{q}_{AI}$ if $X_\nu = H$, and $\eta_L \widehat{q}_{AI}$ if $X_\nu = L$. Every possible contract is dominated by one of the three wage schemes described in the proposition.

Consider a contract where $w_S > 0$, $\overline{\mu} \leqslant \eta_H \widehat{q}_{AI}$, and $\underline{\mu} \geqslant \eta_L \widehat{q}_{AI}$. Under this contract, the manager approves if $X_\nu = H$ and rejects if $X_\nu = L$. The principal's expected profit is bounded above by:

$$\begin{aligned}
&\mathbb{E}[(\mu_{AM} - c)\mathbf{1}_{\{a=A\}} | \widehat{q}_{AI}] - \mathbb{E}[\mu_{AM} w_S \mathbf{1}_{\{a=A\}} | \widehat{q}_{AI}] - \mathbb{E}[w_R \mathbf{1}_{\{a=R\}} | \widehat{q}_{AI}] \\
=&\theta_\nu \big(\eta_H \widehat{q}_{AI} - c\big) - \frac{\overline{\mu} b}{\overline{\mu} - \underline{\mu}} \\
\leqslant&\theta_\nu \big(\eta_H \widehat{q}_{AI} - c\big) - \eta_H k \,, \qquad \text{(A29)}
\end{aligned}$$

with equality if and only if $w_S = \frac{k}{\widehat{q}_{AI}}$ and $w_R = \eta_H k$.

Similarly, for a contract where $w_S > 0$ and $\overline{\mu} > \eta_H \widehat{q}_{AI}$, or where $w_S = 0$ and $w_R \geqslant b$, the principal's profit is bounded above by:

$$\theta_\nu \varphi \max\big\{\eta_H \widehat{q}_{AI} - c, 0\big\} - (1 - \theta_\nu \varphi) b - \theta_\nu \varphi b \cdot \mathbf{1}_{\{\eta_H \widehat{q}_{AI} \leqslant c - b\}} \qquad \text{(A30)}$$

with equality if and only if $w_S = 0$ and $w_R = b$. Under this contract, the manager only approves the project when private benefits are present and $\mu_{AM} > c - b$.

For a contract where $w_S > 0$ and $\underline{\mu} < \eta_L \widehat{q}_{AI}$, or where $w_S = 0$ and $w_R < b$, the principal's profit is bounded above by:

$$\theta_\nu \big(\eta_H \widehat{q}_{AI} - c\big) + (1 - \theta_\nu) \varphi \big(\eta_L \widehat{q}_{AI} - c\big) + (1 - \theta_\nu)(1 - \varphi) \max\big\{\eta_L \widehat{q}_{AI} - c, 0\big\} \qquad \text{(A31)}$$

with equality if and only if $w_S = 0$ and $w_R = 0$. Under this contract, the manager only rejects the project when there is no private benefit and $\mu_{AM} < c$. Given the bound for managerial bias (6), the expression (A31) is lower than either (A29) or (A30).

Comparing (A29) and (A30), we obtain the optimal contract stated in Proposition 7. Summarizing the analysis of the manager's decisions, we obtain the manager's action rule outlined in this proposition. ■

Proof of Result 5

When the AI-provided estimate $\widehat{q}_{AI}$ satisfies $\widehat{q}_{AI} \in \left[\frac{c}{\eta_H} + \frac{\eta_H k - (1-\theta_\nu\varphi)b}{\eta_H\theta_\nu(1-\varphi)}\,, 1\right]$, the principal's expected profit is

$$\theta_\nu(\eta_H \widehat{q}AI - c) - \frac{\eta_H b}{\eta_H - \eta_L}.$$

If instead $\widehat{q}_{AI} \in \left[0,\ \frac{c}{\eta_H} + \frac{\eta_H k - (1-\theta_\nu\varphi)b}{\eta_H\theta_\nu(1-\varphi)}\right)$, the expected profit is $-b$, which is bounded below by

$$\theta_\nu \max\{\eta_H \widehat{q}_{AI} - c, 0\} - \frac{\eta_H b}{\eta_H - \eta_L}.$$

Thus, under the optimal AI-contingent contract, the principal's profit satisfies

$$\pi_{B,AM}^{\text{Contingent}} > \theta_\nu \int_0^1 \max\{\eta_H \widehat{q}_{AI} - c, 0\}\, dF(\widehat{q}_{AI}) - \frac{\eta_H b}{\eta_H - \eta_L}.$$

Applying the argument from Proposition 3, which shows that for the convex function $\max \eta_H q - c, 0$, more precise AI information $\widehat{q}AI$ yields a higher expectation than less precise information X_q, we obtain

$$\begin{aligned}\pi_{B,AM}^{\text{Contingent}} &> \theta_\nu \int_0^1 \max\{\eta_H \widehat{q}_{AI} - c, 0\}\, dF(X_q) - \frac{\eta_H b}{\eta_H - \eta_L} \\ &= \theta\nu(\eta_H \zeta_H - c) \cdot \theta_q - \frac{\eta_H b}{\eta_H - \eta_L} \geqslant \pi_{B,M}.\end{aligned}$$

The principal's profit under the optimal AI-contingent contract surpasses the no-AI benchmark. ■