Delegation and Blackboxes

Abhishek Ray
Costello College of Business, George Mason University
Email: aray8@gmu.edu

Cesar Martinelli
Department of Economics, George Mason University
Email: cmarti33@gmu.edu

February 11, 2026

Abstract

Artificial intelligence (AI)-aided bureaucratic processes impact delegation costs, leading to complex incentive misalignments that productivity metrics overlook. We consider such delegation problems using a tractable principal-agent model in which biased agents can use 'blackbox' AI technology to execute delegated tasks. That is, the principal sets wages to motivate a biased agent who may use AI to produce higher-quality but error-prone outputs with known error distributions. Through comparative statics across five model specifications (wages contracted on effort or output quality, agent technical expertise, principal's affinity potential for non-delegated self-produced outputs, and variable monitoring costs) we establish four core results: (1) AI-aided task execution can require greater effort than traditional methods; (2) agent's technology awareness rents remain strictly positive and typically decrease as the agent's algorithm aversion decreases, except under hidden effort; (3) agent's AI use creates fundamental regret-return trade-offs where preference alignment are sacrificed for monetary returns; (4) wage dynamics reveal the principal adjusts compensation primarily to account for algorithm aversion likelihood rather than execution capability. These findings address critical policy challenges in bureaucratic AI deployment, where inherently biased humans using biased algorithms compound accountability gaps that frameworks—the EU's binding risk-based governance versus U.S. voluntary sectoral approaches—struggle to reconcile amid organizational capacity disparities and persistent questions of algorithmic accountability.

Acknowledgments: We thank Costello College of Business, George Mason University for their support.

Submitted to *Management Science*
manuscript

Delegation and Blackboxes

Abhishek Ray
Costello College of Business, George Mason University, Fairfax, VA 22033, aray8@gmu.edu
Cesar Martinelli
Department of Economics, George Mason University, cmarti33@gmu.edu

Artificial intelligence (AI)-aided bureaucratic processes impact delegation costs, leading to complex incentive misalignments that productivity metrics overlook. We consider such delegation problems using a tractable principal-agent model in which biased agents can use 'blackbox' AI technology to execute delegated tasks. That is, the principal sets wages to motivate a biased agent who may use AI to produce higher-quality but error-prone outputs with known error distributions. Through comparative statics across five model specifications (wages contracted on effort or output quality, agent technical expertise, principal's affinity potential for non-delegated self-produced outputs, and variable monitoring costs) we establish four core results: (1) AI-aided task execution can require greater effort than traditional methods; (2) agent's technology awareness rents remain strictly positive and typically decrease as the agent's algorithm aversion decreases, except under hidden effort; (3) agent's AI use creates fundamental regret-return trade-offs where preference alignment are sacrificed for monetary returns; (4) wage dynamics reveal the principal adjusts compensation primarily to account for algorithm aversion likelihood rather than execution capability. These findings address critical policy challenges in bureaucratic AI deployment, where inherently biased humans using biased algorithms compound accountability gaps that frameworks—the EU's binding risk-based governance versus U.S. voluntary sectoral approaches—struggle to reconcile amid organizational capacity disparities and persistent questions of algorithmic accountability.

Key words: policy, delegation, agency, contracts, bureaucracy

1. Introduction

The use of artificial intelligence (AI) in organizational decision-making and bureaucratic processes has been broadly regarded as a transformative boost to productivity. However, the conventional wisdom that AI systems (or, technology in general) serve as straightforward substitutes for human labor (see, e.g., Autor et al. 2003) overlooks critical trade-offs in output quality, human behavioral responses, and the complementarity requirements between human and machine decision-making. Recent research reveals that incorporating AI into delegation mechanisms creates complex incentive misalignment, exposes inherent biases in bureaucratic processes, and requires sophisticated human oversight (Balasubramanian et al. 2022, Lawrence et al. 2023). In this paper, we address these problems using a simple, tractable model of delegation based on Alonso and Matouschek (2007, 2008).

Drawing on recent research (e.g., Acemoglu 2024), our model extends empirical insights that delve into AI productivity improvements and its interaction with implementation constraints and limited task scope.

Our analysis focuses on the delegation mechanism in bureaucratic settings. Delegating decision-making authority to humans with access to AI systems introduces agency frictions distinct from those in traditional organizational hierarchies. On the one hand, principals (i.e., politicians) cannot restrict or *punish* an AI tool's algorithmic processes for producing below-par outputs, making screening and disciplining difficult when AI's alignment with policy goals is questionable (Chen et al. 2024, Roehl and Crompvoets 2025). Such misalignments mean that deploying AI does not eliminate delegation costs—it transforms them into problems of algorithm monitoring, verification, and override decision-making. On the other hand, human-AI collaborative decision-making exhibits path dependencies, in which initial algorithm performance shapes subsequent organizational learning. Studies in organizational decision-making demonstrate that when decision-makers incorporate AI without complementary organizational practices (e.g., training programs, feedback mechanisms, and decision authority structures), they fail to achieve algorithm appreciation and superior performance (Shrestha et al. 2019). Thus, quality of outcomes in AI-augmented decisions increasingly depends on organizational context rather than algorithmic precision.

Our work also considers human reluctance to incorporate algorithmic information (*algorithm aversion*) within organizational contexts (Dietvorst et al. 2015). Weitzner (2024) provides a rational microfoundation for this phenomenon: by ignoring high-quality algorithmic forecasts, workers might indicate competence by following or overriding an algorithm. This reputational concern leads to equilibria where workers exhibit algorithm aversion despite knowing that the AI performs better, thereby reducing the organizational value of AI investments. Further, Horowitz and Kahn (2024) identifies a Dunning-Kruger effect in AI adoption, where individuals with the lowest AI experience show algorithm aversion, intermediate experience leads to automation bias (over-reliance), and only the highest levels of expertise produce appropriate integration of human and machine judgments. This non-monotonic relationship implies that merely providing AI tools without developing user sophistication can reduce rather than enhance decision quality (also see, e.g., Brynjolfsson et al. 2025).

More recently, the literature on human-AI complementarity reveals that replacing human decision-making with AI algorithms can impoverish organizational learning by eliminating diverse perspectives and trial-and-error experimentation (Balasubramanian et al. 2022). When AI systems optimize based on historical patterns, they systematically exclude alternative approaches and squeeze decision-making into narrow statistical frameworks (i.e., overfitting). This reduction in cognitive diversity represents a hidden cost of delegation to agents using AI that standard productivity metrics fail to

capture. Both (human) bureaucrats and AI systems exhibit biases, but their interactions within delegation mechanisms create particularly complex equity challenges. For instance, minority citizens may prefer AI-driven policy-making (expecting increased fairness) if they perceive representative officials as biased against them. Still, delegating to algorithms might perpetuate or amplify discrimination if these systems are trained on historical data that reflects those biases (Gaozhao et al. 2024).

Table 1 Real-World Examples of Political Delegation to Bureaucracies with Potential AI Use

Delegating Authority	Bureaucratic Agency	Delegated Task	AI Usage Status
U.S. Congress	USCIS (U.S. Citizenship and Immigration Services)	Visa application processing and document verification	AI-powered Evidence Classifier automates document categorization; ImmigrationOS aggregates data for risk profiling (Kalmykov 2026)
State Legislatures	State Welfare Agencies	SNAP (food stamps) eligibility determination and fraud detection	AI and automated data mining systems used; Michigan system flagged 48,000 fraud cases (93% false positives) (Wykstra 2020)
Wisconsin Legislature	Wisconsin Department of Corrections and Courts	Criminal sentencing risk assessment	COMPAS algorithm used to assess recidivism risk; used in sentencing Eric Loomis to six years in prison (2013) (Cox 2017)
Virginia Legislature (HB 6001)	Virginia DMV	Customer service and day-to-day operations	AI integration funded to improve customer experience (Hunton Andrews Kurth LLP 2025)
U.S. Congress	Social Security Administration	Social Security benefit enrollment and eligibility	Executive orders directed removal of in-person interviews; AI increasingly used for digitization (U.S. Department of Health and Human Services 2024)

Consequently, we use our analysis approach to argue that conventional AI-incorporation narratives in delegation often overlook the actual problem: *inherently biased humans using AI to improve outputs when executing delegated tasks*. Research on AI's labor-market impact emphasizes that workers' productivity gains depend on creating new tasks and on developing complementarities between human judgment and machine capabilities (Acemoglu and Restrepo 2019b,a). Organizations face significant barriers to developing these complementarities, including the need for AI product managers fluent in both business and technology, embedded cross-functional teams, and accountable to the senior leadership (Berruti et al. 2025). The bureaucratic context poses particular challenges for developing human-AI delegation mechanisms. Public sector organizations often lack resources for continuous technical training. Further, bureaucracies are typically risk-averse, which discourages experimentation with new technologies, and face political constraints on algorithmic decision-making

that private firms do not encounter (Lawrence et al. 2023). These institutional factors create persistent gaps between AI's technical capabilities and bureaucracies' organizational capacity to deploy those capabilities effectively.

Moreover, the assumption that AI reduces skill requirements by automating complex tasks is contradicted by evidence. George et al. (2024) argues that over-reliance on automation leads to skill erosion and loss of critical thinking capacity—a particular concern in bureaucratic settings where institutional knowledge and contextual judgment prove essential for handling edge cases and unforeseen circumstances. That is, partial automation weakens human capabilities and creates a new dependency on technology that lacks intuition, empathy, and contextual awareness. Thus, quality trade-offs emerge from delegating tasks and are amplified by human behavioral responses, including incentive misalignment, algorithm aversion, and innate biases. Bias persists and can potentially amplify through AI systems, while debiasing capacity depends on organizational conditions that are frequently absent in bureaucratic contexts. Consequently, effective AI use in executing delegated tasks requires sophisticated human skills in algorithm oversight, collaborative decision-making, and contextual judgment—skills unevenly distributed across and within most bureaucratic organizations.

We consider these problems using an extended-form game that models a principal-agent relationship in which the principal must decide whether to delegate a task to a (*biased*) agent who may employ "blackbox" AI technology to produce output. Extending Prendergast (2003), Delfgaauw and Dur (2007), the principal sets wages as an incentive to mitigate any misalignment and shirking; yet the agent can choose to accept or reject proposed wages. Following Ke et al. (2024), Ludwig et al. (2024), who model AI as a 'black box', the output from AI systems, while higher in quality than non-AI output, still suffers from random errors. The principal and agent know the error distribution. However, the principal does not know if the agent is algorithm-averse or inclined to use AI for the task. The only tool the principal can use to incentivize efforts to achieve expected task quality is a wage, contracted on effort (base model follows efficiency wage theory (Jirjahn 2006)) or on output quality (moral hazard, Section 5.1). The model captures key tensions in modern organizational and bureaucratic settings: the trade-off between control and delegation, information asymmetries regarding agent behavior and technology use (i.e., rent-seeking vs. rent-extraction per Definitions 1 and 2), and the reliability-efficiency trade-off inherent in AI-aided effort provision.

Our analysis reveals several important insights. First, we show that the effort required for AI-aided task execution may exceed that for traditional (non-AI) methods, challenging conventional assumptions about technology-aided productivity improvements. Second, technology awareness rents—the informational advantage enjoyed by agents who understand AI capabilities and their use—are strictly positive across all model specifications. Yet, these rents increase with the principal's target output quality but decrease as the agent's rent-seeking increases (i.e., the agent's biases diverge from the

principal's preferences). Further, given the agent's technical expertise (Section 5.2) and principal *affinity potential* (Section 5.3) extensions, technology awareness rents decrease monotonically as the agent's algorithm aversion decreases; however, under hidden effort (Section 5.1), impact on such rents becomes non-monotonic and depends crucially on the agent's marginal effort costs and target output quality; notably, the agent's technical expertise does not affect these rents. Still, the principal's affinity potential—the principal's intrinsic valuation of self-produced output—amplifies these rents by increasing the principal's stake in task outcomes. Finally, under variable monitoring costs, an increase in these costs further reduces technology awareness rents, as monitoring frictions further highlight the agent's informational advantage.

Third, our analysis reveals nuanced outcomes in output quality and agent payoffs. In the base model, AI use uniformly increases output quality, with quality gaps expanding as principal's target rises but contracting as agent-principal bias misalignment increases; the agent experiences higher monetary returns from AI use but simultaneously incurs greater regret costs from compromising their innate biases, creating a fundamental trade-off between financial gains and preference alignment. Under hidden effort, traditional methods can sometimes match or exceed AI-aided quality; the relationship between quality gaps and target outputs becomes non-monotonic; and, while the regret-return trade-off persists, increasing agent expertise may attenuate regret costs and partially compensate for bias-related disutility. Under the principal's affinity for self-produced outputs, traditional output quality may again rival AI-aided results; quality gaps respond positively to target increases but negatively to both bias misalignment and principal affinity strength; and the persistent regret-return trade-off remains unmitigated. Finally, under variable monitoring costs, our core regret-returns insights hold: traditional methods can produce superior-quality outcomes, and quality gaps remain monotonically increase with target outputs and decrease with bias misalignment.

Finally, wage dynamics across our base model and extensions reveal how principals use compensation to incentivize delegated task execution amid uncertainty about AI use and agent biases. In the base model and under variable monitoring costs, wages increase with target levels, reflecting heightened suspicion of AI use and demands for quality assurance, but decrease as agent-principal bias misalignment increases, due to lower output quality and principals' weakening belief in the agent's algorithm aversion. Under hidden effort with fixed base pay, variable compensation exhibits parallel dynamics: decreasing with agent bias, increasing with target outputs and suspected AI use, and declining as algorithm aversion beliefs weaken. The agent's technical expertise extension preserves base model insights. In contrast, expertise levels exert no direct influence on equilibrium wages, indicating that compensation adjusts primarily to output expectations and the likelihood of algorithm aversion rather than to execution capability. When the principal exhibits affinity potential for self-produced outputs, base model insights persist with the additional result that stronger principal

affinity increases agent wages, as principals invest more heavily in securing high-quality delegated outputs that align with their intrinsic preferences.

Drawing on these insights, we seek to influence AI policy in bureaucracies. Fundamentally, when biased human agents use AI tools to perform assigned administrative tasks in government agencies, AI systems trained on historically biased data perpetuate and automate inequality by interacting with human decision-makers who exhibit processing biases, algorithm aversion, and misaligned preferences. Worryingly, accountability frameworks have not evolved to keep pace with the speed and scale of such partially automated decisions. The EU's AI Act mandates risk-based governance with human rights and transparency, overseen by the EU AI Office. The U.S. policy prioritizes voluntary frameworks and sectoral flexibility, raising concerns about fragmentation. Organizations deploying AI without verification capabilities often fail, with 63% lacking governance policies despite unequal work delegation. The OECD recommends enablers, guardrails, and engagement methods for responsible AI use. Practical implementation faces many challenges, including administrative capacity disparities across jurisdictions, technological outsourcing that allows public organizations to avoid accountability for discriminatory outcomes, and citizens' perception of reduced organizational control over algorithmic systems, even as courts and regulators hold humans accountable for AI-aided work. Our work aims to address normative questions arising from these challenges.

The rest of the paper is organized as follows. In Section 2, we discuss how our work adds to the growing body of work in blackbox mechanisms and human incentives for adoption of such mechanisms. Then in Section 3 we illustrate our model, the relevant definitions needed to establish analysis of technology awareness rents, bias, algorithm aversion, and associated extensions. In Section 4, we conduct analysis on the base model with efficiency wage implications since efforts are considered contracted. Subsequently, in Section 5, we discuss four extensions: hidden effort, agent's technology expertise, principal's affinity potential, and variable monitoring costs. We bring all the threads of our analysis together in Section 6, where we discuss the limitations of our approach and how it can be extended in the future.

2. Literature Review

The use of artificial intelligence within organizational hierarchies, especially bureaucracies, raises fundamental questions about delegation, contracting, and accountability that existing literature has only partially addressed. The principal-agent framework, long the workhorse of organizational economics (Holmström 1979, Grossman and Hart 1983), has been extended to settings involving AI-aided decision-making. Yet, critical gaps remain regarding how biased human agents using blackbox AI tools alter incentive structures and wage dynamics. Therefore, in this review, we situate our work's contributions within these gaps.

Acemoglu (2024) provides a task-based macroeconomic framework arguing that AI's aggregate productivity effects will be modest over the next decade, estimating total factor productivity gains of roughly 0.07 percentage points annually. Earlier, Acemoglu and Restrepo (2018) formalized the displacement–reinstatement tension, showing that automation reduces labor's task share while new task creation can counterbalance this. More recently, Brynjolfsson et al. (2025) and Dell'Acqua et al. (2023) report micro-level productivity gains of 14–40 percent in specific occupations. However, these studies uniformly treat productivity as the primary metric, overlooking the delegation costs and incentive misalignments that arise when biased agents deploy error-prone AI tools. Our insight that AI-aided task execution can actually require *greater* effort than traditional methods challenges the prevailing assumption in this literature that AI unambiguously reduces worker effort requirements.

Athey et al. (2020) model the allocation of decision authority between humans and AI, showing that a principal may prefer a biased AI to preserve agent effort incentives—a result that resonates with, but does not fully capture, the technology awareness rents identified in the abstract under review. Bockstedt and Buckman (2026) demonstrates that loss aversion shapes delegation to AI assistants, while Dargnies et al. (2026) shows that overconfidence and transparency moderate algorithm aversion in hiring. Greiner et al. (2026) further shows that incentive framing and stakes systematically alter reliance on algorithmic advice. Critically, these papers examine whether agents *use* AI, rather than how the *possibility* of AI use reshapes wage-setting and rent extraction. The finding that technology awareness rents remain strictly positive yet typically decrease with AI use likelihood — except under hidden effort or variable monitoring costs — fills a notable gap between the algorithm aversion and contract theory literature.

The moral hazard literature has long examined effort–compensation tradeoffs (see the review by Georgiadis 2022), including settings with adverse selection and limited liability (Martimort et al. 2025). Carroll (2015) establishes that linear contracts are optimal under robust contracting when the principal has limited knowledge of the agent's action space. Still, this framework does not incorporate AI-aided task execution. Further, these models do not incorporate the distinctive feature that an agent's AI use creates a *regret-return tradeoff* where preference alignment is sacrificed for monetary returns. Köbis et al. (2025) shows experimentally that delegation to AI increases dishonest behavior, with large language models far more likely than human agents to comply with unethical instructions. This empirical finding complements our theoretical insight that principals adjust compensation primarily to account for algorithm-aversion likelihood rather than execution capability—a wage-dynamics mechanism absent from existing contract-theoretic treatments of AI adoption.

On the regulatory front, the EU AI Act (European Parliament and Council of the European Union 2024), which entered into force in August 2024 with a phased compliance timeline through 2027,

establishes a binding risk-based governance architecture requiring conformity assessments for high-risk AI systems. Mahmutovic (2025) analyzes the Act's proactive tools—regulatory sandboxes and data governance mechanisms—while noting the persistent Collingridge dilemma whereby early intervention faces uncertainty and later action confronts entrenchment. The U.S. approach, by contrast, remains voluntary and sector-specific (Congressional Budget Office 2024). Jarrahi and Ritala (2023) applies the principal-agent lens to AI agents more broadly, emphasizing that AI systems' access to vast data creates information asymmetries that existing oversight mechanisms struggle to address. Borch (2022) similarly documents how machine-learning-based automated trading introduces novel knowledge risks and principal-agent problems that human-defined rule systems did not exhibit. The abstract's emphasis on how biased humans using biased algorithms compound accountability gaps speaks directly to what these frameworks struggle to reconcile: organizational capacity disparities mean that the same regulation imposes vastly different practical burdens, a dimension that neither the EU's horizontal approach nor the U.S. sectoral model adequately addresses at the level of individual delegation relationships within bureaucracies.

In essence, while the principal-agent, task-based automation, algorithm aversion, and AI governance literature have each advanced substantially in recent years, they remain largely siloed. Our work bridges these streams by embedding blackbox AI use within a tractable delegation model, yielding comparative statics on effort, rents, regret-return tradeoffs, and wage dynamics that the existing literature has not jointly characterized. Our proposed contributions are particularly timely as policymakers and organizations confront the practical challenge of designing contracts and regulatory frameworks for settings where inherently biased human agents wield increasingly powerful but opaque algorithmic tools.

3. Model

We present a model of delegation in which the agent has personal biases and can choose between two production technologies: a traditional 'non blackbox' method that produces deterministic output quality increasing in effort, and an opaque 'blackbox' method that yields stochastic output, of higher expected quality but at cheaper effort cost. The principal decides whether to delegate the task or perform it directly. On delegation to an agent, the principal designs a contract specifying a minimum quality threshold, a wage schedule, and monitoring costs. The agent, after observing the contract, decides whether to accept and, on acceptance, privately decides whether to use the 'blackbox' (AI) technology and chooses effort levels in either case. Effort costs are quadratic and convex, with AI-aided effort strictly cheaper, creating a shirking (observable effort) or moral hazard (hidden effort) problem where the agent is tempted to substitute cheap, unreliable technology for costly but reliable effort.

The base model considers shirking using efficiency wages and distinguishes between two forms of surplus capture (*rent extraction*) by the agent. Delegation involves information asymmetry regarding

the agent's awareness of and use of AI tools; thus, the agent extracts rent because the principal, unaware of the agent's tool use but aware of the potential for high-quality output, must pay the agent enough to participate. Rent seeking, by contrast, involves costly agent activity that increases captured surplus without creating value, either by increasing misalignment with the principal's objectives or by investing in technology expertise (e.g., AI tool training) that shifts the agent's equilibrium effort. The principal's optimization problem thus involves setting quality expectations and compensation to minimize these information rents while ensuring the agent accepts the contract and produces adequate output, capturing the fundamental tension between providing incentives and insuring against the agent's private use of opaque technology. In the following sections, we formalize the model using functional and parametric assumptions for tractability. To ensure formal exposition, we use 'blackbox' (b) and 'non-blackbox' (nb) as symbols for the agent's AI use and non-AI use when executing delegated tasks.

3.1. Player Payoffs

Consider a two-player delegation game between a principal (P), who owns a task and bears ultimate responsibility for output quality, and an agent (A), to whom the task *may be* delegated. The agent possesses inherent biases ϕ that are misaligned with the principal and, on delegation, makes unobservable technology choices that determine how output is produced. The output is associated with a random variable θ, and while the agent is unaware of θ, the principal knows its distribution. This informational asymmetry, combined with the agent's hidden action in selecting a production technology, gives rise to the central shirking and moral hazard problems that our framework considers. Following Guo and Shmaya (2023), Hajiaghayi et al. (2024), the principal's payoff takes the form

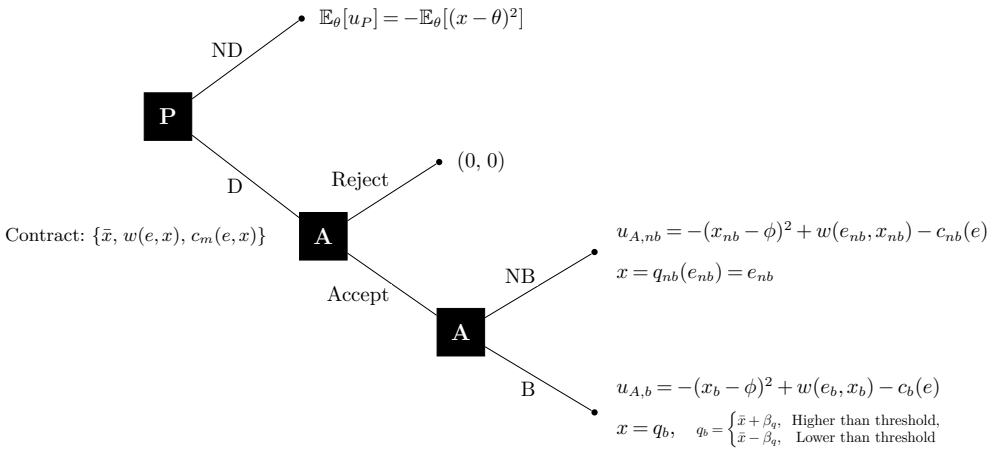


Figure 1 Basic game for delegation with blackbox (i.e., AI systems)

of a *regret-minimization objective*. Under delegation (D), the principal expects an output of quality $\mathbb{E}[\theta] = \bar{x}$ and the expected payoff is given by

$$u_{P,D} = -\mathbb{E}_\lambda\big[\underbrace{(x_i - \bar{x})^2}_{\text{Delegated output regret}} + \underbrace{w(e_i, x_i) + c_m(e_i, x_i)}_{\text{Delegation costs}}\big],$$

where $\theta \sim \mathcal{N}(\bar{x}, \sqrt{\bar{x}}/4)$ denotes the desired quality threshold, $w(\cdot)$ is the wage paid to the agent, and $c_m(\cdot)$ represents monitoring costs borne by the principal. Under non-delegation (ND), the principal performs the task directly and receives expected payoff $u_{P,ND} = -\mathbb{E}_\theta[(x-\theta)^2]$. The agent's utility reflects trading off regret minimization with monetary (wages) returns from exerting efforts as follows:

$$u_A = \underbrace{w(e_i, x) - c_i(e)}_{\text{Effort returns}} - \underbrace{(x_i - \theta)^2}_{\text{Regret}}$$

where $i \in \{b, nb\}$ denotes the technology choice (blackbox or non-blackbox).

Output quality $q_i, i \in \{b, nb\}$ depends critically on the agent's technology choice and efforts. Under nb, output is deterministic and given by $x_{nb} = q_{nb}(e_{nb})$, which is increasing in effort. Under b, output is stochastic: $x_b = q_b$, where q_b is a random variable such that output might surpass expectations (i.e., $\bar{x} + \epsilon$) or fail expectations ($\bar{x} - \epsilon$) with probabilities $\gamma, 1-\gamma$, respectively. Thus, option b offers the agent lower effort costs, since $c_b(e_i) < c_{nb}(e_i)$, but at the expense of unreliable quality. This cost differential creates the fundamental temptation for the agent to substitute cheap, opaque technology for costly but reliable effort.

For analytical tractability, we impose standard functional form assumptions (similar to, e.g., Fanti and Meccheri 2017). The nb output quality is affine in effort, so that $q_{nb} = e_{nb}$. Effort costs are twice differentiable and convex, taking the quadratic form $c_i = \beta_{ci}\, e_i^2/2$, with $\beta_{cb} < \beta_{cnb}$ and $\beta_{ci} \in \mathbb{R}_+$. These linear-quadratic specifications are chosen because they yield closed-form solutions that permit precise comparative statics and clear economic interpretation. When the agent's utility is linear in income and the cost of effort is quadratic, the model delivers tractable expressions in which the agent cares only about the mean and variance of income, not higher moments. This property allows the principal's problem to be expressed transparently as a trade-off between incentive provision and the agent's rent extraction. Moreover, linear contracts are not merely a theoretical convenience but are also empirically relevant, as most real-world incentive schemes—including sales commissions, bonus structures, and performance pay—are approximately linear in form.

The game proceeds in three stages (see Figure 1). In Stage 1, the principal decides whether to delegate the task or to perform it directly. Non-delegation (ND) yields the expected payoff $u_{P,ND} = -\mathbb{E}_\theta[(x-\theta)^2]$, and the game ends. The delegation (D) decision trades off the principal's direct control, constrained by limited information about the agent's blackbox use, against the opportunity of a likely higher quality output. If the principal delegates, the game proceeds to Stage 2, in which the principal

designs and announces a contract specifying three components: a minimum quality threshold $\bar{x} \geq \mathbb{E}[\theta]$ that ensures non-trivial performance expectations; a wage schedule $w(e_i, x)$ that compensates the agent as a function of observed effort; and a monitoring technology with associated cost $c_m(e_i, x)$ that the principal selects to maximize the payoff from delegation.

Upon observing the contract, the agent decides whether to accept or reject based on standard individual rationality. Rejection yields the reservation utility of zero for both parties. The principal believes the agent uses blackbox technology with probability λ implying that the individually rational decision for the agent is to accept if and only if $\mathbb{E}_\lambda[u_{A,i}(\text{effort return})] \geq 0$ for $i \in \{b, nb\}$. If the agent accepts, the game moves to Stage 3, in which the agent privately chooses to use a blackbox (b) or no blackbox (nb) and accordingly, chooses an effort level $e_i \in \mathbb{R}_+$; output is subsequently realized.

3.2. Rent Extraction vs. Rent Seeking

Through our model, we analyze rent extraction and rent seeking as key trade-offs when delegating to agents. Rent extraction refers to the surplus (rent) enjoyed by the agent due to their awareness of and potential use of the blackbox technology, which enables the production of high-quality output without incurring the full cost of the effort required to generate that surplus. Formally, the agent's technology awareness rent under delegation D is defined as the difference between the agent's equilibrium payoff and the payoff the agent would receive under a counterfactual in which the principal could dictate output based only on the prior distribution of the state θ:

Definition 1 **(Technology Awareness Rent)**. *The agent's* technology awareness rent *under delegation D is the difference between the agent's equilibrium payoff and the agent's payoff under a counterfactual in which the principal can dictate x based only on the prior θ:*

$$\mathit{Rent}(D) = \mathbb{E}_\lambda \left[u_A(x_i, \phi, e_i) - u_A(\mathbb{E}[\theta], \phi, e_i) \right]. \tag{1}$$

In other words, the rent *extracted* by the agent represents the surplus in payoff when the principal permits the agent to execute the task with complete discretion over utilizing or not utilizing (with some probability λ) a technology which is essentially blackbox, compared to the scenario in which the principal prescribes the specific output to be generated, which the principal itself knows to be $\bar{x} = \mathbb{E}[\theta]$. From the principal's perspective, these rents represent the difference between the principal's payoff under delegation and the payoff the principal could achieve without delegation, with complete information. The principal's regret minimization objective, when considered as a optimization problem with decision variable $w(e_i, x)$, represents this trade-off.

In contrast, rent seeking refers to costly effort or activity by the agent aimed at increasing captured surplus without creating additional total value. Such activities are welfare-reducing, both through their direct costs and through any additional distortion they induce in the equilibrium wages. In

this model, rent seeking takes two forms: the agent may increase its misalignment (or bias) ϕ with the principal's objectives, or the agent may invest in improving technology use skills or expertise. Formally, if $e_i^* \geq 0$ for $i \in \{b, nb\}$ denote equilibrium effort, then the following defines rent-seeking:

DEFINITION 2 **(Rent-Seeking Activity)**. *Let $e_i^* \geq 0, i \in \{b, nb\}$ denote equilibrium effort. When the agent pursues rent-seeking activity, either the agent's bias ϕ increases; or the agent's technology expertise k increases, thus changing e_i^*.*

Essentially, rent-seeking occurs when either the agent's preference bias ϕ increasingly diverges from the principal's or the agent's expertise in using technology improves (Section 5.2), making it incumbent on the principal to raise wages to incentivize the agent. The key distinction between these two is as follows: *Rent extraction* refers to the surplus left to the agent by the principal's optimal delegation design, which compensates for the agent's information advantage. This is a structural feature of constrained delegation that persists even in the absence of active rent seeking. *Rent seeking*, by contrast, denotes additional surplus pursued through costly (to the agent) effort, representing pure waste from the principal's perspective. Equilibrium rent seeking exists because the marginal return to effort in manipulating bias or improving expertise exceeds the agent's cost. The overall model parameters are in Table 2. We now conduct an equilibrium analysis of the base model in which efforts are contracted. Subsequently, we present extensions to this base model in Section 5.

Table 2 Model Parameters

Parameter	Description
θ	State of the world; distribution known to Principal
ϕ	Agent's bias (misalignment) with Principal's objectives
x	Output produced
$\bar{x}$	Minimum quality threshold set by Principal ($\bar{x} \geq \mathbb{E}[\theta]$)
e_i	Effort exerted by Agent, $i \in \{b, nb\}$
β_b, β_{ci}	Marginal returns (β_b) and marginal cost (β_c) of effort exerted by agent,
$w(e_i, x)$	Wage paid to Agent as function of effort and output
$c_m(e_i, x)$	Monitoring cost incurred by Principal
λ	Principal's belief probability that Agent uses blackbox
q_b	Basic quality from blackbox device
q_{nb}	Basic quality without blackbox; $(2-\alpha)q_b \leq q_{nb} \leq \alpha q_b, \alpha \in (1,2)$
$c_b(e_i)$	Effort cost with blackbox; $c_b(e_i) < c_{nb}(e_i)$
$c_{nb}(e_i)$	Effort cost without blackbox

4. Equilibrium Analysis

We begin the analysis by establishing the equilibrium wage and efforts. Then, we conduct comparative statics on three predominant aspects of the model: wages, rent extraction vs. rent seeking, and output analysis. Each extension in Section 5 focuses on these predominant aspects, along with the wage

dynamics. Note that in the explanations that follow, the term 'blackbox', AI or AI tool(s) have been used interchangeably.

LEMMA 1. *Equilibrium wages* $w^* = \frac{\beta_{cb}(1-\lambda)((\beta_{cnb}+2)\bar{x}-(\beta_{cnb}+4)\phi)}{\beta_{cb}(\beta_{cnb}+3)(1-\lambda)+\lambda(\beta_{cnb}+2)^2}$ *with strictly positive efforts exist for the following region, defined by belief in algorithm aversion* $(1-\lambda)$*, expected output quality* $(\bar{x})$*, and innate agent bias* (ϕ)*:*

$$1-\left(\frac{\beta_{cb}((\beta_{cnb}+4)\bar{x}-(3\beta_{cnb}+8)\phi)}{\beta_{cb}(\beta_{cnb}+4)\bar{x}-\beta_{cb}(3\beta_{cnb}+8)\phi+2\beta_{cnb}(\beta_{cnb}+2)\phi}\right) < 1-\lambda < 1, \text{ with either } \bar{x} \geq 3\phi \text{ or } (2\phi < \bar{x}, (3\beta_{cnb}+8)\phi < (\beta_{cnb}+4)\bar{x}).$$

Proofs are in Appendix EC.1. The following are a few important aspects of this equilibrium. First, note that the equilibrium existence is guaranteed only when the principal's belief in the agent's algorithm aversion (i.e., $1-\lambda$) has a lower bound. That is, only when the principal believes there is some likelihood that the agent will prefer their own efforts rather than using AI to produce output, will the principal be ready to offer wages, given the target quality. Moreover, Lemma 1 shows that when principal-agent misalignment is restricted, strategic delegation models guarantee the existence of an equilibrium and the optimality of the delegation mechanism. Alonso and Matouschek (2008) demonstrate that interval delegation becomes optimal when the agent's bias relative to the principal is sufficiently small, showing that modest preference misalignment permits the principal to design delegation sets that improve upon centralized decision-making while accommodating information asymmetries. Similarly, Pond (2021) establishes conditions under which pooling equilibria sustain agency independence, with such equilibria existing only when political bias and status quo agency bias remain at moderate levels; extreme misalignment between politician and citizen preferences eliminates the conditions necessary for equilibrium maintenance. Lemma 1 essentially extends these insights to our context with bounds on ϕ. We now consider the equilibrium wage and associated efforts analysis.

PROPOSITION 1. *Equilibrium effort levels under delegation need not decrease when the agent uses a black box to execute delegated tasks.*

Proof is in Appendix EC.2. Proposition 1 offers the first surprising finding of our analysis. Recent empirical and theoretical economics research provides no evidence that AI agent adoption necessarily increases worker effort; rather, the consensus suggests AI primarily substitutes for human effort. Noy and Zhang (2023) demonstrates experimentally that ChatGPT substitutes for worker effort rather than complementing skills, allowing workers to produce satisfactory output with reduced time and effort expenditure quickly. Theoretical work on AI in team production reveals a moral hazard mechanism where reliable AI agents—which exert effort without incentive problems—can paradoxically reduce human worker effort as humans use AI to complete tasks (Cheng et al. 2025). While empirical studies document substantial productivity gains of more than 25% from AI assistance

(Brynjolfsson et al. 2025), these gains arise from task automation and restructuring rather than effort intensification, with AI enabling higher output through reduced rather than increased human effort provision.

Proposition 1 offers arguments contrary to such findings. To begin with, when an agent's preferences are misaligned with the principal's objectives (i.e., agent's inherent bias ϕ), the introduction of blackbox (or AI) tools can paradoxically induce the agent to exert *more* effort than in the non-AI scenario through a bias-correction mechanism. The biased agent absent AI, would optimally choose some effort to implement outcomes closer to their ideal point rather than the principal's. The use of AI as an aid fundamentally alters this calculus: if the AI system is trained on or constrained toward the principal's preferences, the agent faces a choice between accepting AI-generated outputs aligned with the principal (but misaligned with the agent's bias) or exerting additional effort to override, modify, or supplement the AI's recommendations to steer outcomes toward their preferred position. This creates an *effort-intensification effect* where bias misalignment becomes *costly* to maintain—the agent must now work harder to align their biased preferences with the final output.

This effect is strongest when AI provides a credible, high-quality baseline that the principal can verify, forcing the biased agent to exert substantial effort either to justify deviations from AI recommendations or to subtly manipulate AI inputs and outputs to achieve their desired bias. Thus, counterintuitively, the very presence of preference misalignment combined with principal-aligned AI can transform the agent's optimal strategy from effort avoidance to effort provision, as the cost of maintaining bias increases while the benefit of appearing compliant with AI-augmented processes rises. Similarly, algorithm aversion among agents can lead to greater effort when using AI. The algorithm-averse agent faces disutility from directly adopting AI-generated outputs and must choose between accepting recommendations that conflict with their preference for human judgment or exerting additional effort to validate, verify, override, or justify deviations from algorithmic suggestions. Again, algorithm aversion and organizational productivity pressures can shift the agent's optimal strategy from moderate effort under full discretion to high effort under AI use, as the psychological cost of ceding judgment to algorithms drives compensatory effort to preserve human agency and expertise in decision-making.

We now consider the how equilibrium wages and the associated surplus for the agent change, given parametric assumptions for bias ϕ, algorithm use likelihood λ, and given the rent extraction vs. seeking framework.

Proposition 2. *The agent always enjoys positive technology awareness rents under delegation. Such rents increase when there is an increase in the principal's expected output target $\bar{x}$, or in the principal's belief in the agent's algorithm aversion $(1-\lambda)$, or when the agent's rent-seeking (ϕ) decreases.*

Proof is in Appendix EC.3. Proposition 2 presents insights on how human agents can extract or seek to enjoy positive rents when executing delegated tasks, possibly using AI. Recent research on AI and delegation has largely overlooked how these technologies fundamentally alter rent-seeking and rent-extraction dynamics in principal-agent relationships characterized by human agents with private information asymmetries. Instead, most have considered AI exclusively *as the* agent and have focused primarily on alignment problems rather than the rent-extraction mechanisms that emerge from algorithmic opacity and human interaction. Specifically, Jarrahi and Ritala (2023) note that "AI agents often have access to vast amounts of data and information processing capabilities that their human principals cannot fully interpret," creating an information asymmetry that amplifies the agent's capacity to extract rents. This algorithmic opacity serves as a rent-seeking tool: agents can strategically exploit advanced processing and information capabilities to justify actions that benefit themselves.

However, Proposition 2 analyzes the rents humans enjoy from AI tools that deliver superior results. Essentially, when a principal delegates tasks to a human agent who can use AI tools, the agent extracts positive technology-awareness rents from the principal's inability to observe or replicate the agent's AI usage decisions. For an expected output target $\bar{x}$, where direct monitoring of whether the agent achieves this target through personal effort, AI assistance, or some combination thereof is costly, the agent's exclusive knowledge of how to effectively use AI tools creates an informational advantage: the principal must compensate the agent for the *claimed* effort required to meet $\bar{x}$, yet cannot distinguish between scenarios where the agent exerts high personal effort versus scenarios where AI substantially reduces the agent's true effort cost. This information asymmetry guarantees strictly positive rents for AI-capable agents, as they receive compensation based on the principal's belief about the effort requirements while privately benefiting from AI-enabled reductions in effort. These technology awareness rents increase monotonically in the principal's output target $\bar{x}$: higher targets necessitate greater compensation under the assumption of high quality human-only production. Yet AI tools allow these agents to meet elevated targets at a lower personal cost, widening the gap between the compensation received and the true effort expended.

Such rents are further augmented when the principal holds stronger beliefs about the agent's algorithm aversion $(1-\lambda)$, where λ represents the principal's belief that the agent likely used AI. As $(1-\lambda)$ increases, the principal increasingly attributes output to human effort rather than algorithmic assistance, leading to higher compensation offers designed to induce participation from a *perceived* algorithm-averse agent. Ironically, an agent who secretly embraces AI tools extracts maximal rents by maintaining the appearance of algorithm aversion, collecting premiums intended to compensate for the higher costs of expected output quality, that the agent does not actually bear. Additionally, technology awareness rents exhibit an inverse relationship with the agent's inherent bias ϕ, which

measures the divergence between the agent's ideal outcome and the principal's preferred outcome. When bias ϕ decreases, the principal perceives greater natural alignment and becomes more willing to delegate discretion with less stringent monitoring, creating expanded opportunities for the agent to substitute AI for personal effort without detection. Lower bias thus enables the agent to maintain the principal's trust while privately leveraging AI to reduce effort costs, extracting rents through the combination of perceived alignment and unobserved technological capability.

Together, these comparative statics reveal that technology awareness rents are crucial to agent's participation and increasing in precisely those parameters—output demands $\bar{x}$, perceived algorithm aversion $(1-\lambda)$, and preference alignment (low ϕ)—that principals might expect would reduce agency costs. We now analyze whether positive rents enjoyed by agents imply that AI use is obvious, and whether AI-aided output will always be better than traditional nb output.

Proposition 3. *While effort returns and output quality from using a blackbox are always higher than those from not using one, regret from using a blackbox is also higher. Moreover, this difference in output quality decreases as the agent engages in rent-seeking.*

Proof is in Appendix EC.4. Recent empirical research reveals a paradoxical tension in human use of AI tools: while AI assistance consistently improves objective task performance and output quality, users frequently experience regret or dissatisfaction due to social stigma and misalignment with intended outcomes. Reif et al. (2025) demonstrates through four experiments with over 4,400 participants that AI use carries significant social penalties in professional settings, with workers perceiving AI-assisted colleagues as less competent and less hard-working despite producing higher-quality outputs; consequently, employees strategically conceal their AI usage to avoid reputational damage, trading measurable performance gains for the appearance of independent work. This stigma effect is particularly pronounced when AI adoption becomes observable to evaluators, reducing task accuracy by 2.7 percentage points as workers underutilize AI tools to preserve professional image (Almog 2025). Beyond social costs, users report dissatisfaction when AI outputs diverge from their original intent. For instance, Wu et al. (2025) find that while human-AI collaboration enhances immediate task performance, it simultaneously undermines intrinsic motivation and increases feelings of boredom, leading workers to experience psychological deprivation when AI-generated content does not align with their creative vision or professional judgment.

Proposition 3 builds on these empirical insights and presents an analytical treatment under delegation. In a delegated task scenario where human agents can use AI, a fundamental tension emerges between objective performance gains and subjective agent utility: AI tools improve measurable output quality while simultaneously generating psychological dissatisfaction when algorithmic outputs diverge from the agent's innate bias (e.g., intended approach or creative vision). Consequently, agents

experience regret not from inferior outcomes but from loss of agency and authenticity in the production process: while AI-generated outputs meet or exceed expected benchmarks, they fail to reflect the agent's intended problem-solving strategy or incorporate domain-specific nuances the agent considers professionally important. This dissatisfaction persists even when AI demonstrably enhances productivity, suggesting that agent utility depends not solely on output quality but also on process alignment between AI recommendations and human intent. The phenomenon reveals an overlooked cost of AI delegation: *principals who evaluate agents purely on performance metrics may inadvertently impose psychological costs on agents* who achieve targets through AI assistance that feels misaligned with their expertise, potentially undermining long-term motivation and retention despite short-term productivity gains.

But does this insight ensure that wages will serve as the appropriate tool to incentivize agent efforts and mitigate any regret in the process? The following proposition sheds light on this question.

PROPOSITION 4. *Equilibrium wages w^* increase with the principal's output target $\bar{x}$ and agent's algorithm aversion $(1-\lambda)$, but decrease with the agent's rent-seeking (ϕ).*

Proof is in Appendix EC.5. In delegation scenarios where principals cannot observe agents' AI usage, equilibrium wage contracts exhibit distinct patterns in response to three key parameter changes: output targets, agent bias, and principals' beliefs about AI use. First, wages increase monotonically with the principal's output target $\bar{x}$ due to a compensation mechanism whereby principals must offer higher wages to induce effort provision for demanding tasks, particularly when they suspect agents may be using AI tools that require additional oversight or validation costs. Recent evidence on the "shadow AI economy" reveals that employees in over 90% of organizations secretly use personal AI tools for work tasks (Almog 2025), creating uncertainty about whether high targets are achieved through human effort or algorithmic assistance; this uncertainty forces principals to compensate based on quality output expectations, driving wages upward with target increases. Second, wages decrease with agent bias ϕ as preference misalignment signals lower expected output quality from the principal's perspective. Ilinov et al. (2024) demonstrate that misalignment between principal and agent preferences creates conflicts of interest where agents make suboptimal decisions relative to the principal's objectives, reducing the value of delegation; consequently, principals discount wages for biased agents to reflect anticipated quality deterioration and the need for costly monitoring to prevent preference-driven deviations from desired outputs. The intuition parallels findings from political economy, where leaders trade off competence versus loyalty when selecting appointees (Ilinov et al. 2024), with bias functioning as an inverse loyalty signal that diminishes the agent's value to the principal.

Third, wages increase with the principal's belief that the agent is algorithm-averse (i.e., as λ increases, $1-\lambda$ decreases). Reif et al. (2025) document that AI use carries significant social penalties in professional settings, with workers perceived as less competent when AI assistance is suspected; this stigma enables principals to justify lower wages for agents believed to use AI, as the principal attributes output quality to algorithmic rather than human capability. The wage penalty for suspected AI use reflects a fundamental shift in the principal's effort inference: traditional contracts assume agents exert costly effort that must be compensated, but when principals believe agents delegate cognitive work to AI systems, the marginal product attributed to human skill declines (Brynjolfsson et al. 2025). Moreover, widespread hidden AI adoption—with surveys showing 30-57% of workers concealing their AI use (Silic et al. 2025)—creates adverse selection dynamics where principals who suspect AI usage reduce wage offers to avoid overpaying for algorithmically assisted output, forcing agents to either accept lower compensation or signal their algorithm aversion through costly effort demonstrations. Thus, the overall possibility of substituting AI for human efforts in delegated tasks presents insights that augment recent empirical insights on human-AI collaboration dynamics in organizations.

We now consider extensions to our base model and highlight the results that differ. Any basic insight or comparative statics result from the base model that is not repeated in the extension implies that the base model insight holds qualitatively.

5. Extensions

In this section we relax certain features of the base model and analyze the resulting equilibrium.

5.1. Hidden Effort

We assume effort is observable based on which the principal designs the contract. We relax this assumption and assume that effort is hidden and the principal contracts on the basis of output quality. Without loss of generality, assume the principal announces wages as $w = p_0 + p_r x_i, i \in \{b, nb\}$, where p_0 is the base fixed wage while p_r is the quality sensitive wage rate. The following proposition highlights how an equilibrium with such a pay structure impacts rents and regret.

Proposition 5. *When the agent's efforts are hidden, equilibrium rents may not always rise with the principal's belief in the agent's algorithm aversion* $1-\lambda$*. Further, blackbox use does not necessarily increase the agent's regret.*

Proof is in Appendix EC.6. When effort provision remains unobservable, and agents possess access to black-box algorithmic tools, the base-model intuition that technology awareness rents increase monotonically with the principal's belief in agent algorithm aversion $(1-\lambda)$ does not hold. While higher perceived algorithm aversion $(1-\lambda)$ initially induces principals to ensure better returns—under the

assumption that agents will exert actual effort for quality output (Reif et al. 2025)—this relationship can reverse when the agent's marginal effort costs are low. That is, when agents are genuinely efficient, blackbox or no blackbox, principals may optimally reduce wages and thus effort returns, thereby eliminating the informational advantage that generates rents. This non-monotonicity arises because strong algorithm aversion beliefs coupled with the agent's efficiency attributes trigger a discrete shift in the principal's optimal contract: rather than continuing to increase wages to compensate a supposedly reluctant and maybe inefficient AI user, the principal redesigns the incentives to reduce dependence on the agent's algorithmic use awareness, effectively closing the rent-extraction channel. That is, the principal now considers that the agent can deliver the expected quality regardless of the agent's awareness of AI tools, thereby mitigating any *moral hazard* arising from hidden effort despite technology awareness.

The agent's regret from black-box use exhibits similar context dependence and does not uniformly increase across all delegation environments. Research shows that while AI use can generate social stigma and professional reputation penalties when observable (Reif et al. 2025), hidden AI adoption in unmonitored settings may produce minimal regret if output quality meets or exceeds expectations and no discovery risk materializes. Regret intensity depends critically on the alignment between AI-generated outputs and the agent's professional judgment. Yet, when efforts are hidden, regret can be mitigated even when algorithmic outputs diverge substantially from the agent's innate biases. In that scenario, the agent can either override AI suggestions (negating effort savings) or accept misaligned recommendations (creating authenticity costs). Still, since wages are now dependent on quality (not effort), regret is mitigated for higher-quality outputs.

These results highlight that neither rents nor regret follow simple monotonic relationships with algorithm aversion beliefs or AI adoption in hidden-effort environments. While moderate algorithm aversion beliefs create rent opportunities through compensatory wage premiums, improved efficiency triggers structural contract changes that eliminate those opportunities. Similarly, agent regret is not an inherent property of black-box use but emerges from the interaction between algorithmic outputs, agent biases, and the observability of AI use. But what happens if the agent is not good at using the technology? Do output quality or agent's regret still show the same behavior? We explore these questions and more in the following extension.

5.2. Technology Expertise

We assume when choosing the technology b, the agent has perfect expertise. But this may not always be true, given the rapid evolution of AI tools. So we relax that assumption in this extension. We consider agent's technology expertise impact the output when using the blackbox such that:

$$x_{b,exp} = \bar{x}k + \epsilon,\ \epsilon \sim U[-1,1], \tag{2}$$

where, as k increases, the agent becomes better at using the blacbox and thus, the *expected* output from blackbox improves. The following proposition the only results that are different qualitatively from the base model.

PROPOSITION 6. *Equilibrium output quality is not always higher using blackbox. Moreover, the quality difference does not always widen when the principal increases expected output quality $\bar{x}$. Finally, while the agent's regret from using a blackbox remains higher than when not using one, it decreases as the expertise k increases.*

Proof is in Appendix EC.7. In delegation scenarios where agents possess imperfect expertise in using blackbox, equilibrium output quality under blackbox use need not dominate output quality under purely human production. When agents lack full technological proficiency—unable to effectively prompt, validate, or integrate AI-generated recommendations—the quality differential between AI-augmented and non-AI outputs depends critically on the expertise threshold relative to task complexity. For tasks requiring nuanced judgment or domain-specific adaptation, an agent with limited AI proficiency may produce lower-quality outputs when using the black box than when relying solely on its own capabilities, as algorithmic suggestions introduce errors the agent cannot detect or correct. This contrasts with the standard assumption in recent empirical work that AI adoption uniformly improves performance (Brynjolfsson et al. 2025): such gains materialize only when agents possess sufficient technological literacy to properly deploy and supervise algorithmic tools, a condition frequently violated in practice as organizations adopt AI faster than workers develop corresponding expertise.

Moreover, the relationship between the principal's output target $\bar{x}$ and the quality gap between AI-aided and non-AI produced outputs shows non-monotonicity under imperfect agent expertise. As the principal raises $\bar{x}$, the agent faces increasing pressure to leverage AI assistance to meet elevated standards; however, if the agent's technological proficiency remains fixed below the expertise threshold required for effective AI supervision, higher targets may actually widen the quality deficit rather than narrow it. This occurs because demanding tasks amplify the consequences of insufficient AI use: the agent that cannot adequately verify or refine complex algorithmic outputs will compound errors as task difficulty increases, whereas the same agent working without AI maintains quality standards consistent with its human capabilities. Consequently, the quality gap may increase, decrease, or remain constant with $\bar{x}$ depending on the agent's expertise level (Noy and Zhang 2023).

Despite lower output quality under imperfect expertise, agents still experience higher regret when using black-box systems than when relying on human judgment alone, reflecting psychological costs from ceding control to opaque algorithms and producing outputs misaligned with their professional standards (Wu et al. 2025). However, this regret differential exhibits diminishing marginal returns to

technological expertise: as agents develop proficiency in AI operation—learning to interpret algorithmic recommendations, identify failure modes, and selectively override unsuitable suggestions—they regain a sense of agency and alignment between AI outputs and their intended approach. The regret function thus decreases in expertise, though it remains bounded away from zero due to irreducible opacity costs inherent in black-box systems that prevent full understanding of algorithmic reasoning. These findings suggest that while training investments can partially mitigate the psychological costs of AI adoption, fundamental limitations in algorithmic transparency ensure that agents will continue to experience residual dissatisfaction even at high levels of expertise.

While the agent's behavioral aspects bring some essential nuance to the discussion, it is not clear what happens when the principal has a vested interest (or special interest) in the outcome of the delegation. In this next extension, we consider this aspect of delegation.

5.3. Principal's Affinity Potential

The base model assumes the principal's sole goal in delegating to agent is the delivery of output quality that meets a certain standard. However, apart from meeting standards, the principal may enjoy an '*affinity potential*' from the outcome – utility gain from implementing their own solution versus delegating to an agent. This would represent a preference premium. This might imply that the expected standard is *unobservable* (from agent's perspective) output expectation. We incorporate that into the model and test its implications. Essentially, the principal's utility is now

$$u_{P,D} = -\mathbb{E}_\lambda\big[\underbrace{(x_i - \bar{x} - g)^2}_{\text{Delegated output regret}} + \underbrace{w(e_i, x_i) + c_m(e_i, x_i)}_{\text{Delegation costs}}\big],$$

where now the principal measures output against the expected quality $\bar{x}$ and g – a personal reference point that the agent is not aware of. Again, we mention only those results that are qualitatively different from the base model.

Proposition 7. *At equilibrium, as the principal's affinity potential increases, the agent's rents and wages increase. Yet, the output quality using blackbox is not guaranteed to be higher than without one.*

Proof is in Appendix EC.8. When the principal's affinity potential increases, equilibrium agent rents and wages must rise to sustain delegation as a viable option. Higher affinity potential signals that the principal derives substantial value from their own approach, raising the opportunity cost of delegation and increasing the principal's reservation utility for maintaining control. To induce the principal to delegate despite this elevated outside option, agents must extract higher informational rents and receive greater compensation to make delegation individually rational. This effect intensifies when agents possess private information about task execution, as principals with high affinity potential

face greater uncertainty about whether delegation actually improves outcomes relative to their own judgment, forcing them to offer premium wages to overcome this information asymmetry and secure agent participation.

The positive relationship between affinity potential and equilibrium rents creates a counterintuitive dynamic when agents employ black-box AI systems: while higher wages might suggest enhanced output quality justifying increased compensation, equilibrium output quality using the black box is not guaranteed to exceed quality without its use. This non-monotonicity arises because agents facing principals with high affinity potential optimize not for absolute quality but for demonstrating sufficient value to preserve delegation arrangements. When agents use blackbox AI tools, they may strategically use algorithms that generate outputs visibly distinct from what the principal would produce independently—emphasizing the value of delegation through differentiation rather than pure quality maximization. Moreover, agents with imperfect AI expertise may exploit the principal's high affinity potential to extract rents by producing outputs that merely differ from (rather than dominate) the principal's preferred approach. The principal, unable to observe actual AI usage, must pay rents based on the expected value of delegation given their high opportunity cost, even when realized AI-augmented quality falls short of non-AI benchmarks.

This equilibrium outcome reveals a fundamental tension in AI-aided delegated task execution and affinity potential: agents capture larger rents precisely when principals are most skeptical about delegation's value, yet this rent extraction need not correlate with quality improvements from black-box use. The mechanism operates through information asymmetry amplification: high affinity potential signals that the principal has strong priors about optimal solutions, making them more uncertain about whether the agent's (AI-assisted or not) outputs genuinely improve upon the principal's approach or merely produce some output of uncertain value (Jarrahi and Ritala 2023). Agents exploit this uncertainty by maintaining the credible threat of non-participation unless compensated their technological capability, regardless of whether the black box actually enhances quality. Consequently, equilibrium wages increase with affinity potential to secure delegation, but output quality comparisons between AI and non-AI scenarios remain ambiguous, determined by the agent's true technological proficiency rather than by the wage premium extracted.

Thus, our analysis suggests that principals with high affinity potential face a similar dual agency problem to the base model: they must pay rents to induce delegation while simultaneously bearing the risk that agents' black-box use substitutes observable differentiation for genuine quality improvements, undermining the efficiency gains delegation is intended to achieve.

5.4. Variable Monitoring Costs

The base model assumes monitoring costs are fixed and standard. However, it can be the case that the principal outsources monitoring to a third-party because of which monitoring is converted to an operational expense. We consider this aspect in this extension.

Proposition 8. *When the principal's monitoring costs increase, the agent's rents decrease. Further, an increase in the principal's belief in the agent's algorithm aversion does not guarantee an increase in rents.*

Proof is in Appendix EC.9. When delegating with variable monitoring costs, increases in the principal's monitoring costs paradoxically reduce rather than increase agent rents, reversing the intuition from classical moral hazard models. When monitoring becomes more expensive, the principal optimally reduces wages and shifts toward contract mechanisms that extract greater surplus ex ante by reducing fixed wages or imposing stricter performance-based compensation. This creates a commitment effect: the principal credibly signals through high-powered incentive contracts that she will not engage in costly ex post verification, forcing the agent to internalize more output risk. Consequently, while higher monitoring costs reduce the principal's ability to prevent shirking or hidden AI use, they simultaneously enable the principal to offer contracts that leave agents with lower informational rents by shifting the participation constraint downward. The agent accepts these less favorable terms because the reduced monitoring threat diminishes the value of private information—the agent can no longer extract rents by strategically exploiting the principal's inability to verify effort, as the principal has preemptively designed contracts that are robust to unobservable actions. This mechanism differs fundamentally from settings where monitoring costs are fixed, in which case agents extract standard informational rents proportional to the difficulty of observation.

Considering the principal's belief in agent algorithm aversion and equilibrium rents, we find non-monotonicity due to offsetting strategic effects. While increased belief in algorithm aversion $(1-\lambda)$ initially raises rents—as the principal offers compensation premiums to induce real effort from a perceived reluctant agent—further increases in $(1-\lambda)$ eventually trigger a shift where the principal abandons typical delegation entirely in favor of higher quality expected output contracts. In such scenarios, the principal reduces rents for suspected outputs that would be of much lower than expected quality. Consequently, the rent function in $(1-\lambda)$ is non-monotonic and potentially decreasing beyond a threshold belief level, contradicting the naive prediction that perceived aversion uniformly increases agent bargaining power.

The interaction between monitoring costs and algorithm aversion beliefs further complicates rents, as these parameters jointly determine the principal's optimal contract design and thereby the agent's surplus extraction opportunities. When monitoring costs are high and algorithm aversion beliefs are

moderate, the principal faces a double bind: she cannot verify AI usage due to prohibitive monitoring expenses, yet she believes the agent requires compensation to produce output as close to expectation as possible. However, this combination does not maximize agent rents, as high monitoring costs constrain the principal's willingness to pay aversion premiums—the principal rationally limits total compensation when she cannot verify whether premiums successfully induce quality outputs. Instead, maximal rents occur at intermediate monitoring costs where the principal retains enough verification capability to make aversion premiums credible incentives, while algorithm aversion beliefs remain sufficiently strong that the principal offers substantial premiums.

Thus, our overall analysis as in Section 4 and in the preceding sections corroborates certain basic, core, insights: (1) AI-aided task execution can require greater effort than traditional methods; (2) agent's technology awareness rents remain strictly positive and typically decrease as the agent's algorithm aversion decreases, except under hidden effort; (3) agent's AI use creates fundamental regret-return trade-offs where preference alignment are sacrificed for monetary returns; (4) wage dynamics reveal the principal adjusts compensation primarily to account for algorithm aversion likelihood rather than execution capability.

6. Conclusion & Future Directions

We develop a tractable principal-agent framework to analyze how black-box AI technology fundamentally alters delegation dynamics when biased agents employ algorithmic tools to execute tasks under incomplete contracts. Our analysis yields four principal insights that challenge conventional wisdom regarding AI-augmented productivity. First, we demonstrate that AI adoption does not necessarily reduce agent effort; indeed, algorithmic assistance can paradoxically increase equilibrium effort provision when agents face algorithm aversion or must validate blackbox outputs. Second, we establish that technology awareness rents—surplus extracted by agents who privately understand how to use AI tools—remain strictly positive across most contracting environments, with these rents exhibiting non-monotonic relationships to algorithm aversion that depend critically on whether effort or output quality is contractible. Third, our model reveals an ever present tension between rent extraction and wages: AI-capable agents sacrifice personal preference alignment and experience psychological regret in exchange for informational rents, creating welfare losses that past work has overlooked. Fourth, we show that principals adjust compensation primarily in response to beliefs about agent algorithm aversion rather than technological capability, generating wage dynamics inconsistent with standard skill-based pay models but consistent with emerging empirical evidence on AI stigma in professional settings.

These findings carry significant implications for bureaucratic AI use, where the compounding effects of human bias and algorithmic bias create accountability gaps that current governance frameworks

inadequately address. The divergence between the European Union's binding risk-based approach and the United States' voluntary sectoral model reflects deeper tensions our analysis exposes: when biased agents control black-box systems, neither ex ante regulation (which cannot anticipate all deployment contexts) nor ex post liability (which cannot penetrate algorithmic opacity) sufficiently aligns incentives. Our comparative statics across monitoring cost regimes suggest that effective governance requires mechanisms that make algorithm aversion costly to maintain while simultaneously reducing the informational advantages agents extract from technological opacity. More broadly, our results indicate that productivity gains from AI adoption in delegation contexts may substantially overstate net welfare improvements when accounting for agents' regret costs, principals' increased compensation burdens to manage algorithm aversion, and the erosion of preference alignment that algorithmic intermediation introduces into principal-agent relationships.

Several extensions of our framework merit investigation. First, dynamic models in which agents learn AI proficiency over time would show how technology awareness rents evolve along adoption curves and whether principals can design contracts that accelerate skill acquisition while minimizing rent extraction during the learning phase. Such models could incorporate experience goods characteristics where agents discover their true algorithm aversion only after initial AI usage, creating option value for experimentation that our static framework cannot capture. Second, introducing endogenous AI quality choices—where agents select among algorithmic tools with varying accuracy-interpretability trade-offs—would generate predictions about which black-box systems agents adopt under different contracting regimes and whether principals can induce transparency through mechanism design. This extension connects to the broader literature on verifiable versus non-verifiable information and could explain empirically observed patterns of AI tool selection in organizational settings.

Third, explicitly modeling the principal's belief formation process regarding agent algorithm aversion would clarify how principals update from observed outputs and whether separating equilibria exist in which AI-embracing agents credibly signal their type. Such signaling models could incorporate costly disclosure of AI usage or voluntary transparency mechanisms that agents employ to command wage premiums for technological sophistication rather than suffer penalties from AI stigma. Fourth, extending the framework to multiple agents would introduce tournament incentives and peer effects that fundamentally alter AI adoption decisions: agents may conceal AI usage to maintain relative performance advantages, or conversely, may ostentatiously adopt AI to signal competence if principals value technological progressiveness. Multi-agent settings also raise questions about optimal team composition when principals can mix algorithm-averse and algorithm-embracing agents to achieve complementarity between human judgment and algorithmic efficiency.

Fifth, incorporating multitasking concerns where AI assistance is more effective for some delegated responsibilities than others would generate predictions about task reallocation and whether principals optimally restrict AI usage to subsets of delegated activities. This extension connects to the literature on job design and could inform debates about which bureaucratic functions should permit algorithmic assistance versus require purely human discretion. Sixth, relaxing the assumption of known error distributions for AI outputs to accommodate Knightian uncertainty or ambiguity aversion would better capture real-world concerns about deploying novel AI systems whose failure modes remain poorly understood. Such models could explain why principals may prohibit AI adoption even when expected output quality exceeds human-only production, as robust contract design under algorithmic ambiguity may favor conservative delegation strategies.

Finally, empirical validation of our theoretical predictions represents a critical research priority. Field experiments that randomly assign AI access to agents performing delegated tasks—with variation in contractibility (effort versus output), monitoring intensity, and principal beliefs about agent algorithm aversion—could directly test our comparative statics regarding effort provision, rent extraction, and wage determination. Survey and audit study methodologies could quantify the magnitude of regret-return trade-offs agents experience and measure how algorithm aversion heterogeneity correlates with AI adoption patterns and compensation outcomes. Linking these microeconomic findings to organizational performance and bureaucratic efficiency would illuminate whether aggregate productivity gains from AI justify the distributional and welfare costs our model identifies, informing evidence-based policy regarding algorithmic governance in public and private sector delegation relationships.

References

Acemoglu, D. (2024). The Simple Macroeconomics of AI. Working Paper 32487, National Bureau of Economic Research.

Acemoglu, D. and Restrepo, P. (2018). Artificial intelligence, automation, and work. In *The economics of artificial intelligence: An agenda*, pages 197–236. University of Chicago Press.

Acemoglu, D. and Restrepo, P. (2019a). Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives*, 33(2):3–30.

Acemoglu, D. and Restrepo, P. (2019b). The wrong kind of AI? Artificial intelligence and the future of labour demand. *Cambridge Journal of Regions, Economy and Society*, 13(1):25–35.

Almog, D. (2025). Barriers to ai adoption: Image concerns at work. *arXiv preprint arXiv:2511.18582*.

Alonso, R. and Matouschek, N. (2007). Relational delegation. *The RAND Journal of Economics*, 38(4):1070–1089.

Alonso, R. and Matouschek, N. (2008). Optimal delegation. *The Review of Economic Studies*, 75(1):259–293.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Balasubramanian, N., Ye, Y., and Xu, M. (2022). Substituting human decision-making with machine learning: Implications for organizational learning. *Academy of Management Review*, 47(3):448–465.

Berruti, F., Hämäläinen, L., Cheta, O., and Anant, V. (2025). When can AI make good decisions? The rise of AI corporate citizens. *McKinsey Insights*.

Bockstedt, J. C. and Buckman, J. R. (2026). Humans' use of ai assistance: The effect of loss aversion on willingness to delegate decisions. *Management Science*, 72(1):323–342.

Borch, C. (2022). Machine learning, knowledge risk, and principal-agent problems in automated trading. *Technology in Society*, 68:101852.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942.

Carroll, G. (2015). Robustness and linear contracts. *American Economic Review*, 105(2):536–563.

Chen, E. O., Ghersengorin, A., and Petersen, S. (2024). Imperfect recall and ai delegation. GPI Working Paper 30-2024, Global Priorities Institute, University of Oxford, Oxford, United Kingdom. Global Priorities Institute Working Paper Series. Dated 26 November 2024.

Cheng, X., Dogan, M., and Yildirim, P. (2025). Artificial intelligence in team dynamics: Who gets replaced and why? *arXiv preprint arXiv:2506.12337*.

Congressional Budget Office (2024). Artificial intelligence and its potential effects on the economy and the federal budget. Technical report, Congressional Budget Office, Washington, DC.

Cox, C. (2017). ¡em¿state v. loomis¡/em¿: Wisconsin supreme court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review*, 130(5).

Dargnies, M.-P., Hakimov, R., and Kübler, D. (2026). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*, 72(1):285–301.

Delfgaauw, J. and Dur, R. (2007). Incentives and workers' motivation in the public sector. *The Economic Journal*, 118(525):171–191.

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*, 1(24-013).

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

European Parliament and Council of the European Union (2024). Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). Text with EEA relevance.

Fanti, L. and Meccheri, N. (2017). A note on managerial delegation with asymmetric and convex costs. *Managerial and Decision Economics*, 38(3):279–284.

Gaozhao, D., Wright, J. E., and Gainey, M. K. (2024). Bureaucrat or artificial intelligence: people's preferences and perceptions of government service. *Public Management Review*, 26(6):1498–1525.

George, A. S., Baskar, T., and Srikaanth, P. B. (2024). The erosion of cognitive skills in the technological age: How reliance on technology impacts critical thinking, problem-solving, and creativity. *Partners Universal Innovative Research Publication*, 2(3):147–163.

Georgiadis, G. (2022). Contracting with moral hazard: A review of theory & empirics. *Available at SSRN 4196247*.

Greiner, B., Grünwald, P., Lindner, T., Lintner, G., and Wiernsperger, M. (2026). Incentives, framing, and reliance on algorithmic advice: An experimental study. *Management Science*, 72(1):302–322.

Grossman, S. J. and Hart, O. D. (1983). An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45.

Guo, Y. and Shmaya, E. (2023). Regret-minimizing project choice. *Econometrica*, 91(5):1567–1593.

Hajiaghayi, M., Mahdavi, M., Rezaei, K., and Shin, S. (2024). Regret analysis of repeated delegated choice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9757–9764.

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91.

Horowitz, M. C. and Kahn, L. (2024). Bending the automation bias curve: A study of human and ai-based decision making in national security contexts. *International Studies Quarterly*, 68(2):sqae020.

Hunton Andrews Kurth LLP (2025). Virginia legislature passes AI bill. Hunton Privacy and Cybersecurity Law Blog. Accessed: February 11, 2026.

Ilinov, P., Matveenko, A., Senkov, M., and Starkov, E. (2024). Optimally biased expertise. *CERGE-EI Working Paper Series*.

Jarrahi, M. H. and Ritala, P. (2023). Rethinking ai agents: A principal-agent perspective. *California Management Review*, 65(3):27–47.

Jirjahn, U. (2006). A note on efficiency wage theory and principal–agent theory. *Bulletin of Economic Research*, 58(3):235–252.

Kalmykov, K. (2026). An overview of the government's AI immigration systems for employers. `https://natlawreview.com/article/`

us-governments-expanding-use-ai-immigration-adjudications-considerations-employers. Accessed: February 11, 2026.

Ke, S., Wu, B., and Zhao, C. (2024). Learning from a black box. *Journal of Economic Theory*, 221:105886.

Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., and Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature*, 646(8083):126–134.

Lawrence, C., Cui, I., and Ho, D. (2023). The bureaucratic challenge to ai governance: An empirical assessment of implementation at u.s. federal agencies. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 606–652, New York, NY, USA. Association for Computing Machinery.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2024). The unreasonable effectiveness of algorithms. *AEA Papers and Proceedings*, 114:623–27.

Mahmutovic, A. (2025). The eu ai act: a proactive framework for comprehensive ai regulation. *International Journal of Law and Information Technology*, 33:eaaf028.

Martimort, D., Poudou, J.-C., and Thomas, L. (2025). Optimal contracts under moral hazard, adverse selection and limited liability. Technical report, HAL.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Pond, A. (2021). Biased politicians and independent agencies. *Journal of Theoretical Politics*, 33(3):279–299.

Prendergast, C. (2003). The limits of bureaucratic efficiency. *Journal of Political Economy*, 111(5):929–958.

Reif, J. A., Larrick, R. P., and Soll, J. B. (2025). Evidence of a social evaluation penalty for using ai. *Proceedings of the National Academy of Sciences*, 122(19):e2426766122.

Roehl, U. and Crompvoets, J. (2025). Inside algorithmic bureaucracy: Disentangling automated decision-making and good administration. *Public Policy and Administration*, 40(2):322–350.

Shrestha, Y. R., Ben-Menahem, S. M., and von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4):66–83.

Silic, M., Silic, D., and Kind-Trüller, K. (2025). From shadow it to shadow ai–threats, risks and opportunities for organizations. *Strategic Change*.

U.S. Department of Health and Human Services (2024). Plan for promoting responsible use of artificial intelligence in automated and algorithmic systems by state, local, tribal, and territorial governments in public benefit administration. Technical report, U.S. Department of Health and Human Services. Accessed: February 11, 2026.

Weitzner, G. (2024). Reputational algorithm aversion. *AEA Annual Meeting Proceedings*.

Wu, S., Liu, Y., Ruan, M., Chen, S., and Xie, X.-Y. (2025). Human-generative ai collaboration enhances task performance but undermines human's intrinsic motivation. *Scientific Reports*, 15(1):15105.

Wykstra, S. (2020). Government's use of algorithm serves up false fraud charges. *Undark Magazine*. Accessed: February 11, 2026.