Corporate AI Backfire and Potential Remedies

Joanne Juan Chen* Brandon Yueyang Han†

September 13, 2024

[Click for the Latest Version]

Abstract

Predictive artificial intelligence (AI) has dramatically transformed corporate decision-making by augmenting human judgment with advanced data analysis and predictive algorithms. This paper explores the potential for AI to backfire in corporate decision-making, particularly in the presence of agency conflicts. While AI enhances managers' ability to predict project success probabilities, it can lead to suboptimal decisions that reduce shareholder profits when managerial incentives are misaligned. The study introduces the concept of "corporate AI backfire" and proposes AI-contingent contracts as a remedy. These contracts align managerial compensation with AI predictions, minimizing the risks of AI-driven inefficiencies. By addressing both the benefits and pitfalls of AI adoption, this research offers a framework for optimizing corporate decision-making in an AI-augmented environment.

Key words: *AI-augmented Decision-making, AI-contingent Contract, Corporate Governance*

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

*Department of Finance, Questrom school of Business, Boston University; Email: joanchen@bu.edu.
†University of Maryland Robert H. Smith School of Business; Email: yhan1@umd.edu.

Unlike traditional methods, artificial intelligence provides more objective and transparent assessments, taking into account the goals of VCs and startup founders. Thanks to precise analysis, AI algorithms are, in some cases, able to identify and draw analysts' attention to promising startups they might have overlooked.

– Elena Volotovskaya, Head of Softline Venture Partners[1]

1 Introduction

Artificial intelligence (AI) has revolutionized corporate decision-making by augmenting human capabilities with advanced analytical tools and data-driven insights. In the rapidly evolving landscape of AI, its applications have become central to various corporate decisions, such as recruitment, loan approvals, and venture capital investments. For example, Unilever utilizes AI for recruitment, leveraging machine learning to predict candidate success by evaluating video interviews based on key personality traits; JPMorgan Chase employs AI for advanced credit analysis and loan approvals; EQT Ventures uses an AI platform named Motherbrain to rank potential investments by analyzing factors such as financials, web traffic, and team composition; Correlation Ventures employs machine learning to evaluate startups' potential by analyzing thousands of financing rounds.

Despite the remarkable advancements in AI technology and its growing role in corporate decision-making, several critical questions remain: Is AI's capability to provide precise information and insights alone sufficient for corporations to fully embrace it without hesitation? Can AI consistently enhance decision-making across various corporate environments, or are there

[1] *How AI Is Helping Venture Capital Investment Decisions.* Forbes, August 21, 2024.

instances where its integration may backfire, potentially reducing profitability due to human factors such as agency conflicts?

This study explores the potential for AI-augmented decision-making to backfire in corporate settings, particularly in the presence of agency conflicts. To our best knowledge, we are the first to identify potential problems in organizations that could arise from the interaction of AI adoption and managerial moral hazard. We investigate how AI, while improving decision-making precision, can sometimes lead to suboptimal investment decisions due to the misalignment of managerial incentives. In response to these challenges, we propose a remedy: compensation contracts that incorporate AI predictions to better align the interests of managers with those of shareholders. Through this framework, we aim to maximize the benefits of AI while avoiding pitfalls associated with its adoption. We also explore alternative remedies and assess their feasibility and robustness in overcoming these challenges.

Specifically, we consider a classical principal-agent setting, where a principal hires a manager to evaluate and make decisions regarding an investment opportunity, such as a loan approval or a venture. The project incurs a cost and has an uncertain outcome. The manager, drawing on his expertise, is able to gather information about the probability of success and potential private benefits. A novel aspect of the model is the introduction of AI, which the principal can provide to the manager. When adopted, AI supplies additional insights into the project's probability of success, enhancing the manager's ability to make well-informed decisions. However, the manager remains indispensable, as AI can only capture and analyze a subset of the factors determining the project's success. These factors are categorized into two groups: the first comprises concrete facts and statistics, which both the manager and AI can analyze – but AI outperforms the manager by processing vast datasets and managing numerous variables beyond human capacity; the second group involves information derived from personal observations, interactions, and nuanced insights, which AI struggles to capture or interpret effectively.

While AI enhances the accuracy of a manager's estimate of a project's success probability, in certain scenarios, the additional information provided by AI can paradoxically lead to reduced profits for shareholders. Consider a scenario in which a manager is evaluating an investment opportunity with the assistance of AI-generated insights. Suppose AI reveals that the project has a moderate chance of success, placing it in the grey zone where private benefits could influence decisions. The manager might decide to proceed with the project, not because it is the optimal choice for the firm, but because the marginal benefits to the manager outweigh the potential costs. In this instance, AI has provided better information, but that information has not translated into a better decision from the perspective of the shareholders. This phenomenon, which we term 'corporate AI backfire', challenges the common assumption that when AI can operate effectively and without errors, it always contributes positively to corporate profitability.

To elaborate on the mechanism, the manager estimates the project's success probability using available information. Without AI, projects are categorized into broad high- or low-success probability groups based on a limited number of characteristics. With AI, the manager incorporates detailed predictions, resulting in a more nuanced estimate that is more likely to fall within a middle range compared to the no-AI scenario. The manager's estimate of success probability plays a central role in his decision-making. He approves the project if the expected payoff from approval, which may include private benefits, exceeds that of rejection; otherwise, he rejects it. The decision rule thus follows a threshold pattern: the project is approved if the estimated success probability is above a high threshold, rejected if below a low threshold, and approved only if private benefits are present when the estimate belongs to a middle range between the two thresholds.

This pitfall is especially concerning given the widespread adoption of AI in corporate environments. As AI technologies continue to evolve and are increasingly relied upon to make critical decisions, the potential for AI-driven decisions to backfire becomes more pronounced. Many organizations are investing heavily in AI systems, under the assumption that more data and

better algorithms will naturally lead to superior outcomes. However, this assumption neglects the complexities introduced by human factors, such as agency conflicts, which can distort the implementation of AI-generated insights. When AI is used to support decision-making in contexts where managerial incentives are not properly aligned, the outcomes can be far from optimal.

To address this issue, we propose a novel approach: designing compensation contracts that explicitly incorporate AI predictions. By aligning managerial incentives with AI-generated insights, firms can mitigate the risks of AI backfire and ensure that the efficiency gains from AI are realized. These AI-contingent contracts would adjust managerial compensation based on the information provided by the AI system, thus reducing the manager's ability to exploit private benefits. For example, if AI predicts a high probability of project success, the manager's compensation could be tied to the outcome of the project, making it less attractive for the manager to pursue a project that offers personal perks but low profitability. On the other hand, if AI predicts a low probability of success, the manager's compensation would be adjusted downward, discouraging approval of unprofitable projects.

In this way, AI-contingent contracts serve to better align managerial decision-making with the goals of shareholders, thereby reducing the likelihood of AI backfiring. Our research provides a theoretical framework for understanding the conditions under which AI adoption in corporate decision-making may fail and offers practical solutions to these challenges. By recognizing that AI is not a panacea and that its integration must be carefully managed to account for human incentives, we can develop more robust governance structures that fully leverage AI's potential while minimizing its risks.

Related Literature

This paper contributes to the extensive literature on moral hazard, observability, and optimal contracting, originally pioneered by Mirrlees (1976) and Holmström (1978, 1979). The novel aspect of this work lies in the inclusion of artificial intelligence (AI), which functions analogously

to a second agent by providing complementary information on the project, but without facing incentive issues. This framework, specific to AI technology, has not been previously explored in the literature. We highlight that the collaboration between AI and the human manager may, in certain cases, backfire and result in suboptimal outcomes for the principal. Inspired by Holmström (1979), which demonstrates that more contractible information about the agent's actions enhances welfare for both the principal and the agent, we introduce the AI-contingent contract, where the manager's contract is contingent on both the project outcome and the information provided by AI. We show that this contract not only improves outcomes compared to situations where AI information is observable but not contractible, but also outperforms cases where AI is not adopted at all.

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 of this literature. Within this context, a particularly relevant work is Acemoglu (2021), in which the author discusses at length the potential economic, political, and social costs of AI adoption. Acemoglu (2021) focuses on the harms induced by unregulated AI itself. Regarding AI and information, the author emphasizes the exploitative effects of AI deployment. For instance, excessive data collection undermines consumer privacy; too much data fosters unfair competition, and can result in behavioral manipulation. Our paper, on the other hand, emphasizes potential shareholder losses arising from additional information provided by AI. We examine the firm-level production relationship and argue that, in the presence of agency conflicts, AI-augmented decision-making can lead to worse outcomes compared to scenarios where AI is not employed. To address this issue, we propose a solution in the form of an AI-contingent contract, offering a novel approach to address the challenges posed by AI adoption.

Research on corporate AI adoption and its implications is still in its early stages, with only a limited number of papers exploring this subject. Theoretical research, such as Weitzner (2024) and Zhong (2024), addresses topics that are distinct from our focus. Weitzner (2024) examines algorithm aversion driven by asymmetric information and reputational concerns. Zhong (2024) investigates the optimal integration of humans and technology when they both can make errors. 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), and Babina et al. (2024), among others. While these studies do not explicitly examine AI backfire, certain aspects of their findings align with ours. For example, 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.

The remainder of the paper is structured as follows. Section 2 introduces the model and presents preliminary analyses. In Subsections 3.1 and 3.2, we derive the optimal contracts and the principal's payoffs in the absence of AI and under AI adoption, respectively. Subsection 3.3 illustrates how corporate AI adoption can backfire and provides sufficient conditions for such backfire to occur. In Section 4, we propose the AI-contingent contract as a remedy and compare it with alternative solutions. Section 5 concludes the paper.

2 The Model

In this section, we present an investment decision-making framework where a principal employs a manager to evaluate and implement projects. We explain how different types of information influence the project's success probability and the comparative advantages of AI and the manager in processing these types of information. In addition, we conduct a preliminary analysis of the

contract structure and managerial decision rules, setting the stage for a deeper investigation into AI's impact on firm performance.

2.1 Model Setup

The principal (she) hires a manager (he) to evaluate investment opportunities and implement projects on her behalf. At time 0, there is an investment opportunity (the "project") that requires an investment cost c and generates an uncertain payoff y at the end of the period. The payoff is 1 if the project succeeds and is 0 if the project fails, i.e., $y \in \{0, 1\}$. The project's probability of success is denoted by p, which is not directly observed. The manager has expertise and can obtain information on p. Therefore, the manager can decide whether to approve or reject the project based on his obtained information. Let a denote the manager's action, where $a \in \{A, R\}$, with A representing approval and R representing rejection.

If the firm adopts predictive artificial intelligence, further information on p will be obtained, and the manager will be able to make the investment decision with the refined information provided by the predictive AI – which we refer to as "AI-augmented decision making".

Incentives and the Contract:

Both players are risk-neutral, and the manager is protected by limited liability. There is a 50% probability that the manager receives a private benefit of b from approving the project. For instance, some projects may offer a manager personal perks (such as luxury travel and entertainment) as part of the business dealings. Similarly, in the context of loan approvals, a loan officer might have side deals or shared interests with certain loan applicants. The private benefit is independent of the project's success probability p. Its potential existence is common knowledge, but the specific state of the world in which it exists is private information known only to the manager after he evaluates the project. We use a random variable B to represent this stochastic realization of private benefit: $B = b$ if the private benefit exists and $B = 0$ otherwise.

The manager's actions and the project payoffs are contractible. Therefore, the contract designed by the principal has a triplet form, (w_A, w_R, w_F), where w_A is the manager's wage if the project succeeds after being approved, w_R is his wage if the project is rejected, and w_F is his wage if the project fails after being approved. The manager acts in the interest of the principal when he is indifferent.

Determinants of the Success Probability:

The project's success probability p is determined by two groups of characteristics. The first group consists of concrete facts and statistics that can be processed and analyzed by the AI. The second group includes information derived from personal observations, interactions, and insights, which can be challenging for the AI to capture or interpret. AI's inability to measure and analyze this second group of characteristics makes the human decision-maker irreplaceable.

To be more specific, the success probability is given by

$$p = \alpha p_1 + (1 - \alpha) p_2 \tag{1}$$

where the first term, αp_1, and the second term, $(1 - \alpha)p_2$, respectively summarize the influence of the first and second group of characteristics on the project's success probability. p_1 and p_2 are independent and identically distributed following $U[0, 1]$. $\alpha \in (0, 1)$ presents the relative importance of the first group. This additive form of Equation (1) indicates that both groups of characteristics are essential and cannot be substituted for one another in determining the probability p.[2] In this paper, we assume $\alpha \geqslant 1 - \alpha$ to highlight the importance of concrete facts and statistics in predicting the project's success probability. The reasoning and main conclusions remain valid even if $\alpha < 1 - \alpha$.

Information Acquisition and Processing:

[2]In this additive formulation, the two groups of characteristics contribute independently to the probability of success. A higher realization of p_1 does not diminish the importance of p_2, since the marginal effect of p_2 on p remains constant. This contrasts commonly used logistic or probit models, where the marginal effect decreases as the predicted probability approaches 0 or 1.

The manager has limited time and mental capacity to process all relevant characteristics. Instead, he focuses his attention on a subgroup of variables X_1 to obtain information about p_1, and X_2 to obtain information about p_2.

In contrast, artificial intelligence can process vast amounts of data and handle numerous variables that may overwhelm a human decision-maker. However, it may struggle to capture other dimensions of information, such as cultural references, moral judgments, and complex interpersonal dynamics. In this context, we consider a well-trained AI model and assume that its inference on the first group of characteristics is accurate; that is, we presume AI can precisely estimate p_1. Characteristics that AI cannot capture or process are summarized in the second group, where the manager is relied upon to estimate p_2.

For analytical convenience, we assume that X_i (for $i = 1, 2$) follows a symmetric binary distribution with outcomes in $\{H, L\}$; and the conditional expectation $\mathbb{E}[p_i|X_i]$ (for $i = 1, 2$) is specified as

$$\mathbb{E}[p_i|X_i] = \begin{cases} q, & \text{when } X_i = H \\ 1 - q, & \text{when } X_i = L \end{cases} \tag{2}$$

where $q \in (1/2, 3/4]$.

In this setting, the outcomes H and L are equally probable, with H representing the case where the mean of p_i's posterior distribution, $\mathbb{E}[p_i|X_i]$, is higher. The parameter q measures the informativeness of variable X_i with respect to the project's success probability p_i. A larger q corresponds to a greater difference between the two posterior means $\mathbb{E}[p_i|H]$ and $\mathbb{E}[p_i|L]$. The upper bound for q is $3/4$, as this posterior mean cannot exceed the mean of the upper half of the unit interval, $[1/2, 1]$.

Assumptions on Parameter Ranges:

To ensure that both groups of characteristics are indispensable in the investment decision, the model assumes that when X_1 and X_2 represent the only available information, the project should be approved only if both provide favorable signals. The following parameter assumptions ensure

this condition is met.

$$\alpha q + (1-\alpha)(1-q) < c < q\,, \tag{3}$$

$$b < \frac{(1-\alpha)(2q-1)}{8q} \cdot \min\Big\{q-c,\, c-\big[\alpha q + (1-\alpha)(1-q)\big]\Big\}\,. \tag{4}$$

The first assumption stipulates that the investment cost c should be within a moderate range. The second assumption requires the private benefit not to be excessively large. Section 3.1 provides a comprehensive explanation and analysis of these assumptions.

2.2 Preliminary Analyses

The manager forms expectations about the project's success probability based on the available information. Let $\mu = \mathbb{E}[p \mid \mathcal{I}]$ denote this expectation, where $\mathcal{I} = \{(X_1, X_2, B)\}$ in the absence of AI, and $\mathcal{I} = \{(p_1, X_1, X_2, B)\}$ when the manager has access to the refined information provided by the predictive AI. In the subsequent sections, we will denote μ in these respective cases as μ_M and μ_{AM}.

The manager approves the project when his expected payoff from approval exceeds rejection:

$$\begin{cases} \mu w_A + (1-\mu) w_F + b \;>\; w_R, & \text{if private benefit exists,} \\ \mu w_A + (1-\mu) w_F \;>\; w_R, & \text{if private benefit does not exist.} \end{cases} \tag{5}$$

He rejects the project if the reverse inequalities hold.

First-step Analysis of the Contract:

The principal sets $w_A \geqslant w_F$ 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_A, w_R, w_F) where $w_F > 0$, an equivalent contract $(w_A - w_F, w_R - w_F, 0)$ achieves the same incentive at a reduced cost. Therefore, in the optimal contract, w_F should be set to 0.

Principal's Payoff Function:

From the above analyses, the principal's realized payoff is $1 - w_A - c$ if the project succeeds after approval, $-c$ if it fails after approval, and $-w_R$ if it is rejected. Therefore, the principal's expected profit is equal to

$$\mathbb{E}\big[\big(p(1-w_A)-c\big)\mathbf{1}_{\{a=A\}}\big] - \mathbb{E}\big[w_R\mathbf{1}_{\{a=R\}}\big]\,, \tag{6}$$

where $\mathbf{1}_{\{a=A\}}$ and $\mathbf{1}_{\{a=R\}}$ are indicator functions for the manager's actions approval A and rejection R, respectively. Applying the law of iterated expectations with respect to the manager's information $\mathcal{I}$, the above expression can be reformulated as one involving $\mu = \mathbb{E}[p|\mathcal{I}]$ instead:

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{E}\big[\big(p(1-w_A)-c\big)\mathbf{1}_{\{a=A\}}\big|\mathcal{I}\big]\Big] - \mathbb{E}\big[w_R\mathbf{1}_{\{a=R\}}\big]\\ =&\mathbb{E}\big[(\mu(1-w_A)-c)\mathbf{1}_{\{a=A\}}\big] - \mathbb{E}\big[w_R\mathbf{1}_{\{a=R\}}\big]\,,\end{aligned} \tag{7}$$

where $\mathbf{1}_{\{a=A\}}$ is $\mathcal{I}$-measurable since the manager's action a is determined by his information.

Manager's Decision Rule:

If $w_A > 0$, the manager's decision rule derived from (5) can be summarized as follows: approve the project when $\mu > \overline{\mu}$; reject the project when $\mu < \underline{\mu}$; and approve the project only if private benefits exist when μ is within the interval $(\underline{\mu}, \overline{\mu})$.[3] The thresholds $\underline{\mu}$ and $\overline{\mu}$ are defined as

$$\underline{\mu} = \frac{w_R - b}{w_A}\,, \quad \overline{\mu} = \frac{w_R}{w_A}\,. \tag{8}$$

The diagram below illustrates the manager's decision rule according to his posterior expectation of the project's success probability, μ.

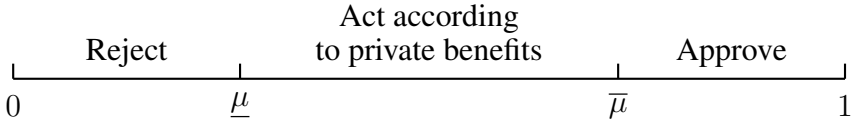


Figure 1: Manager's Decision Rule Following μ

[3] Subsequent analysis of optimal contracts in Section 3 shows that $\overline{\mu} > c$ and $\underline{\mu} < c$. Since the manager acts in the principal's interest when he is indifferent, he approves the project when $\mu = \overline{\mu}$ and rejects it when $\mu = \underline{\mu}$.

Optimal contracts with $w_A = 0$ will only arise in Section 4 when the AI reports are contractible. Therefore, we will reserve the analysis for now and discuss these cases in detail in Section 4.

3 Managerial Decisions and AI Adoption Pitfall

In this section, we model corporate decisions before and after adopting predictive artificial intelligence and demonstrate how AI-augmented decision-making may backfire. Section 3.1 describes the equilibrium without AI. Section 3.2 examines how AI alters the manager's expectations, leading to an equilibrium where the principal balances managerial incentive costs with losses from suboptimal decisions. Section 3.3 compares profits across both equilibriums and identifies conditions under which AI backfire occurs.

3.1 Optimal Contract in the absence of AI

According to previous analyses, the manager makes predictions on p based on X_1 and X_2 in the absence of AI. The manager's expectation in this case, denoted by μ_M, is given by:

$$\mu_M = E[p|X_1, X_2] = \begin{cases} q, & \text{if } X_1 = H \text{ and } X_2 = H, \\ \alpha q + (1-\alpha)(1-q), & \text{if } X_1 = H \text{ and } X_2 = L, \\ \alpha(1-q) + (1-\alpha)q, & \text{if } X_1 = L \text{ and } X_2 = H, \\ 1-q, & \text{if } X_1 = L \text{ and } X_2 = L. \end{cases} \tag{9}$$

These four realizations of μ_M each occur with an equal probability of 1/4, as X_i (for $i = 1, 2$) follows a symmetric binary distribution.[4]

[4]To be more precise, $\mu_M = \mathbb{E}[p|\mathcal{I}] = \mathbb{E}[p|X_1, X_2, B]$ 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 characteristics, X_1 and X_2.

The principal uses w_A and w_R to determine the manager's decision thresholds, $\underline{\mu}$ and $\overline{\mu}$. The objective is to set both thresholds close to c, ensuring that profitable projects are approved and loss-making projects are rejected. However, positioning both thresholds near c results in high incentive costs, as both w_A and w_R increase with the distance between thresholds $\overline{\mu} - \underline{\mu}$:

$$w_A = \frac{b}{\overline{\mu} - \underline{\mu}}\,, \quad w_R = \frac{\overline{\mu} b}{\overline{\mu} - \underline{\mu}}\,, \tag{10}$$

as derived from Equations (8). Consequently, the principal also seeks to maintain some distance between the decision thresholds to manage these incentive costs.

With assumptions (3) and (4), the principal chooses $\overline{\mu} = q$ and $\underline{\mu} = \alpha q + (1-\alpha)(1-q)$ in equilibrium.[5] Setting $\overline{\mu}$ above q would miss profitable investment opportunities, whereas setting $\underline{\mu}$ below $\alpha q + (1-\alpha)(1-q)$ would result in investment losses. Therefore, we determine the optimal wages in equilibrium and can calculate the principal's expected payoff established in Equation (6). This equilibrium is characterized by Lemma 1 below.

Lemma 1 *In the absence of AI,*

1.) the optimal contract is

$$w_A = \frac{b}{(1-\alpha)(2q-1)}, \quad w_R = q w_A = \frac{qb}{(1-\alpha)(2q-1)}, \quad w_F = 0\,;$$

2.) the manager's action rule is

to approve the project if $(X_1, X_2) = (H, H)$*, and to reject the project otherwise;*

3.) the principal's expected profit is

$$\mathbb{E}\big[(\mu_M(1-w_A) - c)\mathbf{1}_{\{a=A\}}\big] - \mathbb{E}\big[w_R \mathbf{1}_{\{a=R\}}\big] = \frac{1}{4}(q-c) - \frac{qb}{(1-\alpha)(2q-1)}\,.$$

In this equilibrium, the manager approves the project when $\mu_M = q$, with a probability of 1/4, and rejects it otherwise. While the manager may sometimes have opportunities to obtain a private benefit, these opportunities do not manipulate/influence the manager's decision-making process.

[5] Please refer to Remark 1 for a detailed examination and explanation.

In other words, the manager's equilibrium actions remain unaffected by the presence of private benefits. The principal's expected profit is given by the difference between the expected project payoff $(q-c)/4$ and the expected wage paid to the manager $qb/(1-\alpha)(2q-1)$.

REMARK 1 ***(Examination and Explanation of Parameter Range Assumptions)***
We impose Assumption (3) to ensure that (i) approving all investments without any selection is not profitable, in line with real-world situations, where profitable investment opportunities are relatively scarce; and (ii) both groups of characteristics are essential in determining the investment decision. Specifically, Requirement (i) excludes the cases where $c < 1/2$*; and Requirement (ii) excludes the following two subcases:* $1/2 \leqslant c < \alpha q + (1-\alpha)(1-q)$*, where the optimal investment decision only requires* $X_1 = H$ *and is independent of* X_2*; and* $c > q$*, where no realization of* μ_M *justifies investment. Assumption (4) limits the extent of the agency conflict, ensuring that in the absence of AI, the manager approves profitable projects (*$\mu_M > c$*) and rejects unprofitable ones (*$\mu_M < c$*) in equilibrium. This scenario is central to our analysis, as it provides a clear contrast to potential AI backfire. In other words, this assumption serves expositional purposes, allowing us to focus on scenarios of greater interest.*

3.2 AI-augmented Decision-making

AI can analyze large volumes of data and handle numerous variables within the first group of characteristics, resulting in a significantly more accurate estimation of p_1 compared to the manager. While we assume for simplicity that AI learns the exact value of p_1, such precision is not strictly necessary. The mechanism and qualitative results of this paper are valid as long as AI's predictions of p_1 are more accurate than those of the manager.

With the assistance of AI, the manager's expectation of p is as follows. [6]

$$\mu_{AM} = \mathbb{E}[p|p_1, X_1, X_2] = \begin{cases} \alpha p_1 + (1-\alpha)q, & \text{if } X_2 = H\,, \\ \alpha p_1 + (1-\alpha)(1-q), & \text{if } X_2 = L\,. \end{cases} \tag{11}$$

Compared to $\mu_M = \mathbb{E}[p|X_1, X_2]$, μ_{AM} incorporates additional information from p_1, leading to a more accurate estimate of the project's success probability, p. As a result, the distribution of μ_{AM} is a mean-preserving spread of μ_M, described by:

$$f(\mu_{AM}) = \frac{dF(\mu_{AM})}{d\mu_{AM}} = \begin{cases} \dfrac{1}{2\alpha}, & \mu_{AM} \in \big[(1-\alpha)(1-q), (1-\alpha)q\big) \\ \dfrac{1}{\alpha}, & \mu_{AM} \in \big[(1-\alpha)q, \alpha + (1-\alpha)(1-q)\big] \\ \dfrac{1}{2\alpha}, & \mu_{AM} \in \big(\alpha + (1-\alpha)(1-q), \alpha + (1-\alpha)q\big] \end{cases} \tag{12}$$

In this context, the principal's problem is to design the contract that maximizes the firm's expected payoff net of the managerial compensation, taking the manager's AI-augmented decision-making process into consideration. Since there is a one-to-one correspondence between (w_A, w_R) and $(\overline{\mu}, \underline{\mu})$ as shown in Equations (10), the principal's problem can be formulated as maximizing the following expression by optimally selecting decision thresholds $(\overline{\mu}, \underline{\mu})$.

$$\begin{aligned} &\max_{(w_A, w_R)} \mathbb{E}\big[(\mu_{AM}(1-w_A) - c)\mathbf{1}_{\{a=A\}}\big] - \mathbb{E}\big[w_R \mathbf{1}_{\{a=R\}}\big] \\ =&\max_{(\overline{\mu},\, \underline{\mu})} \int_{\overline{\mu}}^{1} \big[\mu_{AM}(1-w_A) - c\big] dF(\mu_{AM}) + \int_{\underline{\mu}}^{\overline{\mu}} \left[\frac{1}{2}\big[\mu_{AM}(1-w_A) - c\big] + \frac{1}{2}(-w_R)\right] dF(\mu_{AM}) \\ &\quad + \int_{0}^{\underline{\mu}} (-w_R) dF(\mu_{AM})\,. \end{aligned} \tag{13}$$

The principal's expected profit takes this form because the manager's decision rule is to approve the project with certainty if $\mu_{AM} \geqslant \overline{\mu}$, approve it with a fifty percent probability if $\mu_{AM} \in (\underline{\mu}, \overline{\mu})$, and reject it with certainty if $\mu_{AM} \leqslant \underline{\mu}$, as analyzed in Subsection 2.2.

[6] Similar to the analysis of μ_M in Equation (9), we exclude B from the conditioning information because it is independent of p, p_1, X_1, and X_2.

The above objective function can be decomposed into the following three components:

$$\underbrace{\int_{c}^{1}(\mu_{AM}-c)dF(\mu_{AM})}_{\text{No-agency-conflict Profit}}-\underbrace{\left[\frac{1}{2}\int_{\underline{\mu}}^{c}(c-\mu_{AM})dF(\mu_{AM})+\frac{1}{2}\int_{c}^{\overline{\mu}}(\mu_{AM}-c)dF(\mu_{AM})\right]}_{\text{Loss from Suboptimal Investment Decisions}}$$
$$-\underbrace{\left[\left(\int_{\overline{\mu}}^{1}\mu_{AM}dF(\mu_{AM})+\frac{1}{2}\int_{\underline{\mu}}^{\overline{\mu}}\mu_{AM}dF(\mu_{AM})\right)w_{A}+\left(\frac{1}{2}F(\underline{\mu})+\frac{1}{2}F(\overline{\mu})\right)w_{R}\right]}_{\text{Managerial Wage}}. \quad (14)$$

The first term represents the project's expected profit under a no-agency-conflict benchmark, where there is no private benefit and, thus, no interest misalignment between the principal and the manager. The second term demonstrates the expected loss from suboptimal investment decisions. The manager sometimes makes decisions that are suboptimal from the firm's perspective due to the potential existence of private benefits. In the absence of such benefits, the manager would always approve the project if $\mu_{AM} > c$ and reject it if $\mu_{AM} < c$. However, in this context, when private benefits are present, the manager approves the projects with $\mu_{AM} \in (\underline{\mu}, c)$ that should be rejected; when private benefits are absent, the manager rejects the projects with $\mu_{AM} \in (c, \overline{\mu})$ that should be approved. Each of these two scenarios occurs with a fifty percent probability, resulting in an expected loss compared with the no-agency-conflict case. The third term shows the expected managerial wage across all different scenarios.

The expected managerial wage increases in $\underline{\mu}$ and decreases in $\overline{\mu}$ as shown in Equation (10), while the expected loss from suboptimal investment decisions decreases in $\underline{\mu}$ and increases in $\overline{\mu}$. Therefore, the principal would like to choose $\underline{\mu}$ and $\overline{\mu}$ to minimize the combined cost of wages and losses from the suboptimal investment decisions, in order to maximize Expression (14). Lemma 2 characterizes the results from this optimization.

Lemma 2 *With AI in place,*

1.) the optimal contract is

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

where $\overline{\mu}^*$ and $\underline{\mu}^*$ are implicitly determined by the following system of equations:

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

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 is given by Expression (14) with $(\overline{\mu}, \underline{\mu}) = (\overline{\mu}^*, \underline{\mu}^*)$.*

In this equilibrium, the manager may make suboptimal investment decisions when his expected probability of success, μ_{AM}, falls within the range $(\underline{\mu}^*, \overline{\mu}^*)$, as his actions are driven by private benefits. Specifically, when private benefits are present, the manager approves the projects with $\mu_{AM} \in (\underline{\mu}^*, c)$ that should be rejected; when private benefits are absent, the manager rejects the projects with $\mu_{AM} \in (c, \overline{\mu}^*)$ that should be approved. The principal tolerates these suboptimal decisions because narrowing the gap between these decision thresholds would impose excessively high incentive costs. The corollary below provides a further analysis of the range $(\underline{\mu}^*, \overline{\mu}^*)$.

Corollary 1 *The range $(\underline{\mu}^*, \overline{\mu}^*)$, where private benefits drive the manager's actions, increases with the level of private benefits b. Specifically,*

$$\frac{d\overline{\mu}^*}{db} > 0\,, \quad \frac{d\underline{\mu}^*}{db} < 0\,.$$

Intuitively, incentivizing the manager to make appropriate decisions becomes more costly when the level of private benefits is higher, as the manager is more inclined to pursue these benefits. Consequently, the principal tolerates suboptimal investment decisions over a broader range of $(\underline{\mu}^*, \overline{\mu}^*)$. In contrast, when the level of private benefits is lower, it is less costly to incentivize the manager to make decisions in the best interest of the firm. Therefore, the principal sets a

narrower range of $(\underline{\mu}^*, \overline{\mu}^*)$, with both thresholds, $\underline{\mu}^*$ and $\overline{\mu}^*$, shifting toward the investment cost c from the left and right, respectively. Figure 2 below provides an example.

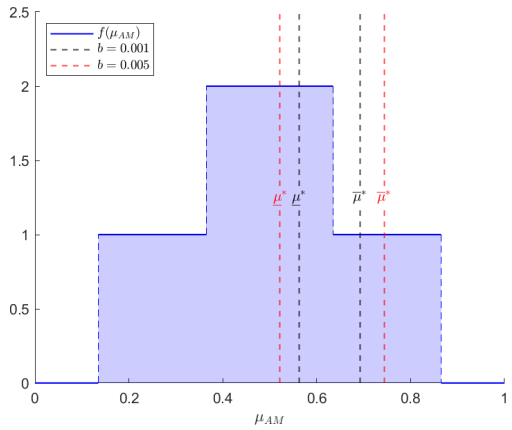


Figure 2: The Range $(\underline{\mu}^*, \overline{\mu}^*)$ Increases With b

The figure depicts the equilibrium decision thresholds $\overline{\mu}^*$, $\underline{\mu}^*$ for private benefits $b = 0.001$ and $b = 0.005$, and the distribution of the manager's expectation $f(\mu_{AM})$. The other parameters for this figure are $\alpha = 0.5$, $q = 0.73$ and $c = 0.61$.

3.3 Corporate AI Backfire

AI enhances the precision of success probability estimates, offering a more accurate evaluation of potential outcomes. Specifically, the new estimate, μ_{AM}, which incorporates AI's insights, provides a more precise assessment of the success probability p compared to the traditional one, μ_M, generated by the manager alone. This improved accuracy in assessing the likelihood of project success can lead to better-informed and more efficient decisions.

For instance, when the signals are mixed, such as when $X_1 = H$ and $X_2 = L$, 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 find p_1 exceptionally high after identifying other factors and statistical patterns. If p_1 is higher than the threshold $[c - (1 - \alpha)q]/\alpha$, the project should be accepted rather than rejected. Conversely, when both X_1 and X_2 are equal to H, the manager perceives the project

as promising. However, if AI uncovers other relevant characteristics indicating that p_1 is below the threshold $[c-(1-\alpha)q]/\alpha$, the project should be rejected despite the manager's positive assessment initially.

In an ideal world without agency conflicts (i.e, there is no private benefit), the adoption of AI would always lead to more efficient decision-making, as AI provides additional information on the project's probability of success. This enables the manager to more accurately distinguish between projects with success probabilities above the investment threshold and those below it. This result is formally established in the following proposition.

Proposition 1 *Without agency conflicts, the principal's expected profit from the project is higher when the manager can access more information. That is, for any information set $\mathcal{I} \subseteq \widetilde{\mathcal{I}}$,*

$$\int_c^1 (\widetilde{\mu}-c)dF(\widetilde{\mu}) \geqslant \int_c^1 (\mu-c)dF(\mu)\,,$$

where $\widetilde{\mu}=\mathbb{E}[p|\widetilde{\mathcal{I}}]$ and $\mu=\mathbb{E}[p|\mathcal{I}]$.

Specifically, in the setting of this model, the principal always benefits from AI-augmented decision-making in the absence of agency conflicts. This increase in the expected profit, termed as the "Efficiency Gain from Information" and denoted by π_{EG}, is always non-negative:

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

When agency conflicts arise – where the manager may pursue private benefits at the expense of the principal's interests – the previously discussed result may no longer apply. With agency conflicts, the corporate adoption of AI can lead to unintended consequences. This potential for AI-augmented managerial decisions to backfire in the corporate environment is the central focus of our paper.

The intuition behind this potential AI backfire is as follows: the adoption of AI alters the manager's expectations regarding p by shifting the distribution from $F(\mu_M)$ to $F(\mu_{AM})$, due to the increased information available. Unlike traditional estimates without AI, μ_M, the AI-

enhanced estimate, μ_{AM}, is more likely to fall within a moderate range of success probabilities. This is because AI generates a more precise estimate of p_1, whereas the manager, relying on the observation of X_1, can only classify p_1 into two categories. As a result, the distribution of μ_{AM} has higher probability densities within this middle range. When μ_{AM} falls within the interval $(\underline{\mu}^*, \overline{\mu}^*)$, specifically, the manager can pursue personal benefits with minimal impact on his wage because the difference between the wage from rejection (w_R) and the expected wage from approval ($\mu_{AM} w_A$) is small in this region. Consequently, the manager can obtain a level of private benefits higher than the wage loss, and he chooses to pursue them at the cost of the principal.

This misalignment of interests ultimately leads to increased losses from suboptimal investment decisions, which can surpass the efficiency gains from AI adoption. As shown in the second term of Equation (14), these losses can be significant and outweigh the efficiency gain from information, π_{EG}. This dynamic underscores how the corporate adoption of AI, despite its ability to enhance decision-making precision, can backfire in the presence of agency conflicts, resulting in outcomes that harm the firm's profitability. Proposition 2 below provides sufficient conditions for AI backfire to occur.

Proposition 2 ***(Existence and Sufficient Conditions for AI Backfire)***

The sufficient conditions for the principal's expected profit to be lower with AI-augmented decision-making are that the Efficiency Gain from Information is not excessively large, and that potential private benefits for the manager remain within a moderate range. Specifically,

$$\begin{cases} \pi_{EG} < \dfrac{c^2}{16\alpha}\left(\dfrac{q}{(1-\alpha)(2q-1)} + \dfrac{1}{2}\right)^{-2}, \\ b \in [\,\underline{b},\, \overline{b}\,], \end{cases}$$

where

$$\pi_{EG} \equiv \int_c^1 (\mu_{AM} - c) dF(\mu_{AM}) - \int_c^1 (\mu_M - c) dF(\mu_M) = \frac{(1+\alpha-2c)^2 + (1-\alpha)^2(2q-1)^2}{8\alpha} - \frac{q-c}{4},$$

and $\underline{b} < \overline{b}$ *are the two positive solutions to the following equation:*

$$\frac{3}{4}\alpha^{-\frac{1}{3}}c^{\frac{2}{3}}\, b^{\frac{2}{3}} - \left(\frac{q}{(1-\alpha)(2q-1)} + \frac{1}{2}\right) b = \pi_{EG}\,. \tag{15}$$

When the Efficiency Gain from Information is exceptionally large, it outweighs the manager's increased ability to exploit private benefits and the corresponding rise in incentive costs. In this case, AI-augmented decision-making consistently results in higher profits for the principal. However, across most parameter ranges where the Efficiency Gain from Information is more modest, AI backfire is likely, particularly when potential private benefits fall within an intermediate range. Below, we provide intuition behind this condition.

In this intermediate range of private benefits, the balance between the efficiency gain from information and the costs associated with incentivizing the manager becomes delicate. When the private benefit, b, is small, the agency conflict is minimal, allowing the principal to set decision thresholds close to the project cost, ensuring that the efficiency gains from AI consistently outweigh any potential losses. Conversely, when b is large, the steep cost of incentivizing the manager makes AI advantageous by allowing the principal to reduce compensation costs through the precise separation of projects with very high versus moderately high success probabilities.

However, when b is at an intermediate level, the agency conflict becomes more pronounced without reaching the extreme inefficiencies of a large b. In this situation, the manager's decisions are increasingly driven by private benefits over a wider range of scenarios. AI, by enhancing the manager's access to information, heightens the manager's ability to exploit these private benefits while still requiring significant incentive compensation. The efficiency gain from AI is insufficient to fully offset the increased incentive costs and losses from more frequent suboptimal investment decisions. Consequently, in these intermediate ranges, AI is more likely to backfire, leading to a net reduction in profits for the principal.

Figure 3 below provides an example of sufficient conditions, demonstrating that the corporate AI backfire may occur when private benefit b is at an intermediate level. In this example, the

sufficient condition for corporate AI backfire is π_{EG} being smaller than the lower bound of the sum of the increase in incentive costs and losses from suboptimal investment decisions.

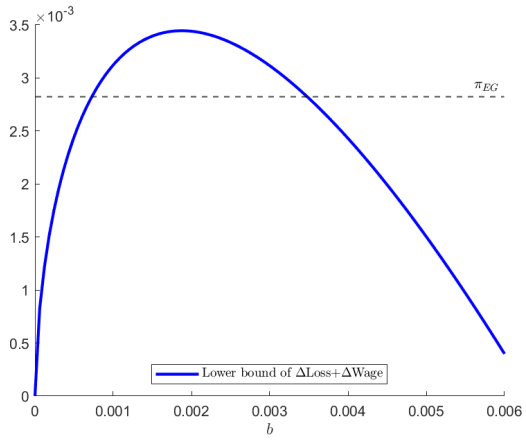


Figure 3: Sufficient Conditions for Corporate AI Backfire, an Example

This figure illustrates the lower bound for the increase in losses from suboptimal decisions and wage costs relative to the no-AI scenario, alongside a horizontal line showing the Efficiency gain from Information. The principal's expected profit is lower with AI-augmented decision-making when the private benefit b lies between the two points where the curve intersects the horizontal line π_{EG}. The parameters are $\alpha = 0.5$, $q = 0.73$, and $c = 0.61$.

4 Potential Remedies

In this section, we discuss potential remedies to the corporate AI adoption pitfall. Section 4.1 proposes an AI-contingent contract that adjusts managerial compensation based on AI-generated information, yielding higher profits across all levels of private benefits compared to the no-AI scenario. Section 4.2 outlines an alternative approach that involves limiting the scope of information AI communicates to the manager.

4.1 AI-contingent Contract

As discussed in the previous section, the backfire can occur because the manager's expectation μ_{AM} may fall within a moderate range, leaving room for suboptimal investment decisions driven by private benefits. Additionally, Under the optimal contract described in Lemma 2, the

manager's decision thresholds, $\underline{\mu}^*$ and $\overline{\mu}^*$, depend on $F(\mu_{AM})$, the distribution of the manager's expectations, but remain unchanged regardless of p_1, the information provided by AI.

Since the information provided by AI is public and verifiable, the principal can leverage it to obtain a more precise distribution, $F(\mu_{AM}|p_1)$, and optimize the following objective in place of Equation (14):

$$\int_c^1 (\mu_{AM}-c)dF(\mu_{AM}|p_1) - \left[\frac{1}{2}\int_{\underline{\mu}}^{c}(c-\mu_{AM})dF(\mu_{AM}|p_1)+\frac{1}{2}\int_c^{\overline{\mu}}(\mu_{AM}-c)dF(\mu_{AM}|p_1)\right]$$
$$-\left[\left(\int_{\overline{\mu}}^1 \mu_{AM}dF(\mu_{AM}|p_1)+\frac{1}{2}\int_{\underline{\mu}}^{\overline{\mu}}\mu_{AM}dF(\mu_{AM}|p_1)\right)w_A+\left(\frac{1}{2}F(\overline{\mu}|p_1)+\frac{1}{2}F(\underline{\mu}|p_1)\right)w_R\right]. \tag{16}$$

Similar to the previous section, the principal aims to minimize the total costs associated with suboptimal investment decisions and managerial compensation, but now this objective is considered for a specific value of p_1. One feasible contract involves selecting $\underline{\mu}$ and $\overline{\mu}$ such that the distribution $F(\mu_{AM}|p_1)$ has minimal probability mass within the open interval $(\underline{\mu}, \overline{\mu})$, thereby diminishing the potential losses from suboptimal decisions, while simultaneously ensuring that the interval is sufficiently wide to effectively mitigate managerial compensation costs. To implement this arrangement, the manager's wages should be contingent on the AI's report, p_1. Therefore, we propose the following "AI-contingent contract" as a remedy to address corporate AI backfire issues.

Proposition 3 *When the wage is contingent on the AI's report,*

1.) The optimal contract is

$$\big(w_A, w_R, w_F\big) = \begin{cases} \big(0,0,0\big), & p_1 \in \left(\frac{c-4b}{\alpha(1+4\varphi)} - \frac{1-\alpha}{\alpha}(1-q)\ ,\ 1\right] \\ \big(\varphi, (\alpha p_1 + (1-\alpha)q)\varphi, 0\big), & p_1 \in \left[\frac{c-3b}{\alpha(1-4\varphi)} - \frac{1-\alpha}{\alpha}q\ ,\ \frac{c-4b}{\alpha(1+4\varphi)} - \frac{1-\alpha}{\alpha}(1-q)\right] \\ \big(0,b,0\big), & p_1 \in \left[0\ ,\ \frac{c-3b}{\alpha(1-4\varphi)} - \frac{1-\alpha}{\alpha}q\right), \end{cases}$$

where $\varphi \equiv \dfrac{b}{(1-\alpha)(2q-1)}$.

2.) The manager's action rule is:

When $X_2 = H$, approve the project with certainty if $\mu_{AM} \geqslant (c-3b)/(1-4\varphi)$, approve it when private benefit exists if $\mu_{AM} \in (c-b, (c-3b)/(1-4\varphi))$, and reject it with certainty if $\mu_{AM} \leqslant c-b$.

When $X_2 = L$, approve the project with certainty if $\mu_{AM} \geqslant c$, approve it when private benefit exists if $\mu_{AM} \in ((c-4b)/(1+4\varphi), c)$, and reject it with certainty if $\mu_{AM} \leqslant (c-4b)/(1+4\varphi)$.

The equilibrium features three distinct wage schemes:

- For intermediate values of p_1, the principal seeks to minimize the likelihood of μ_{AM} falling within the interval $(\underline{\mu}, \overline{\mu})$, while also ensuring that this interval remains as wide as possible to reduce wage costs, as discussed previously. Therefore, the principal optimally sets $\overline{\mu} = \alpha p_1 + (1-\alpha)q$ and $\underline{\mu} = \alpha p_1 + (1-\alpha)(1-q)$. Substituting into Equation (10), we obtain the wage scheme described in the proposition.

- When p_1 is high, investing in the project incurs minimum loss at worst, even if $X_2 = L$. In this case, the principal chooses not to incentivize the manager to reject $X_2 = L$ projects. Instead, the principal sets the wage to zero across all outcomes, and the manager only rejects the project when there is no private benefit and $\mu_{AM} < c$.

- When p_1 is low, the expected profit attainable from investing in the project is minimal, even if $X_2 = H$. Consequently, the principal sets $w_A = 0$ and offers a rejection wage of $w_R = b$ to counterbalance any private benefits derived from project approval. When there is no private benefit, the manager rejects the project irrespective of X_2. When private benefits are present, the manager becomes indifferent and maximizes the principal's profits by approving the project if $\mu_{AM} > c - b$ and rejecting it if $\mu_{AM} \leqslant c - b$.

When p_1 is in the intermediate range, the wage for a successful project, w_A, remains constant, while the rejection wage, w_R, increases with p_1. This outcome may seem surprising, as one

might expect the principal to offer higher rejection wages for projects with low p_1, anticipating greater concern over the manager pursuing private benefits and approving projects with a lower probability of success. However, the underlying mechanism suggests the opposite: decision thresholds $\underline{\mu}$ and $\overline{\mu}$ are reduced for projects with lower success probabilities. Consequently, w_R is lower because it decreases with both decision thresholds.

This AI-contingent contract reduces the loss from suboptimal investment decisions compared to the non-contingent contract in Lemma 2. Furthermore, this contract provides an effective remedy to this corporate AI adoption pitfall, as the principal's profit surpasses the profit without AI for all possible levels of private benefits, as presented in Proposition 4.

Proposition 4 *The principal's profit under the optimal AI-contingent contract is given by*

$$\underbrace{\frac{(1+\alpha-2c)^2+(1-\alpha)^2(2q-1)^2}{8\alpha}}_{\text{No-agency-conflict Profit}}-\underbrace{\frac{b}{4\alpha}\left[8c-2(1-\alpha)-13b-\frac{2\varphi}{b}\left(\frac{(c-3b)^2}{1-4\varphi}-\frac{(c-4b)^2}{1+4\varphi}\right)\right]}_{\text{Loss from Suboptimal Decisions and Wages}}. \tag{17}$$

This profit level exceeds the profit without AI, as stated in Lemma 1, for every possible value of the private benefit b. Moreover, the difference in profits between the AI-contingent contract and the no-AI scenario is greater than the Efficiency Gain from Information, π_{EG}.

The AI-contingent contract enables efficient utilization of AI-generated information, which leads to the profit approaching the level achieved under the no-agency-conflict conditions, as shown by the first expression in Equation (17). In this equilibrium, the total costs from suboptimal investment decisions and managerial wages, represented by the second expression in Equation (17), are lower than $qb/(1-\alpha)(2q-1)$, the managerial compensation cost in scenarios without AI integration. As a result, the AI-contingent contract achieves a profit gain that exceeds the incremental profit derived from AI adoption in the absence of agency conflicts, π_{EG}.

In summary, the AI-contingent contract not only enhances the efficient use of AI information but also significantly reduces costs associated with suboptimal investment decisions and managerial

compensation. These outcomes underscore the efficacy of AI-contingent contracts in resolving agency conflicts and improving overall financial performance.

4.2 Information Design Approach

Another possible approach to address this issue involves limiting the scope of information AI conveys to the manager. Instead of reporting the exact value of probability component p_1, the AI can provide the manager with an imperfect signal of p_1, denoted as S_1. This signal could be a binary indicator that informs the manager whether p_1 exceeds a certain threshold, or a more detailed signal that identifies a precise interval in which p_1 lies, with the interval endpoints related to characteristic X_1. The manager combines X_1 with signal S_1 to form an expectation of p_1, and uses X_2 to form an expectation of p_2. One objective of this design becomes to reduce the likelihood that the manager's expectation, $\alpha\mathbb{E}[p_1|X_1, S_1] + (1-\alpha)\mathbb{E}[p_2|X_2]$, falls within an intermediate range where suboptimal decisions could be taken.

However, this approach may encounter several challenges in real-world applications. First, the manager may struggle to assess how much information in X_1 overlaps with the AI-provided signal S_1, making it difficult to integrate the two pieces of information effectively. For instance, if both X_1 and S_1 suggest that p_1 is high, does this imply that p_1 is sufficiently high to justify proceeding with the project even if $X_2 = L$? Or are these two pieces of information duplicated, meaning the project should still be rejected if the information from group 2 is unfavorable? Similar issues arise when X_1 and S_1 provide conflicting messages.

Furthermore, designing the signals and information structure requires precise knowledge of the manager's belief distribution $f(p_1|X_1)$, which can be challenging for the principal to obtain. Although the principal could learn from the manager's past decisions without AI, these actions only provide information about the conditional expectation $\mathbb{E}[p_1|X_1]$, not the complete distribution, $f(p_1|X_1)$.

In principle, combining information design with the AI-contingent contract could achieve superior results compared to either method alone. However, this approach still faces the same inherent challenges in information design: the principal lacks an accurate distribution to effectively design the signal, while the manager may struggle to integrate the signal with existing information. Careful consideration and adjustments are needed to address these difficulties and optimize its effectiveness.

5 Concluding Remarks

The integration of AI into corporate decision-making processes has undeniably introduced new opportunities for enhancing precision and efficiency. However, this study highlights the critical need to approach AI adoption with a nuanced understanding of potential pitfalls, particularly in the context of agency conflicts. AI-augmented decisions, while capable of improving accuracy, can also lead to unintended backfires when managerial incentives are not aligned with shareholder interests. Our analysis suggests that without carefully designed compensation structures, such as AI-contingent contracts, firms risk falling prey to suboptimal investment decisions driven by private benefits rather than profitability.

The proposed AI-contingent contract framework offers a viable solution, ensuring that AI's predictive power is harnessed effectively while mitigating the risks posed by agency conflicts. This approach not only optimizes decision-making but also safeguards corporate profitability by aligning managerial actions with AI-driven insights. As AI technology continues to evolve and become more integrated into corporate governance, it is essential for firms to develop robust mechanisms that balance technological advancement with the complex realities of human incentives.

Ultimately, the success of AI in corporate environments depends not just on the quality of the algorithms but also on the governance structures in place to manage its application. By

addressing these challenges, firms can fully leverage AI's transformative potential while avoiding the dangers of AI-driven decision-making backfires.

References

Acemoglu, D. (2021). Harms of AI. Technical report, 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. Technical report, National Bureau of Economic Research.

Angelova, V., Dobbie, W. S., and Yang, C. (2023). Algorithmic recommendations and human discretion. Technical report, 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.

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).

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.

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.

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

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

Zhong, H. (2024). Optimal integration: Human, machine, and generative AI.

Appendix

Proof of Lemma 1

When $\overline{\mu} = q$ and $\underline{\mu} = \alpha q + (1-\alpha)(1-q)$, the manager approves the project only if $(X_1, X_2) = (H, H)$ and rejects the project in all other cases. The manager's wages, w_A for approval and w_R for rejection, are determined by Equation (10) and are expressed as follows:

$$w_A = \frac{b}{q-(\alpha q+(1-\alpha)(1-q))} = \frac{b}{(1-q)(2q-1)}\,, \quad w_B = \frac{qb}{q-(\alpha q+(1-\alpha)(1-q))} = \frac{qb}{(1-q)(2q-1)}\,.$$

The principal's expected profit is given by:

$$\begin{aligned}\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] - \mathbb{E}[\mu_M w_A \mathbf{1}_{\{a=A\}}] - \mathbb{E}[w_R \mathbf{1}_{\{a=R\}}] =& \frac{1}{4}(q-c) - \frac{1}{4}qw_A - \frac{3}{4}w_R \\ =& \frac{q-c}{4} - \frac{qb}{(1-\alpha)(2q-1)}\end{aligned}$$

Next, we demonstrate that no other contract achieves this level of profit.

Consider a contract where $w_A > 0$ and $\overline{\mu} > q$, or where $w_A = 0$ and $w_R \geqslant b$. In this case, the manager approves projects with $X_1 = H$ and $X_2 = H$ with a probability of at most $1/2$. According to Equation (4), the principal's expected payoff from the project under this contract is bounded above by:

$$\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] \leqslant \frac{1}{2} \cdot \frac{q-c}{4} < \frac{q-c}{4} - \frac{qb}{(1-\alpha)(2q-1)}\,.$$

Similarly, for a contract where $w_A > 0$ and $\underline{\mu} < \alpha q + (1-\alpha)(1-q)$, or where $w_A = 0$ and $w_R < b$, the manager rejects projects with $X_1 = H$ and $X_2 = L$ with a probability no greater than $1/2$.

$$\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] \leqslant \frac{q-c}{4} - \frac{1}{2} \cdot \frac{\alpha q+(1-\alpha)(1-q)-c}{4} < \frac{q-c}{4} - \frac{qb}{(1-\alpha)(2q-1)}\,.$$

For all remaining contracts, which satisfy $w_A > 0$, $\overline{\mu} \leqslant q$, and $\underline{\mu} \geqslant \alpha q + (1-\alpha)(1-q)$, the principal's expected profit is bounded above by:

$$\mathbb{E}[(\mu_M - c)\mathbf{1}_{\{a=A\}}] - \mathbb{E}[\mu_M w_A \mathbf{1}_{\{a=A\}}] - \mathbb{E}[w_R \mathbf{1}_{\{a=R\}}] = \frac{1}{4}(q-c) - \frac{\overline{\mu} b}{\overline{\mu} - \underline{\mu}} \leqslant \frac{q-c}{4} - \frac{qb}{(1-\alpha)(2q-1)},$$

with equality holding if and only if $\overline{\mu} = q$ and $\underline{\mu} = \alpha q + (1-\alpha)(1-q)$.

Proof of Lemma 2

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

$$\frac{dw_A}{d\overline{\mu}} = -\frac{b}{(\overline{\mu} - \underline{\mu})^2}, \quad \frac{dw_R}{d\overline{\mu}} = -\frac{\underline{\mu} b}{(\overline{\mu} - \underline{\mu})^2}, \quad \frac{dw_A}{d\underline{\mu}} = \frac{b}{(\overline{\mu} - \underline{\mu})^2}, \quad \frac{dw_R}{d\underline{\mu}} = \frac{\overline{\mu} b}{(\overline{\mu} - \underline{\mu})^2}. \tag{A1}$$

Now, differentiating the objective function (14) with respect to $\overline{\mu}$:

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

Using (A1) and the condition $w_R = \overline{\mu} w_A$, and then evaluating at $\overline{\mu} = \overline{\mu}^*$, $\underline{\mu} = \underline{\mu}^*$, we obtain:

$$(\overline{\mu}^* - c) f(\overline{\mu}^*) - \frac{b}{(\overline{\mu}^* - \underline{\mu}^*)^2} \left[2 \int_{\overline{\mu}^*}^{1} \mu_{AM} dF(\mu_{AM}) + \int_{\underline{\mu}^*}^{\overline{\mu}^*} \mu_{AM} dF(\mu_{AM}) + \underline{\mu}^* F(\underline{\mu}^*) + \underline{\mu}^* F(\overline{\mu}^*) \right] = 0. \tag{A2}$$

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

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

Using (A1) and the condition $w_R = \underline{\mu} w_A + b$, and then evaluating at $\overline{\mu} = \overline{\mu}^*$, $\underline{\mu} = \underline{\mu}^*$, we obtain:

$$(c - b - \underline{\mu}^*) f(\underline{\mu}^*) - \frac{b}{(\overline{\mu}^* - \underline{\mu}^*)^2} \left[2 \int_{\overline{\mu}^*}^{1} \mu_{AM} dF(\mu_{AM}) + \int_{\underline{\mu}^*}^{\overline{\mu}^*} \mu_{AM} dF(\mu_{AM}) + \overline{\mu}^* F(\underline{\mu}^*) + \overline{\mu}^* F(\overline{\mu}^*) \right] = 0. \tag{A3}$$

Thus, the equilibrium thresholds $\overline{\mu}^*$ and $\underline{\mu}^*$ satisfy the conditions in (A2) and (A3), and are implicitly defined by this system of equations.

Proof of Corollary 1

Let $g_1(\overline{\mu}^*, \underline{\mu}^*, b)$ represent the left hand side of (A2) and $g_2(\overline{\mu}^*, \underline{\mu}^*, b)$ represent the left hand side of (A3). The derivatives of $g_1(\overline{\mu}^*, \underline{\mu}^*, b)$ and $g_2(\overline{\mu}^*, \underline{\mu}^*, b)$ satisfy:

$$\begin{aligned}
&\frac{\partial g_1}{\partial \overline{\mu}^*} > 0\,, \qquad \frac{\partial g_1}{\partial \underline{\mu}^*} < 0\,, \qquad \frac{\partial g_1}{\partial b} < 0\,, \\
&\frac{\partial g_2}{\partial \overline{\mu}^*} > 0\,, \qquad \frac{\partial g_2}{\partial \underline{\mu}^*} < 0\,, \qquad \frac{\partial g_2}{\partial b} < 0\,, \\
&\frac{\partial g_1}{\partial \overline{\mu}^*}\frac{\partial g_2}{\partial \underline{\mu}^*} - \frac{\partial g_1}{\partial \underline{\mu}^*}\frac{\partial g_2}{\partial \overline{\mu}^*} < 0\,.
\end{aligned} \tag{A4}$$

Applying the implicit function theorem,

$$\begin{bmatrix} \dfrac{d\overline{\mu}^*}{db} \\ \dfrac{d\underline{\mu}^*}{db} \end{bmatrix} = - \begin{bmatrix} \dfrac{\partial g_1}{\partial \overline{\mu}^*} & \dfrac{\partial g_1}{\partial \underline{\mu}^*} \\ \dfrac{\partial g_2}{\partial \overline{\mu}^*} & \dfrac{\partial g_2}{\partial \underline{\mu}^*} \end{bmatrix}^{-1} \begin{bmatrix} \dfrac{\partial g_1}{\partial b} \\ \dfrac{\partial g_2}{\partial b} \end{bmatrix}\,. \tag{A5}$$

Substituting (A4) into (A5), we obtain:

$$\frac{d\overline{\mu}^*}{db} > 0\,, \quad \frac{d\underline{\mu}^*}{db} < 0\,.$$

Proof of Proposition 1

From the law of iterated expectations, we have:

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

The distribution of $\widetilde{\mu}$ is formed by spreading out portions of μ's probability density or probability mass function while leaving the mean unchanged. In other words, the distribution $F(\widetilde{\mu})$ is a mean-preserving spread of $F(\mu)$.

Now, consider the function $\max \mu - c, 0$, which is convex in μ. By Jensen's Inequality,

$$\begin{aligned}\int_c^1 (\widetilde{\mu} - c) dF(\widetilde{\mu}) &= \int_0^1 \max\{\widetilde{\mu} - c, 0\} dF(\widetilde{\mu}) \\ &\geqslant \int_0^1 \max\{\mu - c, 0\} dF(\mu) \\ &= \int_c^1 (\mu - c) dF(\mu)\,.\end{aligned}$$

Substituting in $\widetilde{\mu} = \mu_{AM}$ and $\mu = \mu_M$, we obtain $\pi_{EG} \geqslant 0$.

Proof of Proposition 2

A lower bound for the loss from suboptimal investment decisions is given by:

$$\begin{aligned}\frac{1}{2}\int_{\underline{\mu}}^{c} (c - \mu_{AM}) dF(\mu_{AM}) + \frac{1}{2}\int_c^{\bar{\mu}} (\mu_{AM} - c) dF(\mu_{AM}) &\geqslant \frac{1}{2} \cdot \frac{1}{2\alpha} \cdot \frac{1}{2}[(c - \underline{\mu})^2 + (\bar{\mu} - c)^2] \\ &\geqslant \frac{(\bar{\mu} - \underline{\mu})^2}{16\alpha}\,. \qquad \text{(A6)}\end{aligned}$$

The first inequality holds because the density function $f(\mu_{AM})$ is bounded below by $1/2\alpha$, and the second inequality is a result of the algebraic inequality $x^2 + y^2 \geqslant (x + y)^2/2$.

The expected managerial wage is bounded below by:

$$\begin{aligned}&\left[\int_{\bar{\mu}}^1 \mu_{AM} dF(\mu_{AM}) + \frac{1}{2}\int_{\underline{\mu}}^{\bar{\mu}} \mu_{AM} dF(\mu_{AM})\right] w_A + \left[\frac{1}{2}F(\underline{\mu}) + \frac{1}{2}F(\bar{\mu})\right] w_R \\ >& \left[\int_{\bar{\mu}}^1 \overline{\mu} dF(\mu_{AM}) + \frac{1}{2}\int_{\underline{\mu}}^{\bar{\mu}} \underline{\mu} dF(\mu_{AM})\right] w_A + \left[\frac{1}{2}F(\underline{\mu}) + \frac{1}{2}F(\bar{\mu})\right] \overline{\mu} w_A \\ >& \int_0^1 \frac{\underline{\mu} + \overline{\mu}}{2} dF(\mu_{AM}) \cdot w_A \\ =& \left(\overline{\mu} - \frac{\overline{\mu} - \underline{\mu}}{2}\right) \cdot \frac{b}{\overline{\mu} - \underline{\mu}} \\ >& \left(\frac{c}{\overline{\mu} - \underline{\mu}} - \frac{1}{2}\right) b\,. \qquad \text{(A7)}\end{aligned}$$

Combining (A6) with (A7), we obtain a lower bound for the total cost from suboptimal decision and wage:

$$\begin{aligned}\frac{(\bar{\mu}-\underline{\mu})^2}{16\alpha}+\left(\frac{c}{\bar{\mu}-\underline{\mu}}-\frac{1}{2}\right)b =& \frac{(\bar{\mu}-\underline{\mu})^2}{16\alpha}+\frac{1}{2}\frac{cb}{\bar{\mu}-\underline{\mu}}+\frac{1}{2}\frac{cb}{\bar{\mu}-\underline{\mu}}-\frac{b}{2}\\ \geqslant& 3\cdot\sqrt[3]{\frac{1}{16\alpha}(\bar{\mu}-\underline{\mu})^2\cdot\frac{1}{2}\frac{cb}{\bar{\mu}-\underline{\mu}}\cdot\frac{1}{2}\frac{cb}{\bar{\mu}-\underline{\mu}}}-\frac{b}{2}\\ =&\frac{3}{4}\alpha^{-\frac{1}{3}}c^{\frac{2}{3}}b^{\frac{2}{3}}-\frac{b}{2}\,.\end{aligned}$$

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

$$\frac{3}{4}\alpha^{-\frac{1}{3}}c^{\frac{2}{3}}b^{\frac{2}{3}}-\left(\frac{q}{(1-\alpha)(2q-1)}+\frac{1}{2}\right)b\,. \tag{A8}$$

The maximum of this expression is

$$\frac{c^2}{16\alpha}\left(\frac{q}{(1-\alpha)(2q-1)}+\frac{1}{2}\right)^{-2}\,.$$

Equation (15) has two positive solutions when pi_{EG} is smaller than the above maximum. Let $\underline{b}$ and $\bar{b}>\underline{b}$ denote these two solutions. The expression (A8) has a single peak and stays above π_{EG} throughout the interval $[\underline{b},\bar{b}]$. As a result, the principal's expected profit is lower with AI-augmented decision-making within this range $b\in[\underline{b},\bar{b}]$.

Proof of Proposition 3

For a given value of p_1, the manager's expectation μ_{AM} takes the value of $\alpha p_1+(1-\alpha)q$ if $X_2=H$, and $\alpha p_1+(1-\alpha)(1-q)$ if $X_2=L$. Every possible contract is dominated by one of the three wage schemes described in the proposition.

Consider a contract where $w_A>0$, $\bar{\mu}\leqslant\alpha p_1+(1-\alpha)q$, and $\underline{\mu}\geqslant\alpha p_1+(1-\alpha)(1-q)$. Under this contract, the manager approves if $X_2=H$ and rejects if $X_2=L$. The principal's expected

profit is bounded above by:

$$\begin{aligned}
&\mathbb{E}[(\mu_{AM}-c)\mathbf{1}_{\{a=A\}}|p_1] - \mathbb{E}[\mu_{AM}w_A\mathbf{1}_{\{a=A\}}|p_1] - \mathbb{E}[w_R\mathbf{1}_{\{a=R\}}|p_1] \\
=&\frac{1}{2}\big(\alpha p_1+(1-\alpha)q-c\big) - \frac{\overline{\mu}b}{\overline{\mu}-\underline{\mu}} \\
\leqslant&\frac{1}{2}\big(\alpha p_1+(1-\alpha)q-c\big) - (\alpha p_1+(1-\alpha)q)\varphi\,, \qquad \text{(A9)}
\end{aligned}$$

with equality if and only if $w_A = \varphi$ and $w_R = (\alpha p_1 + (1-\alpha)q)\varphi$.

Similarly, for a contract where $w_A > 0$ and $\underline{\mu} < \alpha p_1 + (1-\alpha)(1-q)$, or where $w_A = 0$ and $w_R < b$, the principal's profit is bounded above by:

$$\frac{1}{2}\big(\alpha p_1+(1-\alpha)q-c\big) + \frac{1}{4}\big(\alpha p_1+(1-\alpha)(1-q)-c\big) + \frac{1}{4}\max\big\{\alpha p_1+(1-\alpha)(1-q)-c, 0\big\} \qquad \text{(A10)}$$

with equality if and only if $w_A = 0$ and $w_R = 0$. Under this contract, the manager only rejects the project when there is no private benefit and $\mu_{AM} < c$.

For a contract where $w_A > 0$ and $\overline{\mu} < \alpha p_1 + (1-\alpha)q$, or where $w_A = 0$ and $w_R \geqslant b$, the principal's profit is bounded above by:

$$\frac{1}{4}\max\big\{\alpha p_1+(1-\alpha)q-c, 0\big\} - \frac{3b}{4} - \frac{b}{4}\cdot\mathbf{1}_{\{\alpha p_1+(1-\alpha)q\leqslant c-b\}} \qquad \text{(A11)}$$

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

Comparing (A9), (A10), and (A11), we obtain the optimal contract stated in Proposition 3. Summarizing the analysis of the manager's decisions, we obtain the manager's action rule outlined in Proposition 3.

Proof of Proposition 4

The principal's profit under the optimal AI-contingent contract is given by

$$
\begin{aligned}
&\mathbb{E}[(\mu_{AM}-c)\mathbf{1}_{\{a=A\}}]-\mathbb{E}[\mu_{AM}w_A\mathbf{1}_{\{a=A\}}]-\mathbb{E}[w_R\mathbf{1}_{\{a=R\}}]\\
=&\frac{1}{2}\int_{(c-3b)/(1-4\varphi)}^{\alpha+(1-\alpha)q}(\mu_{AM}-c)\frac{1}{2\alpha}d\mu_{AM}+\frac{1}{2}\int_{c-b}^{\alpha+(1-\alpha)q}(\mu_{AM}-c)\frac{1}{2\alpha}d\mu_{AM}\\
+&\frac{1}{2}\int_{(c-4b)/(1+4\varphi)}^{\alpha+(1-\alpha)(1-q)}(\mu_{AM}-c)\frac{1}{2\alpha}d\mu_{AM}+\frac{1}{2}\int_{c}^{\alpha+(1-\alpha)(1-q)}(\mu_{AM}-c)\frac{1}{2\alpha}d\mu_{AM}\\
-&\int_{(c-3b)/\alpha(1-4\varphi)-(1-\alpha)q/\alpha}^{(c-4b)/\alpha(1+4\varphi)-(1-\alpha)(1-q)/\alpha}(\alpha p_1+(1-\alpha)q)\varphi\,dp_1-b\left[\frac{c-b}{\alpha}-\frac{1-\alpha}{\alpha}q+\frac{1}{4}\Big(\frac{c-3b}{\alpha(1-4\varphi)}-\frac{c-b}{\alpha}\Big)\right]\\
=&\frac{1}{8\alpha}\left[(1+\alpha-2c)^2+(1-\alpha)^2(2q-1)^2-\left(c-\frac{c-4b}{1+4\varphi}\right)^2-b^2-\left(\frac{c-3b}{1-4\varphi}-c\right)^2\right]\\
-&\frac{\varphi}{2\alpha}\left[\left(\frac{c-4b}{1+4\varphi}+(1-\alpha)(2q-1)\right)^2-\left(\frac{c-3b}{1-4\varphi}\right)^2\right]-\frac{b}{4\alpha}\left((c-b)+\frac{3(c-3b)}{1-4\varphi}-4(1-\alpha)q\right)\\
=&\frac{(1+\alpha-2c)^2+(1-\alpha)^2(2q-1)^2}{8\alpha}-\frac{b}{4\alpha}\left[8c-2(1-\alpha)-13b-\frac{2\varphi}{b}\left(\frac{(c-3b)^2}{1-4\varphi}-\frac{(c-4b)^2}{1+4\varphi}\right)\right].
\end{aligned}
$$

Using $c<q$ and $\alpha\geqslant 1-\alpha$,

$$
\frac{b}{4\alpha}\left[8c-2(1-\alpha)-13b-\frac{2\varphi}{b}\left(\frac{(c-3b)^2}{1-4\varphi}-\frac{(c-4b)^2}{1+4\varphi}\right)\right]<\frac{b}{4\alpha}\cdot 8q\leqslant\frac{qb}{(1-\alpha)(2q-1)},
$$

the total costs from suboptimal investment decisions and managerial wages are lower than the managerial compensation cost without AI integration, $qb/(1-\alpha)(2q-1)$. Consequently, the difference in profits between the AI-contingent contract and the no-AI scenario is greater than π_{EG}:

$$
\begin{aligned}
&\frac{(1+\alpha-2c)^2+(1-\alpha)^2(2q-1)^2}{8\alpha}-\frac{b}{4\alpha}\left[8c-2(1-\alpha)-13b-\frac{2\varphi}{b}\left(\frac{(c-3b)^2}{1-4\varphi}-\frac{(c-4b)^2}{1+4\varphi}\right)\right]\\
-&\left(\frac{q-c}{4}-\frac{qb}{(1-\alpha)(2q-1)}\right)\\
>&\frac{(1+\alpha-2c)^2+(1-\alpha)^2(2q-1)^2}{8\alpha}-\frac{q-c}{4}\\
=&\pi_{EG}\,.
\end{aligned}
$$