From Production to Delegation: AI Adoption with Endogenous Restriction Risk

Vikram K Suresh
University of Cincinnati, Department of Computer Science
krishnvv@mail.uc.edu

Abstract

I examine how AI adoption reshapes human capital through skill substitution from production to AI-dependent management capabilities. Workers face endogenous restriction risk that increases with aggregate AI adoption, creating a coordination failure: individuals over-invest in AI-dependent skills without internalizing systemic fragility. Incorporating image concerns reveals how social stigma initially delays adoption but generates rapid cascades once norms shift, amplifying vulnerability. The decentralized equilibrium features excessive AI reliance and insufficient investment in robust fallback skills relative to the social optimum.

Keywords: Skill Substitution; AI Adoption; Labor Delegation; Restriction Risk

JEL Codes: J24, O33, D62, D83

I Introduction

"I've never felt this much behind as a programmer," wrote AI researcher Andrej Karpathy on X (formerly Twitter) in December 2025. "The profession is being dramatically refactored as the bits contributed by the programmer are increasingly sparse and between." He described the emergence of "a new programmable layer of abstraction to master involving agents, subagents, their prompts, contexts, memories, modes, permissions, tools, plugins, skills, hooks" and "need to build an all-encompassing mental model for strengths and pitfalls of fundamentally stochastic, fallible, unintelligible and changing entities." The challenge was no longer writing code, it was managing AI systems that work but cannot be fully understood. Karpathy characterized this transition as a 'magnitude 9 earthquake' rocking the profession, warning workers to 'roll up your sleeves to not fall behind.'[1]

This transformation is part of a broader economic shift driven by the rapid adoption of AI technologies across industries. Recent work by Acemoglu [2024] builds on the

[1] X post by Andrej Karpathy. `https://x.com/karpathy/status/2004607146781278521`, from Decemeber 26, 2025.

task-based approach from Acemoglu and Restrepo [2018] and Acemoglu and Restrepo [2019] and shows automation can complement labor when productivity effects dominate displacement effects, though gains are found to be modest. Noy and Zhang [2023] and Brynjolfsson et al. [2025] find 40% productivity gains from generative AI for professional workers. Shahidi [2025] argue AI agents could enable a 'Coasean singularity' where transaction costs reductions can create new market designs. Yet, Almog [2025] documents workers reduce AI reliance by 14% when usage is observable by supervisors, fearing it signals poor judgement.

Karpathy's observation reveals a shift from production (coding, debugging, understanding mechanisms) to management (prompting, orchestrating AI systems, delegating to opaque agents). I develop two models to analyze this substitution. In the first model, workers choose between production skills θ_P and management skills θ_M that facilitates AI delegation. This is under an endogenous restriction risk $\pi(\Theta_M)$, where Θ_M is the aggregate management skill. The decentralized equilibrium features over-adoption of management skills relative to the social optimum due to the externality. In the second model, workers are subject to Almog [2025]'s worker image concerns, showing how initial stigma delays adoption but creates a sudden surge in management skill acquisition once AI norms shift at workplaces, explaining why the transition feels abrupt.

The paper contributes to the literature on AI adoption and labor market dynamics. The productivity gains from the adoption of AI can come with hidden costs and markets may under-invest in fallback options. Section II presents the basic model of skill substitution with endogenous restriction risk. Section III incorporates the worker image concerns. Section IV concludes with implications for policy.

II Model 1: Skill Substitution with Endogenous Restriction Risk

II.A Framework

A continuum of homogeneous workers $i \in [0, 1]$ live for two periods $t = 0, 1$. In period 0, workers invest in production skills θ_P, representing traditional expertise, and management skills θ_M, representing the ability to delegate tasks and manage AI agents. In period 1, workers produce output using either their own production skills or AI assistance mediated by their management skills.

Skill acquisition costs are assumed to be separable and convex as given by

$$C(\theta_P, \theta_M) = c_P(\theta_P) + c_M(\theta_M),$$

with $c'_P, c'_M > 0$ and $c''_P, c''_M > 0$. Management skills are cheaper to acquire, i.e. $c'_M(\theta) <$

$c'_P(\theta)$ for sufficiently small θ.

At $t = 1$, output is given by

$$y^{\text{manual}} = \theta_P, \qquad y^{\text{AI}} = \theta_M \cdot A - \varepsilon(A),$$

where A is the exogenous quality of AI. The term $\varepsilon(A)$ captures black-box errors that are difficult to identify even with high management skill. We assume $\varepsilon'(A) > 0$, reflecting increasing difficulty in detecting errors in state-of-the-art AI systems.

Workers choose the production method that maximizes output as the following,

$$y = \max\{\theta_P,\, \theta_M \cdot A - \varepsilon(A)\}.$$

There exists an endogenous probability $\pi = \pi(\Theta_M)$ that AI usage is restricted or disrupted in period 1, where

$$\Theta_M = \int_0^1 \theta_{M,i}\, di$$

denotes aggregate management skill, capturing the intensity of AI adoption. The restriction probability satisfies $\pi'(\Theta_M) > 0$ and $\pi''(\Theta_M) \geq 0$, reflecting increasing and convex political, regulatory, and systemic risks associated with widespread AI reliance.[2]

II.A.1 Worker's Problem

Each worker chooses (θ_P, θ_M) at $t = 0$ to maximize expected utility:

$$\max_{\theta_P, \theta_M} (1 - \pi(\bar{\Theta}_M)) \cdot \max\{\theta_P, \theta_M A - \varepsilon(A)\} + \pi(\bar{\Theta}_M) \cdot \theta_P - c_P(\theta_P) - c_M(\theta_M), \quad (1)$$

where $\bar{\Theta}_M$ is taken as given by the individual worker.

We focus on an *AI-assisted equilibrium* in which $\theta_M A - \varepsilon(A) > \theta_P$. We can write the first order conditions to be the following,

For production skills θ_P:

$$\pi(\bar{\Theta}_M) = c'_P(\theta^*_P), \quad (2)$$

since production skills only matter in the restricted state.

For management skills θ_M:

$$(1 - \pi(\bar{\Theta}_M)) \cdot A = c'_M(\theta^*_M). \quad (3)$$

A decentralized equilibrium is a symmetric equilibrium in which $\theta_{P,i} = \Theta^*_P$ and $\theta_{M,i} = \Theta^*_M$ for all i, and consistency requires $\bar{\Theta}_M = \Theta^*_M$.

[2]For example, rising energy consumption, labor displacement visibility, and geopolitical exposure; see `https://thehill.com/policy/technology/5665503-sanders-ai-data-centers/`.

II.A.2 Social Planner's Problem

The social planner chooses aggregate skill levels (Θ_P, Θ_M) to maximize aggregate welfare:

$$\max_{\Theta_P, \Theta_M} (1 - \pi(\Theta_M)) \cdot [\Theta_M A - \varepsilon(A)] + \pi(\Theta_M) \cdot \Theta_P - \mathcal{C}(\Theta_P, \Theta_M), \tag{4}$$

where

$$\mathcal{C}(\Theta_P, \Theta_M) = \int_0^1 [c_P(\theta_{P,i}) + c_M(\theta_{M,i})] \, di = c_P(\Theta_P) + c_M(\Theta_M)$$

under symmetry.

The first-order condition for Θ_M is:

$$(1 - \pi(\Theta_M^{SP})) \cdot A - \pi'(\Theta_M^{SP}) \cdot \left[\Theta_M^{SP} A - \varepsilon(A) - \Theta_P^{SP}\right] = c_M'(\Theta_M^{SP}). \tag{5}$$

The planner internalizes the marginal effect of aggregate AI adoption on the restriction risk probability.

II.A.3 Main Result

Proposition 1 (Over-Investment in AI-Dependent Skills)**.** *In the AI-assisted equilibrium, there is an over-investment in AI-dependent management skills and under-investment in traditional production skills compared to the socially optimum amounts,*

$$\Theta_M^* > \Theta_M^{SP} \qquad \textit{and} \qquad \Theta_P^* < \Theta_P^{SP}.$$

Proof. From the planner's first-order condition (5),

$$(1 - \pi(\Theta_M^{SP})) \cdot A = c_M'(\Theta_M^{SP}) + \pi'(\Theta_M^{SP}) \cdot \underbrace{[\Theta_M^{SP} A - \varepsilon(A) - \Theta_P^{SP}]}_{>0 \text{ in AI equilibrium}}.$$

The right-hand side exceeds the first-order condition (3) by the strictly positive externality term $\pi'(\Theta_M)$. Since $c_M'' > 0$, this implies $\Theta_M^{SP} < \Theta_M^*$.

From (2), a lower Θ_M reduces $\pi(\Theta_M)$, implying a higher optimal Θ_P under the planner than in the decentralized equilibrium. □ □

II.B Interpretation and Implications

Individual workers therefore are likely to ignore the negative externality from their AI adoption. They perceive the immediate benefits of their contribution to the aggregate management skill Θ_M as negligible. This results in an equilibrium where the economy is overly reliant on AI systems which are susceptible to disruptions. As the quality of AI A improves, the incentive to invest in AI management skills grows which can exacerbate

the externalities and create a wider gap between socially optimum and decentralized outcomes.

III Model 2: Image Concerns and Adoption Dynamics

III.A Motivation

While Model 1 explains over-investment in AI-dependent skills in a static setting, it does not capture the dynamic adoption patterns observed in practice. Almog [2025]'s finds that workers reduce AI reliance when usage is observable, indicating that image concerns play a significant role in adoption decisions. Meanwhile, Karpathy's description of the AI transition as a sudden "magnitude 9 earthquake" suggests a rapid shift rather than a gradual change. By allowing workers to care about both output and perceived competence, Model 1 is extended to incorporate observable AI use and image concerns, showing how initial stigma delays adoption but generates rapid adoption once norms shift.

III.B Image Concerns and Utility

Production is identical to Model 1. After output is realized, both output y_i and AI usage $u_i \in \{0, 1\}$ are observable to evaluators (e.g., managers, peers, or clients), who form judgments about worker competence. Manual production is associated with a higher perceived judgment J^{high}, while AI-assisted production is associated with a lower perceived judgment J^{low}, where $J^{\text{high}} > J^{\text{low}}$.

Worker utility incorporates both output and anticipated image concerns as defined by the following,

$$U_i = y_i - \beta(t) \cdot \mathbf{1}\{u_i = 1\}\Delta J,$$

where $\Delta J \equiv J^{\text{high}} - J^{\text{low}} > 0$, and $\beta(t)$ captures the prevalence of image concerns at time t. Workers internalize this anticipated image penalty when choosing whether to rely on AI.

III.B.1 Social Norm Dynamics

The weight placed on image concerns declines over time as social norms change over time,

$$\beta(t) = \beta_0 e^{-\lambda t},$$

where $\beta_0 > 0$ represents the initial stigma associated with AI usage and $\lambda > 0$ captures the speed of norm evolution. Early in the transition, high stigma discourages AI adoption despite productivity gains. As norms shift and $\beta(t)$ falls, image costs become lower, eventually allowing productivity considerations to dominate.

A worker chooses AI-assisted production if and only if,

$$\theta_{M,i}A - \varepsilon(A) - \theta_{P,i} > \beta(t)\Delta J.$$

Critical threshold. Define the critical image weight at which worker i is indifferent between AI-assisted and manual production,

$$\beta^*(i) \equiv \frac{\theta_{M,i}A - \varepsilon(A) - \theta_{P,i}}{\Delta J}. \tag{6}$$

Worker i adopts AI whenever $\beta(t) < \beta^*(i)$.

III.B.2 Adoption Dynamics and Cascades

To isolate social norm dynamics from selection effects, we focus on a symmetric equilibrium with homogeneous workers, implying a common threshold β^*. As $\beta(t)$ declines. The economy passes through three distinct phases,

Phase 1: Stigma-dominated regime ($\beta(t) > \beta^*$). Image concerns dominate productivity gains, and workers avoid AI despite higher output. Relative to the AI-assisted equilibrium of Model 1, investment remains tilted toward production skills.

Phase 2: Tipping point ($\beta(t^*) = \beta^*$). At time t^*, image costs and productivity gains exactly offset. Small shifts in norms generate a rapid increase in AI adoption, this produces an adoption cascade over a short time interval.

Phase 3: Post-norm-shift regime ($\beta(t) < \beta^*$). AI adoption becomes widespread. Investment shifts toward AI-dependent management skills, with $\Theta_M(t)$ converging toward the AI-assisted equilibrium characterized in Model 1. As traditional production skills decline, the economy becomes increasingly susceptible to AI restriction or disruption shocks.

III.C Interpretation and Implications

Incorporating image concerns reveals how social norms shape the timing and the weaknesses of AI-driven productivity gains. Initial stigma slows adoption and preserves traditional production skills. Once norms shift, rapid adoption generate excessive reliance

on AI and underinvestment in fallback capabilities, magnifying the welfare costs of AI restrictions or failures.

IV Discussion and Conclusion

IV.A Relation to Existing Work

The results presented in this paper add to the growing literature on the economic effects of AI adoption by understanding skill substitution and subsequent externalities rather than task-level productivity alone. Acemoglu [2024] studies the macroeconomic impact of generative AI and finds modest aggregate productivity gains over the next decade. This paper shows, AI driven skill substitution can generate hidden costs that standard productivity accounting misses. Shahidi [2025] highlight how AI reduces transaction costs and enables new market designs. This paper complements their findings by showing decentralized adoption can amplify externalities associated with political, regulatory, and systemic risk. Almog [2025] documents that workers reduce AI usage when it is observable, pointing to the importance of image concerns. Such frictions can be welfare-improving by slowing adoption and preserving traditional skills while policy responses adjust.

Recent experimental evidence reveals the critical role of AI implementation design. Bastani et al. [2024] find that students using GPT-4 tutoring performed 127% better on practice problems but 17% worse on unassisted exams, direct evidence of skill erosion when users simply ask AI for answers rather than engage with the problem. Conversely, Brynjolfsson et al. [2025] document that customer support workers who actively engaged with AI suggestions retained performance gains during system outages, even after three months without AI assistance. Passively relying on AI disrupts skills, actively engaging likely preserves them. The externality operates when workers substitute θ_M for θ_P without maintaining underlying capability, as shown by Bastani et al. [2024] in the educational setting.

IV.B Policy Implications

Individual workers and firms do not internalize the increased risk associated with aggregate AI dependence. A Pigouvian tax that discourage excessive reliance on AI-dependent skills can be adopted as a policy. Taxes that are proportional to the marginal contribution of AI adoption to restriction risk,

$$\tau^* = \pi'(\Theta_M^{SP}) \cdot [\Theta_M^{SP} \cdot A - \varepsilon(A) - \Theta_P^{SP}],$$

or, equivalently, subsidies that encourage investment in traditional production skills. In practice, implementation is complicated by the fact that AI disruptions may come from multiple sources including political, geopolitical, energy-related, or technical. This suggests a menu approach to policy, combining incentives for traditional skill development with energy pricing, strategic reserves of human capital, and supply-chain fortification.

IV.C Conclusion

Rapid AI adoption can deliver substantial productivity gains, but can create systemic susceptibility through endogenous skill depletion. As individuals cannot internalize the risk of their skill substitution, markets tend to under-invest in fallback capacity. As AI adoption increases surpassing the stigma, the economy may appear highly productive yet perform worse than a no-AI baseline when AI access is disrupted. Long-run outcome depends on coordinated choices about skill investment, AI deployment, and supporting institutions. Future work could explore heterogeneity among workers and firms, and policy designs in more detail.

References

Daron Acemoglu. The simple macroeconomics of AI. (32487), May 2024. URL `http://www.nber.org/papers/w32487`.

Daron Acemoglu and Pascual Restrepo. The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542, 2018.

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

David Almog. Barriers to ai adoption: Image concerns at work. Working Paper, 2025. URL `https://arxiv.org/abs/2511.18582`.

Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Ozge Kabakcı, and Rei Mariman. Learning vs. knowing: An experimental analysis of AI-assisted decision-making. SSRN Working Paper 4895486, 2024. URL `https://ssrn.com/abstract=4895486`.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025. doi: 10.1093/qje/qjae044.

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

Peyman Shahidi. The Coasean singularity: AI agents and the future of markets. Working Paper, 2025. URL https://www.nber.org/papers/w34468.