Toward a Bad Job Economy: AI Adoption, Agency Costs, and Job Design*

Matthias Fahn, Jin Li, and Chang Sun†

Faculty of Business and Economics, The University of Hong Kong
HKU Centre for AI, Management and Organization

September 2026

Abstract

We study how AI affects compensation and job design when performance depends on workers' non-contractible effort. In a principal–agent model with limited liability, AI reduces effort costs but more for low than for high effort. This raises the incentive cost of sustaining high effort and can induce firms to replace high-wage, high-effort good jobs with low-wage, low-effort bad jobs, even when good jobs create more surplus. As a result, AI can lower wages, reduce worker welfare, and even depress profits. In a search-and-matching extension with endogenous outside options, these forces are amplified, reinforcing a bad-job economy and potentially reducing employment.

JEL Codes: D86, J41, O33, L23.

Keywords: artificial intelligence; incentives; agency costs; job design; random matching; labor contracts

*We thank Robert Akerlof, Ricardo Alonso, Steffen Altmann, Zsófia Bárány, Daniel Barron, Jean de Bettignies, Yeon-Koo Che, Yi Chen, Wouter Dessein, Luis Garicano, Jie Gong, Anna Gumpert, Anders Humlum, Uta Schönberg, Heiner Schumacher, Dirk Sliwka, and Eduard Talamas, as well as audiences at seminars at Frankfurt, Harvard Business School, the National University of Singapore, Peking University, Ulm, and the University of Hong Kong, and at the 2nd Asian Conference on Organizational Economics (HKU), the 11th Berkeley-Padova-Paris Organizational Economics Conference (UC Berkeley), the 2025 CESifo-TransforM Workshop on the Economic and Societal Impacts of AI (Munich), and the workshop "The Economics and Business of Artificial Intelligence" (HKUST).

†Matthias Fahn: mfahn@hku.hk; Jin Li: jli1@hku.hk; Chang Sun: sunc@hku.hk.

1 Introduction

A question of growing importance is how artificial intelligence reshapes work. Much of the existing literature studies this question through a task-based lens, asking which tasks are automated, which remain human-performed, and how these changes affect wages and employment (cf. Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018a). While this perspective is valuable, it places less emphasis on a distinct but equally important margin: the incentive problems that arise within jobs. In many production settings, performance depends crucially on costly, non-contractible, effort. Firms therefore face not only a technological problem of allocating tasks between humans and machines, but also an organizational problem of designing jobs and compensation schemes that sustain effort.

This paper studies how AI affects job design through this incentive channel. Our central argument is that AI can increase agency costs even when it raises individual productivity. The reason is that modern AI often makes it substantially easier to achieve acceptable or satisfactory performance, while leaving the incremental effort required for especially reliable, careful, or contextually appropriate performance comparatively less affected (Dell'Acqua et al., 2024; Brynjolfsson et al., 2025; Kanazawa et al., 2025; Chen et al., 2025). In other words, AI reduces the cost of effort, but it may reduce the cost of low effort more than the cost of high effort. This widening of the effort-cost gap weakens incentives for sustained high effort and raises the cost to firms of implementing it.

Once this incentive effect is taken into account, AI can alter not only compensation but the structure of jobs themselves. When inducing high effort becomes more expensive, firms may optimally redesign jobs away from roles that rely on discretion, judgment, verification, and responsibility for quality, and toward roles organized around narrower, more standardized, and merely satisfactory performance. We refer to the former as good jobs and the latter as bad jobs. In the language of organization theory, the difference is less about the horizontal allocation of tasks than about the vertical organization of the work that remains: bad jobs are more vertically specialized, while good jobs are more vertically enlarged (Mintzberg, 1979). The concern raised by AI is therefore not only that some human tasks may be automated, but also that the jobs that remain may be redesigned in ways that reduce effort, wages, and worker rents.

We formalize this idea in a principal-agent model with limited liability. A worker chooses between high and low effort, where high effort guarantees high output while low effort yields high output only probabilistically. Because effort is not contractible, implementing high effort may require leaving rents to the worker. This creates a tradeoff for the firm between maximizing relationship value and minimizing rent payments. Depending on the worker's outside option, the optimal contract can therefore take one of two qualitatively different forms. When outside options are sufficiently high, the firm offers a good job that induces high effort and may pay rents. When outside options are low, inducing high effort becomes too costly relative to the rents required, and the firm instead offers a bad job characterized by low effort, low pay, and no rent.

We model AI as reducing the worker's effort costs, with a larger reduction for low effort than for high effort. This is consistent with evidence that AI mainly accelerates the early stages of production

and generates larger productivity gains for lower-baseline workers in writing, customer support, and related prediction and assistance tasks (Chen et al., 2025; Noy and Zhang, 2023; Brynjolfsson et al., 2025; Kanazawa et al., 2025). Under this assumption, AI raises the incentive cost of sustaining high effort, with two immediate consequences. First, even when the firm continues to offer a good job, the compensation required to preserve incentives can increase. Second, the firm may respond by switching from a good job to a bad job. AI thus expands the set of environments in which bad jobs are privately optimal for firms. Put differently, when satisfactory performance becomes easier to achieve, firms may reorganize work around that lower performance threshold rather than preserve jobs built around high discretionary effort.

Importantly, however, this job design result does not rely on the full strength of the assumption that AI reduces low-effort costs more than high-effort costs. For the shift from good jobs to bad jobs, it is enough that AI makes low effort sufficiently cheaper relative to high effort. Indeed, the shift can occur even when AI reduces high-effort costs more than low-effort costs, provided that this reduction is not so large as to offset the increased attractiveness of low effort.

The shift from good jobs to bad jobs has implications for both worker welfare and firm performance. It lowers worker welfare because it eliminates the rents associated with the former. But firm profits need not rise. Although moving to a bad job reduces rent payments, it also lowers the value created by the employment relationship. As a result, AI can generate cases in which both workers and firms are worse off in equilibrium, despite the direct productivity gains that AI provides at the individual level. The paper therefore highlights a distinction between technological efficiency and organizational performance: technologies that make workers more productive in a mechanical sense may nonetheless worsen equilibrium outcomes once firms adjust incentives and job design.

We then show that these forces become stronger in a market setting where workers' outside options are endogenous. To do so, we embed the contracting problem in a random matching model. In this environment, a decline in profits from good jobs reduces vacancy creation, weakens job-finding prospects, and thereby lowers workers' outside options. This feedback further increases the attractiveness of bad jobs relative to good jobs. The labor market can therefore amplify the direct organizational effect of AI and, in some cases, shift the economy from a good-job equilibrium to a bad-job equilibrium. Then, employment may fall as well, in which case the consequences for workers are especially severe: jobs become harder to find, and those that remain offer worse conditions than before.

Three main implications follow from our analysis. First, it suggests that the consequences of AI depend not only on the technology itself, but also on the organizational response. This offers a different perspective on debates about "the right kind of AI" (Acemoglu, 2024). Policy discussions often emphasize steering innovation toward human-complementary technologies that create new tasks. Our analysis shows, instead, that the *same* AI technology can lead to very different outcomes depending on how firms organize work and incentives around it.

Second, our mechanism points to a different labor-market risk than the one emphasized in most of the task-based literature. There, workers are harmed when technology takes over tasks and

reduces demand for their skills (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018b, 2019); by contrast, when productivity rises in tasks that workers continue to perform, wages tend to rise absent displacement or task-price effects (Acemoglu et al., 2025). In our framework, however, workers may continue performing the task and still end up with a worse job. When AI makes acceptable output easier to produce, firms may stop paying for discretion, responsibility, and sustained effort and instead reorganize work into narrower, more standardized roles. Emerging evidence from online labor markets is consistent with this possibility: Siddiq and Zhang (2026) find that, in more AI-exposed Upwork categories following the release of ChatGPT, labor demand became less closely tied to workers' human-capital signals and more closely tied to price, with demand shifting toward lower-priced workers. In this sense, AI can generate a modern form of deskilling (Braverman, 1974) even without eliminating the worker from production.

Third, the move toward bad jobs need not merely redistribute surplus from workers to firms; it can destroy value and leave both sides worse off. This distinguishes our mechanism from the common view that, even if AI harms some workers, aggregate gains remain positive. In our setting, lower incentive costs need not offset the decline in relationship value. This, in turn, points to a role for policy: measures that sustain workers' outside options do not merely redistribute, but can improve efficiency by helping keep the economy in the high-effort, high-value regime.

Related Literature

Our paper contributes to several literatures on how AI shapes work and the future of organizations.

AI and productivity. A growing empirical literature documents substantial performance gains from modern AI tools in knowledge work and creative tasks. Field and experimental evidence shows that generative AI can raise productivity and quality in writing and knowledge-worker settings (Noy and Zhang, 2023; Dell'Acqua et al., 2024; Brynjolfsson et al., 2025), and can also improve outcomes in domains such as marketing content creation (Hartmann et al., 2025) and task performance in specific occupations (Kanazawa et al., 2025). These findings are consistent with the idea that AI can disproportionately lower the personal effort required to achieve satisfactory outcomes. At the same time, evidence suggests that translating these individual-level gains into organization-wide performance is often difficult. A report from MIT (Challapally et al., 2025) finds that 95% of organizations see no measurable profit-and-loss impact from AI investments, while Humlum and Vestergaard (2025) find only modest early labor-market effects of AI chatbots in the aggregate. Consistent with this, a survey by the Centre for AI, Management and Organization at the University of Hong Kong (Fahn et al., 2026) finds that while many firms are piloting customer-facing and operational use cases, only a small minority have scaled them to measurable profit impact, and nearly half report returns below expectations. Across these sources, managers point to organizational and execution barriers—rather than purely technical constraints—as the main obstacles to realizing AI's potential. Our analysis speaks directly to this gap between individual productivity gains and limited organizational returns: when AI makes "good enough" output cheaper, firms may face higher

agency costs, which can mitigate or even reverse the gains from adoption.

Motivation and "mediocrity traps". Relatedly, our mechanism is closely related to recent concerns that generative AI can weaken motivation and reduce the return to sustained effort (Chen et al., 2025; Lee et al., 2025; Lin, 2025; Acemoglu et al., 2026). One strand of this literature stresses that human effort remains essential even in AI-assisted production. In particular, Lin (2025) shows that realized gains depend crucially on continued human adaptation: users systematically adjust prompts across model upgrades and iteratively refine them in response to prior outputs, so AI performance continues to rely on human judgment and effort. A different strand emphasizes that AI may nevertheless weaken incentives to exert such effort. Chen et al. (2025) shows that AI can flatten the quality frontier and generate a "mediocrity trap", while Lee et al. (2025) documents self-reported reductions in cognitive effort among knowledge workers using generative AI. More broadly, Acemoglu et al. (2026) study how highly capable AI systems may erode incentives for human knowledge acquisition. We complement this literature by embedding the effect of AI on effort costs in a tractable contracting framework. In our model, even when human effort remains valuable, AI can change its relative cost so that firms optimally redesign compensation and jobs in ways that reduce effort and worsen equilibrium job quality.

AI and the informativeness of signals. Our paper is closely related to an emerging literature showing that generative AI can reduce the informativeness of observable signals. Galdin and Silbert (2025) study how generative AI reduces the value of signals such as cover letters in job applications. Cui et al. (2025) show that AI-assisted cover-letter writing increases textual tailoring and callbacks while weakening the mapping from cover letters to applicant quality. Nejad et al. (2025) demonstrate that LLMs improve cover-letter quality, especially for weaker applicants, without necessarily improving interview outcomes. Cowgill et al. (2026) show more generally that AI can either worsen or improve screening accuracy depending on whose signals improve most. Our paper applies a similar idea after hiring rather than before hiring: in our model, AI makes performance measures less informative about underlying effort, which raises the rents needed to implement high effort under limited liability.

AI, organizations, delegation, and skill formation. A growing theoretical literature studies how AI changes organizational design itself. Athey et al. (2020) analyze whether decision authority should be assigned to a human agent or to AI, emphasizing the trade-off between the alignment of AI decisions and the need to preserve human initiative. Itoh and Morita (2025) study delegation in a three-stage process of information acquisition, project choice, and execution, and characterize how authority and AI interact when the same agent both acquires information and implements decisions. Cheng et al. (2025) analyze AI adoption in sequential teams with peer monitoring, showing that replacing some workers with AI changes the incentive constraints of the others and therefore the optimal deployment of AI across positions. Ide and Talamàs (2025) and Ide and Talamàs (2026) study how AI reshapes hierarchical knowledge-work organizations and the distribution of gains,

while Ide and Talamàs (2024) shows that labor-market consequences depend on which knowledge dimensions AI improves. A related strand studies skill formation and training inside organizations: Garicano and Rayo (2025) analyze apprenticeship viability when AI performs an increasing share of junior tasks; Bárány and Koren (2026) study how AI can shrink teams and crowd out junior learning opportunities; and Ide (2025) examines how automating early-career tasks can weaken the intergenerational transmission of tacit knowledge. Relative to this literature, our paper studies a different organizational friction, namely the problem of eliciting discretionary non-contractible effort, and how AI can lead firms to redesign jobs away from high-discretion work and toward standardized satisfactory-performance roles.

Task-based technological change. We also complement the task-based literature on technological change and AI, which emphasizes how technology reallocates tasks between labor and machines and how those reallocations affect wages, employment, and inequality. Recent task-based work emphasizes that AI often changes the expertise content and bundling of the tasks that remain human-performed rather than simply automating whole jobs. Autor and Thompson (2025) argue that the labor-market effects of automation depend on whether it increases the expertise required for remaining non-automated tasks. Althoff and Reichardt (2026) study task-specific technical change and comparative advantage; Agrawal et al. (2026) analyze how AI can enhance worker productivity without automating tasks; Maasoum and Lichtinger (2026) emphasize the interaction between productivity, expertise, and effective labor supply; Demirer et al. (2026) analyze how AI can automate chains of tasks and thereby redefine jobs; and Freund and Mann (2026) study job transformation and specialization under AI. Our contribution adds a within-task incentive channel to this literature: even when the human role and the task allocation are unchanged, AI can alter the effort–output mapping in ways that make high effort harder to sustain and lower job quality.

General equilibrium, private information, and competition. Finally, our paper connects to foundational work on competitive equilibrium with private information, as well as to the literature on competition and incentives. Several papers by Prescott and Townsend show how competitive equilibrium and welfare analysis can be extended to economies with private information, including moral-hazard environments, and their later work interprets firms as multi-agent contracts traded in general equilibrium (Prescott and Townsend, 1984a,b, 2006). Schmidt (1997) and Raith (2003) study how product-market competition affects incentive provision for managers. Relative to this literature, our paper emphasizes the role of labor-market competition: endogenous outside options and vacancy creation, rather than rivalry in the output market, determine the incentive cost of sustaining high effort.

2 Model Setup

We now introduce our baseline model and, at the end of the section, briefly discuss several key assumptions underlying the setup. One risk-neutral principal ("firm," "she") can hire one risk-

neutral agent ("worker," "he"). The agent chooses his non-contractible effort level $e \in \{L, H\}$, which is associated with effort costs c^e, with $c^H > c^L > 0$. While an effort level of $e = L$ reflects the "minimal effort" the agent would supply without additional incentives (e.g., due to intrinsic motivation or because this level is verifiable), high effort $e = H$ corresponds to working hard and exerting discretionary effort.

Output $Y \in \{0, y\}$ is either high ($Y = y > 0$) or low ($Y = 0$). The agent's effort determines the probability of achieving high quality. If the agent chooses $e = H$, high output $Y = y$ is obtained with probability 1. If the agent chooses $e = L$, high output is obtained with probability q and low output with probability $1 - q$, with $0 < q < 1$.

We assume

$$(1 - q)y - \left(c^H - c^L\right) > 0,$$

thus high effort is efficient.

Compensation and Payoffs The agent receives a wage $w(Y)$, i.e., he is paid $w(y)$ if $Y = y$, and $w(0)$ if $Y = 0$. Thus, the principal's profit is

$$\begin{aligned} \pi^H =& y - w(y) \\ \pi^L =& q\,(y - w(y)) - (1 - q)w(0) \end{aligned}$$

The agent's utility from working for the principal is

$$\begin{aligned} u^H =& w(y) - c^H \\ u^L =& qw(y) + (1 - q)w(0) - c^L. \end{aligned}$$

If the agent does *not* work for the principal, he receives an outside option utility $\bar{u} \geq 0$, which is shaped by labor-market conditions such as the intensity of competition or the generosity of unemployment benefits. While we treat the outside option as exogenous in the first part of the paper, we endogenize it in a random matching model in Section 5. We also assume the principal's outside option to be sufficiently low that hiring the agent is always profitable.

Finally, we assume that the agent is protected by *limited liability*, thus $w(Y) \geq 0$.

Maximization Problem We derive a subgame perfect equilibrium that maximizes the principal's profits π: The principal's chooses $(e^*, w(0), w(y))$ to maximize

$$\begin{aligned} \max_{w(\cdot),\, e^*} \quad & \pi = \mathbb{E}\big[Y - w(Y) \mid e^*\big] \\ \text{s.t.} \quad & e^* \in \arg\max_e \left\{\mathbb{E}\big[w(Y) \mid e\big] - c^e\right\} && \text{(IC)} \\ & \mathbb{E}\big[w(Y) \mid e^*\big] - c^{e^*} \geq \bar{u} && \text{(PC)} \\ & w(y) \geq 0, \quad w(0) \geq 0 && \text{(LL)} \end{aligned}$$

The Implications of AI We assume that AI enhances the agent's productivity, reducing c^L to c^L_{AI} and c^H to c^H_{AI}, with $c^L_{AI} < c^L$ and $c^H_{AI} \leq c^H$, but that it reduces low-effort cost by more than high-effort cost, i.e.,

$$0 \leq c^H - c^H_{AI} < c^L - c^L_{AI},$$

which implies

$$c^H_{AI} - c^L_{AI} \in \left(c^H - c^L,\, c^H\right].$$

This assumption is based on recent evidence that the adoption of AI is associated with a "baseline-boosting" pattern in which the tool disproportionately lowers the effort required to reach acceptable performance, while leaving the incremental effort needed for excellent/reliably correct performance comparatively high (Noy and Zhang, 2023; Brynjolfsson et al., 2025; Chen et al., 2025; Kanazawa et al., 2025). Note that the assumption that AI reduces low-effort costs more than high-effort costs is stronger than what is needed for our main job-design result in Proposition 1. We will show below that the same result can also arise when AI reduces high-effort costs more than low-effort costs, provided that this additional reduction is not too large.

Finally, we assume that

$$(1-q)y - \left(c^H - c^L_{AI}\right) > 0.$$

This condition is stronger than the requirement that high effort remain efficient, for which $(1-q)y - \left(c^H_{AI} - c^L_{AI}\right) > 0$ would suffice, but it allows us to focus on the central mechanism of how AI affects job design. In Appendix A, we also derive results for the case $(1-q)y \leq c^H - c^L_{AI}$, while maintaining the assumption that high effort remains efficient. The main qualitative results, especially the effect of AI on firms' job design, continue to hold there as well.

2.1 Discussion of Assumptions

Before deriving equilibrium outcomes, we discuss three key assumptions of our setup.

Effort costs rather than output gains We model AI as lowering effort costs, with a larger reduction at low effort. This delivers the paper's central mechanism, as derived below: when AI makes satisfactory performance easier to reach, the cost gap between high and low effort widens, so sustaining high effort becomes more expensive under limited liability. An alternative formulation would let AI raise the probability of a high outcome, with a larger effect under low than under high effort. This would yield very similar qualitative results, because it would again weaken incentives for high effort (holding compensation fixed) and raise the firm's agency costs. However, it would also introduce a direct positive effect on output, thus modeling AI through effort costs keeps the mechanism transparent.

Binary effort We restrict effort to two levels, $e \in \{L, H\}$. However, our logic extends to settings with continuous effort and convex effort costs, where effort raises the probability of a high outcome and AI reduces effort costs by less at higher effort levels. In such environments, AI still compresses

the return to additional effort and can therefore raise the cost of implementing higher effort when limited liability binds.

Technology Choice Our baseline setup assumes a single job in which the firm chooses compensation, and the worker chooses effort. An alternative formulation lets the firm choose between two distinct technologies, or, equivalently, two job designs, before the employment relationship begins. One technology is more standardized: it is designed so that acceptable output can be produced with minimal worker input, corresponding in our framework to the case in which the firm implements low effort. The other technology relies on activities that are costly and non-contractible, such as worker discretion, judgment, and care, and, therefore, corresponds to the case in which the firm implements high effort.

Under this interpretation, adopting the standardized technology is equivalent to offering a bad job: the firm structures the production process so that the worker's role is narrow and largely procedural, and incentive pay is unnecessary. Adopting the discretion-intensive technology is equivalent to offering a good job: the worker retains meaningful responsibility, and the firm must design compensation to sustain high effort. Our analysis then implies that AI, by disproportionately lowering the cost of achieving satisfactory output, makes the standardized technology relatively more attractive. The firm is therefore more likely to adopt the technology that economizes on worker effort even when the discretion-intensive technology would generate greater total surplus. We choose our single-job formulation because it captures the same economic forces in a simpler way. The key results are identical in both setups.

3 Outcomes Without AI

3.1 Approach: Costs of Implementing Low vs. High Effort

The principal chooses between inducing high effort and accepting low effort. We follow the standard Grossman-Hart method for moral hazard problems with discrete effort: first derive the least-cost contract that implements low effort, then the least-cost contract that implements high effort, and finally compare the resulting profits to select the optimal contract. To induce *low effort*, the principal can offer an output-independent wage that merely satisfies the participation constraint:

$$w(y) = w(0) = \bar{u} + c^L.$$

With this contract, the worker's utility is equal to his outside option, $u^L = \bar{u}$.

To induce *high effort*, the principal must design a wage schedule that makes high effort incentive-compatible and at the same time satisfies the agent's participation constraint, i.e.,

$$w(y) - c^H \geq qw(y) + (1-q)w(0) - c^L \tag{IC}$$

$$w(y) - c^H \geq \bar{u}, \tag{PC}$$

must hold. Moreover, limited liability immediately gives $w(0) = 0$, therefore (IC) becomes

$$w(y) \geq \frac{c^H - c^L}{1-q}, \tag{IC}$$

and the optimal wage schedule is the minimal $w(y)$ that satisfies both constraints, as we describe in the following lemma.

Lemma 1 *If the principal wants to implement high effort, the optimal contract satisfies the following:*

$$w(0) = 0, w(y) = max\left\{\frac{c^H - c^L}{1-q}, \bar{u} + c^H\right\}.$$

The worker earns a rent if

$$\bar{u} < \frac{c^H - c^L}{1-q} - c^H \equiv \bar{u}^B.$$

For $\bar{u} < \bar{u}^B$, the rent received by the agent increases in q. Intuitively, as low effort becomes more likely to produce high quality, it becomes harder for the principal to distinguish and reward high effort. Moreover, in this range the wage does not depend on the agent's outside option. This rent reflects the *agency cost* of incentivizing high effort when the agent's incentive constraint pins down $w(y)$. Because the agent can obtain $w(y)$ with positive probability even under low effort, he must be compensated enough to resist this temptation. Put differently, the agent's private cost of choosing high rather than low effort is $\left(c^H - c^L\right)$, whereas the principal's cost of providing incentives for high effort is $\left(c^H - c^L\right)/(1-q)$. The difference between these two amounts,

$$\frac{q\left(c^H - c^L\right)}{(1-q)},$$

is the agency cost of implementing high effort. When $\bar{u}$ is small, limited liability prevents the principal from extracting the resulting rent through negative payments in the low-output state, so it remains with the agent; in this range, neither profits nor utility depend on $\bar{u}$. By contrast, when the agent's outside option is sufficiently large so that $\bar{u} + c^L$—the minimum total payment needed to secure participation—exceeds the agency cost $q(c^H - c^L)/(1-q)$, the principal can induce high effort while keeping the agent's utility equal to his outside option.

3.2 Optimal Contract

Next, we derive the profit-maximizing contract as a function of the agent's outside option $\bar{u}$. Using the least-cost implementations of low and high effort from above, profits under low effort are

$$\pi^L = qy - \bar{u} - c^L,$$

while profits under high effort are

$$\pi^H = y - \max\left\{\frac{c^H - c^L}{1-q},\, \bar{u} + c^H\right\}.$$

When $\bar{u} \leq \bar{u}^B$, we have $\pi^H = y - \frac{c^H - c^L}{1-q}$, in which case

$$\pi^H \geq \pi^L \;\;\Leftrightarrow\;\; \bar{u} \geq \frac{c^H - (2-q)c^L}{1-q} - (1-q)y \;\equiv\; \bar{u}^A.$$

Since $\bar{u}^B > \bar{u}^A$ follows from $(1-q)y > c^H - c^L$, we obtain the following lemma.

Lemma 2 *There exists $\bar{u}^A$ such that the following holds:*

(i) When $\bar{u} \leq \bar{u}^A$, the principal implements low effort and sets wages $w(0) = w(y) = \bar{u} + c^L$.

(ii) When $\bar{u} > \bar{u}^A$, the principal implements high effort and sets wages

$$w(0) = 0, w(y) = max\left\{\frac{c^H - c^L}{1-q},\, \bar{u} + c^H\right\}.$$

The worker earns a rent if $\bar{u} < \bar{u}^B = \frac{qc^H - c^L}{1-q}$, where $\bar{u}^B > \bar{u}^A$.

In general, the principal faces a trade-off between maximizing total surplus (which calls for high effort) and extracting as much rent as possible from the agent. As discussed above, this trade-off—and hence the optimal contract—depends not only on the value of high output and on the agency costs (captured by q), but also on the size of the agent's outside option.

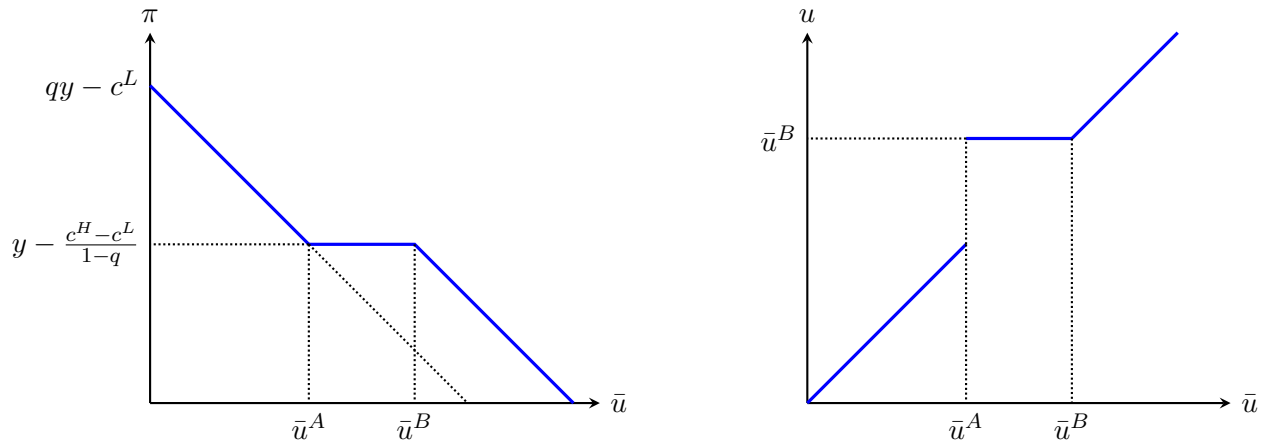
When the outside option is high enough that high effort can be implemented while keeping the agent's utility at his outside option, this is clearly optimal. If implementing high effort requires granting a rent, high effort remains optimal only when the value of high quality is large relative to the agency costs. By contrast, when the outside option is small and the value from high quality is low compared to these costs, inducing low effort is optimal and the agent earns no rent.[1]

3.3 Interpretation: "Good Jobs" versus "Bad Jobs"

The cutoff $\bar{u}_A$ naturally lends itself to an interpretation in terms of job design. When $\bar{u} \leq \bar{u}_A$, the firm offers what we may call bad jobs: the contract induces low effort, pays a low wage $w \leq \bar{u}_A + c_L$, and leaves the worker with no rent. By contrast, when $\bar{u} > \bar{u}_A$, the firm offers good jobs: the contract induces high effort, pays a high wage $w \geq \bar{u}_B + c_H > \bar{u}_A + c_L$, and, for $\bar{u} < \bar{u}_B$, grants the worker a positive rent. The worker's utility therefore jumps discretely at $\bar{u}_A$, illustrating the transition from low-effort, low-pay bad jobs to high-effort, better-paid good jobs. The following figure plots firm profits (a) and worker utility (b) as functions of $\bar{u}$.

[1] Note that it may occur that $\bar{u}_A, \bar{u}_B < 0$, in which case one or both regions are empty under our assumption that $\bar{u} \geq 0$.

Figure 1: Optimal profits and worker utility across outside options.



(a) Firms' profits

(b) Workers' utility

Notes. In panel (a), the solid blue line indicates profits under the optimal contract. The dashed segment to the right of $\bar{u}^A$ represents the profits that would be obtained if bad jobs were offered for $\bar{u} > \bar{u}^A$, while the dashed segment to the left of $\bar{u}^A$ continues the blue line and shows the profits that would obtain if good jobs were offered for $\bar{u} \leq \bar{u}^A$. In panel (b), the solid blue line indicates the worker's utility under the optimal contract.

This distinction between good and bad jobs maps naturally onto the concept of vertical specialization in organizational design (Mintzberg, 1979, Chapter 4). Horizontal specialization concerns the breadth of a job: how many different tasks the worker performs. Vertical specialization concerns its depth: the degree of discretion, judgment, and responsibility the worker retains over how the work is carried out. In our framework, the shift from good jobs to bad jobs is primarily a change in vertical specialization. Bad jobs are vertically specialized: the worker executes a prescribed, satisfactory-performance role with little discretion. Good jobs are vertically enlarged: the worker is expected not only to carry out the activity but also to supply judgment, initiative, care, and responsibility for quality.

Thus, "good" and "bad" in our framework refer not to the intrinsic content of the task, nor primarily to the breadth of the task bundle, but to the depth of the human role within the job. Precisely because good jobs grant the worker more discretion over how the work is carried out, they also require incentives to ensure that this discretion is used to sustain high effort rather than to settle for merely satisfactory performance. We return to this interpretation and its connection to broader debates on workplace deskilling after deriving the main results in the next section.

4 Outcomes With AI

We first derive a benchmark in which effort is contractible. Then, the principal can require $e = H$ and pay $w = c^H + \bar{u}$ conditional on the agent exerting high effort. The agent's utility then equals $\bar{u}$ and profits are $\pi = y - c^H - \bar{u}$, i.e., are equal to the total net surplus. AI only reduces the cost term c^H,so $\Delta u = 0$, $\Delta w = c^H_{AI} - c^H \leq 0$ and $\Delta \pi = c^H - c^H_{AI} \geq 0$. Thus, with verifiable effort, profits are

only affected by the reduced individual effort costs, and AI unambiguously raises profits.

Now, recall that AI lowers the agent's effort costs, and does so more for low effort than for high effort. This forces the firm to adjust its contract because keeping the pre-AI scheme is generally bad for profits: If high effort was implemented before AI, existing incentives may no longer be strong enough. Even when high effort remains optimal for the worker without changing the contract, most of the efficiency gains would then be captured by the worker rather than the firm.

In the following, we therefore distinguish two types of responses. First, *compensation design*, where the principal adjusts the pay scheme while holding job design (good vs. bad jobs) fixed. Second, *job design*, where the principal decides whether the worker remains in a good or bad job, or is moved between them in response to AI.

Before doing that, we note that, with AI the structure of the optimal contract is identical to the one described in Lemma 2, only the cutoffs shift right. This is collected in the following Lemma:

Lemma 3 *With AI, there exist values $\bar{u}^B_{AI} > \bar{u}^A_{AI}$, where $\bar{u}^B_{AI} > \bar{u}^B$ and $\bar{u}^A_{AI} > \bar{u}^A$.*

4.1 Compensation Design

Bad jobs When the firm offers a bad job, it can cut compensation to $w_{AI} = \bar{u} + c^L_{AI}$, so profits rise to $\pi^L_{AI} = qy - \bar{u} - c^L_{AI}$. In other words, with bad jobs, AI lowers wages and raises profits, while the agent's utility remains at his outside option.

Good jobs With good jobs, the effect of AI on the optimal contract is twofold. First, AI raises the value created, which is good for the firm. Second, it makes this value harder to capture, because maintaining incentive compatibility now requires paying a higher wage to the worker:

$$w_{AI}(y) \geq \frac{c^H_{AI} - c^L_{AI}}{1-q} > \frac{c^H - c^L}{1-q},$$

so the agency cost of incentivizing high effort increases.

Put differently, AI makes it easier to do the right thing, but even easier to do the wrong thing. Under the pre-AI compensation scheme, the temptation to exert low effort becomes stronger, so the firm must raise wages to preserve incentives when the outside option is small. This may even reduce profits, as the following Lemma shows.

Lemma 4 *If the firm continues to offer good jobs with AI, then there exists a threshold $\bar{u}^{B*}$, such that the following holds:*

(i) When $\bar{u} \leq \bar{u}^{B*}$, AI decreases profits.

(ii) When $\bar{u} > \bar{u}^{B*}$, AI increases profits.

Intuitively, if the participation constraint $u \geq \bar{u}$ determines the agent's compensation (so the incentive constraint (IC) is slack both before and after AI), then AI raises profits because the principal can lower the wage $w(y)$, i.e., AI reduces the cost of implementing high effort.

If instead (IC) binds both before and after AI, profits fall: higher agency costs force the principal to increase wages to preserve incentive compatibility. In the intermediate case, where (PC) determines wages before AI but (IC) binds afterward, the effect on profits is ambiguous and depends on which force dominates.

4.2 Job Design

If the principal offers a good job before AI, there are two margins along which she can adjust. First, she can modify compensation when the agent's outside option is low in order to preserve the good job, as described in the previous section. Second, she may find it optimal to switch from a good to a bad job; this latter adjustment is characterized in the following proposition.

Proposition 1 *AI favors bad jobs.*

(i) If the firm offers a bad job before AI, it continues to offer a bad job after AI.

(ii) If the firm offers a good job before AI, it switches to offering a bad job if $\bar{u} \leq \bar{u}_{AI}^{A}$.

The proposition follows directly from the upward shift in the cutoff, $\bar{u}^A \to \bar{u}_{AI}^A$ (provided $\bar{u}_{AI}^A > 0$). This shift implies that, for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right]$, the principal offers a good job before AI but a bad job after AI. This is because, with higher agency costs and a weak outside option, the trade-off between maximizing firm value and limiting rents to the agent becomes more severe. Conversely, if $\bar{u} > \bar{u}_{AI}^A$, a good job is offered both before and after the introduction of AI, while if $\bar{u} \leq \bar{u}^A$, a bad job is offered throughout.

Note that the assumption that AI lowers low-effort costs more than high-effort costs is not necessary for Proposition 1. Since

$$\bar{u}_A^{AI} - \bar{u}_A = \frac{(2-q)(c^L - c_{AI}^L) - (c^H - c_{AI}^H)}{1-q},$$

AI expands the range of outside options for which firms offer bad jobs whenever

$$(2-q)(c^L - c_{AI}^L) > c^H - c_{AI}^H,$$

which can hold even when AI reduces high-effort costs more than low-effort costs, provided that the reduction in high-effort costs is not too large.

Proposition 1 identifies a channel through which AI can worsen job outcomes that is different from standard task-substitution mechanisms. In the existing literature, AI harms workers primarily by automating tasks they previously performed, reducing demand for their skills (Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2018b, 2019). By contrast, in the standard task-based view, productivity gains in tasks that remain human-performed tend to raise wages, absent displacement or task-price effects (Acemoglu et al., 2025). Our model shows that even if the tasks do not change, AI can lower job quality by changing the incentive structure. When satisfactory output becomes easier

to achieve, sustaining high effort becomes costlier under limited liability, and the firm optimally shifts to a narrower, lower-discretion role.

Empirically, this channel generates predictions that differ from those of the standard task-based framework. First, our model predicts that, for a sufficiently low outside option, incentive-based pay will decrease even for workers that appear to do the same tasks as before. Second, because of the standardization of technology, the performance differences and the pay differences within the job are likely to narrow as well.

To summarize, our results predict a shift toward bad jobs, especially when workers' outside options are low, for example because alternative jobs are scarce or unemployment benefits are limited. This suggests governments can help sustain good jobs under AI by strengthening outside options, highlighting a complementarity between the social safety net and the viability (and even improvement) of high-effort, high-wage jobs.

4.3 Consequences of AI-Induced Job Design

We now study the consequences of an AI-induced shift from good to bad jobs for worker utility and firm profits, before turning to AI adoption.

Proposition 2 *Suppose AI induces a shift from a good job to a bad job. Then the worker's utility weakly decreases.*

Moreover, there exists a cutoff $\bar{u}^{A*} \in \left(\bar{u}^A, \bar{u}^A_{AI}\right)$ *such that:*

(i) if $\bar{u} \leq \bar{u}^{A*}$*, the shift from a good to a bad job increases the firm's profit;*

(ii) if $\bar{u}^{A*} < \bar{u}$*, the shift from a good to a bad job decreases the firm's profit.*

The intuition for the utility result is straightforward. A shift from a good to a bad job typically eliminates the worker's rent: under a bad job, the wage only needs to satisfy the participation constraint, whereas under a good job, incentive compatibility may force the firm to leave a rent to the worker. Hence, utility falls whenever the good job previously paid a rent. The only exception is the case in which the economy moves from a good job *without* rent to a bad job (which can occur only if $\bar{u}^B \leq \bar{u}^A_{AI}$); then the worker's utility equals $\bar{u}$ both before and after, so utility is unchanged.

For firms, the effect is more nuanced.[2] A shift from a good to a bad job has two countervailing consequences. On the one hand, the value generated by the relationship falls because effort declines. On the other hand, if the good job previously involved a rent, this rent disappears when the firm switches to a bad job, increasing the firm's share of the surplus.

If the good job paid no rent (which is possible only if $\bar{u}^B \leq \bar{u}^A_{AI}$), the second channel is absent: the wage is pinned down by the participation constraint both before and after, so the shift only lowers relationship value and profits must fall. If the good job did pay a rent, however, the elimination of this rent works in the opposite direction. For relatively low $\bar{u}$, the reservation wage in the bad job is

[2] In Appendix A, we characterize the effects of AI on profits and utility for all values of $\bar{u}$, not only for the range in which AI induces a shift from good to bad jobs.

sufficiently low that the rent saving outweighs the loss in value, so profits rise. If $\bar{u}$ is relatively high, the wage required to sustain the bad job is high as well, reducing the gain from rent elimination. Beyond the cutoff $\bar{u}^{A*}$, the loss in value dominates, and profits fall.

Propositions 1 and 2 together point out a dire possibility: AI can simultaneously shift firms toward bad jobs and make both parties worse off. We now discuss in more detail the implications of our mechanism.

First, the result can be understood as a form of incentive-driven deskilling. A long tradition in the sociology of work, originating with Braverman (1974), argues that firms have a systematic incentive to simplify and standardize jobs in order to reduce their dependence on workers' skill and discretion. Our model provides a precise economic rationale for this pattern in the context of AI. When AI lowers the cost of achieving satisfactory output, the agency cost of sustaining discretionary effort rises relative to its benefit. The firm responds by narrowing the worker's role and increasing vertical specialization: work is reorganized away from judgment, initiative, and responsibility for quality. Consistent with this logic, Siddiq and Zhang (2026) find that, following the release of ChatGPT, human-capital signals became less important for labor demand in more AI-exposed Upwork categories, while price became more important. Moreover, the premium for workers with stronger human-capital signals declined and demand shifted toward lower-priced workers, which the authors interpret as evidence of AI-induced labor commoditization. Although their evidence concerns the market valuation of workers rather than within-firm job design, it is in line with the broader implication of our mechanism that, when AI standardizes satisfactory output, demand becomes less sensitive to worker-specific skill and discretion and more sensitive to cost.

Second, the welfare result in Proposition 2 stands in contrast with the common narrative that even if AI may hurt the workers, its overall gain is positive. Because the shift to bad jobs can be lose-lose, value is destroyed. Such destruction of value calls for the role of government intervention. As we will discuss later, policies that sustain workers' outside options do not merely redistribute, but may improve efficiency by keeping the economy in the high-effort, high-value regime.

AI Adoption The fact that the switch from a good job to a bad job can leave both parties worse off raises the question of why a firm would adopt AI in the first place. Indeed, adoption would not be optimal in this range if it were fully under managerial control. In practice, however, many tools—especially ubiquitous chatbots—are difficult to restrict or monitor, so firms must take worker AI use as given and respond by redesigning contracts and jobs. This is particularly true in settings where AI can be used "quietly" as a drafting or idea creation tool and its use is hard to observe, such as content generation and marketing roles, customer support, HR and recruiting, and software development or data analysis. Consistent with this, a recent survey (Challapally et al., 2025) reports that only about 40% of firms purchase official LLM subscriptions, yet employees in over 90% of those firms already use personal AI tools for work. Importantly, in our model the worker is always better off with AI *holding the compensation scheme fixed*, simply because effort becomes less costly; any decline in equilibrium utility reflects the firm's optimal adjustment of compensation or job

design.

For $\bar{u} \leq \bar{u}^{A*}$ within the good-job-to-bad-job region, AI raises the firm's profits while (weakly) reducing the worker's utility, i.e., it generates a conflict of interest. When adoption is under managerial control (rather than driven by unavoidable worker use, as discussed above)—as is often the case for systems beyond simple chatbots—the firm adopts AI *whenever it can decide unilaterally*. Regulation, labor-representation rules, and the need for tacit employee cooperation can, however, create effective worker veto power and thereby alter this trade-off. In many European systems, employee representation provides such veto points over workplace technologies; for example, German works councils have codetermination and consultation rights under the Works Constitution Act (Federal Republic of Germany, 2001), while in France collective economic redundancies must be accompanied by a "Plan de sauvegarde de l'emploi" (République française, 2017).

Importantly, the profit gain in this conflict-of-interest range is smaller than the worker's utility loss, so compensating the worker to secure cooperation is not profitable.

5 General Equilibrium: Search, Matching and AI

In this section, we embed the earlier contracting problem into a general equilibrium model with search and matching (a detailed description of the model and additional results can be found in Appendix A). In this setup, we show that endogenizing workers' outside options $\bar{u}$—by interpreting them as the value of unemployment in a frictional labor market—leaves our main qualitative results unchanged, while also yielding additional insights on the impact of AI on job design and employment.

5.1 The Model and Equilibrium

We consider a continuous-time economy populated by a fixed mass of workers L and an endogenous mass of firms N. The labor market is characterized by search frictions, where the formation of productive matches is governed by a matching function $\mu(U, V)$. Here, U represents the number of unemployed workers and V denotes the number of open vacancies posted by firms. The matching function is assumed to be homogeneous of degree one. We define $\theta = V/U$ as labor market tightness, which determines the job-finding rate $f(\theta)$ and the vacancy-filling rate $s(\theta) = f(\theta)/\theta$. Matches separate exogenously at rate λ.

Firms enter the market by paying a flow vacancy-posting cost κ. Once a match is formed, production occurs according to the time-invariant contract $w(Y)$ and the effort choice $e \in \{L, H\}$. The firm's flow profit π is the residual of output minus the wage paid. We assume a discount rate r for all agents. In equilibrium, the number of firms N is determined by a free-entry condition, which dictates that firms will continue to post vacancies until the expected value of a vacancy is driven to zero.

A worker's welfare is determined by their state. While matched, a worker receives a wage and incurs an effort cost $c(e)$. While unemployed, the worker receives a flow utility $b > 0$, which represents unemployment benefits, the value of leisure, or home production. The critical departure

in this section is the endogenization of the worker's outside option $\bar{u}$. In our baseline, $\bar{u}$ was a fixed reservation utility; here, it is the flow value of unemployment (rV^U). This value is endogenous because it depends on the equilibrium contract (the expected utility from future employers) and the job-finding rate, $f(\theta)$. In this setup, the contracts offered by firms determine the value of being in the labor force, while the aggregate state of the labor market, i.e., the outside option, determines which contracts are optimal. Similar to our baseline model, we assume that b is sufficiently small so that matches are profitable enough and firms are willing to post vacancies in equilibrium.

In equilibrium, the market settles into different regimes based on the unemployment benefit, b, which plays a similar role as the exogenous outside option, $\bar{u}$, in the baseline model. When b is high, firms are willing to offer "good jobs" – contracts that implement high effort and may provide workers with information rents. When b is low, firms find it more profitable to offer "bad jobs" – low-effort contracts in which workers receive no rents and are indifferent between working and unemployment.

Different from the baseline equilibrium, multiple equilibria exist when b is in an intermediate range in this general equilibrium setup. The economy might settle into either a "good jobs" or "bad jobs" state. For example, a "good job" equilibrium can be self-fulfilling: if all firms offer high-effort contracts with rents, the outside option stays high, which in turn makes it easier for each individual firm to provide incentives. Similarly, when all firms offer low-effort contracts, the outside option is low, making it optimal for an individual firm to do so as well. Finally, there is a "mixed" equilibrium where a fraction of firms offer good jobs, and the others offer bad jobs, though such an equilibrium is not stable.

The existence of these distinct regimes and multiple equilibria implies that job quality is often a matter of coordination and equilibrium selection, where industry norms or "corporate culture" determine which outcome prevails. In the following analysis, we assume that the "good job" equilibrium, which produces high-quality output and maximizes the total social surplus per match, prevails whenever multiple equilibria exist. We next analyze how the introduction of AI can affect the regimes and collapse the "good job" equilibrium entirely.

5.2 AI, Job Design and Employment

We model the introduction of AI as reductions in the effort costs as in the baseline model. We assume that AI reduces low-effort costs more than high-effort costs, i.e., $c^L - c^L_{AI} > c^H - c^H_{AI} \geq 0$. We maintain the assumption we made in Section 2 that AI does not reduce the low-effort cost too much, i.e., $(1-q)y > c^H - c^L_{AI}$.

The impact of AI on job design is similar to that in the baseline model. Intuitively, AI increases information rents and makes it harder for firms to implement high efforts. Given our assumption that the "good job" equilibrium prevails when multiple equilibria exist, there is a minimum unemployment benefit required to sustain good jobs. We denote this cutoff in the baseline equilibrium as $\underline{b}$ and that in the equilibrium after AI adoption as $\underline{b}_{AI}$. To disentangle the partial equilibrium (PE) agency cost effect from the general equilibrium feedback, we also define $\underline{b}^{PE}_{AI}$ as the counterfactual cutoff computed at AI costs (c^H_{AI}, c^L_{AI}) while holding market tightness fixed at its pre-AI level θ^*. We have

the following lemma:

Lemma 5 *AI adoption increases the minimum unemployment benefit required to sustain good jobs. Moreover, the general equilibrium feedback through the job-finding rate amplifies this increase:*

$$\underline{b}_{AI} > \underline{b}_{AI}^{PE} > \underline{b}.$$

When firms offer good jobs, the value of output is y regardless of whether AI is used. Since AI increases workers' rents, it reduces firm profits from good jobs. In contrast, firms have higher profits from bad jobs because AI reduces the cost of exerting low effort. This has a direct effect on firms' switching from good to bad jobs, captured by the gap $\underline{b}_{AI}^{PE} > \underline{b}$. In addition, the general equilibrium amplifies this effect through endogenous vacancy creation and outside options. Lower firm profits from good jobs lead to fewer vacancies and a lower job-finding rate. Since employment in good jobs with positive rents offers strictly higher flow utility than unemployment, a lower job-finding rate reduces the value of unemployment, thereby reducing the worker's outside option. This makes it even harder for firms to offer good jobs, raising $\underline{b}$ further beyond $\underline{b}_{AI}^{PE}$. Therefore, when the unemployment benefit is between $\underline{b}_{AI}^{PE}$ and $\underline{b}_{AI}$, a good job equilibrium could be maintained after AI adoption if we ignored general equilibrium effects, but not once we account for the change in the job-finding rate.[3]

In contrast to the baseline model with one worker and one firm, the introduction of the labor market equilibrium allows us to discuss the impact of AI on employment, i.e., the number of matches, M. In the next proposition, we show that beyond making firms switch from good to bad jobs, AI can also lead to lower employment.

Proposition 3 *When $(1-q)y - (c_{AI}^H - c_{AI}^L)$ is sufficiently small, there exists a cutoff $b^{A*} < \underline{b}_{AI}$. As long as $b \in (b^{A*}, \underline{b}_{AI})$, AI adoption leads firms to switch from good to bad jobs, and the total employment decreases.*

The proposition identifies the starkest possible outcome of AI adoption: not only does job quality deteriorate, but total employment falls as well. The intuition is straightforward. AI makes good jobs more expensive to sustain by raising information rents, pushing firms to switch to bad jobs. Under our parameter restrictions, bad jobs are also less profitable than good jobs were before AI, so firms post fewer vacancies. Therefore, workers face a doubly adverse labor market — the jobs available are worse, and there are fewer of them.

6 Conclusion

This paper studies how modern AI affects incentives and job design when performance depends on non-contractible effort and workers are protected by limited liability. Our central assumption is that AI lowers the personal cost of reaching satisfactory performance more than it lowers the cost

[3] In the proof of Lemma 5, we provide analytical characterizations of the changes in $\underline{b}$ and other useful thresholds.

of sustained high effort. In that case, AI raises the incentive payments needed to induce high effort and can increase agency costs even as it raises individual productivity. Firms may therefore replace high-wage, high-value good jobs with low-wage, low-value bad jobs. In our framework, the main risk from AI is therefore not only labor displacement, but also a reorganization of work toward standardized satisfactory-performance jobs.

Embedding this contracting problem into a frictional labor market shows that these forces can be amplified in general equilibrium. When AI reduces profits in the high-effort regime, vacancy creation falls, job-finding rates decline, and workers' outside options weaken, which can push the economy across the threshold where firms switch to bad jobs. The consequences of AI thus depend not only on technical capabilities, but also on organizational responses and on expectations about how other firms redesign jobs. This helps explain why the same technology can support either a high-wage equilibrium or a bad-job equilibrium.

Our baseline assumes take-it-or-leave-it offers by the firm, consistent with evidence that many firms possess wage-setting power rather than facing perfectly competitive labor supply (Card, 2022; Manning, 2021). Under that assumption, AI adoption involves a trade-off between value creation and rent extraction, and conflicts of interest between firms and workers can be substantial. If workers instead have positive bargaining power, the adoption decision internalizes more of the joint surplus, because the firm can no longer appropriate the full gain from reducing rents. Stronger worker bargaining power would therefore reduce the wedge between privately profitable adoption and worker welfare and make good-job outcomes more likely. This also suggests an important role for unions and other institutions of worker representation: by increasing workers' share of surplus and giving them a collective voice over adoption, they may affect not only how the gains from AI are divided, but also whether the economy settles in a good-job or a bad-job equilibrium.

A second implication concerns product-market and demand-side effects. AI may initially raise the value of output by improving quality, allowing especially early adopters to earn higher profits that partly offset weaker incentives. Over time, however, we would argue that competition, diffusion, and entry are likely to erode such a value premium. Quality gains are partly a public margin that rivals can imitate, whereas agency costs remain a private margin that firms continue to bear. We therefore view our mechanism as especially relevant for the medium to long run, once AI capabilities are widely available and the key question becomes how firms organize jobs and incentives around them. A richer model incorporating macroeconomic implications could also generate broader demand-side effects: if AI pushes the economy toward low-rent, low-wage jobs, the resulting decline in labor income may dampen spending. Together with competitive pressure passing some productivity gains through to prices, this may generate deflationary tendencies.

Overall, the paper suggests that the labor-market consequences of AI are not determined solely by whether a task can be technologically automated. They also depend on how the remaining human effort is incentivized once AI makes "good enough" performance cheaper and firms respond by redesigning jobs and compensation. While we have emphasized the case in which workers continue to produce output directly, the same logic also applies more broadly to settings in which

AI generates a first-pass output and human workers remain responsible for checking, refining, and taking responsibility for quality. In such environments, AI may not eliminate the need for labor, but instead transform it into a more oversight- and verification-intensive role whose success still depends on sustaining costly effort, and therefore on incentives.

Finally, our results also indicate that policies play an important role: sustaining a good-job AI equilibrium may require not only better technology, but also labor-market institutions and organizational practices that maintain outside options and preserve the profitability of high-effort work. Without such support, AI may not eliminate human work, but it can still shift the economy toward a bad-job equilibrium of low wages, weak incentives, and standardized human roles.

References

Acemoglu, Daron, "The World Needs a Pro-Human AI Agenda," Project Syndicate November 2024. Accessed 2025-12-21.

__ **and David Autor**, "Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, eds., *Handbook of Labor Economics*, Vol. 4B, Elsevier, 2011, pp. 1043–1171.

__ **and Pascual Restrepo**, "Artificial Intelligence, Automation and Work," SSRN Scholarly Paper ID 3098384, Social Science Research Network, Rochester, NY January 2018.

__ **and** __ , "The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment," *American Economic Review*, 2018, *108* (6), 1488–1542.

__ **and** __ , "Automation and New Tasks: How Technology Displaces and Reinstates Labor," *Journal of Economic Perspectives*, 2019, *33* (2), 3–30.

__ **, Dingwen Kong, and Asuman Ozdaglar**, "AI, Human Cognition and Knowledge Collapse," NBER Working Paper 34910, National Bureau of Economic Research 2026.

__ **, Fredric Kong, and Pascual Restrepo**, "Tasks at Work: Comparative Advantage, Technology and Labor Demand," in "Handbook of Labor Economics," Vol. 6, Elsevier, 2025, pp. 1–114.

Agrawal, Ajay K., John McHale, and Alexander Oettl, "Enhancing Worker Productivity Without Automating Tasks: A Different Approach to AI and the Task-Based Model," NBER Working Paper 34781, National Bureau of Economic Research 2026.

Althoff, Lukas and Hugo Reichardt, "Task-Specific Technical Change and Comparative Advantage," Technical Report January 2026.

Athey, Susan C., Kevin A. Bryan, and Joshua S. Gans, "The Allocation of Decision Authority to Human and Artificial Intelligence," *AEA Papers and Proceedings*, 2020, *110*, 80–84.

Autor, David and Neil Thompson, "Expertise," *Journal of the European Economic Association*, 2025, *23* (4), 1203–1271.

Bárány, Zsófia L. and Miklós Koren, "Broken Ladders: AI, Teamwork, and the Dynamics of Skill Formation in the Workplace," February 2026. Working paper. First version: May 2025.

Braverman, Harry, *Labor and Monopoly Capital: The Degradation of Work in the Twentieth Century*, Monthly Review Press, 1974.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond, "Generative AI at Work," *Quarterly Journal of Economics*, May 2025, *140* (2), 889–942.

Card, David, "Who Set Your Wage?," *American Economic Review*, 2022, *112* (4), 1075–1090.

Challapally, Aditya, Chris Pease, Ramesh Raskar, and Pradyumna Chari, "State of AI in Business 2025 Report," Technical Report 2025. Industry research report.

Chen, Yvonne Jie, Jie Gong, Jin Li, and Zibo Zhao, "Better Technology, Worse Motivation: GenAI's Mediocrity Trap," *Working Paper*, 2025.

Cheng, Xienan, Mustafa Dogan, and Pinar Yildirim, "Artificial Intelligence in Team Dynamics: Who Gets Replaced and Why?," NBER Working Paper 34259, National Bureau of Economic Research September 2025.

Cowgill, Bo, Pablo Hernández-Lagos, and Nataliya Langburd Wright, "Does AI Cheapen Talk? Theory and Evidence From Global Entrepreneurship and Hiring," *IZA Discussion Paper*, 2026, *18442.* Earlier version circulated in 2024 as a Columbia Business School Research Paper.

Cui, Jingyi, Gabriel Dias, and Justin Ye, "Signaling in the Age of AI: Evidence from Cover Letters," *arXiv preprint arXiv:2509.25054*, 2025.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," Working Paper 24-013 2024.

Demirer, Mert, John J. Horton, Nicole Immorlica, Brendan Lucier, and Peyman Shahidi, "Chaining Tasks, Redefining Work: A Theory of AI Automation," NBER Working Paper 34859, National Bureau of Economic Research 2026.

Fahn, Matthias, Roger Chung, Pascal Hua, Jie Gong, George Dickinson, and Jin Li, "Deloitte-HKU AI Adoption Index 2026: The Paradox of Promise and Performance," Technical Report, HKU Centre for AI, Management and Organisation and Deloitte China January 2026. Accessed 2026-04-02.

Federal Republic of Germany, "Works Constitution Act (Betriebsverfassungsgesetz – BetrVG)," Official English translation provided by the Federal Ministry of Labour and Social Affairs (BMAS), published via Gesetze im Internet 2001. Full citation on source page: Works Constitution Act of 25 September 2001 (Federal Law Gazette I, p. 2518), as amended by Article 1 of the Act of 19 July 2024 (Federal Law Gazette 2024 I p. 248). Accessed 2025-12-23.

Freund, Lukas B. and Lukas F. Mann, "Job Transformation, Specialization, and the Labor Market Effects of AI," CESifo Working Paper 12072, CESifo 2026.

Galdin, Anaïs and Jesse Silbert, "Making Talk Cheap: Generative AI and Labor Market Signaling," *arXiv preprint arXiv:2511.08785*, 2025. Job Market Paper.

Garicano, Luis and Luis Rayo, "Training in the Age of AI: A Theory of Career Viability," Technical Report 20634, CEPR Discussion Paper September 2025. Revised March 2026; earlier circulated as "Training in the Age of AI: A Theory of Apprenticeship Viability".

Hartmann, Jochen, Yannick Exner, and Samuel Domdey, "The power of generative marketing: Can generative AI create superhuman visual marketing content?," *International Journal of Research in Marketing*, March 2025, *42* (1), 13–31.

Humlum, Anders and Emilie Vestergaard, "Large Language Models, Small Labor Market Effects," NBER Working Paper 33777 May 2025.

Ide, Enrique, "Automation, AI, and the Intergenerational Transmission of Knowledge," Technical Report 20940, CEPR Discussion Paper December 2025.

__ **and Eduard Talamàs**, "The Turing Valley: How AI Capabilities Shape Labor Income," Technical Report 19415, CEPR Discussion Paper August 2024. Revised January 2026.

__ **and** __ , "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy*, 2025, *133* (12), 3762–3800.

__ **and** __ , "The Impact of AI on Global Knowledge Work," *Journal of Monetary Economics*, 2026, *157*, 103876.

Itoh, Hideshi and Kimiyuki Morita, "The Allocation of Decision Authority in Three-Stage Decision Processes with Applications to Artificial Intelligence in Organizations," December 2025. Working paper, December 15, 2025.

Kanazawa, Kyogo, Daiji Kawaguchi, Hitoshi Shigeoka, and Yasutora Watanabe, "AI, Skill, and Productivity: The Case of Taxi Drivers," *Management Science*, 2025. Forthcoming. Earlier version: NBER Working Paper No. 30612.

Lee, Hao-Ping (Hank), Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson, "The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers," in "Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems" CHI '25 Association for Computing Machinery New York, NY, USA April 2025, pp. 1–22.

Lin, Sijie, "Learning to Prompt: Human Adaptation in Production with Generative AI," Technical Report, Working Paper 2025. Version December 2025.

Maasoum, Seyed Mahdi Hosseini and Guy Lichtinger, "Generative AI, Expertise, and Inequality: A Race Between Productivity and Scarcity," Working Paper, SSRN 2026.

Manning, Alan, "Monopsony in Labor Markets: A Review," *ILR Review*, 2021, *74* (1), 3–26.

Mintzberg, Henry, "Design of Positions: Job Specialization," in "The Structuring of Organizations: A Synthesis of the Research," Englewood Cliffs, NJ: Prentice-Hall, 1979, chapter 4.

Nejad, Kian Abbas, Giuseppe Musillo, Till Wicker, and Niccolò Zaccaria, "Labor Market Signals: The Role of Large Language Models," Working Paper, SSRN 2025.

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *Science*, July 2023, *381* (6654), 187–192.

Prescott, Edward C. and Robert M. Townsend, "General Competitive Analysis in an Economy with Private Information," *International Economic Review*, February 1984, *25* (1), 1–20.

__ and __ , "Pareto Optima and Competitive Equilibria with Adverse Selection and Moral Hazard," *Econometrica*, January 1984, *52* (1), 21–45.

Prescott, Edward Simpson and Robert M. Townsend, "Firms as Clubs in Walrasian Markets with Private Information," *Journal of Political Economy*, August 2006, *114* (4), 644–671.

Raith, Michael, "Competition, Risk, and Managerial Incentives," *American Economic Review*, August 2003, *93* (4), 1425–1436.

République française, "Code du travail: Article L1233-61 (Plan de sauvegarde de l'emploi)," Légifrance (official portal for French law) 2017. Version en vigueur depuis le 22 décembre 2017; modifié par Ordonnance n°2017-1718 du 20 décembre 2017, art. 1. Accessed 2025-12-23.

Schmidt, Klaus M., "Managerial Incentives and Product Market Competition," *The review of economic studies*, 1997, *64* (2), 191–213.

Siddiq, Auyon and Niuniu Zhang, "Human Capital, AI, and Labor Commoditization," June 2026.

A Appendix – General Results

First, we describe the consequences of AI on profits for all values of the outside option.

Proposition A-1 *Suppose* $(1-q)y > c^H - c^L_{AI}$. *Then there exist* $\bar{u}^{A*}$, $\bar{u}^{B*}$ *such that:*

(i) When $\bar{u} \leq \bar{u}^{A*}$, *AI increases the firm's profit.*

(ii) When $\bar{u}^{A*} < \bar{u} < \bar{u}^{B*}$, *AI decreases the firm's profit.*

(iii) When $\bar{u} \geq \bar{u}^{B*}$, *AI increases the firm's profit.*

If $(1-q)y \leq c^H - c^L_{AI}$, *AI increases profits for all* $\bar{u}$.

Profits can decline with AI when outside options fall within an intermediate range. In this range, *organizational production costs increase even though personal effort costs have gone down.* To identify the conditions under which profits rise or fall, we examine all possible moves within or between the regions defined in Lemma 2 and 3:

- Region 1: Bad job (low effort, no rent)
- Region 2: Good job (high effort, rent)
- Region 3: Good job (high effort, no rent).

If outside options are either sufficiently high or sufficiently low so that the system remains in Region 3 or Region 1, then profits rise. In these cases, no rent is paid to the agent either before or after AI, and optimal effort remains unchanged. The only effect is the agent's reduced personal cost of effort, which increases profits. (Note that the latter scenario may not apply if the relevant cutoff values are negative.)

If the system remains in Region 2 (rent is paid both before and after AI), profits fall. The higher agency cost of sustaining high effort dominates, and organizational production costs increase.

Profits also fall when the system transitions from Region 3 to Region 1 (which is possible if $\bar{u}^B \leq \bar{u}^A_{AI}$). In this case, no rent is paid before or after AI, but effort falls because high agency costs make it unprofitable to sustain high effort. The principal still captures the entire surplus, but the overall relationship value declines with quality.

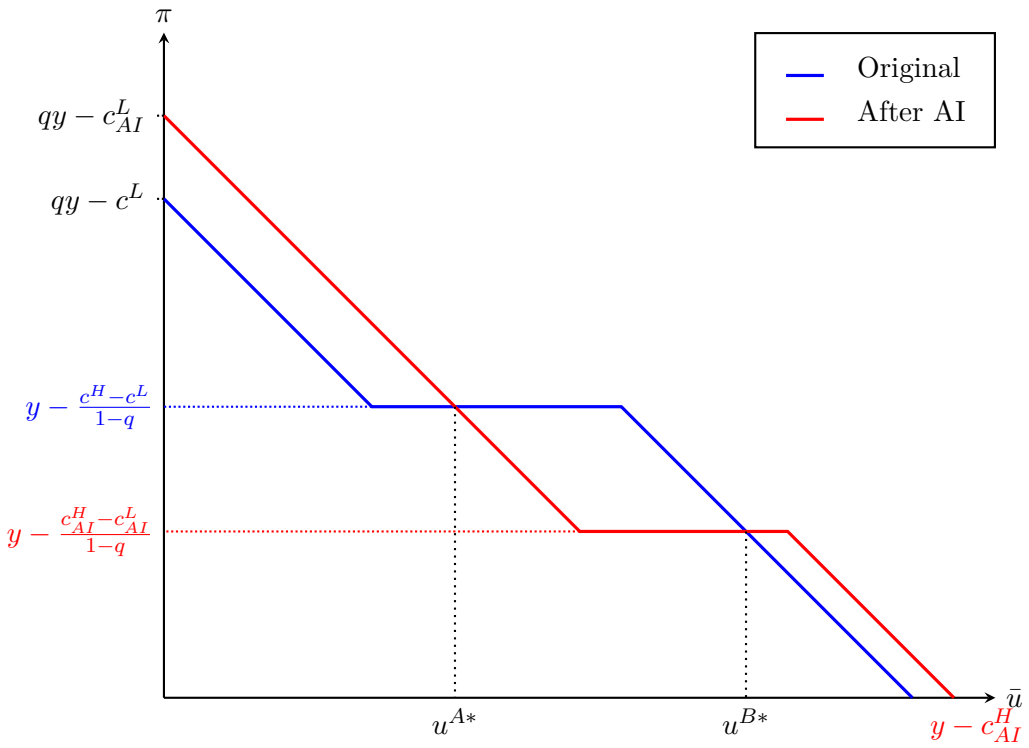
When AI induces a transition from Region 3 to Region 2 or from Region 2 to Region 1, the effect on profits can go in either direction. A move from Region 3 to Region 2 introduces rents for the agent. This reduces profits, but lower effort costs raise them. Profits increase if the outside option exceeds the cutoff $\bar{u}^{B*}$, and fall if it does not. A move from Region 2 to Region 1 eliminates agent rents, which raises profits, but lowers implemented quality, which reduces them. Profits increase if the outside option is below the cutoff $\bar{u}^{A*}$, and decrease otherwise.

Note that the relative position of the cutoff values depends on whether $\bar{u}^A_{AI} < \bar{u}^B$ or $\bar{u}^A_{AI} > \bar{u}^B$. In the former case, $\bar{u}^{B*}$ lies between $\bar{u}^B$ and $\bar{u}^B_{AI}$, while $\bar{u}^{A*}$ lies between $\bar{u}^A$ and $\bar{u}^A_{AI}$. In the latter case, $\bar{u}^{B*}$ lies between $\bar{u}^A_{AI}$ and $\bar{u}^B_{AI}$, while $\bar{u}^{A*}$ lies between $\bar{u}^A$ and $\bar{u}^B$.

If $(1-q)y \leq c^H - c^L_{AI}$, profits increase for all values of $\bar{u}$. Since we also assume $(1-q)y \geq c^H_{AI} - c^L_{AI}$, this case corresponds to situations with relatively small y and a highly effective AI.

Figure A-1 displays profits with and without AI, for the case in which a profit reduction happens for intermediate values of the outside option.

Figure A-1: Firm Profit Function with and without AI



Utility Next, we show how AI affects the agent's utility for all values of his outside option.

Proposition A-2 *The following holds:*

(i) When $\bar{u} \leq \bar{u}^A_{AI}$, the worker's utility weakly decreases with AI

(ii) When $\bar{u} > \bar{u}^A_{AI}$, the worker's utility weakly increases with AI.

Generally, the agent's utility either increases or remains unaffected by AI, except in the case where the agent received a rent prior to AI but the introduction of AI makes high effort unprofitable. In that case, the principal switches to low effort, and the agent's payoff falls to the outside option.

A.1 Attitudes Towards AI

Here, we show when interests between firm and worker are aligned for all $\bar{u}$

Corollary A-1 *The following holds:*

1. ***Win-Win:*** *When* $\bar{u} \in \left(\bar{u}^{B*}, \bar{u}^{B}_{AI}\right)$*, AI strictly increases the firm's profit and the agent's utility.*

2. ***Lose-Lose:***

   (a) *Suppose* $(1-q)y > c^H - c^L_{AI}$ *and* $\bar{u}^B > \bar{u}^A_{AI}$*: When* $\bar{u} \in \left(\bar{u}^{A*}, \bar{u}^{A}_{AI}\right)$*, AI strictly reduces the firm's profit and the agent's utility.*

   (b) *Suppose* $(1-q)y > c^H - c^L_{AI}$ *and* $\bar{u}^B \leq \bar{u}^A_{AI}$*: When* $\bar{u} \in \left(\bar{u}^{A*}, \bar{u}^{B}\right)$*, AI strictly reduces the firm's profit and the agent's utility.*

For relatively high outside options, interests are fully aligned: the relationship moves from a good job without rent to a good job with rent, and the productivity gain from the reduction in c^H_{AI} dominates the loss from reduced rent extraction. Note that even for $\bar{u} > \bar{u}^B_{AI}$, i.e., to the right of the win–win region, the agent is not worse off, since his utility remains at $\bar{u}$. Thus, whenever AI adoption raises the principal's profits, the agent is at least weakly better off as well.

As laid out in the main part, both parties are worse off when the relationship switches from a good job with rents to a bad job, and the loss in value from lower effort outweighs both the full rent extraction in the bad job and the reduction in low-effort cost c^L_{AI}.[4]

Next, we describe when conflicts of interests emerge.

Corollary A-2 *The following holds:*

1. ***Pro-Worker:*** *Suppose* $(1-q)y > c^H - c^L_{AI}$*. When* $\bar{u} \in \left(\bar{u}^{A}_{AI}, \bar{u}^{B*}\right)$*, AI reduces the firm's profit but increases the agent's utility.*

2. ***Pro-Firm:***

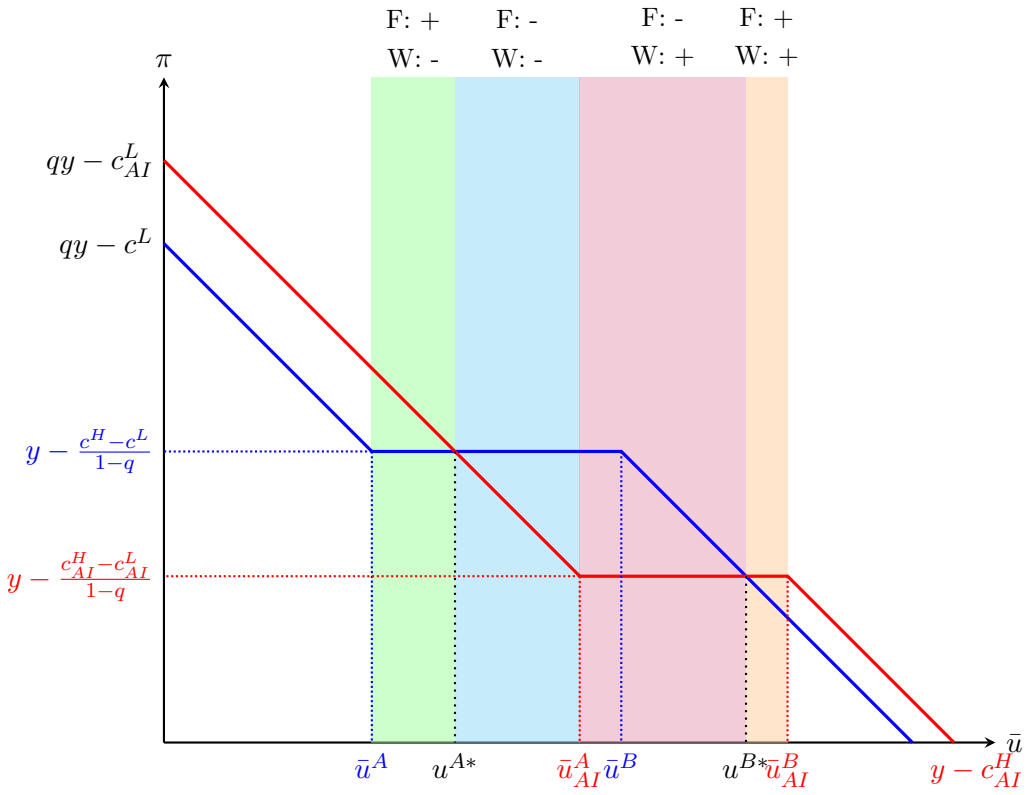
   (a) *Suppose* $(1-q)y > c^H - c^L_{AI}$*: When* $\bar{u} \in \left(\bar{u}^{A}, \bar{u}^{A*}\right)$*, AI strictly increases the firm's profit and reduces the agent's utility. Within this range the profit gain is smaller than the utility loss.*

   (b) *Suppose* $(1-q)y \leq c^H - c^L_{AI}$*: When* $\bar{u} \in \left(\bar{u}^{A}, \bar{u}^{B}\right)$*, AI strictly increases the firm's profit and reduces the agent's utility. Within this range the profit gain exceeds the utility loss.*

Figure A-2 displays all potential cases stated in the Corollaries.

[4] Note that when $(1-q)y \leq c^H - c^L_{AI}$, this region is empty, because in that case AI increases profits for all values of the outside option.

Figure A-2: Firm Profit Function with and without AI.



A.2 Effectiveness of AI: Comparative Statics Results

We now provide a systematic analysis of how the effectiveness of AI—i.e., the differential effects of reductions in c^H_{AI} and c^L_{AI}—shapes equilibrium outcomes.

Our first lemma investigates how reductions in c^H_{AI} versus c^L_{AI} change the sign and magnitude of AI's impact on equilibrium outcomes. Using profits as an example, we analyze how the relative sizes of c^H_{AI} and c^L_{AI} affect $\Delta\pi \equiv \pi^{AI} - \pi^{noAI}$, and delineate the regions in which making AI "better" renders $\Delta\pi$ positive, negative, or unchanged. The same logic extends to the utility and wages.

Lemma A-1 *For $\bar{u} > \bar{u}^A_{AI}$, a smaller c^H_{AI} enhances the impact of AI on profits but worsens its effects on utilities (for utilities no effect for $\bar{u} > \bar{u}^B_{AI}$). For $\bar{u} \leq \bar{u}^A_{AI}$, c^H_{AI} does not affect AI's impact on payoffs.*

For $\bar{u} > \bar{u}^B_{AI}$, a smaller c^L_{AI} has no effect on payoffs. For $\bar{u} \in \left(\bar{u}^A_{AI}, \bar{u}^B_{AI}\right)$, a smaller c^L_{AI} worsens the effects of AI on profits, but enhances its effects on utilities. For $\bar{u} \leq \bar{u}^A_{AI}$, a smaller c^L_{AI} enhances the effects of AI on profits, while there is no effect on utilities.

In general, improvements in high-effort AI are beneficial for profits—but only when the agent's outside option is sufficiently high to keep high effort optimal after AI adoption. By contrast, improvements in low-effort AI (a lower c^L_{AI}) leave profits unchanged if high effort remains optimal post-AI and the agent does not earn a rent. If high effort remains optimal but the agent is granted

a rent, profits fall because the agent captures the efficiency gains. If low effort is optimal post-AI, however, profits rise, as production costs fall without the need to pay rents.

Next, we are interested in how changes in c_{AI}^H and c_{AI}^L affect the size of the ranges in which effort, profits, and utilities move up or down. These comparative-statics results clarify not only the direction of AI's impact but also the breadth of situations in which different adoption outcomes arise.

All of these findings are summarized in the following Lemma.

Lemma A-2 *(i)* ***Effort.*** *AI reduces effort for outside options $\bar{u} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right)$. The length of this range shrinks when c_{AI}^H falls and expands when c_{AI}^L falls.*

(ii) ***Profits.*** *If $(1-q)y > c^H - c_{AI}^L$, AI reduces profits for $\bar{u} \in \left(\bar{u}^{A*},\, \bar{u}^{B*}\right)$. The length of this range shrinks when c_{AI}^H falls and expands when c_{AI}^L falls.*

(iii) ***Utility.*** *AI increases the agent's utility for $\bar{u} \in \left(\bar{u}_{AI}^A,\, \bar{u}_{AI}^B\right)$. The length of this range expands when c_{AI}^H falls and shrinks when c_{AI}^L falls.*

- *(a) For $\bar{u}^B > \bar{u}_{AI}^A$, AI decreases the utility for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right)$. The length of this range shrinks when c_{AI}^H falls and expands when c_{AI}^L falls.*
- *(b) For $\bar{u}^B \leq \bar{u}_{AI}^A$, AI decreases the utility for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right)$. The length of this range is unaffected by c_{AI}^H or c_{AI}^L.*

Online Appendix – Omitted Proofs

Proof to Lemma 3

The value of the cutoffs,

$$\begin{aligned}\bar{u}^B_{AI} =& \frac{qc^H_{AI} - c^L_{AI}}{1-q}\\ \bar{u}^A_{AI} =& \frac{c^H_{AI} - c^L_{AI}\,(2-q)}{1-q} - (1-q)\,y,\end{aligned}$$

follows immediately, as well as that

$$\bar{u}^B_{AI} - \bar{u}^B = \frac{q\left(c^H_{AI} - c^H\right) + c^L - c^L_{AI}}{1-q} > \frac{\left(c^H_{AI} - c^H\right) + c^L - c^L_{AI}}{1-q} > 0$$
$$\bar{u}^A_{AI} - \bar{u}^A = \frac{c^H_{AI} - c^L_{AI} - c^H + c^L + \left(c^L - c^L_{AI}\right)(1-q)}{1-q} > 0.$$

Proof to Lemma 4

Profits with high effort are $\pi^H = y - \max\left\{\frac{\left(c^H - c^L\right)}{1-q},\, \bar{u} + c^H\right\}$ before AI, and $\pi^H_{AI} = y - \max\left\{\frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q},\, \bar{u} + c^H_{AI}\right\}$ after AI, and the cutoff $\bar{u}^B = \frac{\left(c^H - c^L\right)}{1-q} - c^H$ increases with AI. For the profit difference $\Delta\pi^H \equiv \pi^H_{AI} - \pi^H$, we thus have the following cases:

- $\bar{u} \geq \bar{u}^B_{AI}$: $\Delta\pi^H = c^H - c^H_{AI} > 0$
- $\bar{u} \in \left(\bar{u}^B,\, \bar{u}^B_{AI}\right)$: $\Delta\pi^H = \bar{u} + c^H - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q}$
  - For $\bar{u} \to \bar{u}^B_{AI}$, $\Delta\pi^H = c^H - c^H_{AI} > 0$
  - For $\bar{u} \to \bar{u}^B$, $\Delta\pi^H = \frac{\left(c^H - c^L\right) - \left(c^H_{AI} - c^L_{AI}\right)}{1-q} < 0$
- $\bar{u} \leq \bar{u}^B$: $\Delta\pi^H = \frac{\left(c^H - c^L\right) - \left(c^H_{AI} - c^L_{AI}\right)}{1-q} < 0.$

It follows that profits increase for $\bar{u} > \bar{u}^{B*} \equiv \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} - c^H \in \left(\bar{u}^B,\, \bar{u}^B_{AI}\right)$, and decrease for $\bar{u} \leq \bar{u}^{B*}$.

Proof to Propositions 1, 2, A-1, and A-2 (Complete Characterization of Outcomes With and Without AI)

Here, we compare the situations with and without AI, and derive the consequences for AI on profits, as well as wages and the agent's utility. Before doing that, let us first collect the outcomes for all ranges, separately for the cases without and with AI.

Outcomes

Without AI

1. $\bar{u} \geq \bar{u}^B$

   $e = H$

   $w = \bar{u} + c^H$

   $\pi = y - w = y - \bar{u} - c^H$

   $u = w - c^H = \bar{u}$

2. $\bar{u} \in \left[\bar{u}^A,\, \bar{u}^B\right)$

   $e = H$

   $w = \frac{\left(c^H - c^L\right)}{1-q}$

   $\pi = y - w = y - \frac{\left(c^H - c^L\right)}{1-q} > 0$

   $u = w - c^H = \frac{qc^H - c^L}{1-q} > \bar{u}$

3. $\bar{u} < \bar{u}^A$

   $e = L$

   $w = \bar{u} + c^L$

   $\pi = qy - w = qy - \bar{u} - c^L > y - \frac{c^H - c^L}{1-q} > 0$

   $u = w - c^L = \bar{u}$

With AI

1. $\bar{u} \geq \bar{u}^B_{AI}$

   $e = H$

   $w = \bar{u} + c^H_{AI}$

   $\pi = y - w = y - \bar{u} - c^H_{AI}$

   $u = w - c^H_{AI} = \bar{u}$

2. $\bar{u} \in \left[\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right)$

   $e = H$

   $w = \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q}$

   $\pi = y - w = y - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} > 0$

   $u = w - c^H_{AI} = \frac{qc^H_{AI} - c^L_{AI}}{1-q} > \bar{u}$

3. $\bar{u} < \bar{u}^A_{AI}$

   $e = L$

   $w = \bar{u} + c^L_{AI}$

   $\pi = qy - w = qy - \bar{u} - c^L_{AI} > y - \frac{c^H_{AI} - c^L_{AI}}{1-q} > 0$

   $u = w - c^L_{AI} = \bar{u}$

Consequences of AI

We now describe the consequences of AI, as a function of the agent's outside option. Recall the values of the cutoffs,

$$\begin{aligned}
\bar{u}^B_{AI} =& \frac{qc^H_{AI} - c^L_{AI}}{1-q} \\
\bar{u}^A_{AI} =& \frac{c^H_{AI} - c^L_{AI}(2-q)}{1-q} - (1-q)\,y \\
\bar{u}^B =& \frac{qc^H - c^L}{1-q} \\
\bar{u}^A =& \frac{c^H - c^L(2-q)}{1-q} - (1-q)\,y
\end{aligned}$$

It follows that $\bar{u}^B_{AI} > \bar{u}^B$, $\bar{u}^A_{AI} > \bar{u}^A$, $\bar{u}^B_{AI} > \bar{u}^A_{AI}$, and $\bar{u}^B > \bar{u}^A$. The only ambiguity holds for $\bar{u}^B$ vs. $\bar{u}^A_{AI}$, and we have to distinguish between the cases $\bar{u}^B > \bar{u}^A_{AI}$ and $\bar{u}^B \leq \bar{u}^A_{AI}$. There, note that

$$\begin{aligned}
&\bar{u}^B > \bar{u}^A_{AI} \\
\Leftrightarrow &(1-q)^2\,y + qc^H - c^L > c^H_{AI} - c^L_{AI} - c^L_{AI}(1-q)\,. \\
\Leftrightarrow &\left[(1-q)\,y - c^H + c^L_{AI}\right](1-q) > c^H_{AI} - c^L_{AI} - \left(c^H - c^L\right).
\end{aligned}$$

We have assumed that $c^H_{AI} - c^L_{AI} \in \left(c^H - c^L,\, c^H\right]$. Thus, if this condition is satisfied for $c^H_{AI} - c^L_{AI} = c^H$ (i.e., $c^H_{AI} = c^H$ and $c^L_{AI} = 0$), it holds for all AI levels. If this condition is violated for $c^H_{AI} - c^L_{AI} \to c^H - c^L$, it holds for no AI

This gives the following three cases:

A)

$$c^H \leq (1-q)\,y + c^L_{AI} - \frac{c^L}{(1-q)},$$

then $\bar{u}^B \geq \bar{u}^A_{AI}$ for all c^{AI}.

B)

$$(1-q)\,y + c^L_{AI} - \frac{c^L}{(1-q)} < c^H < (1-q)\,y + c^L_{AI},$$

then there exists some Δc^{AI}, such that

B.1) $c_{AI}^H - c_{AI}^L - \left(c^H - c^L\right) \leq \Delta c^{AI}$: $\bar{u}^B \geq \bar{u}_{AI}^A$

B.2) $c_{AI}^H - c_{AI}^L - \left(c^H - c^L\right) > \Delta c^{AI}$: $\bar{u}^B < \bar{u}_{AI}^A$

C)

$$c^H \geq (1-q)\, y + c_{AI}^L,$$

then $\bar{u}^B < \bar{u}_{AI}^A$ for all c^{AI}. However, note that this constrains c_{AI}^H, since $c_{AI}^H - c_{AI}^L < (1-q)y$ holds by assumption.

In the following, we analyze these cases separately, defining situations a)-e) to describe the regions before and after the introduction of AI. However, since $\bar{u}_{AI}^B$ is the highest cutoff in each case, we have

a) Situation $\bar{u} > \bar{u}_{AI}^B$: Remain in region 3 (good job, no rent)

- High effort remains optimal
- $\Delta w = c_{AI}^H - c^H < 0$
- $\Delta \pi = c^H - c_{AI}^H > 0$
- $\Delta u = 0$

CASES A) and B.1): $\bar{u}^B > \bar{u}_{AI}^A$**:** $(1-q)\, y > \frac{c_{AI}^H - q c^H + c^L - c_{AI}^L (2-q)}{(1-q)}$ We also have $c^H - c_{AI}^L < (1-q)\, y$

b) Situation $\bar{u} \in \left(\bar{u}^B,\, \bar{u}_{AI}^B\right]$: Move from region 3 to 2 (good job, no rent → good job, rent)

- High effort remains optimal
- $\Delta w = \frac{\left(c_{AI}^H - c_{AI}^L\right)}{1-q} - \left(\bar{u} + c^H\right) \gtrless 0$
- $\Delta \pi = c^H - \frac{\left(c_{AI}^H - c_{AI}^L\right)}{1-q} + \bar{u} \gtrless 0$
  $\bar{u} = \bar{u}_{AI}^B$: $\Delta \pi = c^H - c_{AI}^H > 0$
  $\bar{u} = \bar{u}^B$: $\Delta \pi = \frac{c^H - c^L - \left(c_{AI}^H - c_{AI}^L\right)}{1-q} < 0$
- $\Delta u = \frac{q c_{AI}^H - c_{AI}^L}{1-q} - \bar{u} > \frac{q c_{AI}^H - c_{AI}^L}{1-q} - \bar{u}_{AI}^B = 0$

c) Situation $\bar{u} \in \left(\bar{u}_{AI}^A,\, \bar{u}^B\right]$: Remain in region 2 (good job, rent)

- High effort remains optimal
- $\Delta w = \frac{\left(c_{AI}^H - c_{AI}^L\right) - \left(c^H - c^L\right)}{1-q} > 0$
- $\Delta \pi = \frac{\left(c^H - c^L\right)}{1-q} - \frac{\left(c_{AI}^H - c_{AI}^L\right)}{1-q} < 0$
- $\Delta u = \frac{q\left(c_{AI}^H - c^H\right) + c^L - c_{AI}^L}{1-q} \geq \frac{c_{AI}^H - c^H + c^L - c_{AI}^L}{1-q} > 0$

d) Situation $\bar{u} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right]$: Shift from region 2 to 1 (good job, rent → bad job)

- Effort goes down
- $\Delta w = \bar{u} + c_{AI}^L - \frac{\left(c^H - c^L\right)}{1-q} < 0$, since $(1-q)\, y > \frac{c_{AI}^H - q c^H + c^L - c_{AI}^L (2-q)}{(1-q)}$

$$
\begin{aligned}
& \bar{u} + c_{AI}^L - \frac{\left(c^H - c^L\right)}{1-q} \\
<& \bar{u}_{AI}^A + c_{AI}^L - \frac{\left(c^H - c^L\right)}{1-q} \\
=& \frac{c_{AI}^H - c_{AI}^L - \left(c^H - c^L\right)}{1-q} - (1-q)\, y \\
<& \frac{c_{AI}^H - c_{AI}^L - \left(c^H - c^L\right)}{1-q} - \frac{c_{AI}^H - q c^H + c^L - c_{AI}^L\,(2-q)}{(1-q)} \\
=& - c^H + c_{AI}^L < 0.
\end{aligned}
$$

- $\Delta \pi = \frac{\left(c^H - c^L\right)}{1-q} - c_{AI}^L - (1-q)\, y - \bar{u} \gtrless 0$

  $\bar{u}_{AI}^A$: $\Delta \pi = \frac{c^H - c^L - \left(c_{AI}^H - c_{AI}^L\right)}{1-q} < 0$

  $\bar{u}^A$: $\Delta \pi = c^L - c_{AI}^L > 0$
- $\Delta u = \bar{u} - \frac{q c^H - c^L}{1-q} = \bar{u} - \bar{u}^B < 0$
- $\Delta \pi + \Delta u = c^H - c_{AI}^L - (1-q)\, y < 0$, thus it is not profitable to compensate the agent

e) Situation $\bar{u} \leq \bar{u}^A$: Remain in region 1 (bad job)

- Low effort remains optimal
- $\Delta w = c_{AI}^L - c^L < 0$
- $\Delta \pi = c^L - c_{AI}^L > 0$
- $\Delta u = 0$

Conclusion:

- There are 2 cutoffs, $\bar{u}^{A*} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right)$ and $\bar{u}^{B*} \in \left(\bar{u}^B,\, \bar{u}_{AI}^B\right)$, with $\bar{u}^{B*} > \bar{u}^{A*}$, such that:
  - $\bar{u} > \bar{u}^{B*}$: $\Delta \pi > 0$, $\Delta w < 0$
  - $\bar{u}^{A*} < \bar{u} < \bar{u}^{B*}$: $\Delta \pi < 0$
    - $\bar{u}_{AI}^A < \bar{u} < \bar{u}^{B*}$: $\Delta w > 0$
    - $\bar{u}^{A*} < \bar{u} \leq \bar{u}_{AI}^A$: $\Delta w < 0$
  - $\bar{u} < \bar{u}^{A*}$: $\Delta \pi > 0$, $\Delta w < 0$
- Utility
  - remains unaffected for $\bar{u} > \bar{u}_{AI}^B$, even though wage goes down.

- increases for $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right]$, even though w goes down for $\bar{u} \in \left(\bar{u}^{B*},\, \bar{u}^B_{AI}\right]$, it goes up for $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^{B*}\right)$
- decreases for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^A_{AI}\right]$, wage goes down
- remains unaffected for $\bar{u} \leq \bar{u}^A$, even though wage goes down.

CASE B2): $\bar{u}^B \leq \bar{u}^A_{AI}$: $\bar{u}^A_{AI} \geq \frac{qc^H - c^L}{1-q}$, and $(1-q)\,y > c^H - c^L_{AI}$

b) Situation $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right]$: Move from region 3 to 2 (good job, no rent → good job, rent)

- High effort remains optimal
- $\Delta w = \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} - \left(\bar{u} + c^H\right) \gtrless 0$
- $\Delta \pi = \bar{u} + c^H - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} \gtrless 0$

  $\bar{u} = \bar{u}^B_{AI}$: $\Delta \pi = c^H - c^H_{AI} > 0$

  $\bar{u} = \bar{u}^A_{AI}$: $\Delta \pi = c^H - c^L_{AI} - (1-q)\,y < 0$
- $\Delta u = \frac{qc^H_{AI} - c^L_{AI}}{1-q} - \bar{u} > \frac{qc^H_{AI} - c^L_{AI}}{1-q} - \bar{u}^B_{AI} = 0$

c) Situation $\bar{u} \in \left(\bar{u}^B,\, \bar{u}^A_{AI}\right]$: Move from region 3 to 1 (good job, no rent → bad job)

- Effort goes down
- $\Delta w = c^L_{AI} - c^H < 0$
- $\Delta \pi = c^H - c^L_{AI} - (1-q)\,y < 0$
- $\Delta u = 0$

d) Situation $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right]$: Move from region 2 to 1 (good job, rent → bad job)

- Effort goes down
- $\Delta w = \bar{u} + c^L_{AI} - \frac{\left(c^H - c^L\right)}{1-q} \leq \bar{u}^B + c^L_{AI} - \frac{\left(c^H - c^L\right)}{1-q} = c^L_{AI} - c^H < 0$
- $\Delta \pi = \frac{\left(c^H - c^L\right)}{1-q} - c^L_{AI} - (1-q)\,y - \bar{u} \gtrless 0$

  $\bar{u}^B$: $\Delta \pi = c^H - c^L_{AI} - (1-q)\,y < 0$

  $\bar{u}^A$: $\Delta \pi = c^L - c^L_{AI} > 0$
- $\Delta u = \bar{u} - \frac{qc^H - c^L}{1-q} = \bar{u} - \bar{u}^B < 0$
- $\Delta \pi + \Delta u = c^H - c^L_{AI} - (1-q)\,y < 0$, thus it is not profitable to compensate the agent

e) Situation $\bar{u} \leq \bar{u}^A$: Remain in region 1 (bad job)

- Low effort remains optimal
- $\Delta w = c^L_{AI} - c^L < 0$
- $\Delta \pi = c^L - c^L_{AI} > 0$

- $\Delta u = 0$

Conclusion:

- There are 2 cutoffs, $\bar{u}^{A*} \in \left(\bar{u}^A,\, \bar{u}^B\right)$ and $\bar{u}^{B*} \in \left(\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right)$, with $\bar{u}^{B*} > \bar{u}^{A*}$, such that:
  - $\bar{u} > \bar{u}^{B*}$: $\Delta\pi > 0$, $\Delta w < 0$
  - $\bar{u}^{A*} < \bar{u} < \bar{u}^{B*}$: $\Delta\pi < 0$
    - $\bar{u}^A_{AI} < \bar{u} < \bar{u}^{B*}$: $\Delta w > 0$
    - $\bar{u}^{A*} < \bar{u} \leq \bar{u}^A_{AI}$: $\Delta w < 0$
  - $\bar{u} < \bar{u}^{A*}$: $\Delta\pi > 0$, $\Delta w < 0$
- Utility
  - remains unaffected for $\bar{u} > \bar{u}^B_{AI}$, even though wage goes down.
  - increases for $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right]$, even though w goes down for $\bar{u} \in \left(\bar{u}^{B*},\, \bar{u}^B_{AI}\right]$, it goes up for $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^{B*}\right)$
  - remains unaffected for $\bar{u} \in \left(\bar{u}^B,\, \bar{u}^A_{AI}\right]$, even though wage goes down.
  - decreases for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right]$, wage goes down
  - remains unaffected for $\bar{u} \leq \bar{u}^A$, even though wage goes down.

CASE C): $\bar{u}^B \leq \bar{u}^A_{AI}$: $\bar{u}^A_{AI} \geq \frac{qc^H - c^L}{1-q}$, and $(1-q)\,y \leq c^H - c^L_{AI}$

b) Situation $\bar{u} \in \left(\bar{u}^A_{AI},\, \bar{u}^B_{AI}\right]$: Move from region 3 to 2 (good job, no rent → good job, rent)

- High effort remains optimal
- $\Delta w = \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} - \left(\bar{u} + c^H\right) < 0$
- $\Delta\pi = \bar{u} + c^H - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} > 0$, since

$$\begin{aligned} &\bar{u} + c^H - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} \\ >&\bar{u}^A_{AI} + c^H - \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} \\ =&c^H - c^L_{AI} - (1-q)\,y \geq 0 \end{aligned}$$

- $\Delta u = \frac{qc^H_{AI} - c^L_{AI}}{1-q} - \bar{u} > \frac{qc^H_{AI} - c^L_{AI}}{1-q} - \bar{u}^B_{AI} = 0$

c) Situation $\bar{u} \in \left(\bar{u}^B,\, \bar{u}^A_{AI}\right]$: Move from region 3 to 1 (good job, no rent → bad job)

- Effort goes down

- $\Delta w = c_{AI}^L - c^H < 0$
- $\Delta \pi = c^H - c_{AI}^L - (1-q)\, y > 0$
- $\Delta u = 0$

d) Situation $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right]$: Move from region 2 to 1 (good job, rent $\rightarrow$ bad job)

- Effort goes down
- $\Delta w = \bar{u} + c_{AI}^L - \frac{\left(c^H - c^L\right)}{1-q} \leq \bar{u}^B + c_{AI}^L - \frac{\left(c^H - c^L\right)}{1-q} = c_{AI}^L - c^H < 0$
- $\Delta \pi = \frac{\left(c^H - c^L\right)}{1-q} - c_{AI}^L - (1-q)\, y - \bar{u} > 0$, since

$$\begin{aligned} & \frac{\left(c^H - c^L\right)}{1-q} - c_{AI}^L - (1-q)\, y - \bar{u} \\ > & \frac{\left(c^H - c^L\right)}{1-q} - c_{AI}^L - (1-q)\, y - \bar{u}^B \\ = & c^H - c_{AI}^L - (1-q)\, y > 0 \end{aligned}$$

- $\Delta u = \bar{u} - \frac{q c^H - c^L}{1-q} = \bar{u} - \bar{u}^B < 0$
- $\Delta \pi + \Delta u = c^H - c_{AI}^L - (1-q)\, y > 0$, thus possible to compensate the agent.

e) Situation $\bar{u} \leq \bar{u}^A$: Remain in region 1 (bad job)

- Low effort remains optimal
- $\Delta w = c_{AI}^L - c^L < 0$
- $\Delta \pi = c^L - c_{AI}^L > 0$
- $\Delta u = 0$

Conclusion:

- Profits go up, wage goes down for all $\bar{u}$
- Utility
  - remains unaffected for $\bar{u} > \bar{u}_{AI}^B$, even though wage goes down.
  - increases for $\bar{u} \in \left(\bar{u}_{AI}^A,\, \bar{u}_{AI}^B\right]$, even though w goes down
  - remains unaffected for $\bar{u} \in \left(\bar{u}^B,\, \bar{u}_{AI}^A\right]$, even though wage goes down.
  - decreases for $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right]$, wage goes down
  - remains unaffected for $\bar{u} \leq \bar{u}^A$, even though wage goes down.

Q.E.D.

Proof of Corollary A-2.

Suppose $(1-q)y > c^H - c^L_{AI}$. In Proposition A-1, we have shown that AI increases profits unless the agent's outside option is in the "middle range" $\bar{u} \in \left(\bar{u}^{A*}, \bar{u}^{B*}\right)$, with $\bar{u}^{A*} = \frac{\left(c^H - c^L\right)}{1-q} - c^L_{AI} - (1-q)\, y$ and $\bar{u}^{B*} = \frac{\left(c^H_{AI} - c^L_{AI}\right)}{1-q} - c^H$. In Lemma A-2, we have demonstrated that AI decreases the agent's utility for $\bar{u} \in \left(\bar{u}^A, \bar{u}^A_{AI}\right]$ (if $\bar{u}^B > \bar{u}^A_{AI}$) or for $\bar{u} \in \left(\bar{u}^A, \bar{u}^B\right]$ (if $\bar{u}^B \leq \bar{u}^A_{AI}$). Since $\bar{u}^{B*} > \bar{u}^A_{AI}$ and $\bar{u}^{B*} > \bar{u}^B$, the effect of AI on the agent's utility is non-negative at and beyond the upper bound of the "middle range." However, since $\bar{u}^A < \bar{u}^{A*}$, the effect of AI on the agent's utility is negative at and in the left neighborhood of the lower bound of the "middle range." Thus, in the range $\left(\bar{u}^A, \bar{u}^{A*}\right)$, AI increases profits but decreases utility, with $\bar{u}^{A*} - \bar{u}^A = c^L - c^L_{AI}$. In this range, $\Delta\pi + \Delta u = c^H - c^L_{AI} - (1-q)\, y < 0$.

The results for $c^H - c^L_{AI} - (1-q)\, y \geq 0$ follow immediately from Proposition A-1 and Lemma A-2

Q.E.D.

Proof of Lemma A-1.

Here, we compute the partial derivatives of Δw, $\Delta\pi$, Δu with respect to c^H_{AI} and c^L_{AI} for all potential cases, as derived in the proof to Proposition A-1 and Lemma A-2.

a) Situation $\bar{u} > \bar{u}^B_{AI}$: Remain in region 3

- $\frac{\partial \Delta\pi}{\partial c^H_{AI}} = -1 < 0$
- $\frac{\partial \Delta\pi}{\partial c^L_{AI}} = 0$
- $\frac{\partial \Delta u}{\partial c^H_{AI}} = \frac{\partial \Delta u}{\partial c^L_{AI}} = 0$

CASES A) and B.1): $\bar{u}^B > \bar{u}^A_{AI}$

b) Situation $\bar{u} \in \left(\bar{u}^B, \bar{u}^B_{AI}\right]$: Move from region 3 to 2

- $\frac{\partial \Delta\pi}{\partial c^H_{AI}} = -\frac{1}{1-q} < 0$
- $\frac{\partial \Delta\pi}{\partial c^L_{AI}} = \frac{1}{1-q} > 0$
- $\frac{\partial \Delta u}{\partial c^H_{AI}} = \frac{q}{1-q} > 0$
- $\frac{\partial \Delta u}{\partial c^L_{AI}} = \frac{-1}{1-q} < 0$

c) Situation $\bar{u} \in \left(\bar{u}^A_{AI}, \bar{u}^B\right]$: Remain in region 2

- $\frac{\partial \Delta\pi}{\partial c^H_{AI}} = -\frac{1}{1-q} < 0$
- $\frac{\partial \Delta\pi}{\partial c^L_{AI}} = \frac{1}{1-q} > 0$
- $\frac{\partial \Delta u}{\partial c^H_{AI}} = \frac{q}{1-q} > 0$

- $\frac{\partial \Delta u}{\partial c_{AI}^L} = \frac{-1}{1-q} < 0$

d) Situation $\bar{u} \in \left(\bar{u}^A,\, \bar{u}_{AI}^A\right]$: Move from region 2 to 1

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = -1 < 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{\partial \Delta u}{\partial c_{AI}^L} = 0$

e) Situation $\bar{u} \leq \bar{u}^A$: Remain in region 1

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = -1 < 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{\partial \Delta u}{\partial c_{AI}^L} = 0$

CASE B2) and Case C: $\bar{u}^B \leq \bar{u}_{AI}^A$

b) Situation $\bar{u} \in \left(\bar{u}_{AI}^A,\, \bar{u}_{AI}^B\right]$: Move from region 3 to 2

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = -\frac{1}{1-q} < 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = \frac{1}{1-q} > 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{q}{1-q} > 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^L} = \frac{-1}{1-q} < 0$

c) Situation $\bar{u} \in \left(\bar{u}^B,\, \bar{u}_{AI}^A\right]$: Move from region 3 to 1

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = -1 < 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{\partial \Delta u}{\partial c_{AI}^L} = 0$

d) Situation $\bar{u} \in \left(\bar{u}^A,\, \bar{u}^B\right]$: Move from region 2 to 1

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = -1 < 0$
- $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{\partial \Delta u}{\partial c_{AI}^L} = 0$

e) Situation $\bar{u} \leq \bar{u}^A$: Remain in region 1

- $\frac{\partial \Delta \pi}{\partial c_{AI}^H} = 0$
- $\frac{\partial \Delta \pi}{\partial c_{AI}^L} = -1 < 0$

– $\frac{\partial \Delta u}{\partial c_{AI}^H} = \frac{\partial \Delta u}{\partial c_{AI}^L} = 0$

To conclude, these results deliver the following links

- Smaller c_{AI}^L and profits: no effect if $\bar{u} > \bar{u}_{AI}^B$, negative effect if $\bar{u} \in \left(\bar{u}_{AI}^A, \bar{u}_{AI}^B\right)$, positive effect if $\bar{u} \leq \bar{u}_{AI}^A$
- Smaller c_{AI}^L and utility: no effect if $\bar{u} > \bar{u}_{AI}^B$, positive effect if $\bar{u} \in \left(\bar{u}_{AI}^A, \bar{u}_{AI}^B\right)$, no effect if $\bar{u} \leq \bar{u}_{AI}^A$
- Smaller c_{AI}^H and profits: positive if $\bar{u} > \bar{u}_{AI}^A$, no effect if $\bar{u} \leq \bar{u}_{AI}^A$
- Smaller c_{AI}^H on utility: No effect if $\bar{u} > \bar{u}_{AI}^B$, negative effect if $\bar{u} \in \left(\bar{u}_{AI}^A, \bar{u}_{AI}^B\right)$, no effect if $\bar{u} \leq \bar{u}_{AI}^A$.

Q.E.D.

Proof of Lemma A-2.

For effort, the statement follows from

$$\begin{aligned}\frac{\partial \bar{u}_{AI}^A}{\partial c_{AI}^H} &= \frac{1}{1-q} > 0 \\ \frac{\partial \bar{u}_{AI}^A}{\partial c_{AI}^L} &= \frac{-(2-q)}{1-q} < 0.\end{aligned}$$

For profits, note that

$$\bar{u}^{B*} - \bar{u}^{A*} = \frac{c_{AI}^H - q c_{AI}^L}{1-q} - c^H - \frac{\left(c^H - c^L\right)}{1-q} + (1-q)\, y,$$

thus

$$\begin{aligned}\frac{\partial\left(\bar{u}^{B*} - \bar{u}^{A*}\right)}{\partial c_{AI}^H} &= \frac{1}{1-q} > 0 \\ \frac{\partial\left(\bar{u}^{B*} - \bar{u}^{A*}\right)}{\partial c_{AI}^L} &= \frac{-q}{1-q} < 0.\end{aligned}$$

For the agent's utility, note that

$$\bar{u}_{AI}^B - \bar{u}_{AI}^A = -c_{AI}^H + c_{AI}^L + (1-q)\, y,$$

thus

$$\frac{\partial\left(\bar{u}_{AI}^{B}-\bar{u}_{AI}^{A}\right)}{\partial c_{AI}^{H}}=-1<0$$
$$\frac{\partial\left(\bar{u}_{AI}^{B}-\bar{u}_{AI}^{A}\right)}{\partial c_{AI}^{L}}=1>0.$$

Moreover,

$$\frac{\partial\left(\bar{u}_{AI}^{A}-\bar{u}^{A}\right)}{\partial c_{AI}^{H}}=\frac{1}{1-q}>0$$
$$\frac{\partial\left(\bar{u}_{AI}^{A}-\bar{u}^{A}\right)}{\partial c_{AI}^{L}}=\frac{-(2-q)}{1-q}<0.$$

Q.E.D.

A Online Appendix – Detailed Derivations of the General Equilibrium

A.1 Value Functions and Equilibrium

Given the matching function, $\mu(U,V)$, the job finding rate for workers as $\frac{\mu(U,V)}{U}=\mu(1,V/U)\equiv f(\theta)$, where $\theta\equiv\frac{V}{U}$ is the labor market tightness. Similarly, we denote the matching rate for a vacancy as $\frac{\mu(U,V)}{V}=f(\theta)/\theta\equiv s(\theta)$. A higher θ means more vacancies per unemployed worker, which leads to a higher job finding rate $f(\theta)$ and a lower vacancy filling rate $s(\theta)$. We consider a stationary equilibrium where all endogenous variables are constant over time. Let M denote the number of matches between workers and firms. Therefore, the number of unemployed workers is $U=L-M$, and the number of vacancies is $V=N-M$. In the stationary equilibrium, the number of matches satisfies

$$\lambda M=\mu(U,V). \tag{A-1}$$

Each match produces a flow profit of π for the firm, which will be determined by the equilibrium contract and output. Denoting the discount rate as r, values of a vacancy and a filled job as Π^V and Π^J, we can write the value functions (Hamilton-Jacobi-Bellman equations) of firms as follows:

$$r\Pi^J=\pi+\lambda(\Pi^V-\Pi^J),$$
$$r\Pi^V=-\kappa+s(\theta)(\Pi^J-\Pi^V).$$

An unemployed worker receives a flow utility of b. For example, b can represent unemployment benefits or the value of leisure. We do not exclude the possibility that b is negative, representing the disutility of being unemployed. However, similar to the case with an exogenous outside option, we assume that b is not too low such that firms can make positive profits. When matched with a

firm, the worker receives a take-it-or-leave-it offer, which takes the form of $w(Y)$, where Y is the output produced by the match. We assume that the firm commits to the same contract until the match is separated and do not allow for renegotiation or dynamic contracts. The flow utility of an employed worker exerting effort e is therefore $w(Y|e) - c(e)$. Denoting the optimal effort choice as e^*, the values of being unemployed (V^U) and employed V^E can be written as follows:

$$rV^U = b + f(\theta)(V^E - V^U),$$
$$rV^E = \mathbb{E}[w(Y)|e^*] - c(e^*) + \lambda(V^U - V^E).$$

The worker's participation constraint requires that $V^E \geq V^U$. Combining with the second equation above, it is equivalent to $\mathbb{E}[w(Y)|e^*] - c(e^*) \geq rV^U$. Compared to the model with an exogenous outside option, the worker's outside option $\bar{u}$ is now endogenized as the flow value of unemployment rV^U. This is the only difference from the baseline contracting problem described in Section 2. The incentive compatiblity constraint remains unchanged.

A general equilibrium of the model is defined as follows:

Definition 1 (General Equilibrium) *A general equilibrium consists of the number of matches M, vacancies V, unemployed workers U, firms N, contracts $w(Y)$, effort level e^*, and values Π^J, Π^V, V^E, V^U such that*

1. *The contract $w(Y)$ and effort level e^* solve the contracting problem described in section 2 with the endogenous outside option rV^U.*
2. *The values Π^J, Π^V, V^E, V^U satisfy the HJB equations above.*
3. *The number of matches satisfies $M = L - U = N - V$.*
4. *The labor market clearing condition (A-1) holds.*
5. *The free entry condition holds: $\Pi^V = 0$.*

A.2 Equilibrium without AI

We are now ready to characterize the equilibrium without AI in the following proposition.

Proposition A-3 (Equilibrium under random matching) *There exist values $b^A > \underline{b}$ such that:*

(i) When $b \leq \underline{b}$, firms implements low effort $e^* = L$. Workers receive no rent and are indifferent between employment and unemployment.

(ii) When $b \geq b^A$, firms implement high effort $e^* = H$.

(iii) When $b \in (\underline{b}, b^A)$, there are multiple equilibria:

*(a) **bad jobs only**: no firm implements effort, and workers earn no rent.*

(b) ***good jobs only***: *all firms implement high effort, and workers earn rents.*

(c) ***mixed***: *a fraction of firms implement high effort, and the rest implement low effort. Firms are indifferent between implementing high and low efforts. Workers earn rents from the high-effort firms and are indifferent between employment and unemployment when matched to low-effort firms.*

The equilibria can be ranked by firm profits/values and total employment M as follows:

$$\textit{Bad jobs only} > \textit{Good jobs only} = \textit{Mixed}.$$

They can be ranked by the value of unemployment and employment as follows:

$$\textit{Good jobs only} > \textit{Mixed} > \textit{Bad jobs only}.$$

Proof of Proposition A-3.

To implement high effort $e = H$, the firm sets $w(0) = 0$ and chooses

$$w(y) = \max\left\{\frac{c^H - c^L}{1-q},\, rV^U + c^H\right\}.$$

To implement low effort $e = L$, the firm sets $w(Y) = rV^U + c^L$, regardless of output Y.

We consider two cases, depending on which term is larger in the expression of $w(y)$ when implementing high effort.

1. $\frac{c^H - c^L}{1-q} \leq rV^U + c^H$. In this case, the worker's flow utility of employment equals $w(y) - c^H = rV^U$. Subsituting into the HJBs of V^U, V^E, we obtain $V^U = V^E = b/r$. We define the cutoff

$$b^B \equiv \frac{qc^H - c^L}{1-q}. \tag{A-2}$$

The initial inequality condition is equivalent to $b \geq b^B$.

Firm's flow profits under high and low efforts become

$$\pi(e = H) = y - w(y) = y - rV^U - c^H,$$
$$\pi(e = L) = qy - w(0) = qy - rV^U - c^L$$

For firms to adopt incentive contracts, we need $\pi(e = H) \geq \pi(e = L)$, i.e.,

$$(1-q)y \geq c^H - c^L,$$

which always holds in our setup.

2. $\frac{c^H - c^L}{1-q} > rV^U + c^H$. This further implies that, when implementing efforts, workers' flow utility

is

$$\mathbb{E}[w(Y)|e=H]-c(e=H)=\frac{c^H-c^L}{1-q}-c^H=b^B.$$

The initial inequality thus implies that $b^B > rV^U$. Combined with HJBs of V^U, V^E, we have $V^E > V^U$ and $b < b^B$.

Firm profits become

$$\pi(e=H)=y-w(y)=y-\frac{c^H-c^L}{1-q},$$
$$\pi(e=L)=qy-w(0)=qy-rV^U-c^L.$$

We have three possible cases:

(a) All firms adopt incentive contracts. In this case,

$$V^U=\frac{(r+\lambda)b+f(\theta)b^B}{r(r+\lambda+f(\theta))}.$$

The condition that no single firm wants to switch to no incentive contracts is

$$y-\frac{c^H-c^L}{1-q}>qy-rV^U-c^L.$$

We define the threshold

$$b^A\equiv\frac{c^H-(2-q)c^L}{1-q}-(1-q)y, \tag{A-3}$$

and the above condition is equivalent to $rV^U > b^A$. Substituting the solution of V^U, we have a new threshold for b:

$$b>\underline{b}\equiv\frac{r+\lambda+f(\theta)}{r+\lambda}b^A-\frac{f(\theta)}{r+\lambda}b^B. \tag{A-4}$$

Since $b^A < b^B$, it is straightforward that $\underline{b} < b^A$.

(b) No firm implements incentive contracts. In this case, $w(e=L)-c^L = rV^U$, which further implies that $V^E = V^U = b/r$. For firms not to implement high efforts, we need $\pi(e=H) < \pi(e=L)$, i.e.,

$$y-\frac{c^H-c^L}{1-q}<qy-b-c^L\Rightarrow b<b^A.$$

(c) Firms are indifferent between implementing efforts and not. We solve for the fraction, γ, of firms that implement efforts. Due to random matching, workers matched to high-effort firms with probability γ and earn flow utility of b^B. With probability $1-\gamma$, they are

matched to low-effort firms and earn flow utility of rV^U. The expected flow utility of employment is therefore

$$\gamma b^B + (1-\gamma) r V^U.$$

For firms to be indifferent between implementing efforts and not, we need $\pi(e=H) = \pi(e=L)$:

$$y - \frac{c^H - c^L}{1-q} = qy - rV^U - c^L \Rightarrow b^A = rV^U.$$

Substituting back into the equation for the flow utility of employment and the solution of V^U, we obtain

$$\gamma = \frac{(r+\lambda)(b^A - b)}{f(\theta)\,(b^B - b^A)}. \tag{A-5}$$

One can verify that $\gamma \in (0,1)$ when $b \in (\underline{b}, b^A)$.

We now compare the three equilibria when $b \in (\underline{b}, b^A)$. The firms' flow profits are

$$\pi_{\text{good-job}} = \pi_{\text{mixed}} = y - \frac{c^H - c^L}{1-q}, \quad \pi_{\text{bad-job}} = qy - b - c^L.$$

When $b < b^A$, we have $\pi_{\text{bad-job}} > \pi_{\text{good-job}} = \pi_{\text{mixed}}$.

The equilibrium market tightness is determined by the zero-profit condition for posting a vacancy, $\Pi^V = 0 \Leftrightarrow s(\theta)\pi = (r+\lambda)\kappa$. Since $s(\theta)$ is decreasing in θ, we must have

$$\theta_{\text{bad-job}} > \theta_{\text{good-job}} = \theta_{\text{mixed}}.$$

Using the steady-state condition, and the formula for the job finding rate, we have

$$\lambda M = \mu(U,V) = f(\theta)U = f(\theta)(L-M).$$

Since $f(\theta)$ is increasing in θ, we must have

$$M_{\text{bad-job}} > M_{\text{good-job}} = M_{\text{mixed}}.$$

Next, we consider workers' utilities. For unemployed workers, we have

$$V^U_{\text{good-job}} = \frac{(r+\lambda)b + f(\theta^*)b^B}{r(r+\lambda+f)},$$
$$V^U_{\text{bad-job}} = \frac{b}{r},$$
$$V^U_{\text{mixed}} = \frac{b^A}{r}.$$

For employed workers, we have

$$V^E_{\text{good-job}} = \frac{\lambda b + (r + f(\theta^*))b^B}{r(r + \lambda + f(\theta^*))},$$
$$V^E_{\text{bad-job}} = \frac{b}{r},$$
$$V^E_{\text{mixed}} = \frac{(r + f(\theta^*))b^A - rb}{rf(\theta^*)}.$$

In the expressions of $V^U_{\text{good-job}}$, $V^E_{\text{good-job}}$ and V^E_{mixed}, the market tightness θ^* refers to that corresponding to the profit level $\pi = y - \frac{c^H - c^L}{1-q}$. One can verify that

$$V^U_{\text{good-job}} > \frac{(r + \lambda)\underline{b} + f(\theta^*)b^B}{r(r + \lambda + f(\theta^*))} = \frac{b^A}{r} = V^U_{\text{mixed}}$$
$$V^U_{\text{mixed}} = \frac{b^A}{r} > \frac{b}{r} = V^U_{\text{bad-job}},$$
$$V^E_{\text{good-job}} > \frac{\lambda\underline{b} + (r + f(\theta^*))b^B}{r(r + \lambda + f(\theta^*))} = \frac{(r + f(\theta^*))b^A - r\underline{b}}{rf(\theta^*)} > V^E_{\text{mixed}},$$
$$V^E_{\text{mixed}} = \frac{(r + f(\theta^*))b^A - rb}{rf(\theta^*)} > \frac{b}{r} = V^E_{\text{bad-job}},$$

where we have used the expression of b^A_{u0} in equation (A-4) to derive the first equality in the first and third lines. ■

The equilibrium structure is similar to the one with an exogenous outside option. In general, firms are more likely to implement high efforts when the unemployment benefit, b, thus the endogenous outside option V^U, is high. A new feature is the existence of multiple equilibria when b is in an intermediate range. The outside option, V^U, depends on the prevailing contracts offered by firms in the market. When more firms offer incentive contracts and implement high efforts, workers receive higher wages and earn rents, which raises V^U. This in turn makes it easier for firms to implement high efforts. Therefore, there exists a self-fulfilling good-jobs-only equilibrium where all firms implement high efforts, and workers earn rents. Conversely, when no firm implements effort, workers receive low wages and earn no rent, leading to a self-fulfilling bad-jobs-only equilibrium. Finally, there also exists a mixed equilibrium where a fraction of firms implement high efforts, and the rest implement low efforts. Firms are indifferent between implementing high and low efforts. Workers earn rents from the high-effort firms and are indifferent between employment and unemployment when matched to low-effort firms.

A.3 Proofs of Lemma 5 and Proposition 3

Proof of Lemma 5. To analyze the change in $\underline{b}$ after AI adoption, we first consider the change in the cutoffs b^A and b^B. Applying equation (A-3), we have

$$b^A_{AI} - b^A = \frac{c^H_{AI} - c^L_{AI} - c^H + c^L + \left(c^L - c^L_{AI}\right)(1-q)}{1-q} > 0.$$

Applying equation (A-2), we have

$$b^B_{AI} - b^B = \frac{q\left(c^H_{AI} - c^H\right) - \left(c^L_{AI} - c^L\right)}{1-q} > \frac{\left(c^H_{AI} - c^H\right) - \left(c^L_{AI} - c^L\right)}{1-q} > 0.$$

One can further show that

$$b^A_{AI} - b^A - (b^B_{AI} - b^B) = c^H_{AI} - c^L_{AI} - c^H + c^L > 0. \tag{A-6}$$

Finally, applying the formula for $\underline{b}$ in equation A-4, we have

$$\underline{b}_{AI} - \underline{b} = b^A_{AI} - b^A + \frac{f(\theta^*_{AI})}{r+\lambda}(b^A_{AI} - b^B_{AI}) - \frac{f(\theta^*)}{r+\lambda}(b^A - b^B),$$

where θ^* and θ^*_{AI} are the equilibrium labor market tightness corresponding to the profit levels under incentive contracts with positive worker rents, i.e., $\pi = y - \frac{c^H - c^L}{1-q}$ and $\pi_{AI} = y - \frac{c^H_{AI} - c^L_{AI}}{1-q}$. Since $\pi_{AI} < \pi$, we have fewer vacancies created and $\theta^*_{AI} < \theta^*$, and thus $f(\theta^*_{AI}) < f(\theta^*)$.

We define $\underline{b}^{PE}_{AI}$ by setting $\theta^*_{AI} = \theta^*$ in the above equation. Therefore,

$$\underline{b}^{PE}_{AI} = \underline{b} + b^A_{AI} - b^A + \frac{f(\theta^*)}{r+\lambda}\left[b^A_{AI} - b^A - (b^B_{AI} - b^B)\right] > \underline{b}.$$

The difference between $\underline{b}_{AI}$ and $\underline{b}^{PE}_{AI}$ becomes

$$\underline{b}_{AI} - \underline{b}^{PE}_{AI} = \frac{f(\theta^*_{AI}) - f(\theta^*)}{r+\lambda}\left(b^A_{AI} - b^B_{AI}\right).$$

Since $f(\theta^*_{AI}) < f(\theta^*)$ and $b^A_{AI} - b^B_{AI} < 0$, we must have $\underline{b}_{AI} > \underline{b}^{PE}_{AI}$, i.e., the general equilibrium feedback further increases the cutoff. ■

Proof of Proposition 3. As analyzed in the proof of Proposition A-3, firm profits are $qy - b - c^L$ when $b < b^A$ and firms implement no efforts, $y - \frac{c^H - c^L}{1-q}$ when $b \in [\underline{b}, b^A]$ and firms implement efforts, and $y - b - c^H$ when $b \geq b^B$. These expressions are identical to those for exogenous outside options. Since we maintain the same assumption that $(1-q)y - (c^H - c^L_{AI}) > 0$, we can apply the results in Proposition 2, i.e., there is a cutoff b^{A*} (corresponding to $\bar{u}^{A*}$ in the exogenous outside option case), such that at $b = b^{A*}$, firms' profits from a bad job after AI adoption are the same as those from a good job before AI adoption. We have solved for this cutoff in the proof of

the corollary:

$$b^{A*} = \frac{c^H - c^L}{1-q} - c^L_{AI} - (1-q)\, y.$$

Taking the difference between $\underline{b}_{AI}$ and b^{A*}, we have

$$\begin{aligned}\underline{b}_{AI} - b^{A*} &= b^A_{AI} + \frac{f(\theta^*_{AI})}{r+\lambda}(b^A_{AI} - b^B_{AI}) - \frac{c^H - c^L}{1-q} + c^L_{AI} + (1-q)\, y \\ &= \frac{c^H_{AI} - c^L_{AI} - \left(c^H - c^L\right)}{1-q} - \frac{f(\theta^*_{AI})}{r+\lambda}\left[(1-q)y - (c^H_{AI} - c^L_{AI})\right].\end{aligned}$$

Note that $f(\theta^*_{AI})$ is bounded by $f(\theta^*)$ from above. Therefore, when $(1-q)y - (c^H_{AI} - c^L_{AI})$ is sufficiently small, we must have $\underline{b}_{AI} > b^{A*}$.

For $b \in (b^{A*}, \underline{b}_{AI})$, firms' profits from a bad job after AI adoption are higher than those from a good job before AI adoption. However, since $b < \underline{b}_{AI}$, profits from a good job after AI adoption is even lower. Therefore, firms switch from good jobs to bad jobs after AI adoption. To see the change in employment levels, note that the value of a vacancy has to be zero in the steady-state equilibrium:

$$\Pi^V = 0 \Rightarrow s(\theta)\pi = (r+\lambda)\kappa.$$

Since $s(\theta)$ decreases in θ, a decline in π implies a drop in θ. Moreover, we have

$$\lambda M = \mu(U,V) = f(\theta)(L-M).$$

A drop in θ means a smaller $f(\theta)$, thus lower M. The number of matches, or equivalently the employment level (good jobs), drops after AI adoption. ■